



Leximancer Manual

Version 4

© 2011

Table of Contents

SECTION 1. INTRODUCTION TO LEXIMANCER	4
WHAT IS LEXIMANCER?	4
THEORY: CONTENT ANALYSIS	8
WHAT IS CONTENT ANALYSIS?	8
Types of Content Analysis	9
INTERESTED IN LEARNING MORE ABOUT CONTENT ANALYSIS?	10
SECTION 2. THE CONCEPT MAP	10
Theory: Concepts and Conceptual Mapping in Leximancer.....	10
Concept Seed Words	11
Concept Learning	12
The Initial Display.....	13
Themes	14
Concepts	18
Buttons in the header above the Concept Map	21
The Concept Cloud	27
REPORT TABS	29
Themes tab	29
Concepts tab	30
Thesaurus	34
Pathway tab	36
Query	39
Summary tab	43
SECTION 3: CREATING AN AUTOMATIC/EXPLORATORY MAP.....	44
CREATING AN AUTOMATIC CONCEPT MAP	44
Supported File Types	44
Desktop Installations	44
Leximancer Portal Accounts	45
CREATING A NEW FOLDER AND PROJECT	47
New Folder	47
New Project	48
THE MAIN LEXIMANCER USER INTERFACE:	50
Using the Web Crawler.....	55
SECTION 4: CREATING A MANUALLY ADJUSTED MAP.....	62
2A. TEXT PROCESSING	63
STOPWORD REMOVAL	68
2B. CONCEPT SEEDS SETTINGS	74
3. THESAURUS GENERATION:	78
3A. PRACTICAL: CONFIGURING CONCEPT EDITING	79
USING TAGS.....	86
THE AUTOMATIC SENTIMENT LENS.....	88

Configuring Sentiment Lens.....	88
3B. GENERATING THE THESAURUS.....	92
CONCEPT PROFILING	96
4. RUN PROJECT:.....	99
4A. EDITING COMPOUND CONCEPTS	99
4B. CONCEPT CODING.....	109
KILL CONCEPTS AND REQUIRED CONCEPTS.....	111
CLASSIFICATION SETTINGS.....	112
4C. PROJECT OUTPUTS.....	116
GENERATING THE CONCEPT MAP	117
TOPICAL VERSUS SOCIAL MAPPING.....	118
CONFIGURING THE INSIGHT DASHBOARD REPORT.....	125
DATA EXPORTS	131
SECTION 5: EXAMPLE ADVANCED TECHNIQUES.....	138
1. MANUAL CONCEPT SEEDING	139
CONFIGURING MANUAL CONCEPT SEEDING	140
2. PROFILING	144
CONFIGURING CONCEPT PROFILING	146
3. PROFILING USING TAG CATEGORIES	148
Discrimination of Categories based on Semantics	153
5. ANALYSING TRANSCRIPTS	160
6. ANALYSING SPREADSHEET DATA	167
PRACTICAL: ANALYSING SPREADSHEET DATA.....	167

Applications of Leximancer

Application	Type of Text	Output Options	Possible Projects
Basic Text Analysis	Any non-protected text: Word Docs, PDF, online content (html), .txt, .xml, etc.	Visual, via the Concept Map; Report, via the Insight Dashboard; Statistical, via Leximancer data exports.	Communication research; Analysis of speeches over time; Blog analysis.
Coding Open-ended Surveys	Customer feedback data; Call Centre data; Qualitative research spreadsheet data.	Statistical, via Leximancer data exports; Link open-ended questions to metadata.	Employee satisfaction survey; Net Promoter Score analysis.
Site and Archive Concept Navigation	Electronic content; Litigation evidence (electronic form)	Profile concepts being investigated; Concept co-occurrence data.	Legal e-discovery; Alternative to manually maintained site maps.
Media Analysis	Electronic media articles	Profile of company or issue	Competitor analysis; Online opinion analysis.
Customer Relationship Management (CRM)	Communication from customers	Current issues and concerns of customers	Policy and campaign development
Academic Research	Any	Concept Map; Statistical output.	History, Literature, Media Studies, Sociology, Politics...

- If you would like to see Leximancer in action, example maps can be found under on the Science page of the Leximancer website <https://www.leximancer.com/>.

Theory: Content Analysis

- *This section of the manual is for those wishing to understand more about the theoretical underpinnings of Leximancer;*
- *More practical, instructional chapters are to follow.*

What is Content Analysis?

Content analysis is a research tool used for determining the presence of words or concepts in collections of textual documents. It is used for breaking down the material into manageable categories and relationships in order to quantify and analyze text.

Once extracted, these measurements can be used to make valid inferences about the ideas contained within the text (such as the presence of propaganda), properties of the writer or speaker (such as his or her psychological state), the audience to which the material is presented, or properties of the culture of the time in which the material was written.

Content analysis is an important research methodology as it can be used to analyse any form of verbal communication from written to spoken forms. As text documents tend to exist over long periods of time, the technique can be used to extract valuable historical and cultural insights.

As content analysis can be performed on numerous forms of data ranging from political speeches and open-ended interviews to newspaper articles and historical documents, it is invaluable to many researchers. Such uses include:

- historical analysis of political speeches
- detecting the existence and level of propaganda
- coding surveys that ask open-ended questions
- determining the psychological state of the writers
- assesses textual content against measures (e.g. censoring)
- assess cultural differences in populations.

Types of Content Analysis

In general, approaches to content analysis fall into two major categories: conceptual analysis and relational analysis.

In **conceptual analysis**, documents are measured for the presence and frequency of concepts. Such concepts can be words or phrases, or more complex definitions, such as collections of words representing each concept. One of Leximancer's main features is that it can automatically extract its own dictionary of terms for each document set using this information. That is, it is capable of inferring the concept classes that are contained within the text, explicitly extracting a thesaurus of terms for each concept. This approach also relieves the user of the task of formulating their own coding scheme.

Relational analysis, by contrast, measures how such identified concepts are related to each other within the documents. Leximancer measures the co-occurrence of concepts found within the text, automatically extracts this information, and represents the information visually for comparison. By doing so it displays the main relationships between concepts.

One of the strengths of the Leximancer system is that it conducts both forms of analysis, measuring the presence of defined concepts in the text as well as how they are interrelated. The following sections describe

Leximancer's method for extraction of these concepts and their interrelationships.

Interested in Learning More about Content Analysis?

If you are interested in learning more about the issues relating to content analysis and the various techniques that are used, we recommend the following book: **Weber, R.P. (1990) Basic Content Analysis. Newbury Park, Calif.: Sage Publications, 2nd ed.**

Section 2. The Concept Map

Theory: Concepts and Conceptual Mapping in Leximancer

Concepts in Leximancer are collections of words that generally travel together throughout the text. For example, a concept building may contain the keywords mill, warrant, tower, collapsed, etc.

These terms are weighted according to how frequently they occur in sentences containing the concept, compared to how frequently they occur elsewhere. Sentences are tagged as containing a concept if enough accumulated evidence is found.

Terms are weighted so the presence of each word in a sentence provides an appropriate contribution to the accumulated evidence for the presence of a concept. That is, a sentence (or group of sentences) is only tagged as containing a concept if the accumulated evidence (the sum of the weights of the keywords found) is above a set threshold.

Thesaurus Concept ▲	Word	Score ▼
area	building	6.77
assessment	mill	6.02
basement	warrant	5.71
building	<i>Grovepark Street</i>	5.49
<i>Calor</i>	tower	5.42
case	collapsed	4.85
coating	chimney	4.74
companies	stability	4.67
company	pitched	4.6
correct	rectangular	4.53
corrosion		

Aside from detecting the overall presence of a concept in the text, the concept definitions are also used to determine the frequency of co-occurrence between concepts. This co-occurrence measure is what is used to generate the concept map.

Concept Seed Words

In Leximancer, the definition of each concept (i.e. the set of weighted terms) is automatically learned from the text itself. Concept seed words represent the starting point for the definition of such concepts, with each concept definition containing one or more seeds.

They are called seeds as they represent the starting point of the concept, with more terms being added to the definition through learning. Occasionally, more appropriate central terms may be discovered, pushing the seeds away from the centre of the concept definition.

Leximancer automatically identifies concept seeds by looking for words that most frequently appear in the text. Alternatively, the user can manually provide seed words.

Concept Learning

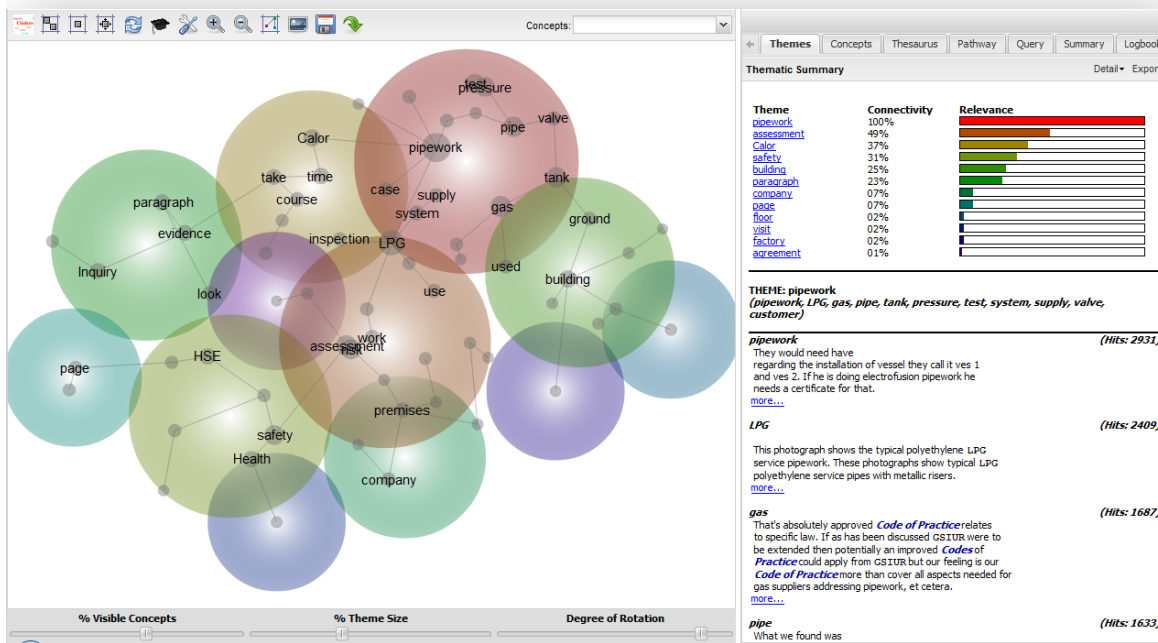
Leximancer begins with a set of seed words, as defined above. During the learning process, words highly relevant to the seed are continuously updated, and eventually form a thesaurus of terms for each concept.

Apart from adding highly relevant words to a concept, Leximancer may also add words that are negatively correlated with the concept (i.e. words that rarely appear in sentence blocks containing the concept and frequently appear elsewhere).

The aim of concept learning is to discover clusters of words which, when taken together as a concept, maximise the relevancy of all the other words in the document.

The Concept Map

Once Leximancer has run the learning process and developed a list of concepts contained in the text, and their relationship to each other, the information is presented via the Concept Map.



Understanding the Concept Map

The Concept Map is divided into two sections: a visual display of concepts and their relationships to each other on the left; and report tabs on the right for interacting with the concept map.

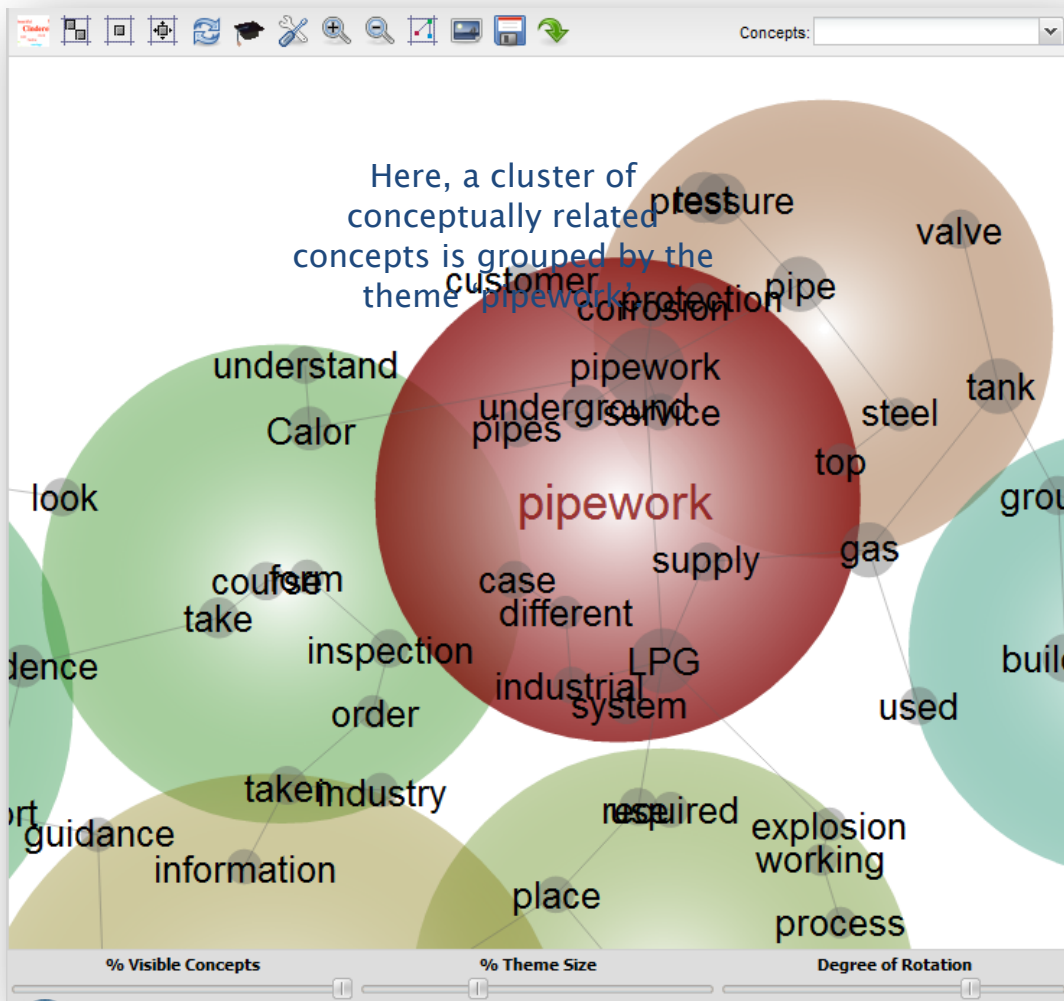
The Initial Display

When the map first opens, the top 50% of concepts are visible on the map. These are the concepts that appear most frequently in the text, and those that are most-connected to other concepts on the map.

Use the % Visible Concepts slider (beneath the map) to change the number of concepts visible on the map. Moving the slider all the way to the left hides all the concepts, and moving it all the way to the right reveals all the concepts.

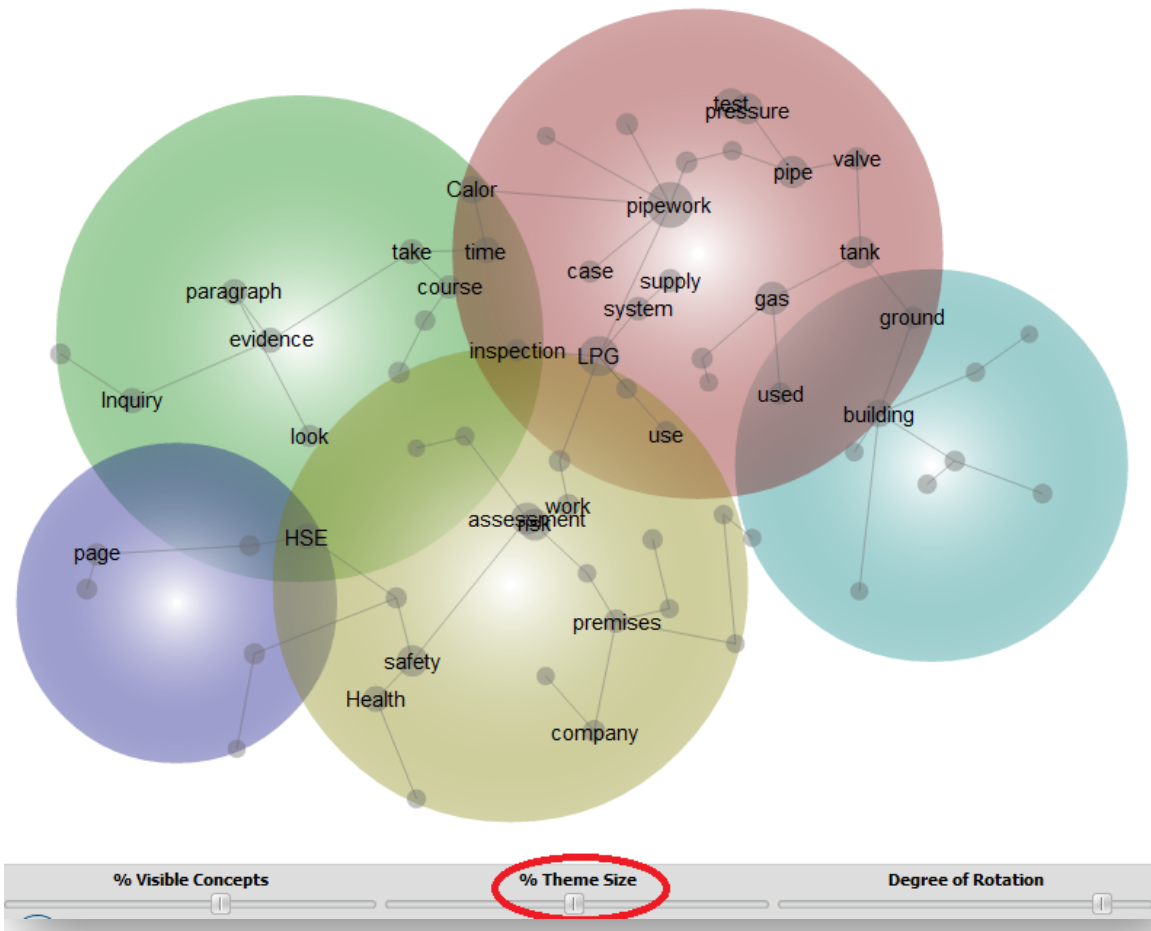
Themes

The concepts are clustered into higher-level 'themes' when the map is generated. Concepts that appear together often in the same pieces of text attract one another strongly, and so tend to settle near one another in the map space. The themes aid interpretation by grouping the clusters of concepts, and are shown as coloured circles on the map:



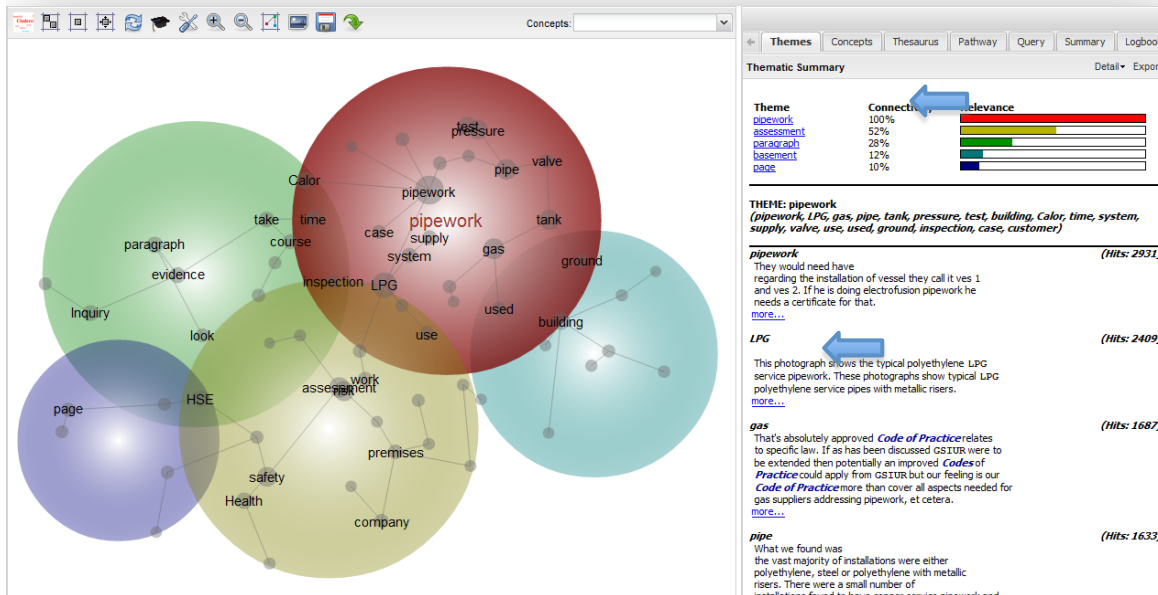
The themes are heat-mapped to indicate importance. This means that the 'hottest' or most important theme appears in red, and the next hottest in orange, and so on according to the colour wheel.

When the map first opens, the Theme Size is set to 33%, but you can move the Theme Size slider beneath the map to adjust the grouping of concepts on the map. Move the slider to the right to make fewer, broader themes, and move it to the left to make more, tighter themes:



When the map first opens, the tab on the right presents a Summary of the Themes. A bar chart ranks the most important themes relative to one another, and beneath that the concepts visible within each theme are listed. A list of representative excerpts is included for each theme, so that you can read some examples quickly to understand how and why the concepts in that theme appear together in the text. Hover your mouse over the More button to see the query syntax used to return each excerpt within a theme, and click the more button to read further examples.

In the screen shot below, the Pipework theme (shown as a red circle on the map) contains concepts such as pipework, LPG and gas. Excerpts linking these concepts are shown in the Theme Summary in the right-hand tab:



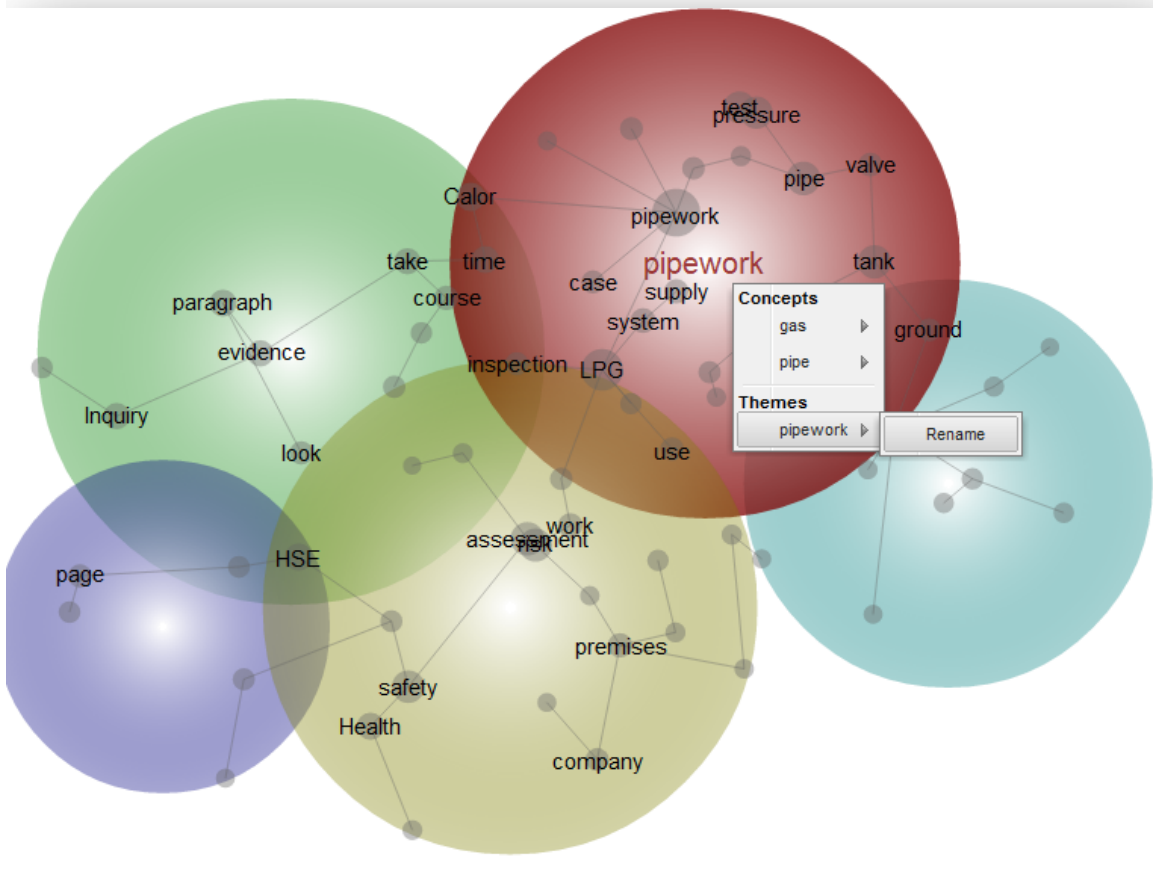
If you adjust the size of the theme circles using the slider beneath the map, the Themes Summary updates to represent the new groups you have created on the map.

You can make all the themes disappear from the map by moving the Theme Size slider all the way to the left (0%).

If you hover your mouse over a theme circle on the map, the name of that theme will appear.

Initially, each theme takes its name from the most frequent and connected concept within that circle.

You can change the names of the themes if you right click on the map near the theme name. A list of nearby concepts and themes will appear. Hover your mouse over the concept or theme of interest to get an option to Rename it:



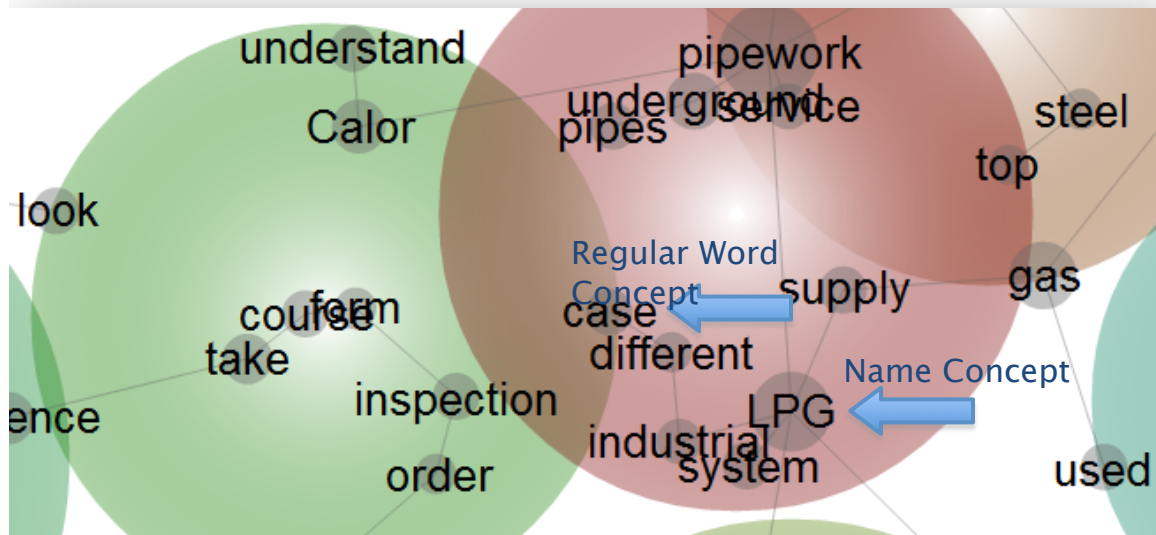
You can make all the theme names visible permanently on the map by clicking the Map Settings (crossed spanner and screwdriver) button in the header above the map and tick Theme Names Always Visible.

Concepts

The Concept Map contains the names of the main concepts that occur within the text. These are shown as grey labels on the map.

Concepts written with an upper case first letter as assumed to represent name-like (proper noun) concepts. These often include the names of people or locations, and appear with a capital first letter on the map. Examples below include LPG and Calor. All other word-like concepts appear in lower case on the map, and refer to objects, actions and so on.

Concept Display



The frequencies with which the name- and word-like concepts appear in the text are also listed separately in the Concepts tab on the right of the map:

Name
Concept

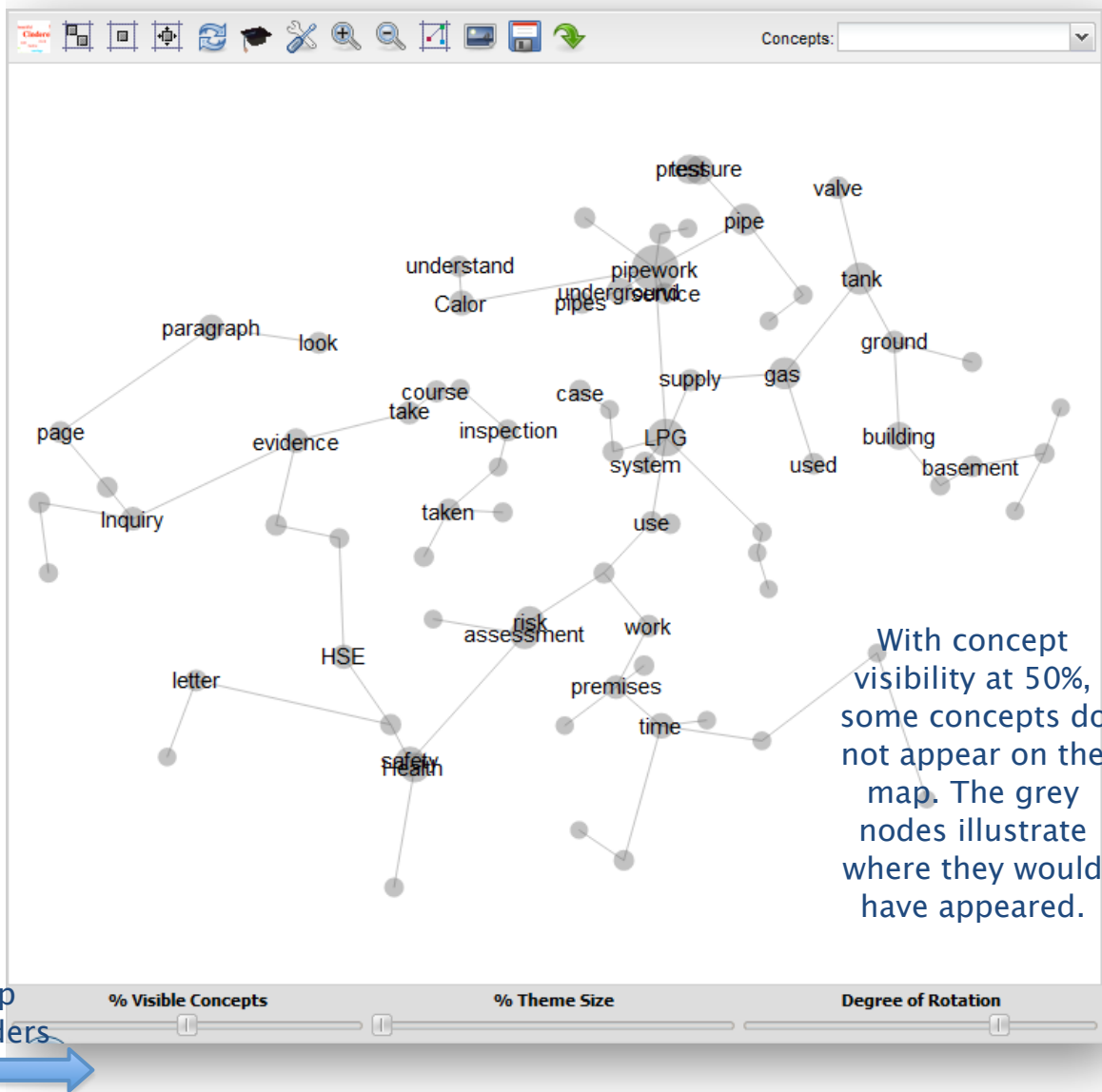


Regular
Word
Concept



Ranked Concepts				Export
Name-Like	Count	Relevance		
LPG	1652	76%	<div style="width: 76%;"></div>	
<i>Health</i>	1080	50%	<div style="width: 50%;"></div>	
<i>Calor</i>	873	40%	<div style="width: 40%;"></div>	
<i>Inquiry</i>	852	39%	<div style="width: 39%;"></div>	
HSE	811	37%	<div style="width: 37%;"></div>	
ICL	562	26%	<div style="width: 26%;"></div>	
Word-Like	Count	Relevance		
pipework	2170	100%	<div style="width: 100%;"></div>	
tank	1534	71%	<div style="width: 71%;"></div>	
gas	1246	57%	<div style="width: 57%;"></div>	
pipe	1165	54%	<div style="width: 54%;"></div>	
safety	1108	51%	<div style="width: 51%;"></div>	
time	1052	48%	<div style="width: 48%;"></div>	
paragraph	1047	48%	<div style="width: 48%;"></div>	
assessment	1001	46%	<div style="width: 46%;"></div>	
risk	1001	46%	<div style="width: 46%;"></div>	
building	970	45%	<div style="width: 45%;"></div>	
pressure	935	43%	<div style="width: 43%;"></div>	
test	926	43%	<div style="width: 43%;"></div>	
evidence	874	40%	<div style="width: 40%;"></div>	
page	839	39%	<div style="width: 39%;"></div>	
take	735	34%	<div style="width: 34%;"></div>	
letter	719	33%	<div style="width: 33%;"></div>	
statement	716	33%	<div style="width: 33%;"></div>	
premises	683	31%	<div style="width: 31%;"></div>	
look	642	30%	<div style="width: 30%;"></div>	
document	627	29%	<div style="width: 29%;"></div>	
underground	625	29%	<div style="width: 29%;"></div>	
system	593	27%	<div style="width: 27%;"></div>	
valve	582	27%	<div style="width: 27%;"></div>	

The brightness of a concept's label reflects its frequency in the text. The brighter the concept label, the more often the concept is coded in the text.



You can reveal hidden concepts by moving the % Visible Concepts slider (underneath the map) to the right. To reveal the most important concepts in order, move the slider far left then slowly drag the pointer to the right.

Buttons in the header above the Concept Map



Hover your mouse over a button above the map to see its name.

Starting from the left, they include:

The Concept Cloud button reveals a graphical alternative to the concept map. The Concept Cloud will be explained in a separate section to follow.

The Switch to Social Network (Gaussian) button causes the concept to be reclustered using a Gaussian algorithm. Clicking the same button again returns you to the original view, created using a topical (Linear) clustering algorithm.

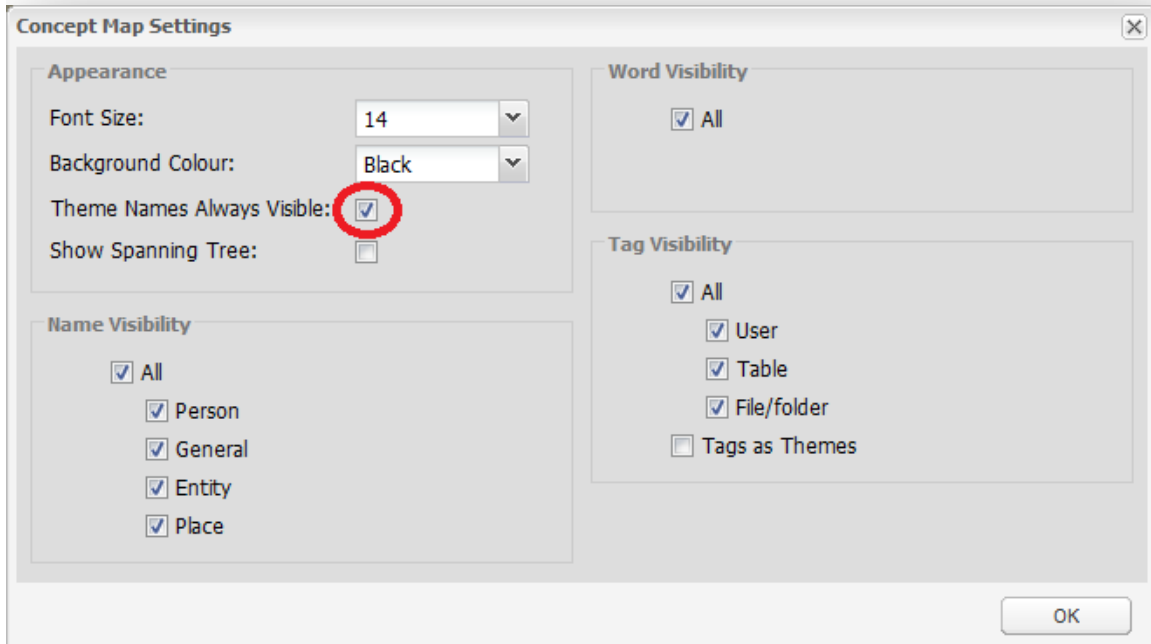
The Center Map button simply centers the map image in the screen space.

The Rest Map to Original View button returns the map to the way it looked when it first opened.

The Recluster Map button scatters the concepts randomly in the map space initially, then uses a clustering algorithm to allow the concepts to attract one another once more so as to lay the map out on screen.

The Cluster Map button allows the concepts more iterations of attracting one another to settle in stable locations on the map (without randomising them first).

The Map Settings button allows you to change various visual aspects of the concept map. Clicking the crossed spanner and screwdriver button above the map opens this interface:

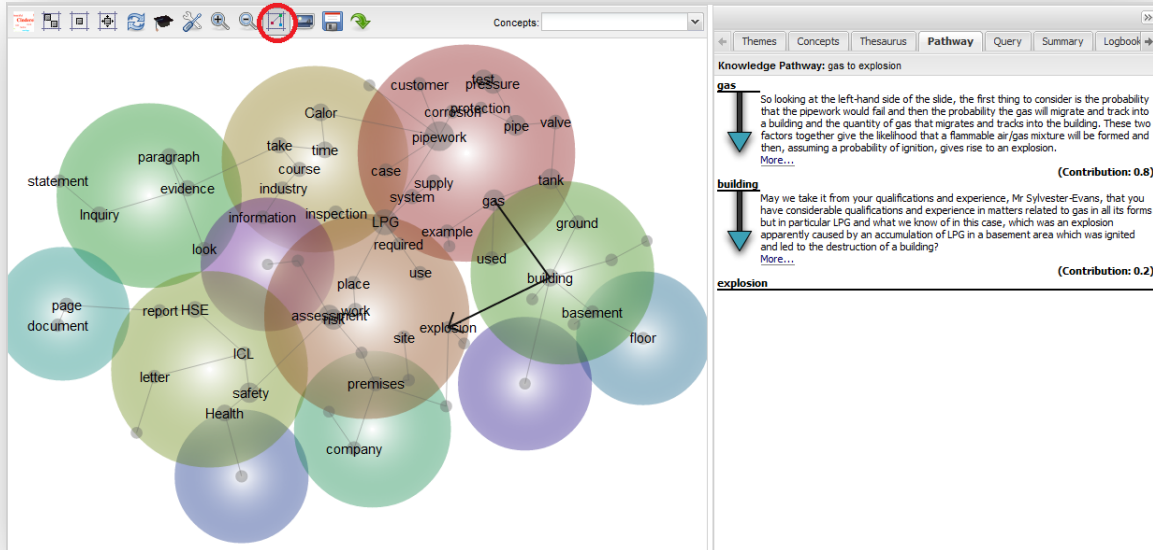


The Map Settings interface allows you to increase the Font Size of labels on the map, and to change the background colour from white to black.

It also allows you to make the theme names always visible on the map by ticking the Themes Names Always Visible button.

You can choose whether to show the spanning tree on the map. The spanning tree appears as a grey network of connections between concepts (like a spider web) beneath the concept network. It shows the most-likely connections between concepts (like a road map of highways), but there are other (less-strong) connections between concepts (like backstreets).

click on a concept on the map, then select another, Leximancer will indicate the strongest (lowest-cost) pathway between those two concepts on the map:



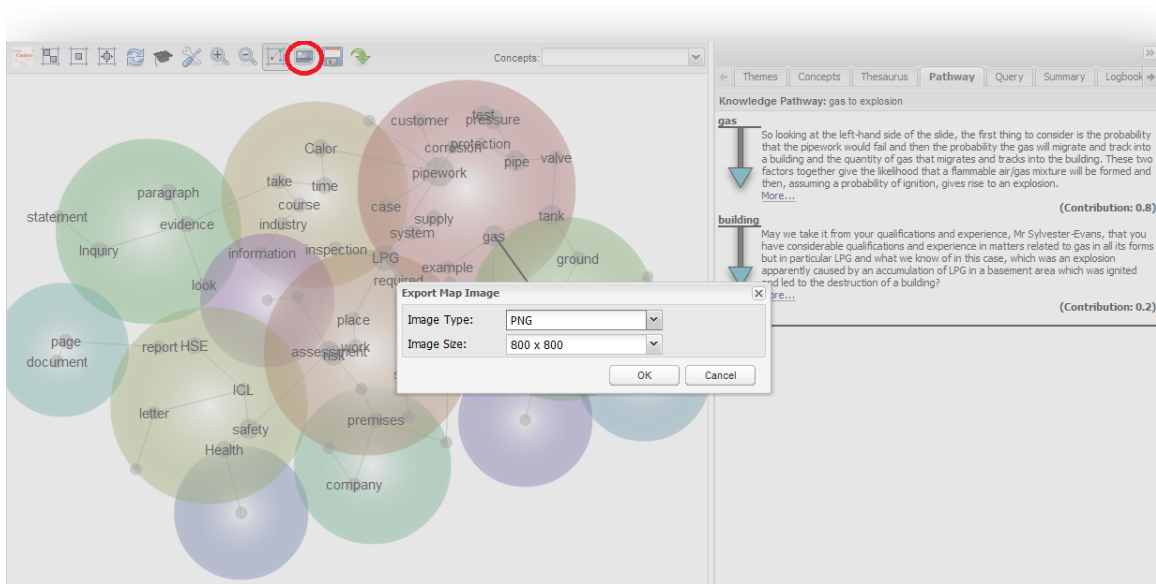
A black line appears on the map to indicate which other concepts might be bypassed in order to move from the gas to the explosion concepts in this example. The tab on the right shows the probability of each leg of the path, and presents an excerpt of text linking the two concepts involved in each leg.

The pathways are intended to tell stories emerging from the text, and focus on indirect connections between concepts on the map.

Toggling the pathways mode off allows you to click on more than one concept on the map without a pathways being drawn between them.

Clicking the same button again returns you to the pathways mode.

The Export Map button allows you to take a picture of the concept map:



You can choose the file type and resolution of the image.

Note that pop-ups must be enabled in the browser so that the map image can appear in a new window. The image can then be saved to local disk.

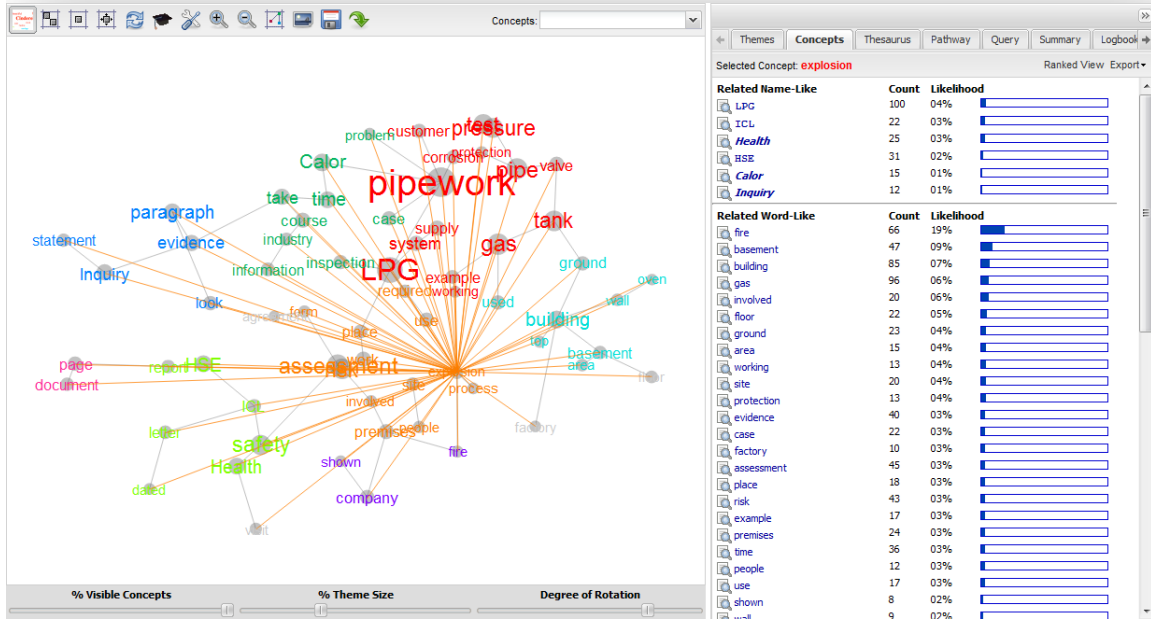
The Save Map button lets you save the current map configuration (in case you have changed theme names etc.) so that you can return to it later. Simply enter a name for the map image, and click OK.

The Load Map button lets you reload a map configuration saved previously using the Save Map button. Simply select the name of the map image you require and click Load.

The Concepts Dropdown List on the right above the map is an alphabetised concept finder. It allows you to locate (and select) a concept of interest in the map space. This is useful where you know the name of a concept of interest, but cannot see it on the map easily.

(blue, green), denote the least relevant. The font size of each concept's label denotes its frequency in the text.

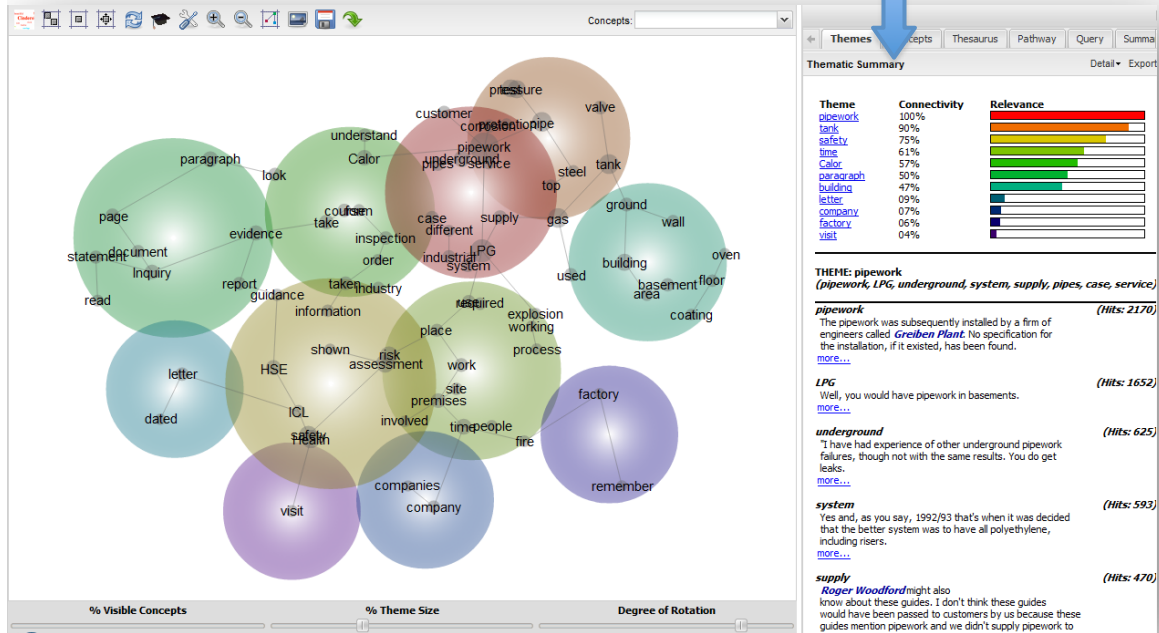
The Concept Cloud is fully interactive. It behaves like the concept map, in that you can click on a concept (or tag) to select it and see the list of related concepts in the right-hand tab:



Report Tabs

The right-hand window contains several report tabs that allow for further interaction with the Concept Map. Each tab represents a different way to interact with the map and explore the results.

Themes tab

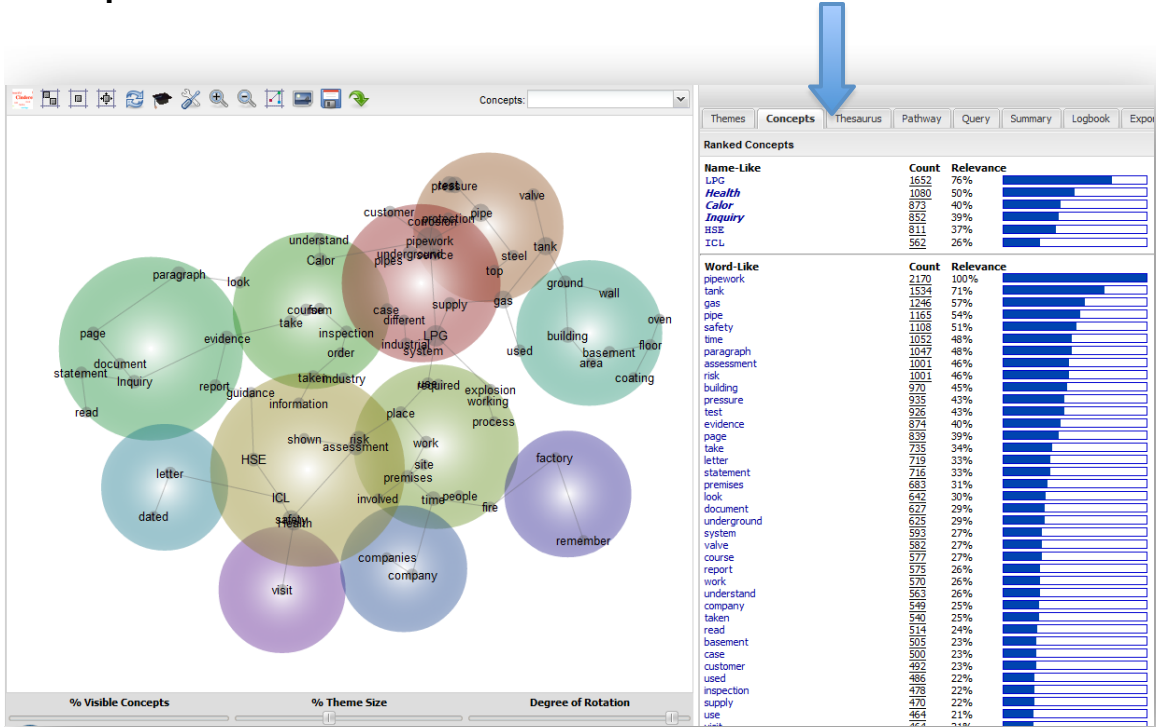


As mentioned previously, themes are the coloured circles that group clusters of concepts. The themes are heat-mapped, meaning that hot colours (red, orange) denote the most important themes, and cool colours (blue, green), denote those less important.

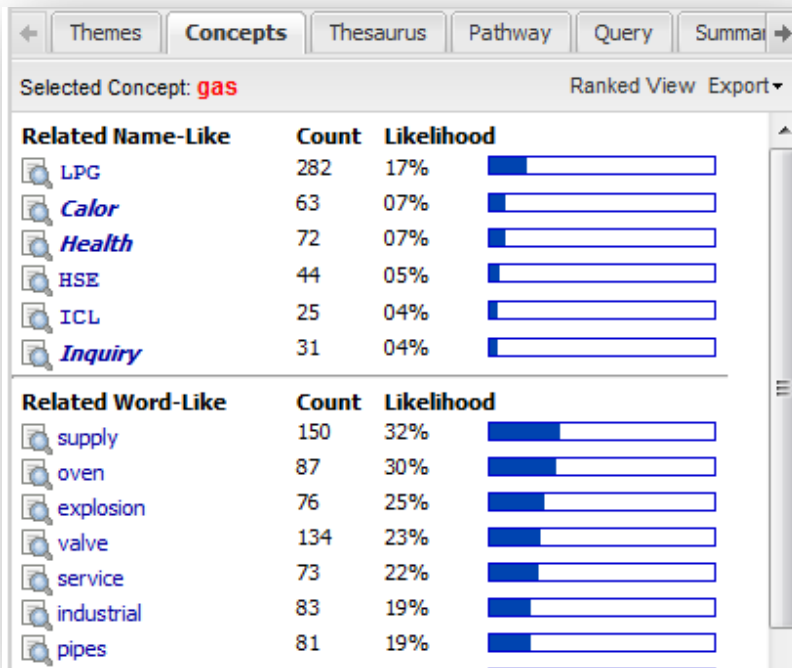
The Thematic Summary includes a 'connectivity' score to indicate the relative importance of the themes. Below this the name of the most important theme, followed by a list of the concepts contained within that theme. Examples of text containing each of the constituent concepts in the theme are then listed beneath. Hover your mouse over the 'more'

button to see the syntax of the query run to return that piece of text. Also notice the 'Hits' score to the right showing how many text excerpts matches each query.

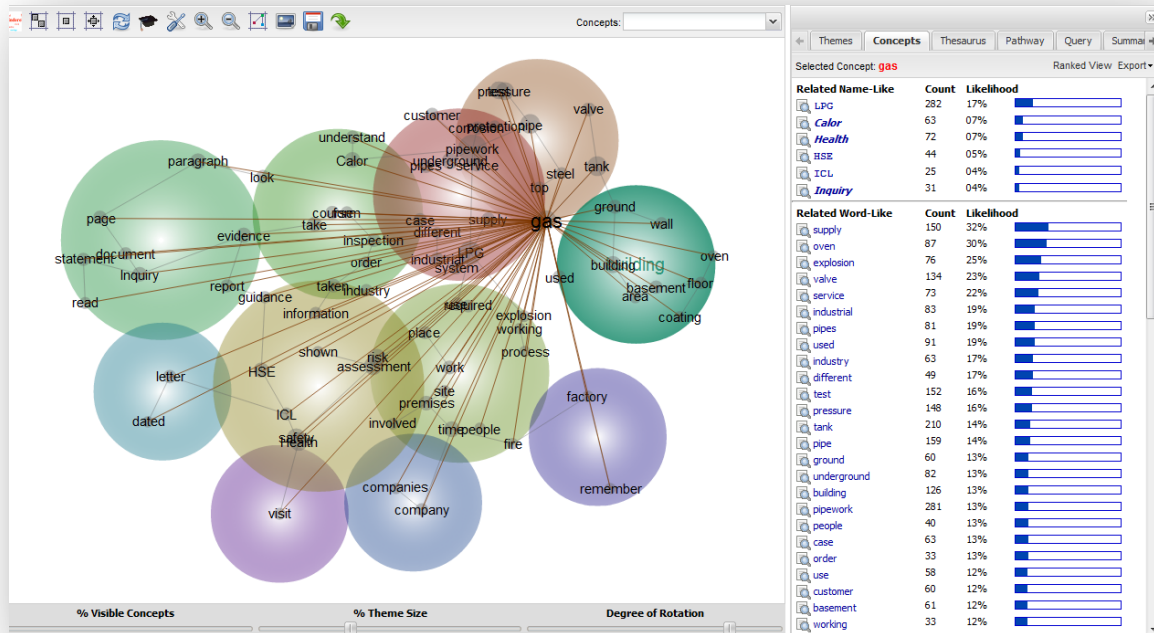
Concepts tab



Clicking the Concepts tab displays a list of name-like and word-like concepts, ranked by their frequency of occurrence in the text. Clicking on a concept in this list reveals its connections with other concepts:



This information may also be accessed and displayed visually by clicking on a concept on the map itself. The brightness of the ray indicates the strength of relationship (co-occurrence) between the concepts. A ranked list of the related name- and word-like concepts is displayed automatically in the Concepts tab on the right:



From the related concepts lists, you can browse the locations in the document where concepts co-occur by clicking in the Browse button (the magnifying glass icon).

Related Name-Like	Count	Likelihood
LPG	282	17%
Calor	63	07%
Health	72	07%
HSE	44	05%
ICL	25	04%
Inquiry	31	04%

Related Word-Like	Count	Likelihood
supply	150	32%
oven	87	30%
explosion	76	25%
valve	134	23%
service	73	22%
industrial	83	19%
pipes	81	19%

Browsing extracts automatically takes you to the Query tab in the right-hand panel, and displays instances where the concepts of interest co-occur in the text.

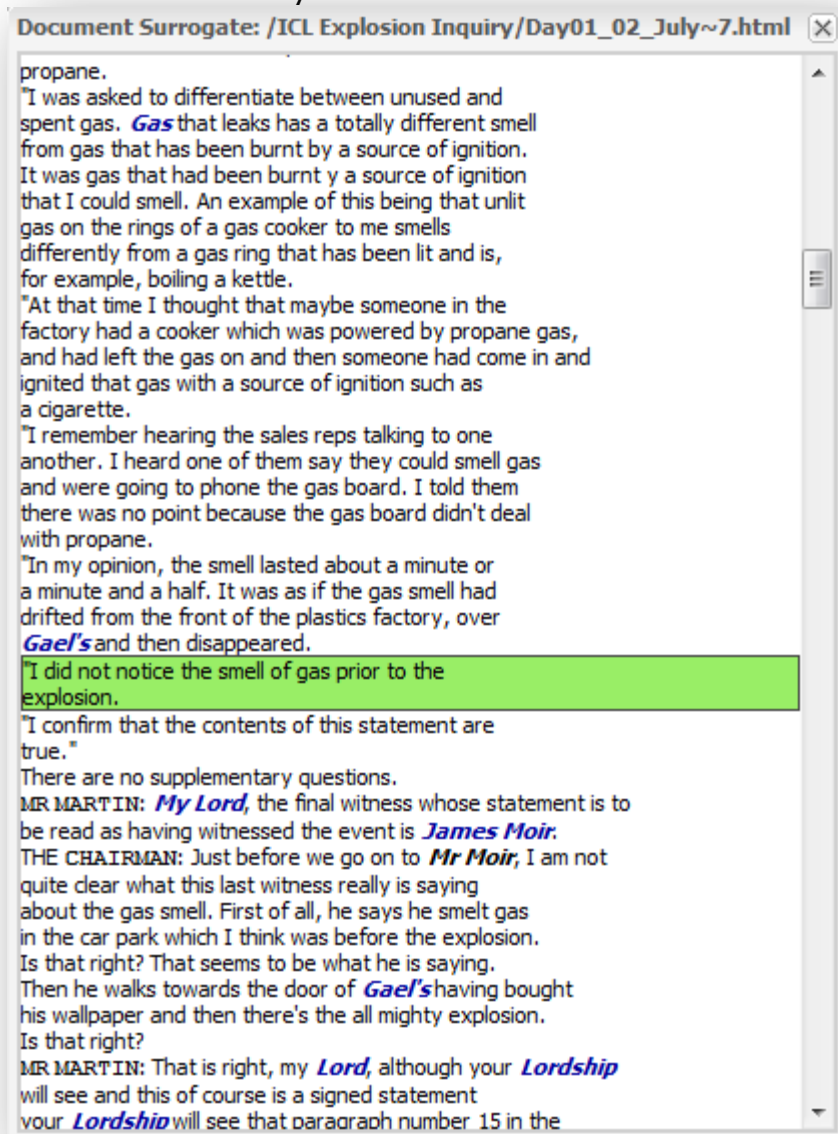
WORD:gas AND WORD:explosion Search

Export Page Export All Log All

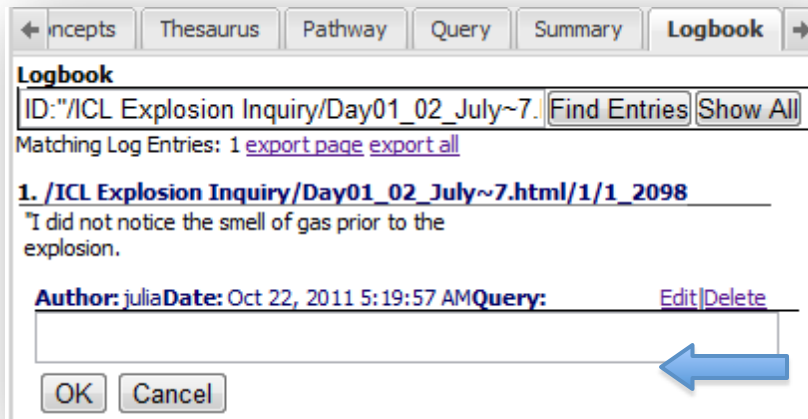
Result

- [/ICL Explosion Inquiry/Day01_02_July~7.html 1_2098](#) Add to Log | Full Text
"I did not notice the smell of gas prior to the explosion.
- [/ICL Explosion Inquiry/Day01_02_July~7.html 1_2102](#) Add to Log | Full Text
THE CHAIRMAN: Just before we go on to *Mr Moir*, I am not quite clear what this last witness really is saying about the gas smell. First of all, he says he smelt gas in the car park which I think was before the explosion.
- [/ICL Explosion Inquiry/Day01_02_July~7.html 1_2145](#) Add to Log | Full Text
"While waiting on the police to inform us of the escape route, I smelt a very strong smell of *Calor* or propane gas. I was worried about the smell as a spark could cause another explosion.
- [/ICL Explosion Inquiry/Day02_03_July~2.html 1_360](#) Add to Log | Full Text
So it was an incident involving, what, an explosion of LPG gas?
- [/ICL Explosion Inquiry/Day08_15_July~7.html 1_2013](#) Add to Log | Full Text
I did not notice any gas smells on the Monday before the tragedy or on the day of the tragedy. I find it strange that on the date of the tragedy, that immediately before the explosion, I never smelt a gas leak.

Click on Full Text to read any extract in context:

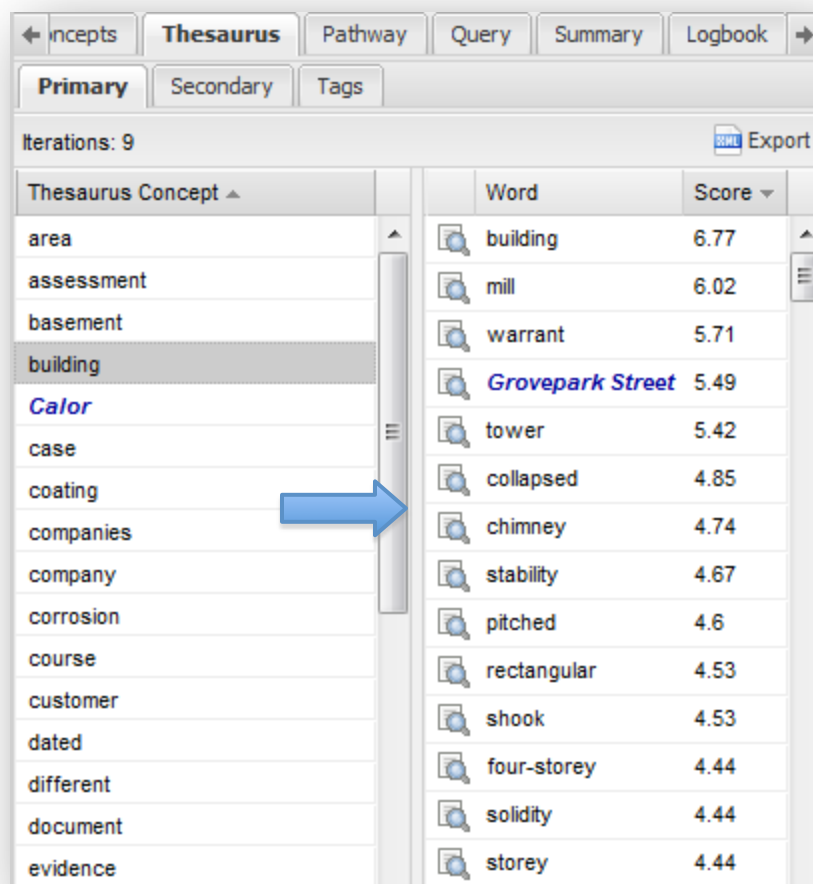


From the Query Results, click on Add to Log to add an extract of interest to the LogBook for export or reporting. When an extract is logged, the Add to Log button changes to a View Log option. Click View Log to review the list of extracts in the LogBook, then select Edit to add your own notes about an excerpt:



Thesaurus

The Thesaurus tab displays a list of your concepts, the number of iterations performed by Leximancer on that concept, and a ranked list of the thesaurus words that define and describe each concept.



Click on a concept in the alphabetical list on the left to reveal the list of Leximancer found to be associated with that concept on the right. The thesaurus list also shows the relevancy weightings associated with each indicative word. The iterations count (top left) tells you the number of times the corpus was reread and coded with evolving concept definitions before a stable classification result was achieved.

You can click on the Evidence button (the magnifying lens icon) to the left of a thesaurus item to browse text excerpts where that thesaurus term appears as evidence for a concept of interest:

← Incepts Thesaurus Pathway **Query** Summary Logbook

WORD:building AND WTERM:collapsed Search

Export Page Export All Log All

Result

[/ICL Explosion Inquiry/Day01_02_July~6.html 1_1944](#) Add to Log | Full Text
By that time the building had totally collapsed.

[/ICL Explosion Inquiry/Day08_15_July~4.html 1_985](#) Add to Log | Full Text
"I did not know the building that collapsed on 11th May 2004. I had never visited the premises."

[/ICL Explosion Inquiry/Day01_02_July~3.html 1_670](#) Add to Log | Full Text
"**Tony** and I went outside but initially we could not see anything for the dust. Once the dust settled, I saw that the main building had collapsed.

[/ICL Explosion Inquiry/Day07_11_July~5.html 1_1370](#) Add to Log | Full Text
Anyway, that is where you came out of the **Stockline** building and to your right would be, perhaps not exactly as it is shown here but similar, the remains, such as they were, of the collapsed mill building?

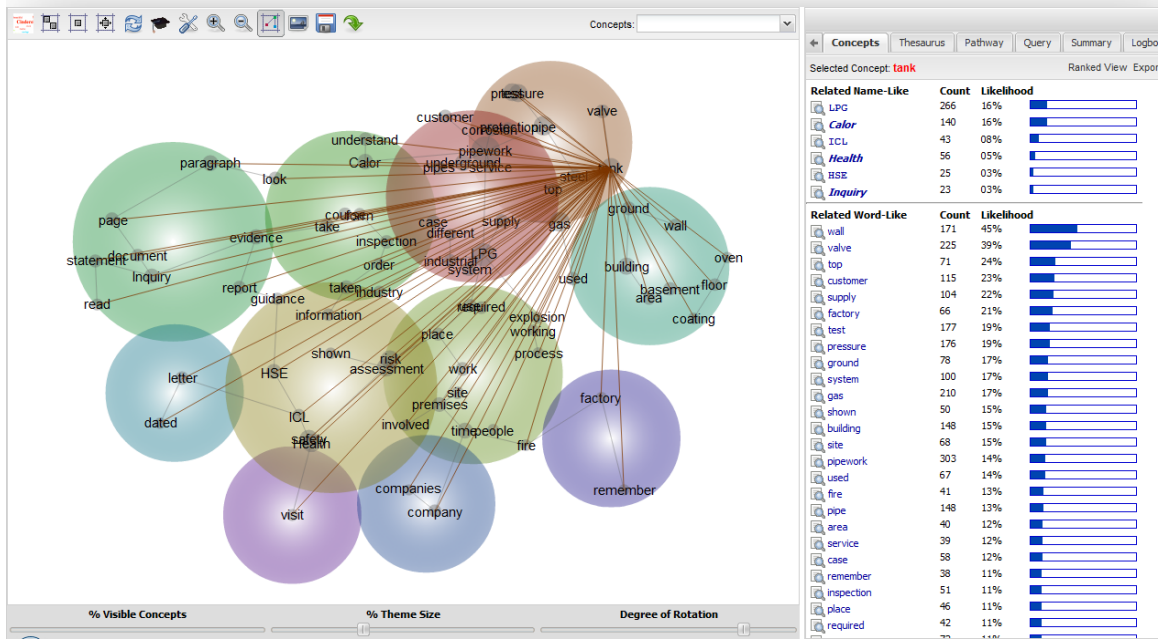
[/ICL Explosion Inquiry/Day01_02_July~7.html 1_2075](#) Add to Log | Full Text
"As I came out of **Gael's** door, I saw that that building, which I now know as a plastics factory, had collapsed. From where I was standing I could only see the end of the factory nearest to **Gael's** shop.

[/ICL Explosion Inquiry/Day21-13_November~3.html 1_430](#) Add to Log | Full Text
I should perhaps say that, if I understand the evidence correctly, there was no evidence before this **Inquiry** that it was surprising that this building collapsed given the force of the explosion.

[/ICL Explosion Inquiry/Day16-24_October~4.html 1_885](#) Add to Log | Full Text
The basic question I would like to ask you first, **Mr Neafe**, is we know as a result of the tragic events the building almost entirely collapsed and was destroyed as a building and clearly a very large quantity, as I recall, 5,500 tonnes of material, we have seen photographs.

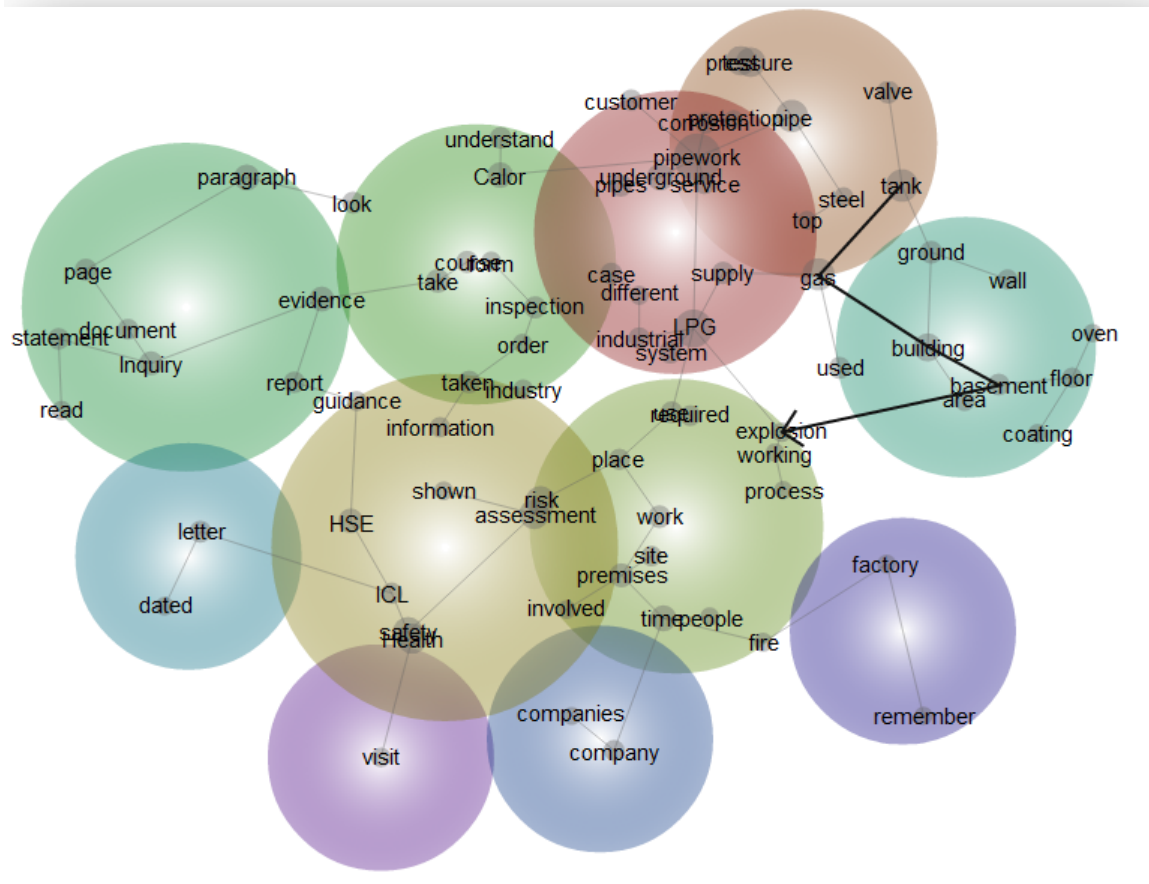
Pathway tab

The Pathway tab displays and describes the most likely relationship chain between two concepts. It allows you navigate the most likely path in conceptual space from a start concept to an end concept. To create a pathway you must first enable Pathway Mode using the relevant button above the map. Then select a concept on the map:



Now, clicking another concept will illustrate the pathway between them, along with example text. The relationships between concepts are best thought of as correlations, though the text segments describing the relationship may define a direction for cause.

The example below creates a pathway between the ‘tank’ concept and the ‘explosion’ concept:




When you identify a pathway on the map, a list of text segments describing the legs of the path appears automatically in the right-hand panel. This list is like a narrative of text segments that are relevant to the legs (sections between concepts) of the path.


The ‘start’ concept appears at the top of the list, and the associated text segment explains the link between this and the next related concept in the pathway. This pattern continues until the final text segment linked to the ‘end’ concept is listed.


The link between ‘tank’ and ‘explosion’ revealed the following connections:


← Concepts Thesaurus **Pathway** Query Summary Logbo →

Knowledge Pathway: tank to explosion

tank
 After the explosion I went over to the scene and I saw a big green and gold coloured gas tank. It had been the noise coming from the tank that had alerted me to it.
[More...](#) **(Contribution: 0.48)**

gas
 If an incident occurred at a bulk storage tank, in particular a leak, a greater distance between the tank and the potential ignition source gives greater opportunity for the gas to disburse and for any flammable concentration to diminish. Equally, if there is leakage at the vessel that caught fire the further away a building was from the tank the less risk that someone will be harmed by a flammable event at the tank or that the event could escalate.
[More...](#) **(Contribution: 0.28)**

building
 The results of the investigation and an explanation for the explosion and subsequent collapse will be given in the evidence of Dr Stuart Hawksworth from the HSL. Dr Hawksworth and his colleagues were able to establish that the metal chequer-plate floor of the Despatch Department had buckled upwards by force from below exerting a pressure of 7.06 tonnes per square metre.
[More...](#) **(Contribution: 0.18)**

basement
 The lower level of the original surface of the yard was revealed and was ascertained that the elbow which had failed that the adjacent pipework had been above the original surface of the level of the yard and that it had been buried when the level of the yard was raised. Although the pipework within the building had been broken and distorted in the explosion it was established that there was a continuous run of underground LPG pipework from the bulk storage tank in the yard into the basement beneath the Despatch Department.
[More...](#) **(Contribution: 0.05)**

explosion

Query

We have already discussed one of Leximancer's query functions: querying one concept against another:

← Concepts Thesaurus Pathway **Query** Summary Logbo →

WORD:pipework AND WORD:explosion Search

Export Page Export All Log All

But Leximancer allows for increased query specificity. Entering the following query terms in the query search bar:

NAME:[concept]

- searches for your specified name concept

WORD:[concept]

- searches for your specified regular word concept

TAG:[file, folder, or tag]

- searches for a pre-defined tag in your data

WTERM:[word]

- searches for a regular keyword (different to a concept)

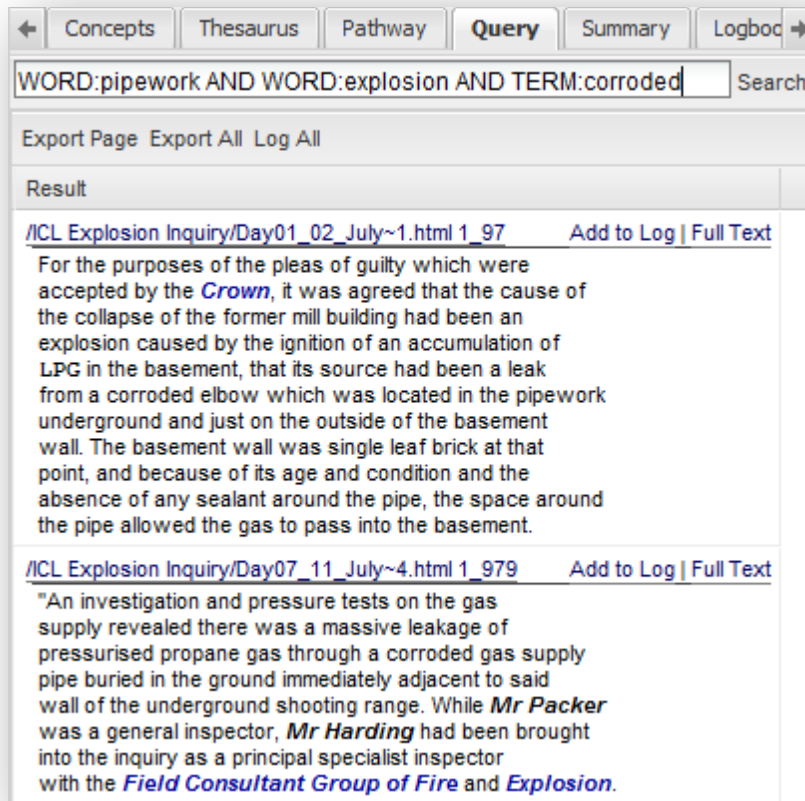
NTERM:[word]

- searches for a name keyword (different to a concept)

TERM:[word]

- equivalent to (WTERM:[word] OR NTERM:[word])

These terms can be used in conjunction to search for co-occurrences of any number of specified concepts, tags, and / or keywords. For example, you could search for excerpts that mention the keyword 'corroded' with the concepts 'pipework' and 'explosion' by entering this syntax into the Query box:



There are also other, more specialized, query searches available:

ITAG:[file, folder, or tag]

- searches for your specified ignored tag. For example, a spreadsheet column with more than 500 values is classified as an ignored tag. Although ignored by Leximancer, you are still able to use it in your query to narrow results.

SECTION:[section number]

- searches for the section number of the text block

There are also some additional new tools to assist with query searches.

These are query parsers that allow for more complex queries:

NAME:[partial concept]*

- adding a star to the end of a concept search will cause the query to be applied to the letters present and any derivatives. You must

provide at least one character for this search. **NAME:*** alone will not produce results.

- For example, a query of **NAME:environment*** will search for environment, environments, environmental, etc.

NAME:[concept] OR NAME:[concept]^2

- searches for either concept specified, but with the latter concept to be considered more important
- for example, a query of **NAME:greenhouse OR NAME:emissions^2** will search for both 'greenhouse' and 'emissions' concepts, but 'emissions' concepts will be considered more important.

+NAME:[concept] +WORD[concept]

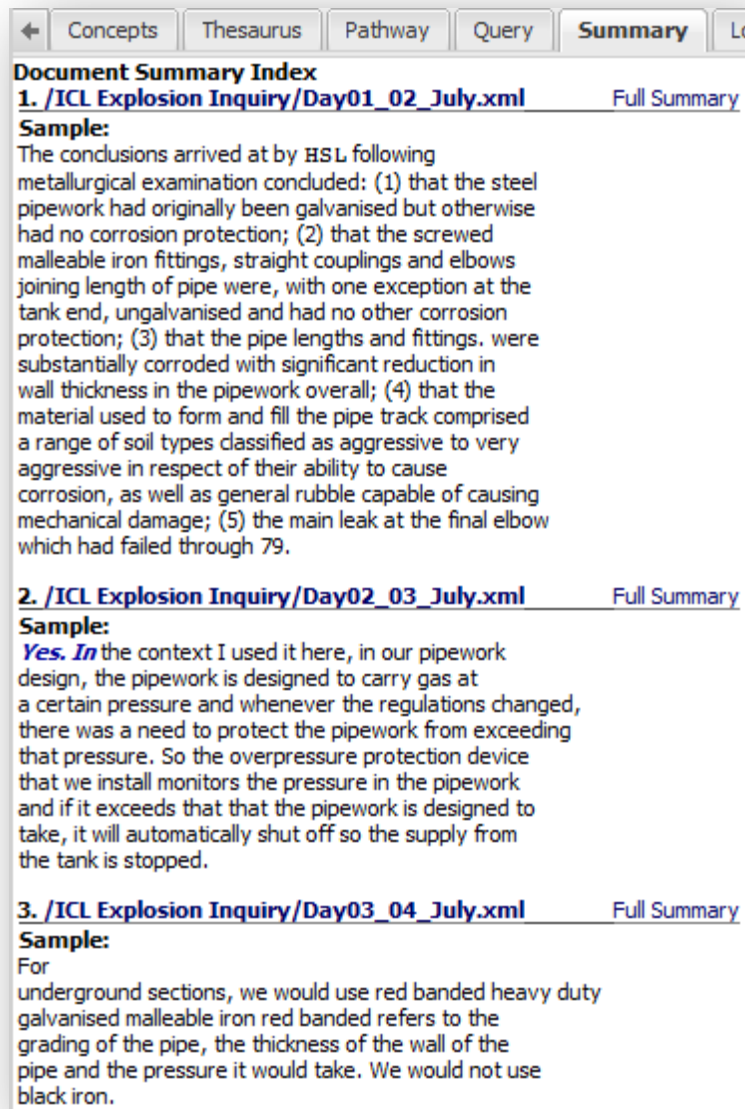
- this is shorthand code for the search **NAME:[concept] AND WORD:[concept]**

+NAME[concept] +WTERM:[concept]

- searches for the name concept co-occurring with a keyword

Summary tab

The Summary tab displays extracts containing the most important concepts discovered from the text. The list contains characteristic text segments that illustrate the relationships between key concepts:



The screenshot shows a software interface with a tabbed menu at the top containing 'Concepts', 'Thesaurus', 'Pathway', 'Query', 'Summary', and 'Lo'. The 'Summary' tab is active, displaying a 'Document Summary Index'. The index lists three entries, each with a 'Full Summary' link:

- 1. /ICL Explosion Inquiry/Day01_02_July.xml** Full Summary
Sample:
The conclusions arrived at by HSL following metallurgical examination concluded: (1) that the steel pipework had originally been galvanised but otherwise had no corrosion protection; (2) that the screwed malleable iron fittings, straight couplings and elbows joining length of pipe were, with one exception at the tank end, ungalvanised and had no other corrosion protection; (3) that the pipe lengths and fittings were substantially corroded with significant reduction in wall thickness in the pipework overall; (4) that the material used to form and fill the pipe track comprised a range of soil types classified as aggressive to very aggressive in respect of their ability to cause corrosion, as well as general rubble capable of causing mechanical damage; (5) the main leak at the final elbow which had failed through 79.
- 2. /ICL Explosion Inquiry/Day02_03_July.xml** Full Summary
Sample:
Yes. In the context I used it here, in our pipework design, the pipework is designed to carry gas at a certain pressure and whenever the regulations changed, there was a need to protect the pipework from exceeding that pressure. So the overpressure protection device that we install monitors the pressure in the pipework and if it exceeds that that the pipework is designed to take, it will automatically shut off so the supply from the tank is stopped.
- 3. /ICL Explosion Inquiry/Day03_04_July.xml** Full Summary
Sample:
For underground sections, we would use red banded heavy duty galvanised malleable iron red banded refers to the grading of the pipe, the thickness of the wall of the pipe and the pressure it would take. We would not use black iron.

When you have finished exploring the Concept Map project tabs, close the Map Explorer tab using the Close button (X) in the top right-hand corner.

Section 3: Creating an Automatic/Exploratory Map

The aim of this chapter is to give you an overview and introduction to using Leximancer. In this chapter, you will be performing an exploratory analysis of a data set comprised of online media about climate change.

Exploratory maps involve minimal input from the user and are the starting off point for analysis using Leximancer. They are a means of gaining an overview of your data before adjusting Leximancer settings for more tailored results. Manually adjusted maps will be explored in the next chapter.

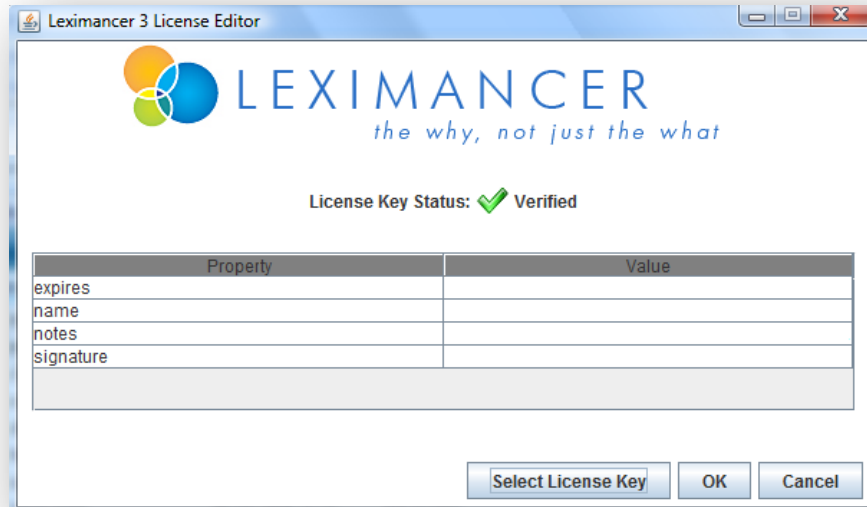
Creating an Automatic Concept Map

Supported File Types

Currently .doc, .docx, .pdf, .html, .xhtml, .htm, .txt, .xml, rtf, .tsv and .csv files are supported. If there are other formats of text data you wish to process, consider pre-converting them into plain text (.txt) first.

Desktop Installations

If you are working with a desktop version of the software, click on the desktop shortcut, or select Leximancer 4 from your Start menu. The first time you run Leximancer, you will be prompted to enter your licence key. If you have downloaded your licence key file, clicking Select Licence Key allows you to navigate to the location where it's saved. Click Open to load the key information, and then OK to start the program:



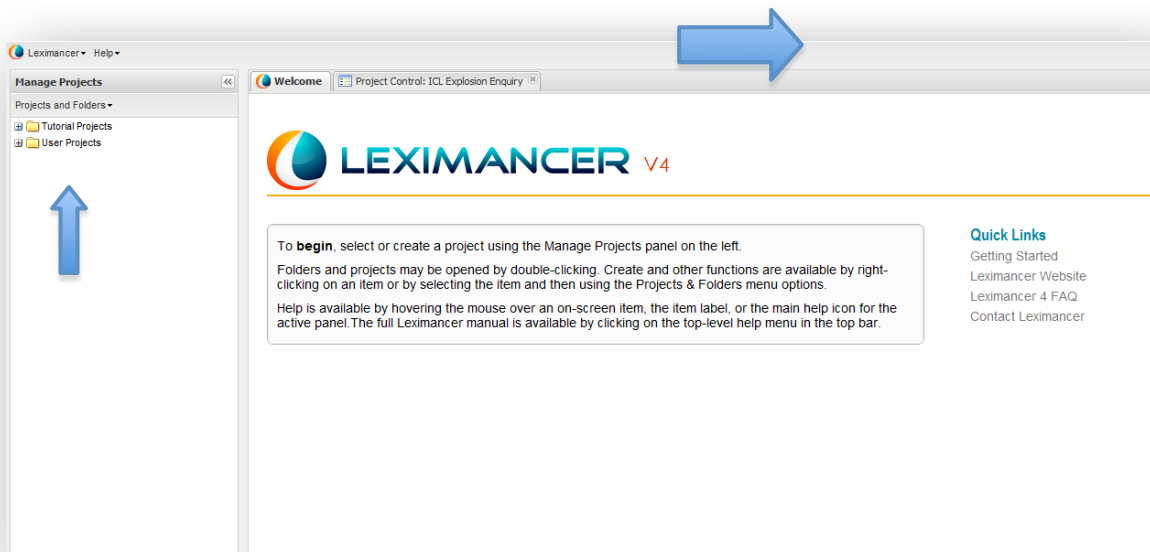
A Leximancer icon will appear in your system tray when the programme is running. To exit the application, right click on the icon and select Exit Leximancer.



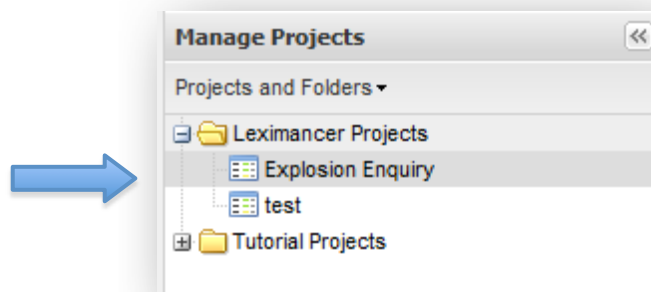
Leximancer Portal Accounts

If you are using the online Leximancer portal, open your internet browser (for example Mozilla Firefox or Internet Explorer) and navigate to the Leximancer 4 url (www.leximancer.com/lexiportal/). Login using your username and password to start a session.

When the program starts, you're presented with the Manage Projects interface:



Two top-level folders are visible: Tutorial Projects (a folder containing generic example projects that come pre-loaded with the software), and Leximancer / User Projects (a folder to house your projects). Click on the plus sign adjacent to a folder to expand the tree and view any folders or files within. In this case, there is a project called 'Explosion Enquiry' inside the Leximancer Projects folder:

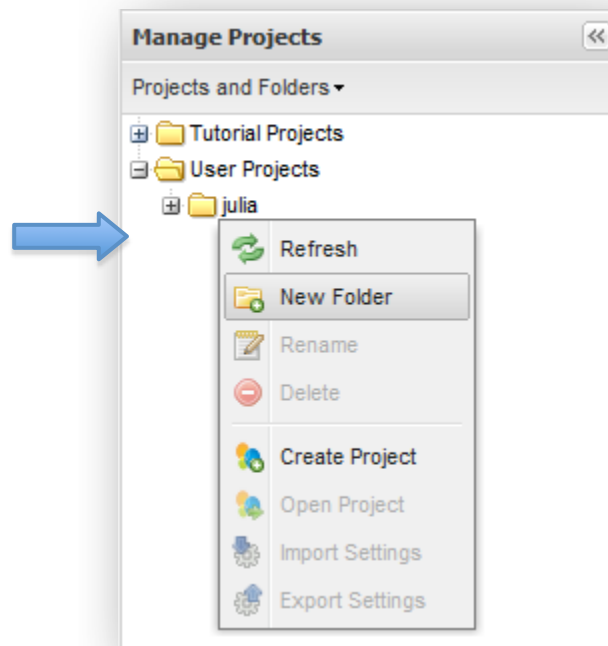


- Double clicking on a project's name opens that project.
- Right clicking on a project allows you to Rename or Delete an existing project.
- Right clicking on Leximancer / User Projects, then selecting Create Project allows you to create your own new project.
- Create a hierarchy to organise your projects by housing New Projects in New Folders.

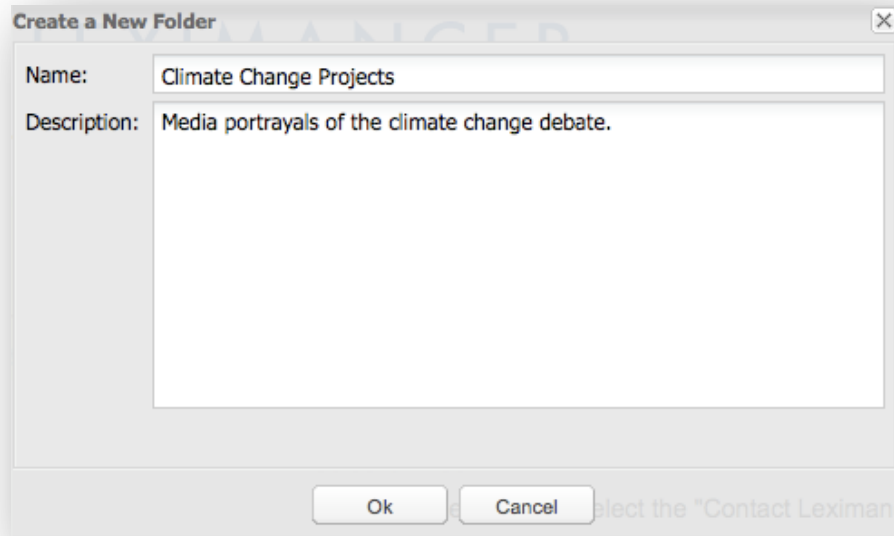
Creating a New Folder and Project

New Folder

Right click on Leximancer Projects (Desktop) folder, or open the User Projects folder and then the folder by your name (Portal), and select the New Folder option. You can then name (and optionally describe) the New Folder.



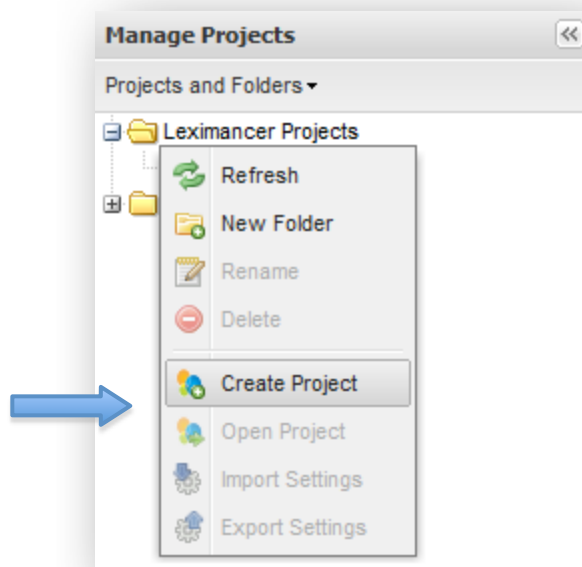
As a project file or folder is created using this name, try to use typical conventions for naming files (for instance, avoid including characters such as ".", "\", "*", or "/").



After you have named the folder, click 'OK'. This creates an empty project folder by this name under Leximancer / User Projects.

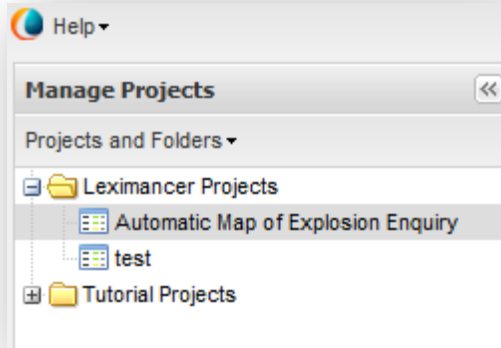
New Project

Right click on your new folder under Leximancer / User Projects, and select 'Create Project' to name (and optionally describe) a new project.

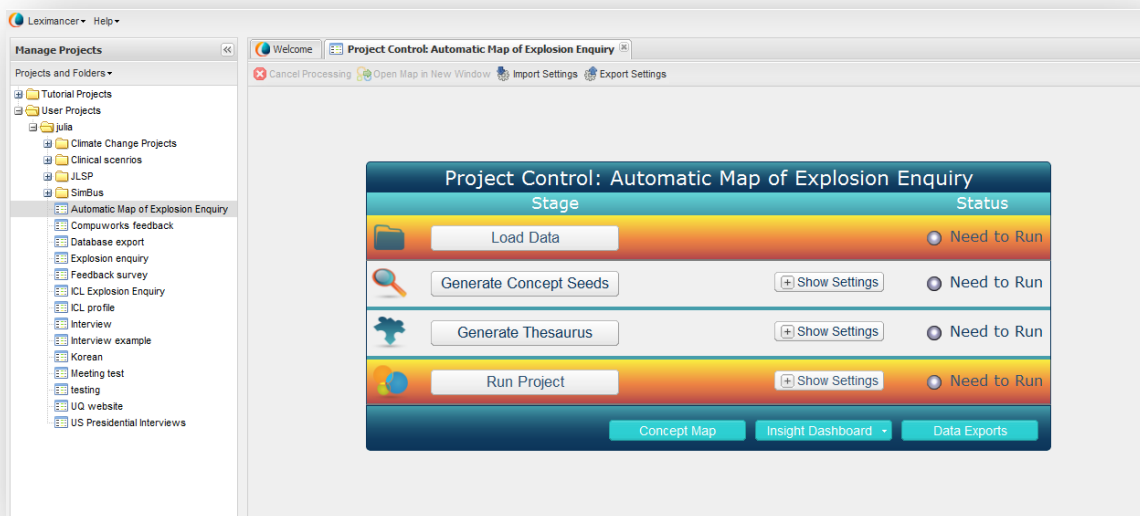


It helps to be descriptive when naming your projects, so that later, when you have multiple projects, you don't need to open them to check their

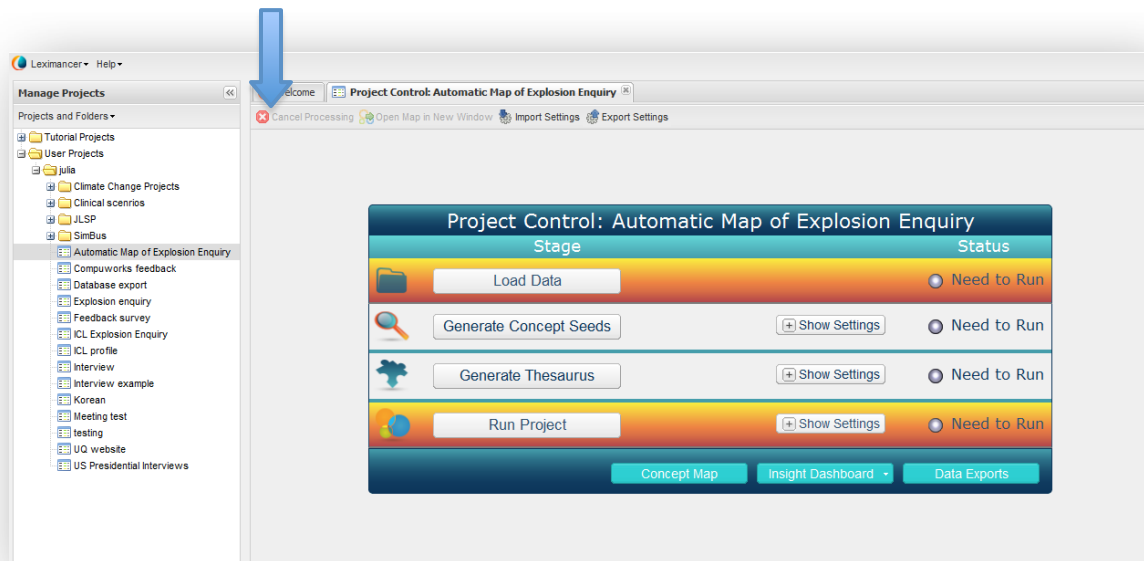
content. As we are creating an Automatic Map, I will name the project to reflect that:



This opens the newly-created project, and displays the main user interface in the right hand panel:



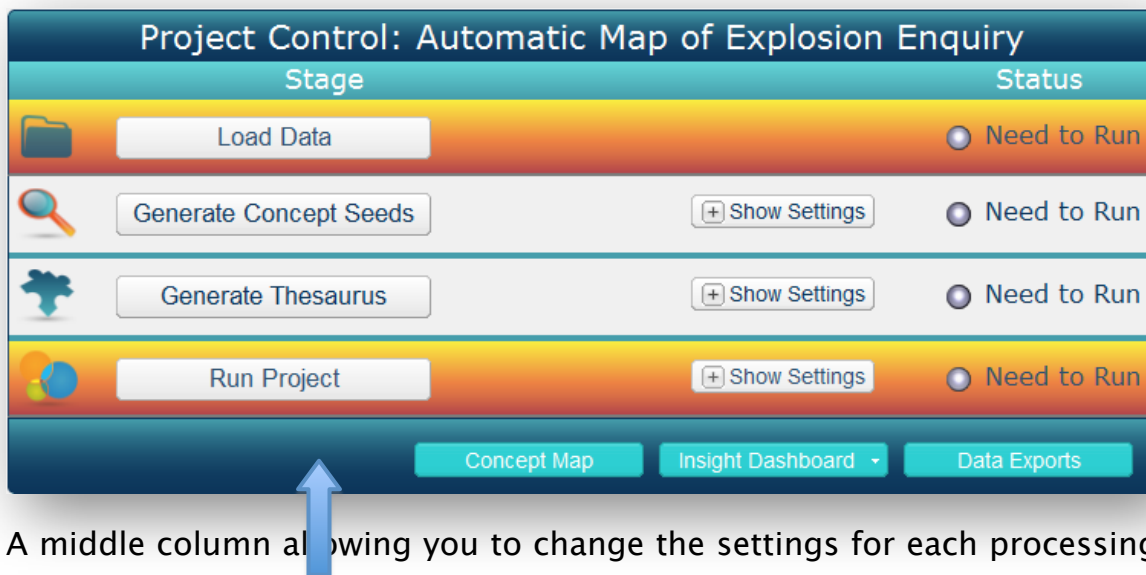
You can open several Leximancer projects at once. When you've opened one or more projects, you can also collapse the Project Selection interface using the arrows:



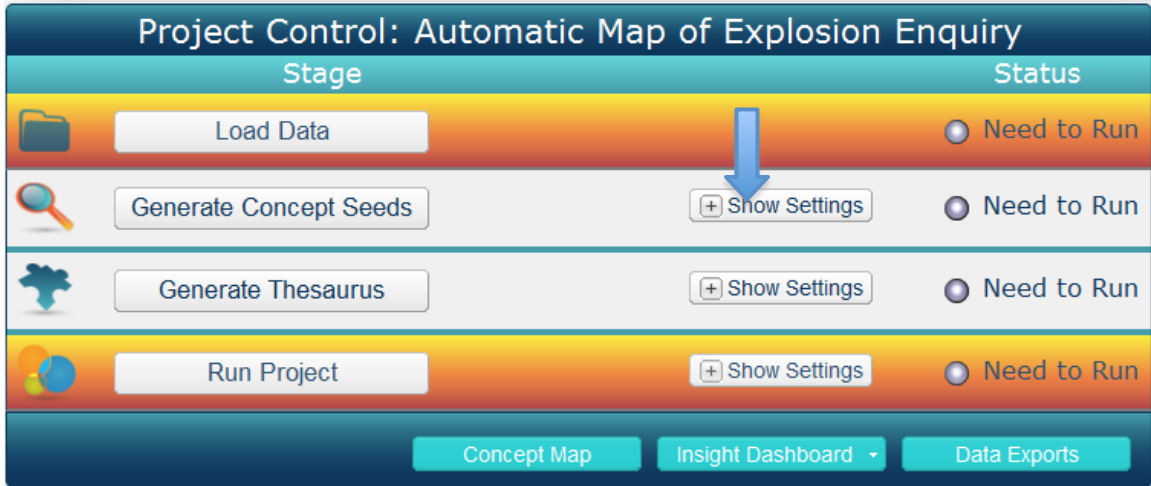
Having created a project, you are now ready to use the Main Leximancer User Interface.

The Main Leximancer User Interface:

The user interface is a flow diagram showing Leximancer's phases of processing to generate a conceptual map. The Project Control Panel consists of a left hand column that gives the name of the processing stage:



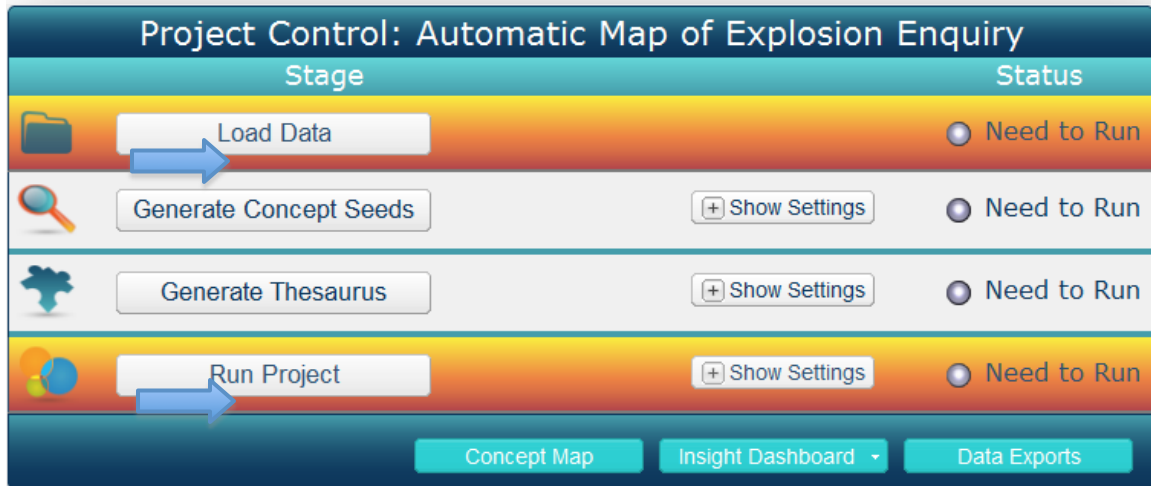
A middle column allowing you to change the settings for each processing stage:



And a status column that keeps track of the completion of each stage of processing:



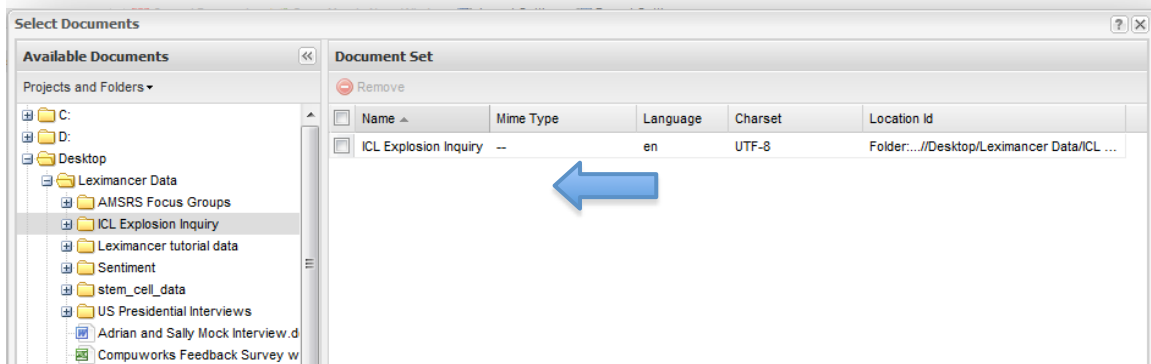
For an Automatic Map, we only need to complete the **Load Data** step, and then click **Run Project**. Both of these buttons appear in orange in the main user interface:



When you click Load Data,

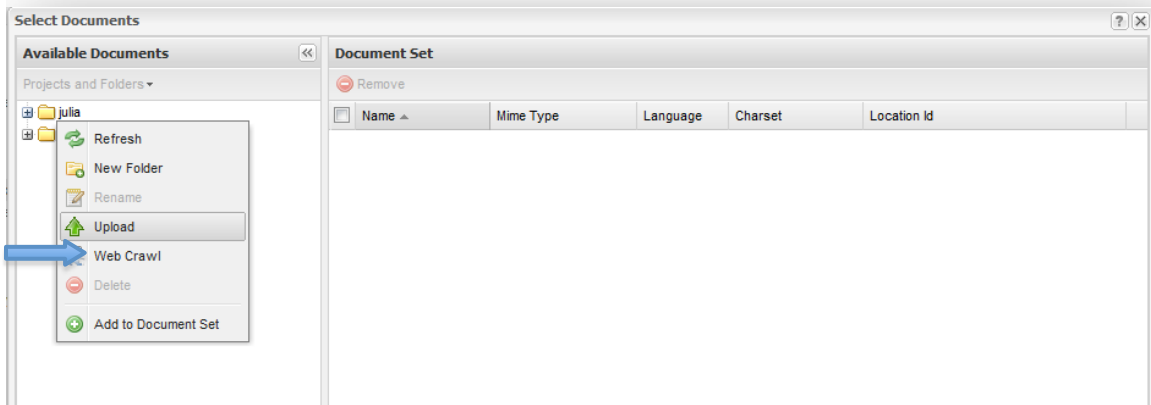
If you are using Leximancer **Desktop**:

- You will see your local drives listed in the Available Documents panel on the left.
- Expand the drive directories to see the files and folders within them.
- Drag and drop the desired files and folders into the Document Set area on the right to link them to the project, then click OK.



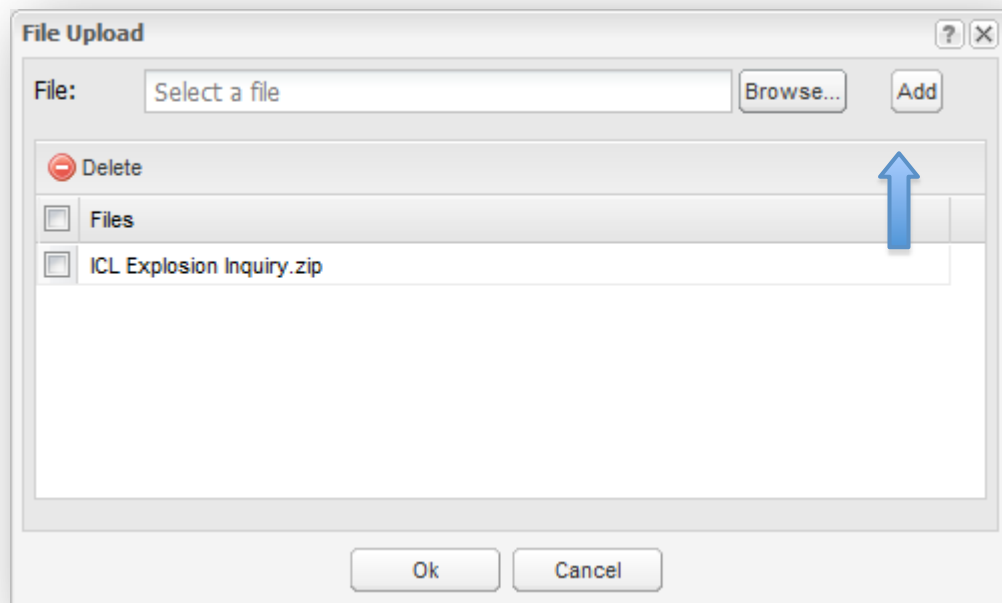
If you are using Leximancer **Server or Portal**:

- Right click on your user data folder (usually your name or company) in the Available Documents panel on the left, and select Upload.



- Browse to locate the data on your local machine. If you wish to upload multiple files at once, place them in a zipped folder before selecting them for uploading. The files and folders will be extracted from the zipped archive automatically on upload to Leximancer.

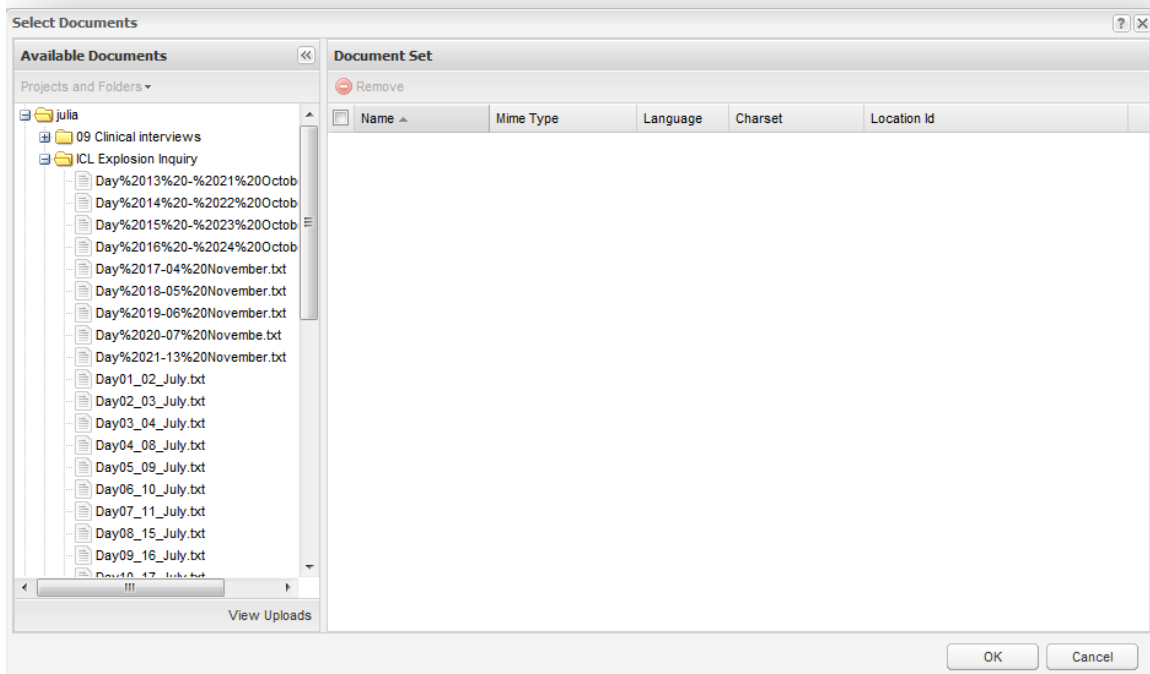
In the File Upload window, select the documents you wish to analyse and click Add. Once all the documents you want have been selected, click OK to begin uploading:



In this case, we have uploaded a single zipped folder containing a transcript for each day of a hearing after an explosion at a plastics factory.

There is some feedback in the lower left of screen to let you know that your files are uploading, and when it is complete.

Once the data is uploaded, you can expand the parent folder to see the files and folders within:



You could choose to analyse just one of the documents within a folder (e.g., a single day's transcript) by dragging and dropping an individual file into the Document Set area on the right. Alternatively you can drag the whole parent folder into the Document Set area on the right to analyse all its contents.

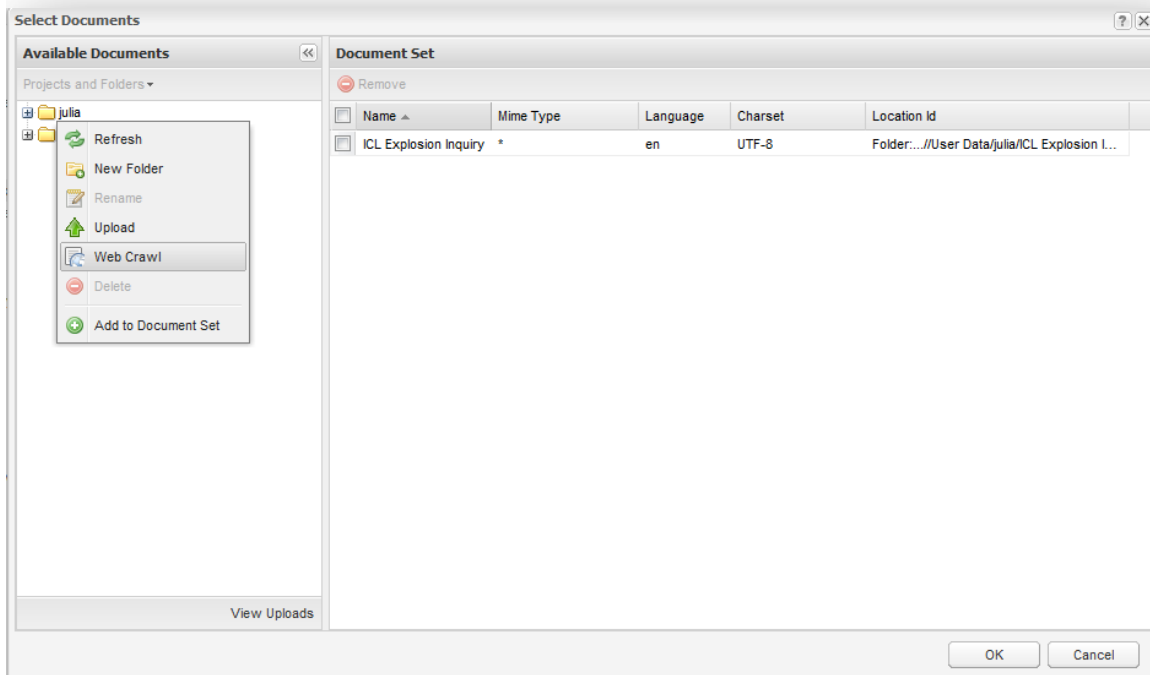
Selecting the parent folder instructs Leximancer to analyse all the files and folders within. This allows you to analyse multiple files and folders at

once, and facilitates comparisons between groups of documents using the software's automatic tagging options.

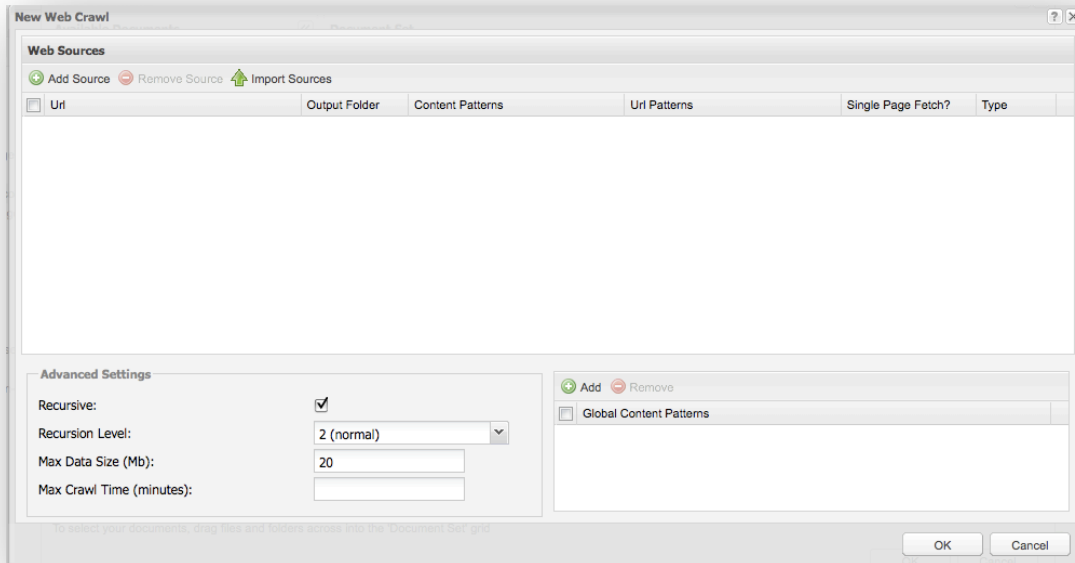
For Sever and Portal Users:

Using the Web Crawler

If you have arranged to use the web crawling facility, after clicking Load Data, right click on the folder by your name (or your company's name), and click web crawl:



The following interface will appear:

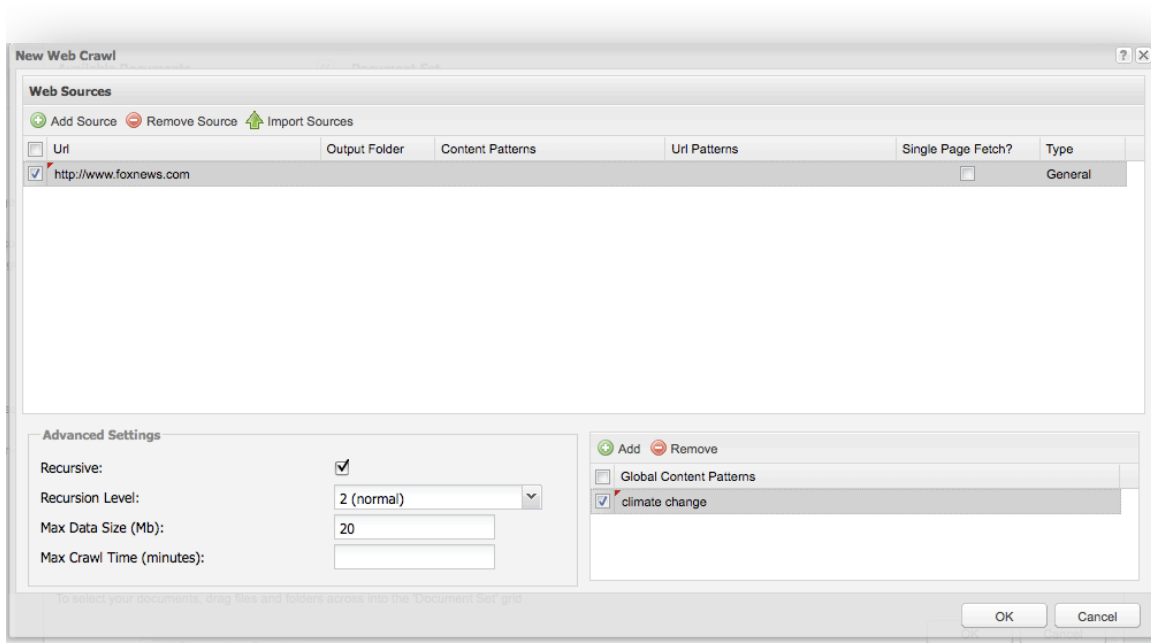


Provide a name for the web crawl, then press 'Add Sources' and type in the Url(s) you are interested in analysing. Blog and review sites tend to work best.

Tick the box to the right to do a Single Page Fetch if you only want the text on that page to be retrieved.

In the Advanced Settings you can set the Level of Recursion. This defaults to 2, allowing the crawling to follow links from the page to 2 levels down. You can also increase the maximum data size if you want to crawl a large amount of textual data.

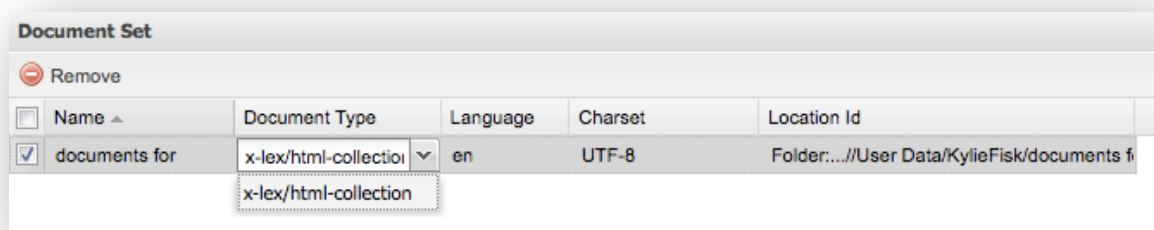
Enter keywords into the Specific or Global Content Pattern boxes to focus the crawl on pages that mention particular words.



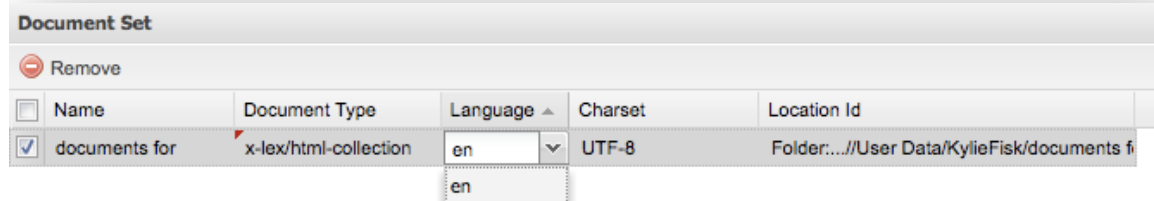
The web crawl will begin when you click Ok. When it is complete, you can drag and drop the folder of crawled text into the Document Set area, and then run the project as usual.

For All users, once you have moved some files into the Document Set area:

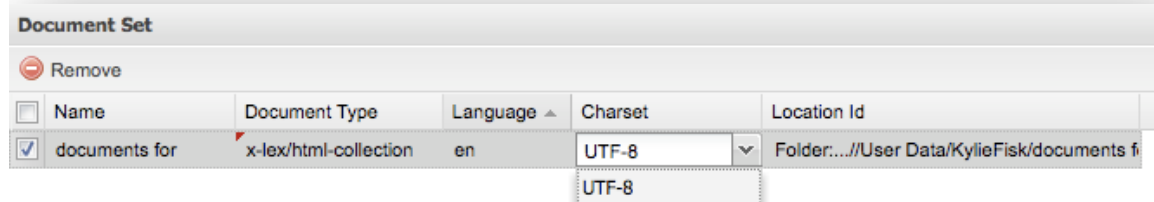
You have the option at this stage to specify the type of document(s) you wish to analyse. The current example uses online data, so .html was selected. Selection of data type helps Leximancer to be more sensitive to the idiosyncrasies of that data type (for example, html artifacts from online documents).



You may also specify the language of your documents:

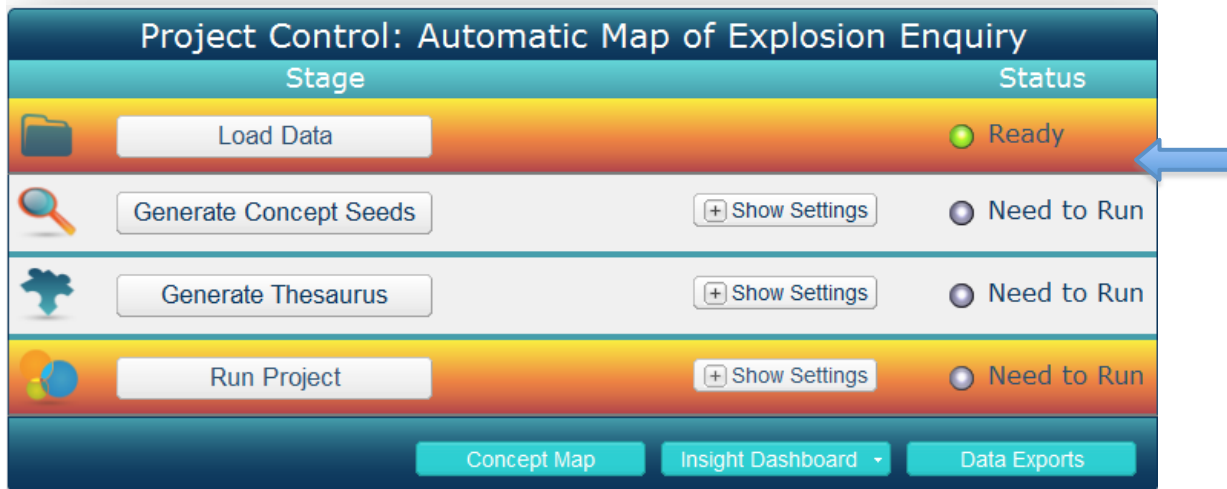


And the type of Character encoding used:

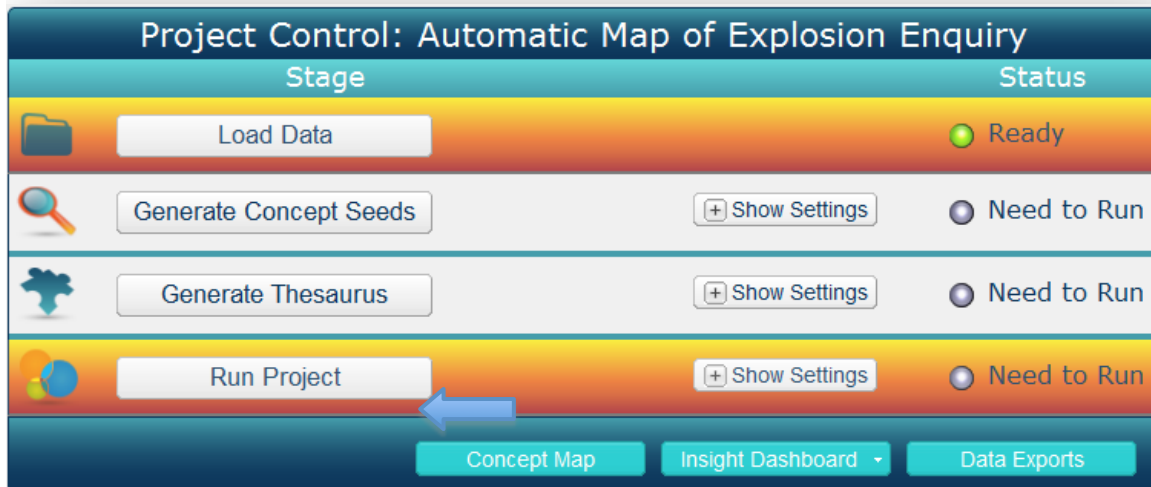


Finally, click OK to link the data to the current project.

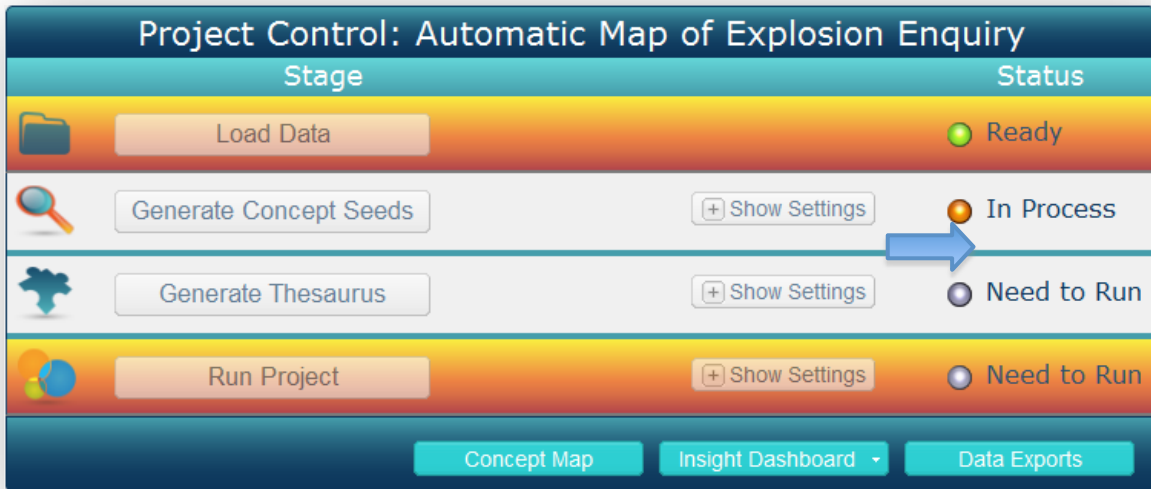
This returns you to the main user interface, where the Status of the Load Data stage will be green, as this step has now been completed:



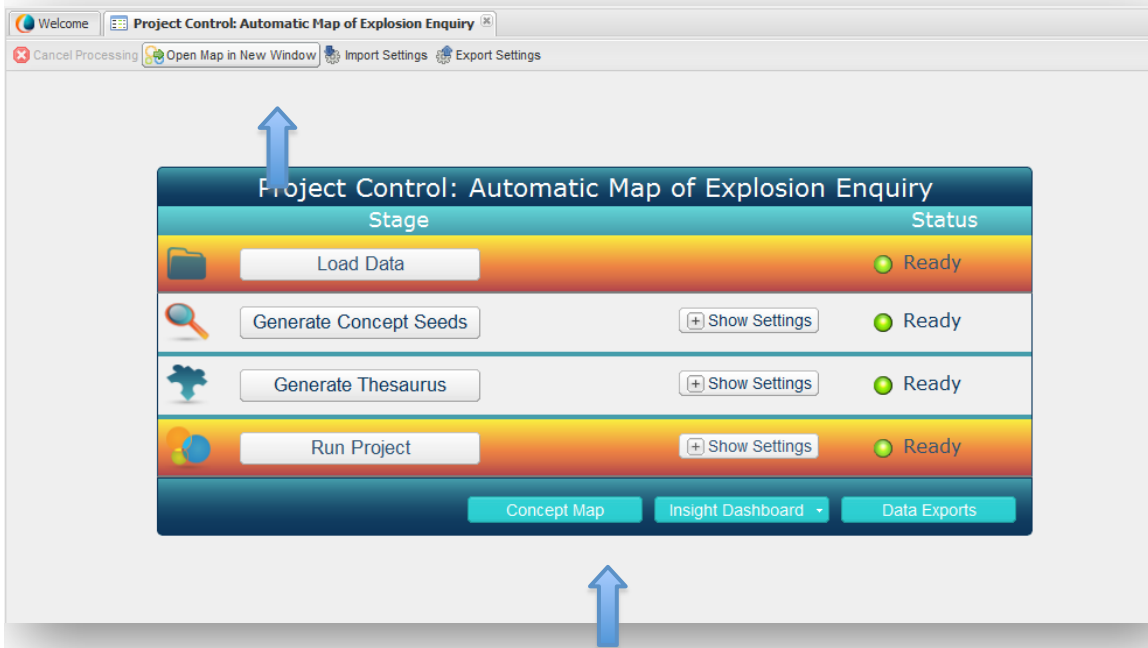
To perform an automatic or exploratory analysis, click the Run Project button in orange at the bottom of the Project Control Panel. This will run the project using default settings.



While the stages are running, they will flash orange and say 'In Process'. Once they are completed, the status nodes turn to green in colour and display the word Ready. Progress information is also visible at the bottom left of the screen:



When all phases of processing are complete, either click the Concept Map button at the bottom on the main Control Panel, or click Open Map in New Window (in the upper left of screen to open a larger map in a new window).



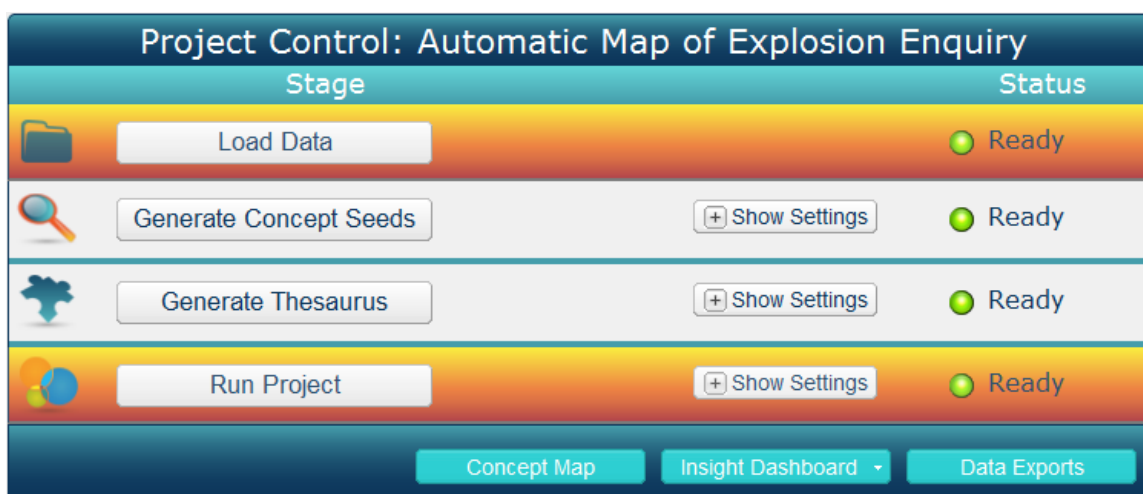
The Main User Interface also includes:

- A Close tab button (x) that allows you to close the current project (current settings are saved) to exit or load other projects;
- A Cancel Processing button that stops Leximancer processing, wherever it is up to at that time;
- An option to Import Settings or Export Settings, in which you may save the settings of your project, or apply settings from an earlier project;

Section 4: Creating a Manually Adjusted Map

The manual will now discuss the ways in which the user may alter the settings of their project to create a concept map. This section will follow the sequence of the stages as represented in the control panel (below).

However, the 'Load Data' stage will not be covered in this chapter. Please refer to the start of the previous chapter, Creating an Automatic/ Exploratory Map: For Beginner Users (page 38) for information on how to start Leximancer and load data.



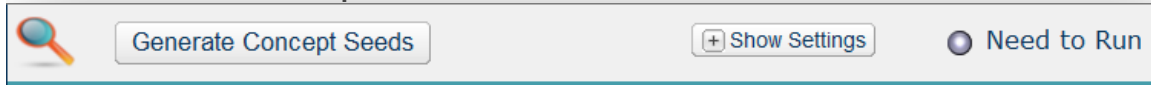
Stages of Processing

1. Load Data:



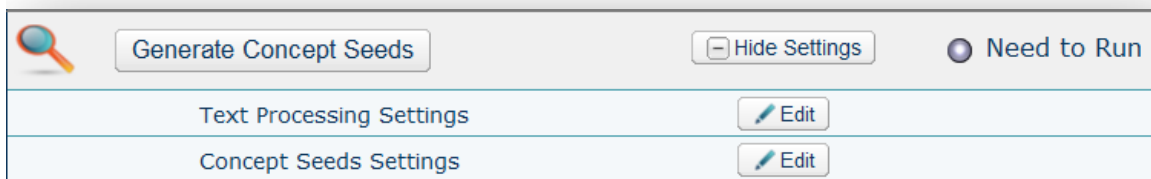
Please refer to the previous chapter (page 38) for information on how to load data in Leximancer.

2. Generate Concept Seeds:



You can run this stage of processing using default settings by clicking the **Generate Concept Seeds** button.

Alternatively this stage can be expanded (using the plus sign next to Show Settings) to reveal two sub-stages within:



There are options to Edit the Text Processing and Concept Seeds Settings.

2a. Text Processing

This is the first phase of processing that is run from Leximancer's main menu. This phase converts the raw documents into a useful format for processing. Preprocessing involves the following steps:

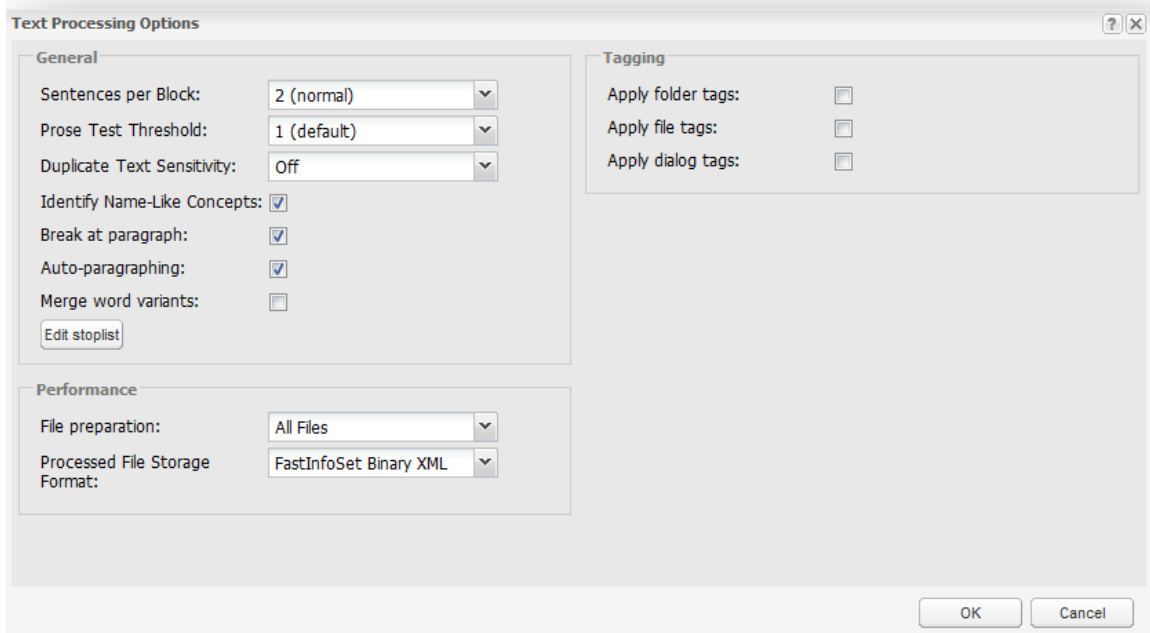
- **Splitting the information into sentences, paragraphs and documents.** These boundaries are important as they generally mark transitions in meaning. The conceptual map of the documents extracted by Leximancer reflects the co-occurrence of distinct concepts. To prevent concepts from being perceived to be related across changes in context (such as across different

documents), the co-occurrence is only measured within (and NOT across) blocks typically containing 2 sentences.

- **Removal of non-lexical and weak semantic information.** Within each sentence, the punctuation is removed along with a collection of frequently occurring words (called the stop-list) that hold weak semantic information (such as the words ‘and’ and ‘of’). Furthermore, for documents extracted from Internet email and news groups, the headers are cleaned up and the non-text attachments are removed.
- **Identifying proper names, including multi-word names.** Often in documents the proper names (such as people, places or company names) depict important entities that should be mapped. For this reason, proper names are extracted as potential concepts. In Leximancer, words are classified as proper names if they start with a capital letter. As every word that starts a sentence falls into this definition, only start-of-sentence words that are not in the predefined stop-list are considered as names.
- **Optional prose test of each sentence.** To remove non-textual material from the text, such as menus and forms in web pages, sentences that are unlikely to be part of the specified language are removed. This is achieved heuristically by removing sentences that contain less than 1 (or 2) of the stop-list words. If processing spoken language, this setting should be turned off.

Practical: Configuring Pre-processing in Leximancer

Clicking on the 'Edit' button for Text Processing reveals the following interface:



Manual Editing Options:

Sentences per Block (1–100): This option allows you to specify the number of sentences comprising each context block (or text segment). A context block is said to contain a concept if the words therein provide sufficient cumulative evidence of its presence.

Commentary: The best value for this parameter depends on the nature of the data. It should almost always be two or three, though in some instances one sentence is sufficient (eg: press release data or verse)

Prose Test Threshold (0–5): The Prose Test feature examines raw text sentences to decide whether they are valid prose from the configured languages. This is achieved by counting the number of stop-words that appear within each sentence. If this number is high, it is likely to be a

sentence from a configured language. This option allows you to specify the number of stop-words that are required for the sentence to be further processed.

Commentary: This feature is good for reports or other prose documents where you don't want to process tables of numbers or lists of words. It is almost essential for web pages or e-mail messages which often contain menus or signatures. This sort of repeated data can potentially contaminate your automatic seeds and machine learning. If the data is not prose from a supported language, or if it is composed of transcribed speech or other colloquial matter which does not obey the rules of prose style, then this feature should be weakened or disabled. You should also disable this if you need to analyse absolutely every bit of text in the data

Duplicate Text Sensitivity (Off|Auto|1-8): This setting suppresses the processing of duplicated text. This option is especially useful when analysing email data or blogs and reviews, where cross-posting and quoting is common.

Identify Name-Like Concept (Yes|No): This setting is important if you would like words that seem to be names (i.e., non-stop words, starting with a capital letter) to be stored as potential concepts.

This setting requires text data which uses upper and lower case, where upper case designates proper names. This doesn't work in many languages, or in some text data where case is missing, but it is very

useful much of the time for tagging proper names. Note that it binds compound names into one token.

Break at Paragraph (Yes|No): This setting is to prevent context blocks from crossing paragraph boundaries. Only if the majority of paragraphs in the text are shorter than 2 sentences should you consider ignoring paragraphs.

Auto-Paragraphing (Yes|No): This setting identifies whitespace, particularly line-breaks and paragraph-breaks, if the document is consistent in its spacing, to identify new paragraphs. If there is a document with no reliable spacing for paragraph boundaries, this setting should be turned off.

Merge Word Variants (Yes|No): This option employs a stemming algorithm to identify the headword for initial thesaurus items. For instance if stemming is turned, the initial thesaurus terms for the stem look in the Concept Seed Editor may include looked and looking. If you don't like the results, you can ungroup the thesaurus items in the seed editor. Lemmatisation is off by default.

File Preparation (All Files|New Files Only): The New Files Only setting allows you to optimise processing by preprocessing only the new files in your data folder (i.e. files that have not already been preprocessed). The All Files option will delete all previously prepared text, and start the preprocessing from scratch.

Commentary: The normal situation involves mapping a fixed set of documents, and in this case you want any previously prepared text cleared out from the map. In case you want to add some new files from a data folder to an existing map, then use this setting to stop

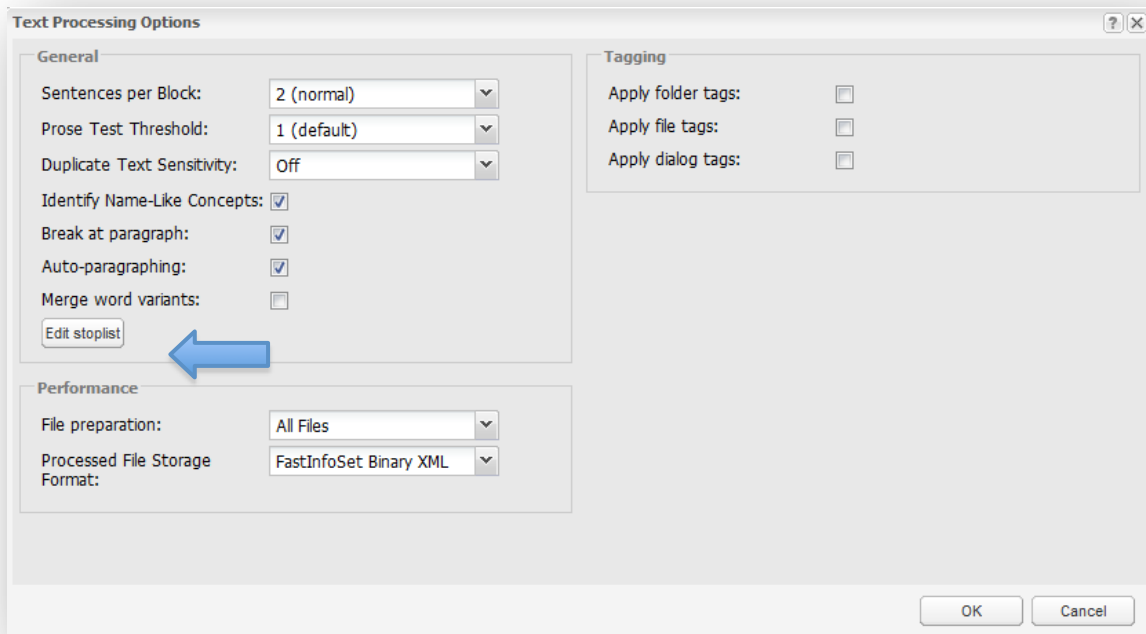
the previously prepared text from being deleted. This will also prevent the importation and preparation of any files in the data folder which have been prepared before.

Stopword Removal

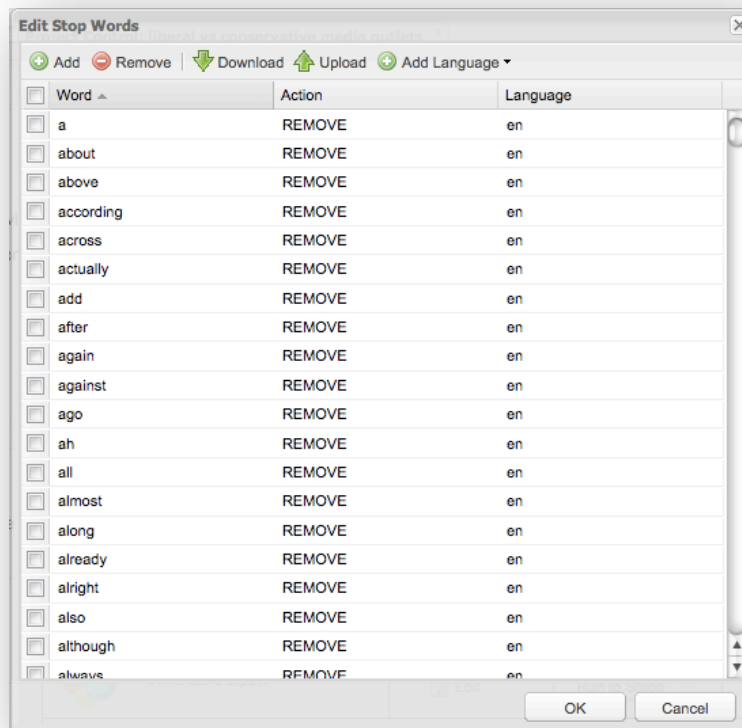
During Preprocessing, words with low semantic-content (meaning) are removed from the text data using a predefined Stopword List. An example of an English Stopword is 'and'. This word occurs frequent in English text, but would not constitute a useful or clearly-defined concept. If you are using an unsupported language, you can update this list (e.g. by translating the contained words into your language). Note that stop words are removed from the text to analysed, and cannot be selected as manual seed words.

Commentary: Stopwords are frequent words in a language that are rather arbitrarily designated as having little semantic meaning. Leaving stop words in the data has an obvious effect on the automatic seed selection. If you leave the stop words in, some will be chosen as automatic concepts. This can be annoying, depending on what you are looking for. The presence of stop words also impacts on the machine learning of thesaurus concepts, since almost everything can be correlated with words such as 'is' or 'and'.

Edit Stopword List (Button): You can check the words that are counted as stop-words by clicking this button:

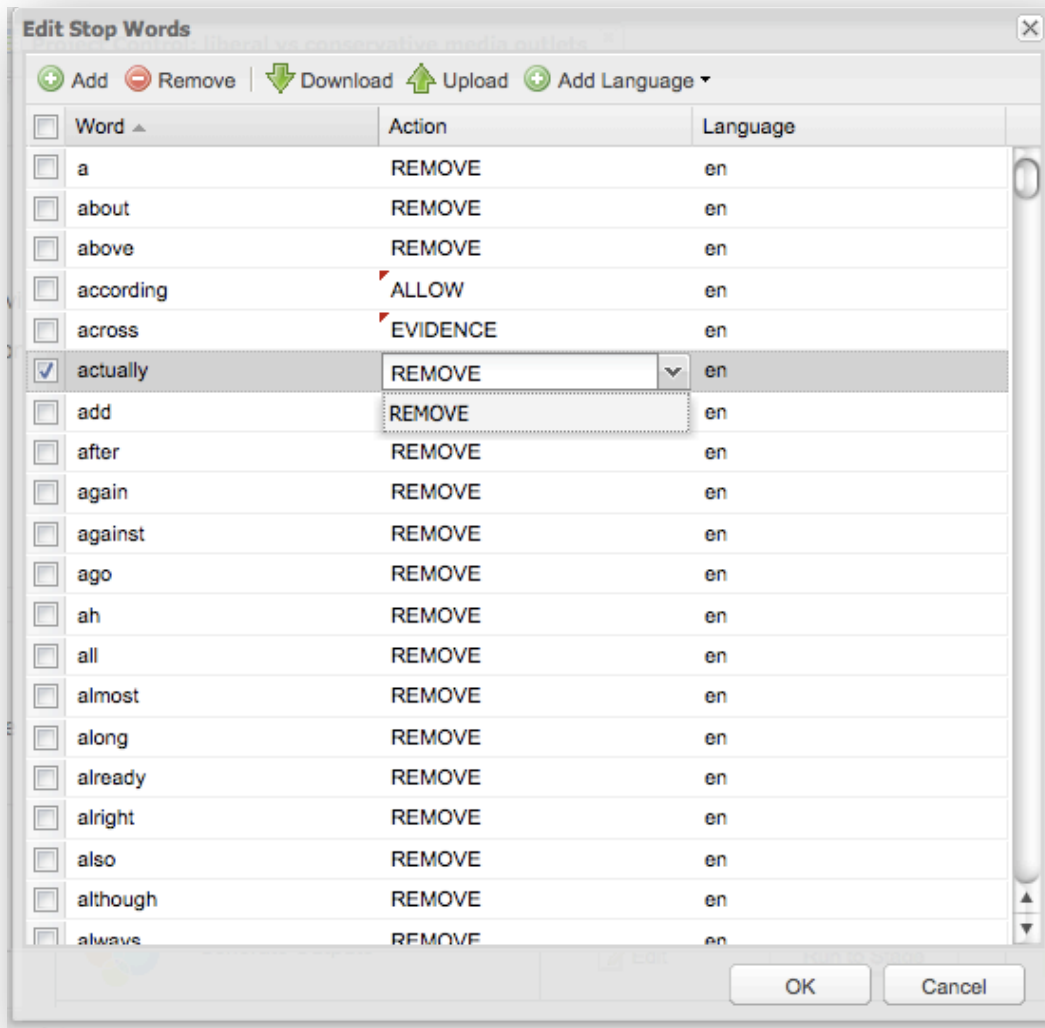


The following interface will appear:



You can browse through this alphabetical stoplist and see if you find any words that you would rather have left in the analysis. If so, click on the Remove option next to the word, and change the status of the word to Allow using the dropdown menu. Set the word's status to Evidence if you

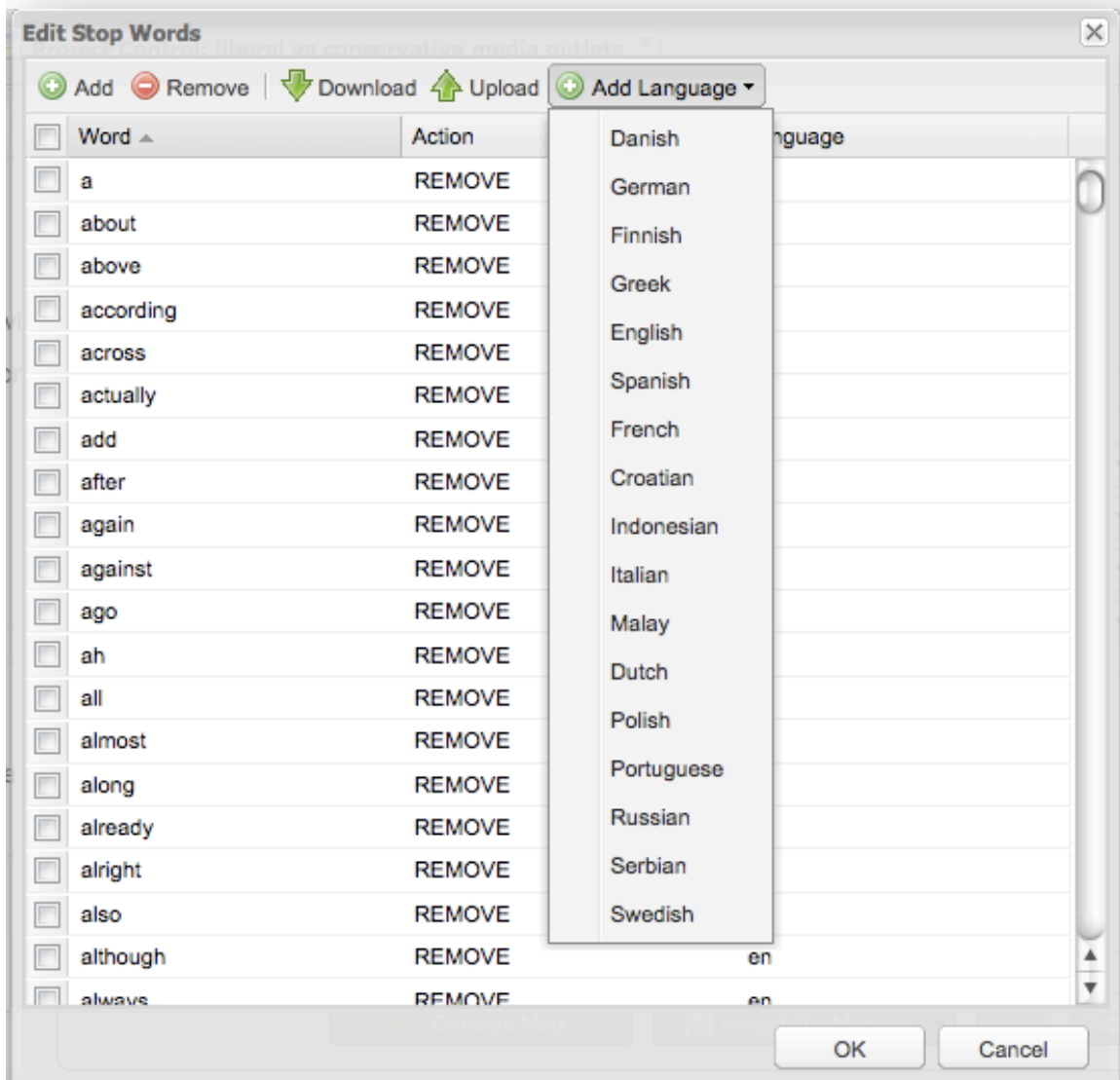
want to allow the word to form part of a concept thesaurus, but not to be considered as a possible concept candidate itself:



You can edit words in the list by clicking on them and typing in the text box that's revealed. You can also Add words to the stoplist, or Remove them from the list. The Download button lets you download the current .xml stoplist file, and the Upload option lets you upload another .xml stoplist file.

The Language column lets you know the language from which particular stopwords come in case you are analysing documents in different languages. You can change this setting so that stopwords are only removed from text in the appropriate language. The language abbreviations in the list are ISO 639 country codes.

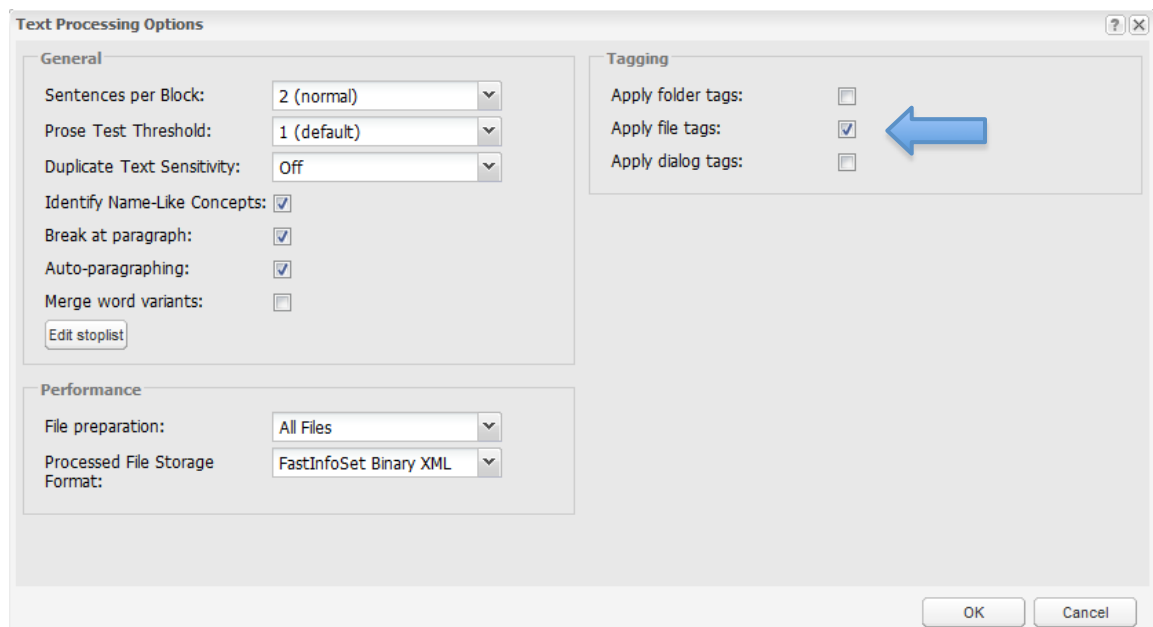
You may also add the stoplist of another language. This includes high-frequency, non-semantic words from a wide selection of languages. Note that the relevant stoplist will automatically load after selection of the language in the earlier document selection window:



Press 'OK' to return to 'Pre-process' box, then click 'OK' to return to the Control Panel.

Tagging Options

Tags are important for comparing different documents based on their conceptual content, for example, different speakers in transcript documents, or for a comparison between different text sources. At this stage in processing, you can instruct Leximancer to pay attention to certain tag categories so that you may analyse them later.



The **Apply Folder Tags (Yes|No)** and **Apply File Tags (Yes|No)** options can cause each part of the folder path to a file, and optionally the filename itself, to be inserted as a tag on each sentence in the file. In our example project, we can use the File Tagging facility to compare the content of the transcripts on different days of the hearing. Source document tags can then be included as concepts on the map.

Commentary: This is a powerful feature that lets you code all the sentences in each document with categorical tags just by placing the files in folders, possibly within other folders etc. The tags can

then be included on the map, among the topical concepts. This is useful for performing document clustering in a semantic context, or for making a discriminant analysis between two categories. For example, if you had a set of political speeches from a debate on some issue, you could give each speech file the name of the politician, and place all the speech files from each political party in a folder named as the name of the party. Then, if you had several sets of these speeches from different years, you could place each year's set of folders in its own folder named with the relevant year. Applying folder and filename tags will then insert the name of each politician as a tag on each sentence, and the name of the containing political party as a separate tag on each sentence, and also the name of each year. When you map this data, you will find a concept for each politician, each party, and each year in the Tags collection. You can then choose which of these dimensions to cluster together on the map, so you could view the issues by year, by party, by politician, by year and party, by politician and party, by year and politician, or all of them co-varying at once if you are adventurous.

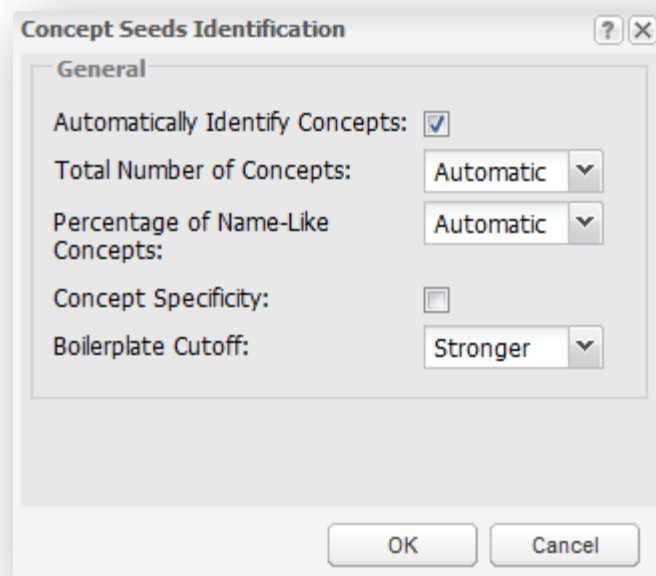
The **Apply Dialogue Tags (Yes|No)** function is designed to utilise speaker labels or headers in the text data. This setting identifies speaker labels that start with upper-case, end with a colon and space, and are located at the start of a line. Each speaker label is appended to the end of every subsequent sentence until a new label is found. This can be useful for analysing focus group or interview transcripts.

2b. Concept Seeds Settings

This is the phase of processing in which seed words are identified as the potential starting points of concepts. Concept seeds represent the starting point for the definition of concepts. They are single words (such as 'violence') that are the potential central keywords of distinct concepts. In this optional phase, the user may indicate whether they want Leximancer to automatically identify seed words (and the configuration of those seed words), or whether the user will manually provide seed words.

Practical: Configuring Concept Seeds Identification

The Concept Seeds Identification settings allow you to choose the number of concepts (if any) that you would like Leximancer to automatically extract from the text. To modify these settings, click 'Edit' on the Concept Seeds Identification in the main interface, and the following dialogue will appear:



Automatically Identify Concepts (Yes|No): Turn off automatic identification of concepts if you would like only concepts that you define yourself on the map.

Total Number of Concepts (Automatic|Number): This sets the number of concepts to be extracted automatically. More diverse content requires more concepts, but less than 100 is recommended. Leaving this setting on Automatic allows Leximancer to extract the naturally emergent number of concepts from the data.

Commentary: The larger the data set, or the more conceptually diverse, the more concepts you should look for. As a rough guide, think of the number of concepts increasing logarithmically with the number of documents. However some data, such as magazine article collections, are very diverse. Note that the selected set of concepts starts at the top of a ranked list, so you should always get the most important and general concepts. You need to decide how deep you want to go into the more specific concepts. Be aware that if you force more concepts than are really found in the data, you can start getting junk concepts from among the noise.

Percentage of Name-Like Concepts (Automatic|Number): This setting allows you to set what proportion of the automatically-extracted concepts should be forced to be names. Leximancer identifies names by looking for words that do not begin a sentence but start with a capital letter.

Commentary: This setting's default is Automatic, which creates a natural mixture of words and names by not forcing any names into

the list. If you are not interested in names at all, you can set this to 0%. Increase this number if you are particularly interested in names.

Concept Specificity (On|Off): Concrete concepts differ from normal concepts in that they are much more specific. They are words that are strongly related to a small number of concepts, as opposed to normal concepts that are strongly linked with a large number of other concepts. Concrete concepts are suitable for creating book indexes, for example. Because they are so specific, however, you will need many more concrete concepts to cover the content of a document (which can be very demanding on resources).

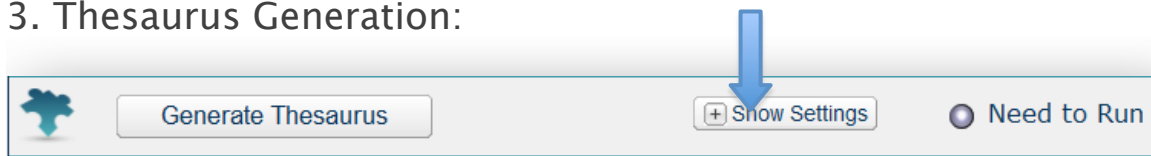
Commentary: Use this with caution. You need lots of memory and lots of concepts to use this option effectively.

Boilerplate Cutoff (Off-Very Strong): Some concepts may be overly specific in the sense that they only appear in a particular textual context. For example, certain words that only appear in repeated addresses, disclaimers or copyright notices may be identified as a specific concept. Increasing the Boilerplate cutoff filters out such overly specific concepts from the automatic list. If this value is set too strong, however, the remaining concepts may become too general.

Commentary: This feature looks for common words that frequently occur in the same context and stops them from being selected as automatic concept seeds. It can be useful for web pages that contain repeated links or menus, but consider that if set too high,

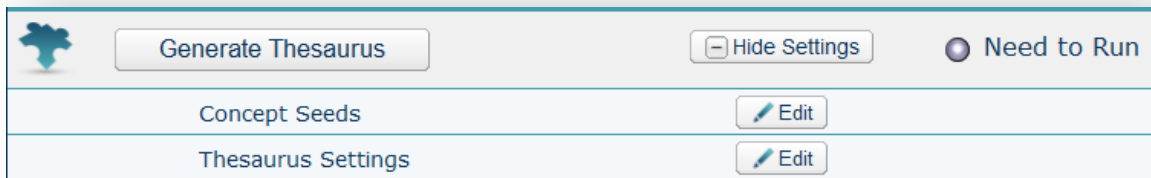
it could also cull concepts you might want. Note that this filter does not remove the words from the text, it just stops them being selected as automatic concepts. You can still use these words as manual seeds.

3. Thesaurus Generation:



You can run this stage of processing using default settings by clicking the Generate Thesaurus button.

Alternatively you can expand this stage (using the plus sign next to Show Settings) to Edit the settings for two sub-stages within:



Starting with the seeds automatically extracted by Leximancer, the Concept Seeds Editing phase (3a) allows users to edit, add or remove concept seeds from the list.

The following phase, Thesaurus Learning (3b), then generates the thesaurus of terms associated with each seed. As mentioned earlier, concepts are collections of correlated words that encompass a central theme. Once such lists of words have been identified for each concept, the concept map can be generated to illustrate the relationships between the concepts in the text.

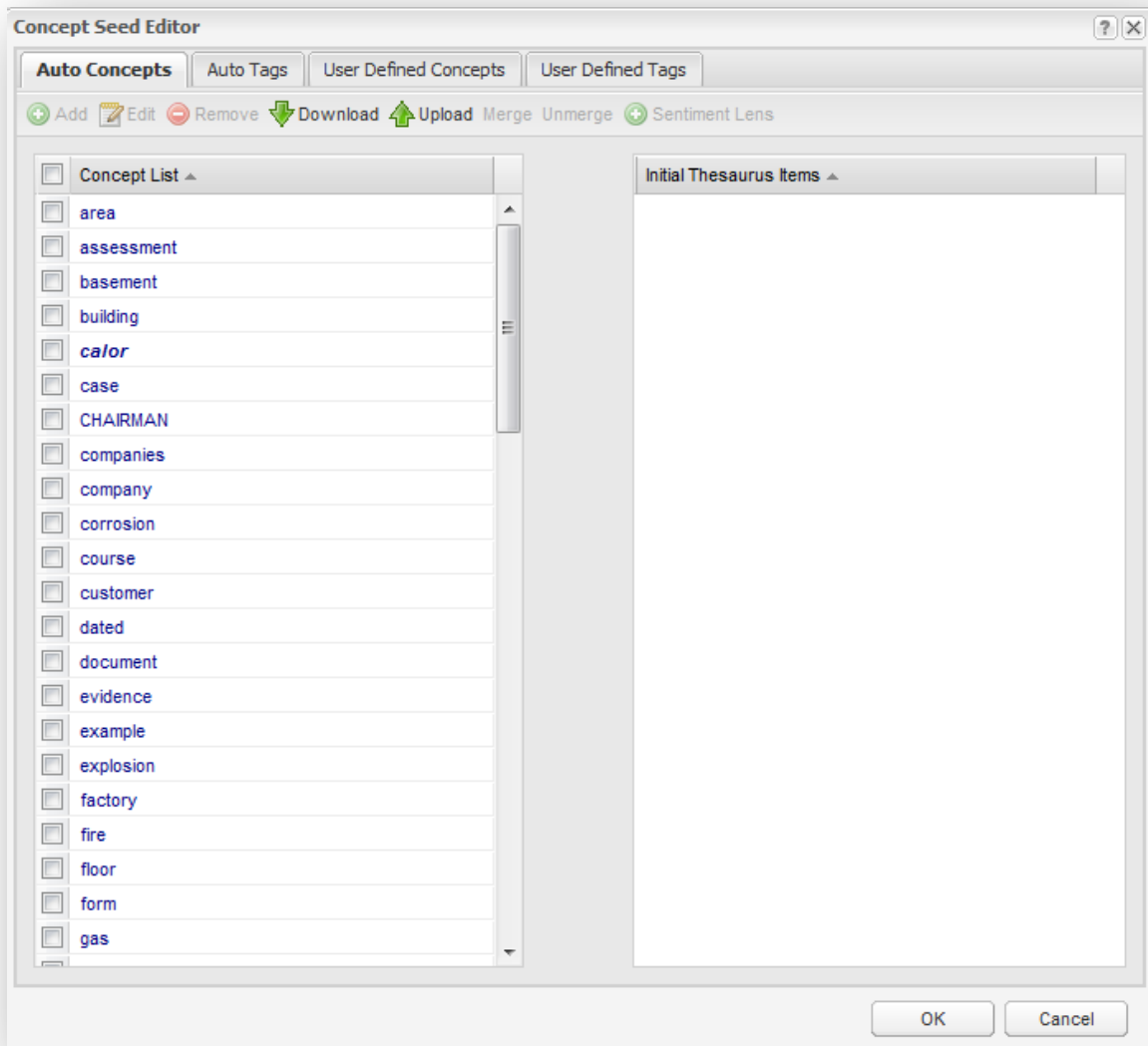
The learning of the thesaurus associated with each concept is an iterative process. Seed words are named as such because they start out as being the central terms of a concept definition – related keywords are collected during learning. During learning, seeds can also be pushed to the periphery if more important terms are discovered.

3a. Practical: Configuring Concept Editing

Clicking on Edit Concept Seeds opens an interface that allows you to edit, add or delete concepts. This is important for a number of reasons:

- automatically extracted maps may contain concepts (such as think and thought) that are similar, or other concepts that are not of interest to you. In the Concept Editing interface you can merge similar-looking concepts into a single concept, or delete concepts that you do not wish to see on the map
- you may wish to create your own concepts (such as violence) that you are interested in exploring, or create categories (such as dog) containing specific instances of terms found in your text (such as hound and puppy).

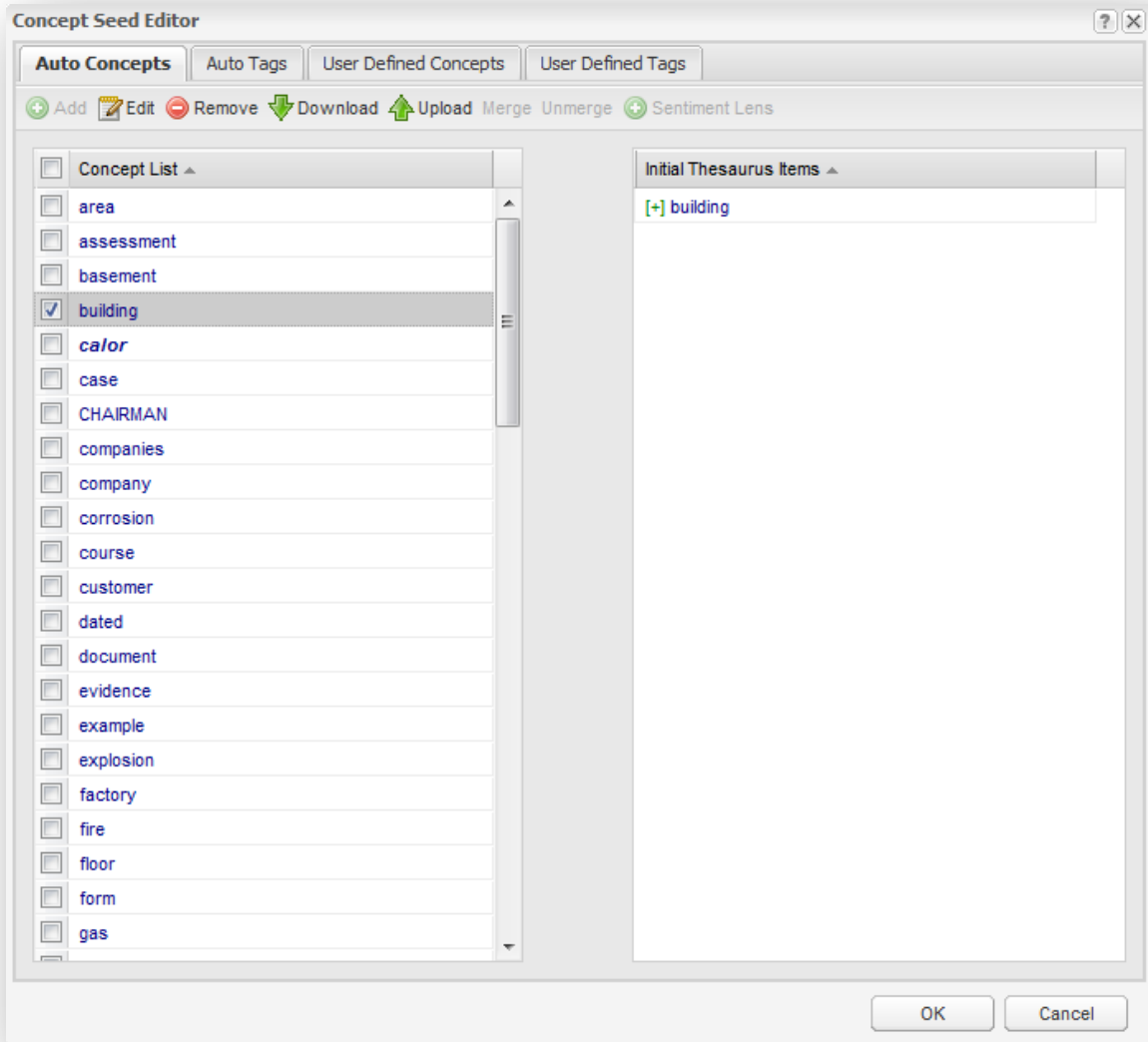
Once the Generate Concept Seeds stage has been run, you can check and modify the discovered concepts by opening the Generate Thesaurus Settings and clicking on Edit Concept Seeds. The following interface will appear:



Here you can edit the concepts extracted automatically by Leximancer in the Auto Concepts tab, and create your own manual concepts in the User Defined Concepts tab.

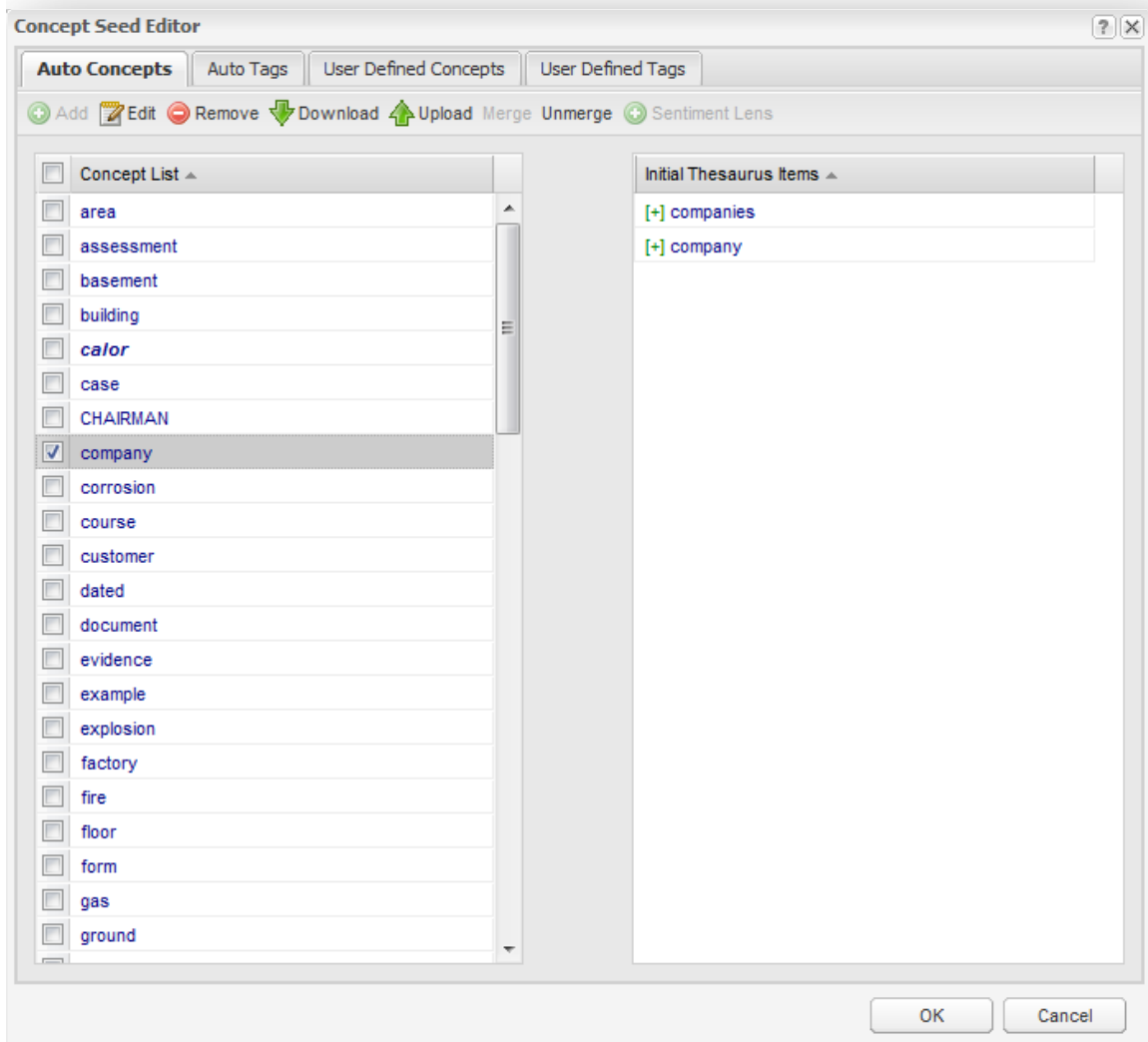
Name-like concepts appear in bold and italicized font in the alphabetical list.

Prior to thesaurus learning, only the central seed term for each concept has been identified. In the dialogue above, clicking to select the concept 'building' reveals a single seed term (identical to the concept name) comprising the thesaurus for this concept at this stage:



Single click on concept seeds to select them. Use the Remove button to remove any unwanted automatic concept seeds. Use the Select All button, or hold down control <ctrl> while clicking, to select multiple items.

You can merge similar concept seeds by selecting two concepts and clicking the 'Merge' button. If you do so, the merged concept takes its name from one of the concept seeds, and the concept then has two thesaurus items. For example, if you merge the concept seeds 'company' and 'companies', the following will result:

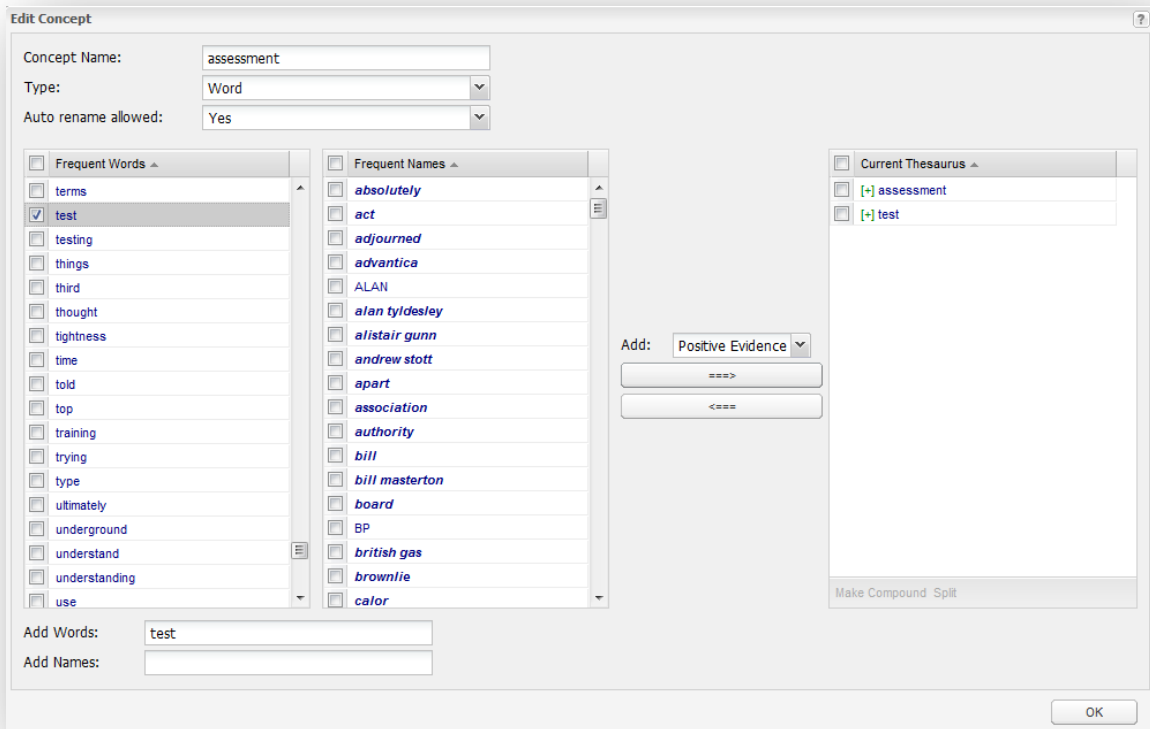


Do not forget to untick the merged item before moving on to work on other concepts. If you do not untick first, the following action will affect the merged item as well.

If you change your mind, you can select the merged concept in the list and use the Unmerge button to separate the two original seeds.

You can also edit automatically extracted concepts. For instance, if you wish to add additional thesaurus items, select the concept from the list and click Edit.

The following dialogue will appear:



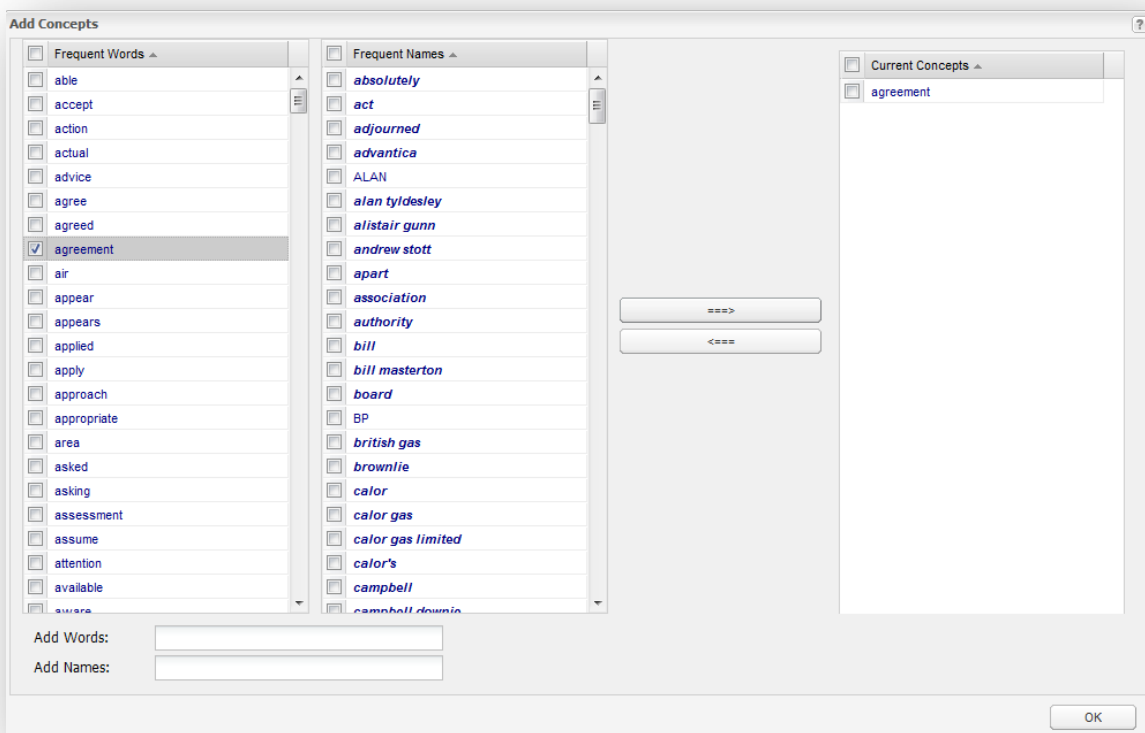
This interface allows you to change the name of the concept, identify it as a name-like or word-like concept, and choose whether to allow Leximancer to rename the concept during subsequent learning.

You can use the arrow buttons to Add or Remove terms related to the concept. Add terms if you believe that there are other words that predict well the presence of the concept in a section of text. You can choose words from the Frequent Words or Frequent Names lists, or enter your own words in the Add Words or Add Names text boxes. You can also identify terms that constitute negative evidence, or evidence that a concept is absent, but use this option with care. Leximancer will

automatically learn the weightings for these words from the text during the Thesaurus Learning phase.

If you wish to create your own concept(s), close this dialogue and return to the Concept Seed Editor interface. Click on the User Defined Concepts tab and then click on Add. This opens the Add Concepts interface, where you can define new concepts yourself.

Name the new concept using the lists of frequent words and names, or type a concept name into the text box. Use the right arrow button to move the name of the new concept over to the Current Concepts list:



Click OK to close this window to see your new concept under the Used Defined Concepts tab in the main Edit Emergent Concept Seeds dialogue. The new user-defined concept can now be edited in a similar fashion to automatic concepts.

If you wish to rerun the project from a prior stage and retain your edits to emergent concepts, you should click OK to save your edits. Then reopen the Edit Concept Seeds interface, and Download your edited list of concepts somewhere on your local drives.

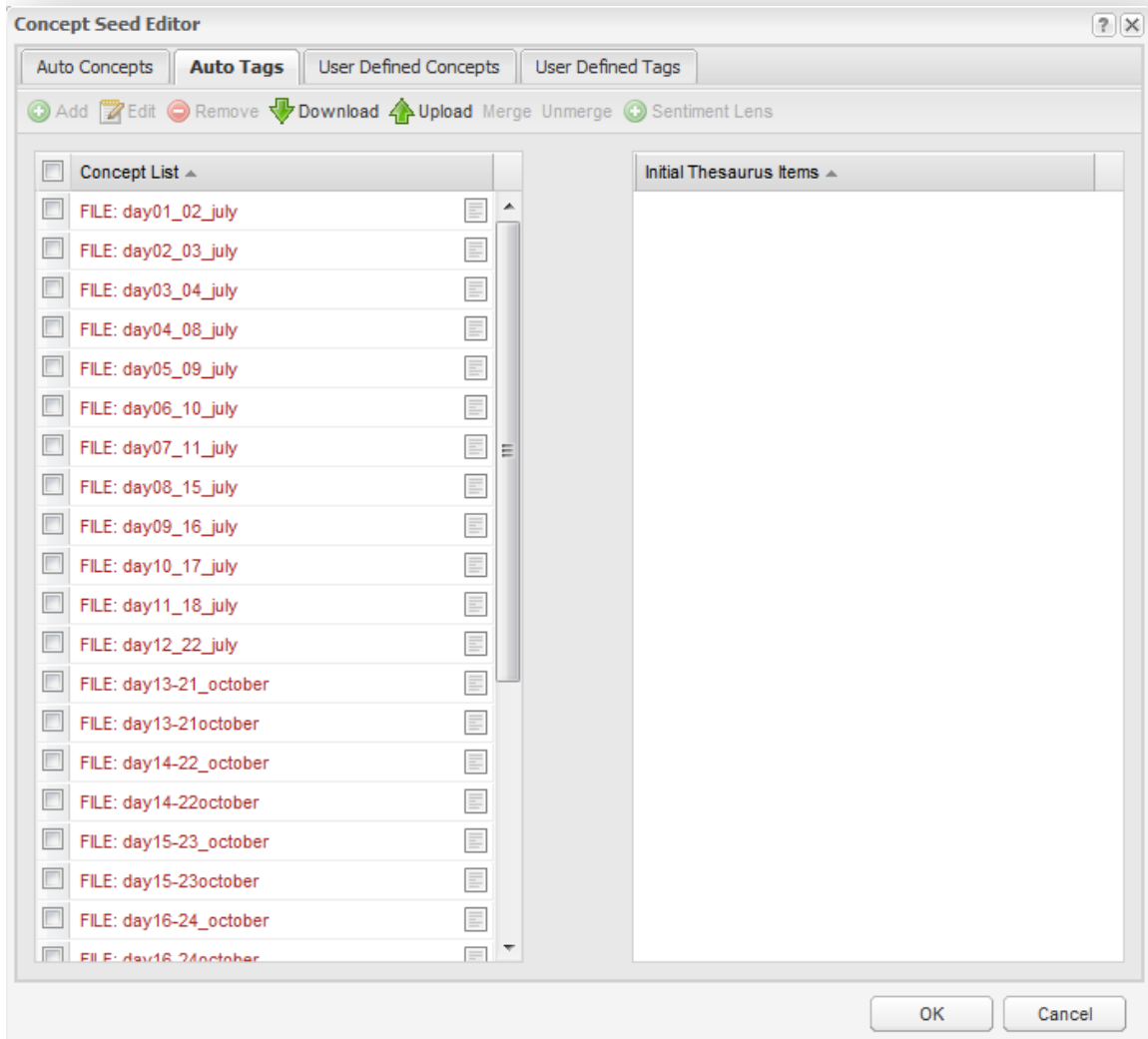
If you run the Generate Concept Seeds stage again, Leximancer will revert to its original list of auto concepts.

If you wish to use your saved list of edited concepts, go into the Edit Concept Seeds interface again, and Upload your saved concepts seeds file in the Auto-concepts tab.

You must save the concept seeds in each of the Auto- and User-defined tabs separately. The separation affords greater control by allowing you to reload individual seeds lists if you wish.

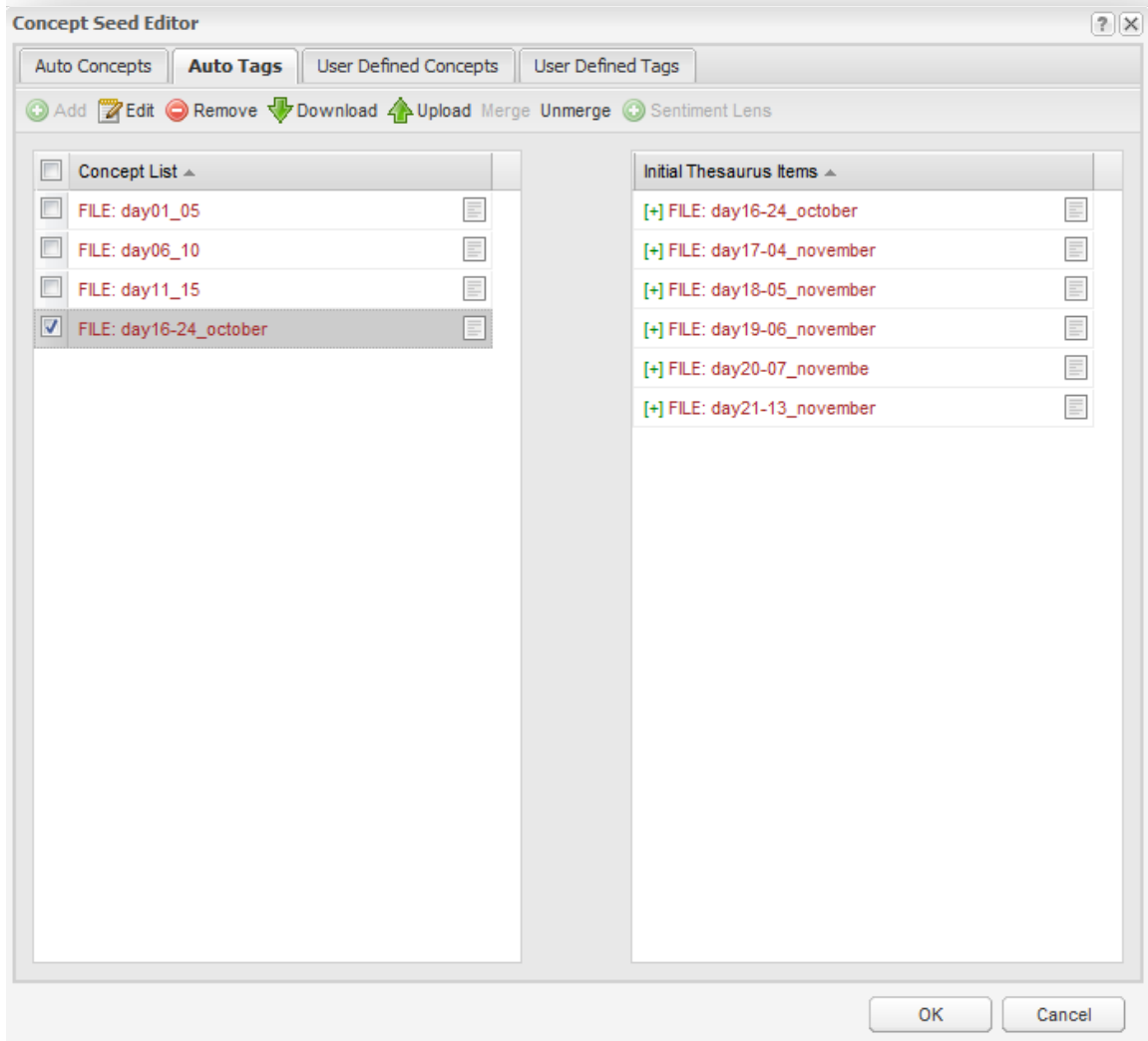
Using Tags

If you have opted to Apply File Tags in the Text Processing settings, then you should see a ‘tag’ representing each of your source documents in the Auto Tags tab in the Edit Concept Seeds interface:



Tag concepts are concepts for which no associated terms will be learned by Leximancer (unless otherwise instructed). They are useful if you want to make comparisons among groups within the data.

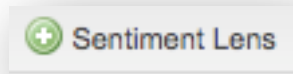
You can aggregate the tags using the Merge button, similar to merging concepts. In this case for example, we could merge the document tags to create 4 weeks of hearing transcripts for comparison:



If you have created Folder Tags, the numbering lets you know in which level of the hierarchy a folder resides (Level 1 is the top level).

You can also create User-Defined Tags to perform a simple keyword search for particular terms of interest.

The Automatic Sentiment Lens

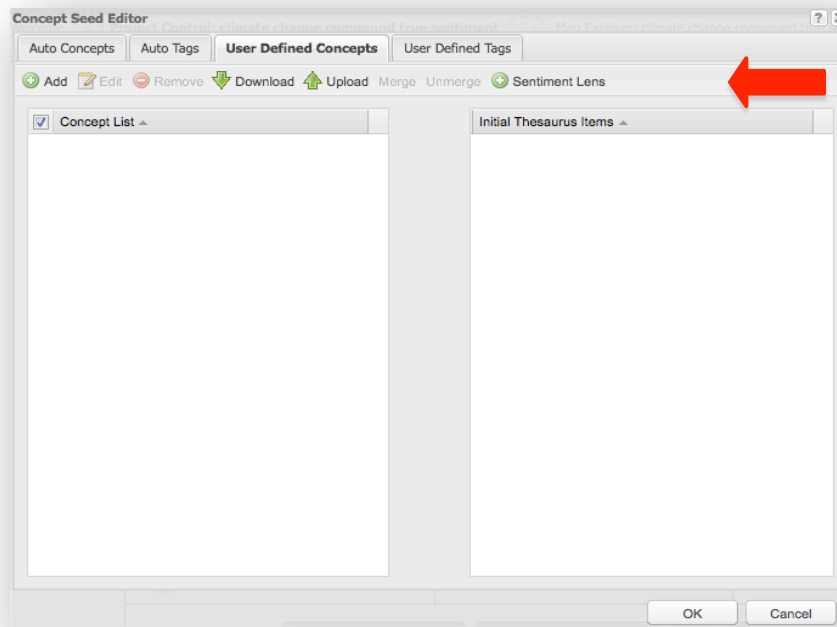


The Sentiment Lens automatically generates insight into positive and negative sentiment in your text. A default set of sentiment concept seeds and their terms are added to your user-defined list.

Sentiment Lens will only apply the sentiment terms that are identified as relevant and used consistently within your document set during processing. Sentiment Lens increases both the ease and accuracy of sentiment analysis

Configuring Sentiment Lens

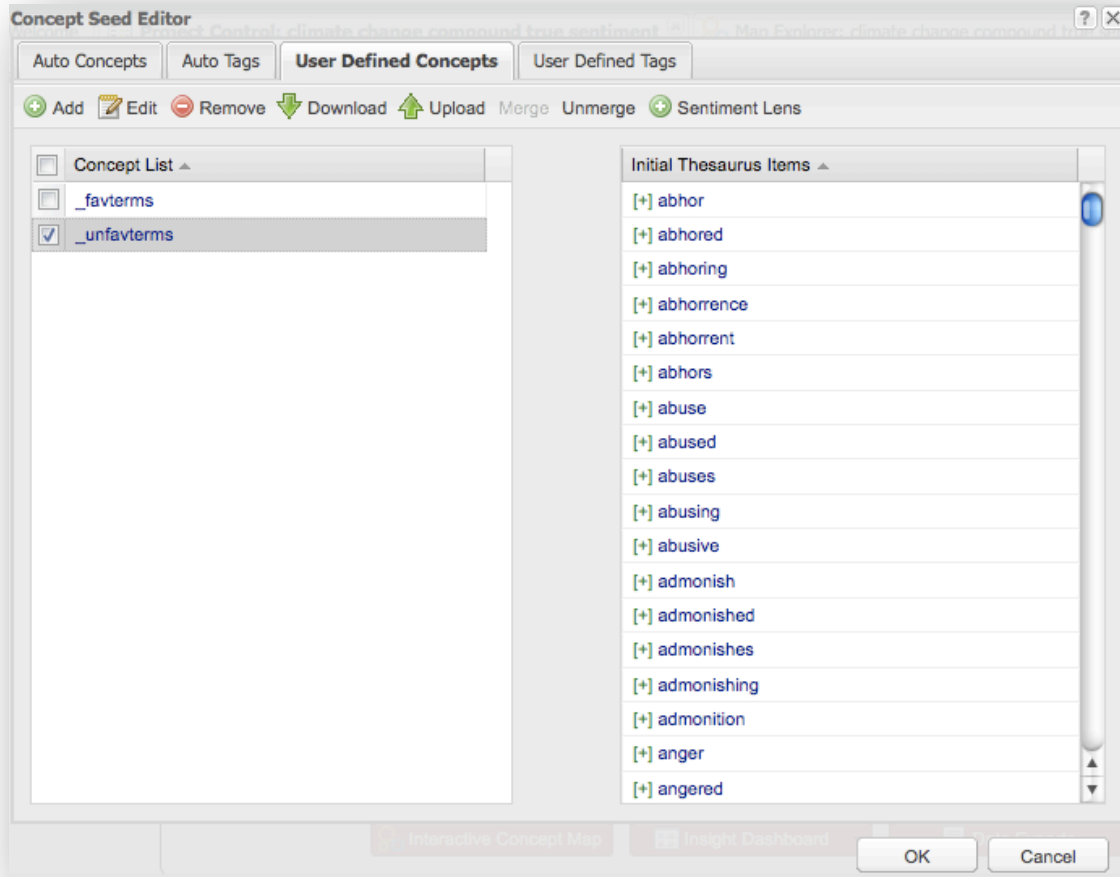
Under the 'User Defined Concepts' tab, you will notice a button on the top right: Sentiment Lens. Clicking this button will merge a pre-defined list of sentiment seeds into your editor:



****Please note:** if you make other changes to your user seeds, including tags, you must make these changes **FIRST** and then save them by clicking 'ok'. **THEN** you may re-enter the 'Edit Emergent Concept Seeds' stage and

use Sentiment Lens. Leximancer will display a warning dialog if you attempt to run without saving or discarding changes.

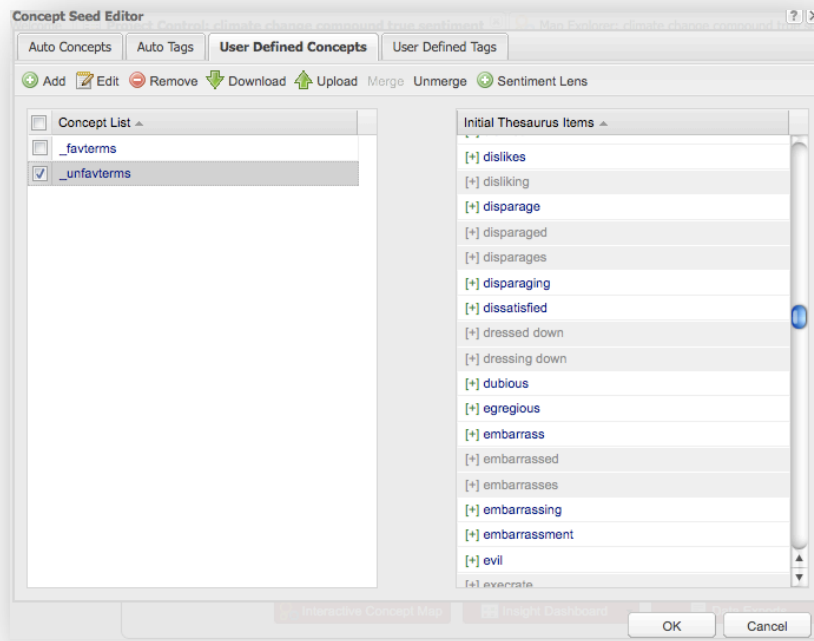
A list of sentiment seeds should appear under the categories ‘_favterms’ and ‘_unfavterms’. Under the ‘User Defined Tags’ tab is ‘_negationterms’.



- ‘_favterms’ is a list of commonly used favourable sentiment terms: approve, best, commend, favour, lauded, praise, great, etc
- ‘_unfavterms’ is a list of negative sentiment terms: abhor, anger, blame, denounce, frown, revolt, etc
- ‘_negationterms’ is a short list of words causing the following word to be negated: not, nothing, etc.

Click ‘Okay’ to return to the control panel.

If you run the Thesaurus Learning stage, and then return to the 'User Defined Concepts' tab in the 'Edit Emergent Concept Seeds' stage, you can observe the effect of Sentiment Lens. Sentiment thesaurus terms that are irrelevant or inconsistent in your text will be grey. Those left coloured are suitable for analyzing sentiment in your text and will be used as thesaurus items to develop sentiment concepts.



Once you reach the Interactive Concept Map, you may also observe new Sentiment terms that have been automatically added to the Thesaurus.

When analysing news articles about climate change for example, the term 'mongering' does not appear in the original seed list under '_unfavterms'. Yet once Sentiment Lens is applied and run, it appears in the list of Thesaurus items for '_unfavterms' (next page):

Initial Thesaurus Items ▲
[+] jam
[+] lack
[+] lament
[+] lamentable
[+] mess
[+] messed
[+] mock
[+] mocked
[+] mocking
[+] mocks
[+] negative
[+] objected
[+] objecting
[+] objection
[+] objections
[+] odious
[+] offence
[+] offensive



Word	Score ▼
disapproval	5.14
egregious	5.14
frustrated	5.14
frustrating	5.14
horror	5.14
objections	5.14
scandals	5.14
difficulty	4.98
fixing	4.98
ill	4.98
shocked	4.98
civics	4.8
countires	4.8
disgusting	4.8
handing	4.8
impeachment	4.8
mongering	4.8



Concept Seed Editor stage
Concept Map stage

Thesaurus tab in

This is because in the climate change literature, the term ‘fear mongering’ is used with negative connotations to describe climate change science. Hence, the automatic Sentiment Lens has picked it up as

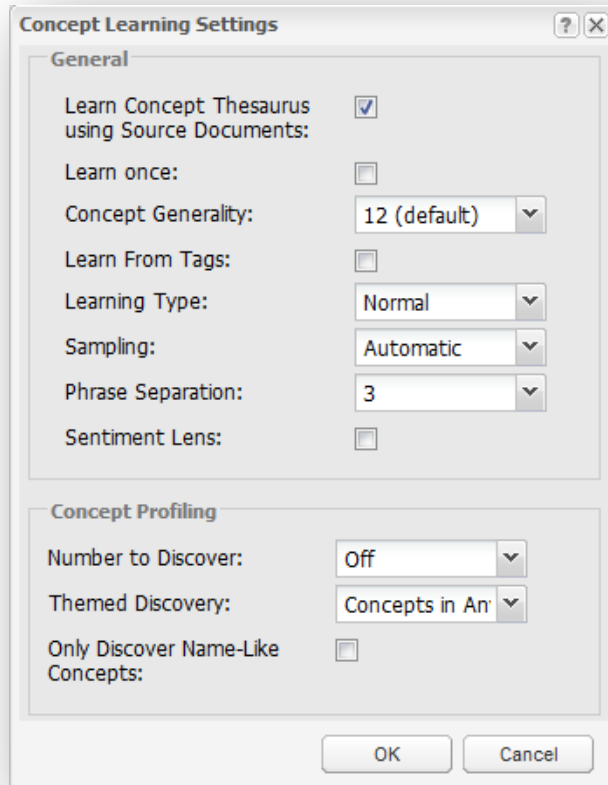
a term that contributes evidence to unfavourable sentiment, even in the absence of the word 'fear' preceding it.

3b. Generating the Thesaurus

The Thesaurus Learning phase generates a thesaurus of terms associated with each concept. Concepts are collections of correlated words that encompass a central theme, for example, the concept client may contain the terms customer, customers, client, clients, subscriber and subscribers. The learning of the thesaurus associated with each concept is an iterative process. Seed words start out as being the central terms of a concept, collecting related keywords over time. Through learning, the seed items can be pushed to the periphery if more important terms are discovered.

Configuring Thesaurus Learning

Clicking on the Thesaurus Settings Edit button reveals the following interface:



The **Learn Concept Thesaurus From Source Documents (Yes|No)** option allows you to turn off the thesaurus learning and prevents Leximancer from adding additional items to the concept definitions. This will result in searches for concepts as keywords, rather than using a weighted accumulation of evidence. This may be essential for data sets shorter than a few pages. In such cases, thesaurus abstraction is less useful due to a smaller vocabulary. Few sensible emergent concepts would be produced.

Concept Generality (1–21): This setting allows you to control the generality of each learned concept. This value is inversely related to the relevancy threshold required for a word to be included in the concept thesaurus. Raising this value will increase the fuzziness and generality of each concept definition by increasing the number of words that will be

included in each concept. After you have run the learning phase, examine the log to see how many iterations of thesaurus learning took place to arrive at the final concept definitions. This number should ideally be between 3 and 9. If the number is more than 9, consider lowering the learning threshold. Conversely, if the number of iterations is very low, consider raising this threshold.

Commentary: This setting controls how easy it is for the thesaurus learning to grow a concept to a broader set of words. The easier it is, the more iterations of learning will occur, the more words will be added to the thesaurus, and the higher the risk of the concept growing into a more general concept. This can result in the concept being renamed or being subsumed into another concept which already exists. If you examine the log file after learning you can monitor this behaviour. If the data does not consist of natural language, you should probably disable thesaurus learning, as described above.

Learn From Tags (Yes|No): You can use this option if you have any tags, either automatically extracted from tables, file or folder names, or speaker tags, or manually entered user tags. Turning on Learn From Tags will treat tags like concept seeds, learning a thesaurus definition for each. This setting is important if you are conducting Concept Profiling (discussed below) where you wish to extract concepts that discriminate between different folders or files (such as extracting what topics segregate Liberal from Labour party speeches).

Learning Type (Normal|Supervised): There are two forms of learning that are supported by Leximancer: Automatic and Supervised. Automatic

is the default behaviour, and in this a concept thesaurus is learned to characterize a list of seed words. For example, the initial seed word 'dog' be included in the thesaurus for the Dog concept, which also contains a collection of other dog-like terms discovered from related text.

Supervised classification, in contrast to Automatic learning, is possible when you have a complete definition of a category (such as a folder tag, speaker label, table category, or human coding tag) already embedded in some training text. In this case, you want to build classifiers that attempt to faithfully match human classification decisions, rather than discover an extended thesaurus from seed words. The learned concept should not include the initial seed item from its thesaurus definition. You are effectively giving Leximancer examples of a concept that you want learned. For example, if you were training the system to learn the concept 'violence', you might write the code 'Violence' in locations of the text that you think are examples of this concept. However, you do not wish the term 'Violence' to be a crucial term that is required to trigger the concept. Instead, a concept will be created, encompassing other discriminating terms from those contexts, that does not include this supervised term in the extracted classifier.

Sampling (Automatic|1-10): Sampling during learning speeds up the learning process by only reading every nth block of text. The automatic setting is normally fine, but you can override this, if necessary, by choosing n.

Commentary: The automatic sampling setting looks at the size of the total text data set to decide an appropriate value. The actual sampling number is increased by 1 for every 15 Mb of text data, so for anything under 15 Mb, sampling of 1 is used, which means

every context block is examined for learning. For 15 to 30 Mb of text data, a sampling of 2 will be used, which means that every second context block is examined for learning. Note that classification never uses sampling, and classifies every context block. Tests have show that classification performance only decreases marginally if the automatic schedule is followed.

Concept Profiling

These settings allow the learning process to discover new concepts that are associated with selected user-defined and automatic concepts. This is useful for profiling concepts or names, for doing discriminant analysis on prior concepts, or for adding a layer of more specific concepts which expand upon a top layer of general concepts. Profiling also allows you to ignore large sections of text that are not relevant to your particular interests.

Once the initial concept definitions have been created, words that are highly relevant to these concepts can be identified as potential seeds for new concepts. For example, if you profile the initial seed 'flowers', a concept definition is grown around this word as usual. Then new concepts are developed from the 'flowers' definition that would produce more specific topics, such as 'roses', 'daffodils', 'petals', and 'bees'.

This process is useful if you are trying to generate concepts that will allow segregation between various document categories. For example, if you are trying to discover differences between good and bad applicants, simply place exemplars of each in two separate folders (one for each type), and Apply Folder Tags in the Preprocessing stage. This will create a

concept class for each folder. In the Concept Editor, **only** retain these folder tags in your Automatic Concepts and Tags lists. Switch on Learn From Tags in the Thesaurus Learning phase, and use the profiling settings described below to extract relevant segregating concepts.

Number to Discover (Off|10–1000): This parameter specifies how many concepts should be profiled or discovered from the pre-defined concepts. More pre-defined concepts normally require more discovered concepts, but less than 100 is recommended. As a guide, select between 3 and 10 discovered concepts per pre-defined concept to give a reasonable coverage.

Themed Discovery (Concepts in ALL| Concepts in ANY| Concepts in EACH): Choose how you want the discovered concepts to be related to the pre-defined concept set: ANY gives the Union (related to concept1 OR concept2 ...), EACH gives the Exclusive Disjunction (XOR: related to concept1 OR concept2 but not both), and ALL gives the INTERSECTION (related to concept1 AND concept2). Choosing the intersection of the concepts will only extract concept seeds that are highly relevant to all or most of the learned concepts. For example, conducting a themed discovery from the concepts sun, surf, and sand may lead to concepts more relevant to beach scenarios than using words relevant to only one of these concepts (ie: the union of the concepts). XOR is designed for strong discrimination of target classifications (ie: finding specific concepts that segregate between the predefined concepts).

Commentary: If the pre-defined concepts surround some theme, such as say beach life or environmental issues, you probably want the discovered concepts to follow the theme, so choose the AND

(intersection) option. If you want to discover concepts that strongly discriminate between the pre-defined concepts or tags, choose the XOR (disjunction) operator.

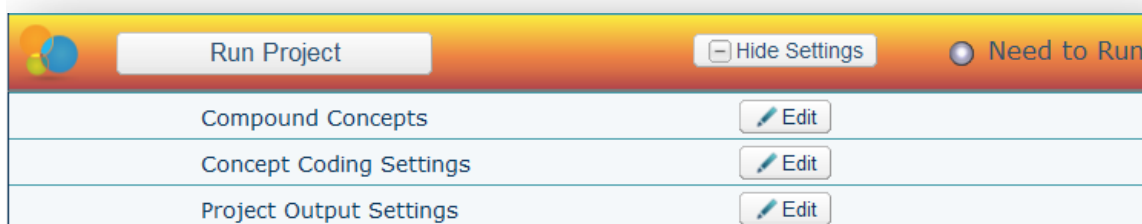
Only Discover Names (Yes|No): This option lets you discover name concepts only when profiling. This is useful for discovering social networks of association.

4. Run Project:



You can run this stage of processing (and all of these preceding it) using default settings by clicking the Run Project button.

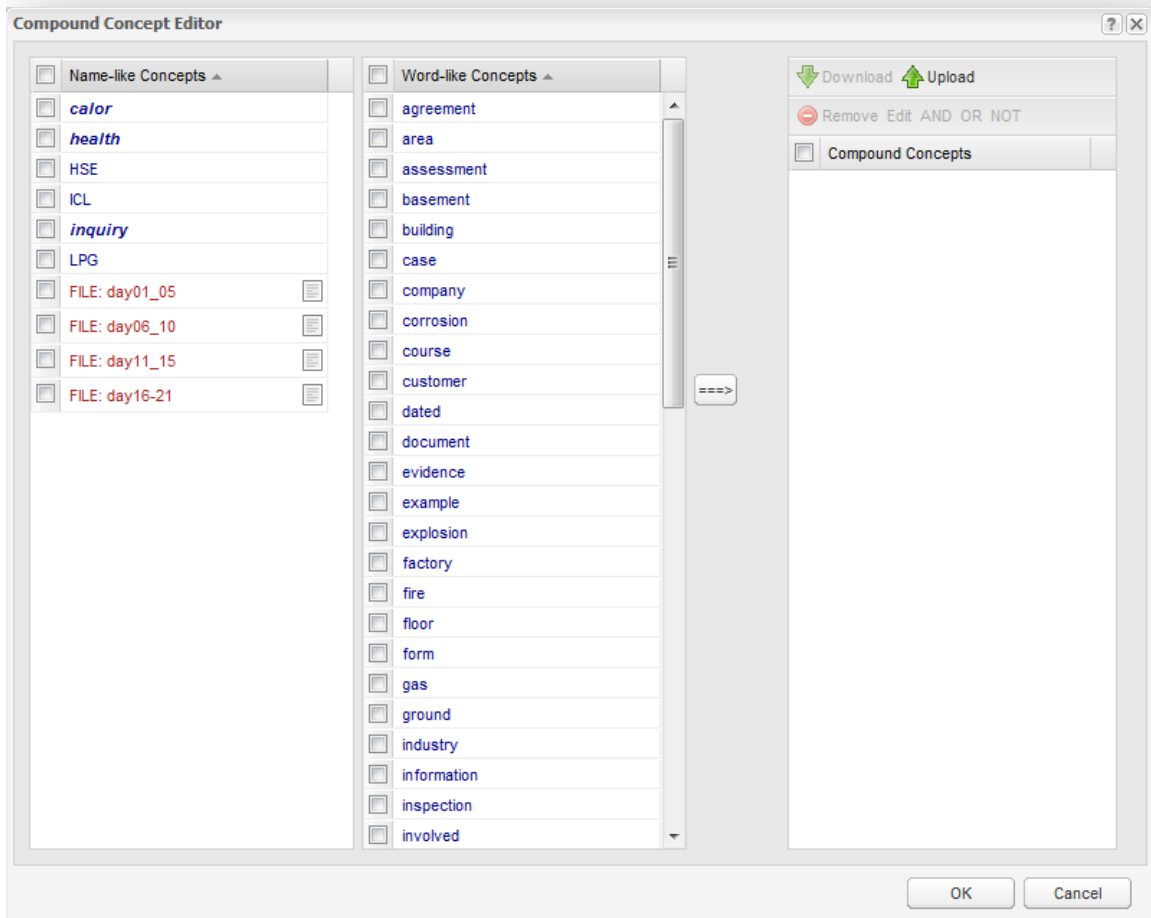
Alternatively you can expand this stage (using the plus sign next to Show Settings) to Edit the settings for three sub-stages within:



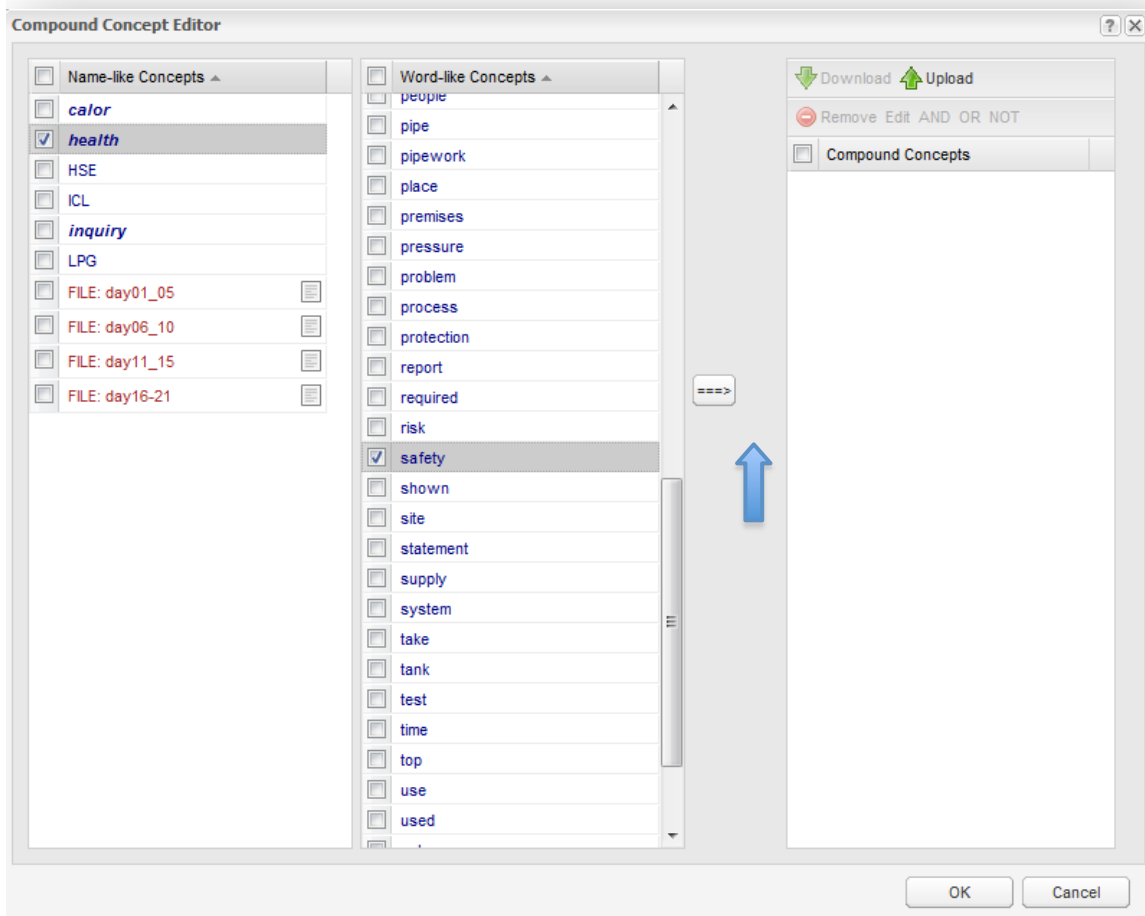
4a. Editing Compound Concepts

Manually compound selected concepts via Boolean operators to obtain deeper and more meaningful analysis.

Clicking 'Edit' at the 'Compound Concepts' stage opens this interface:

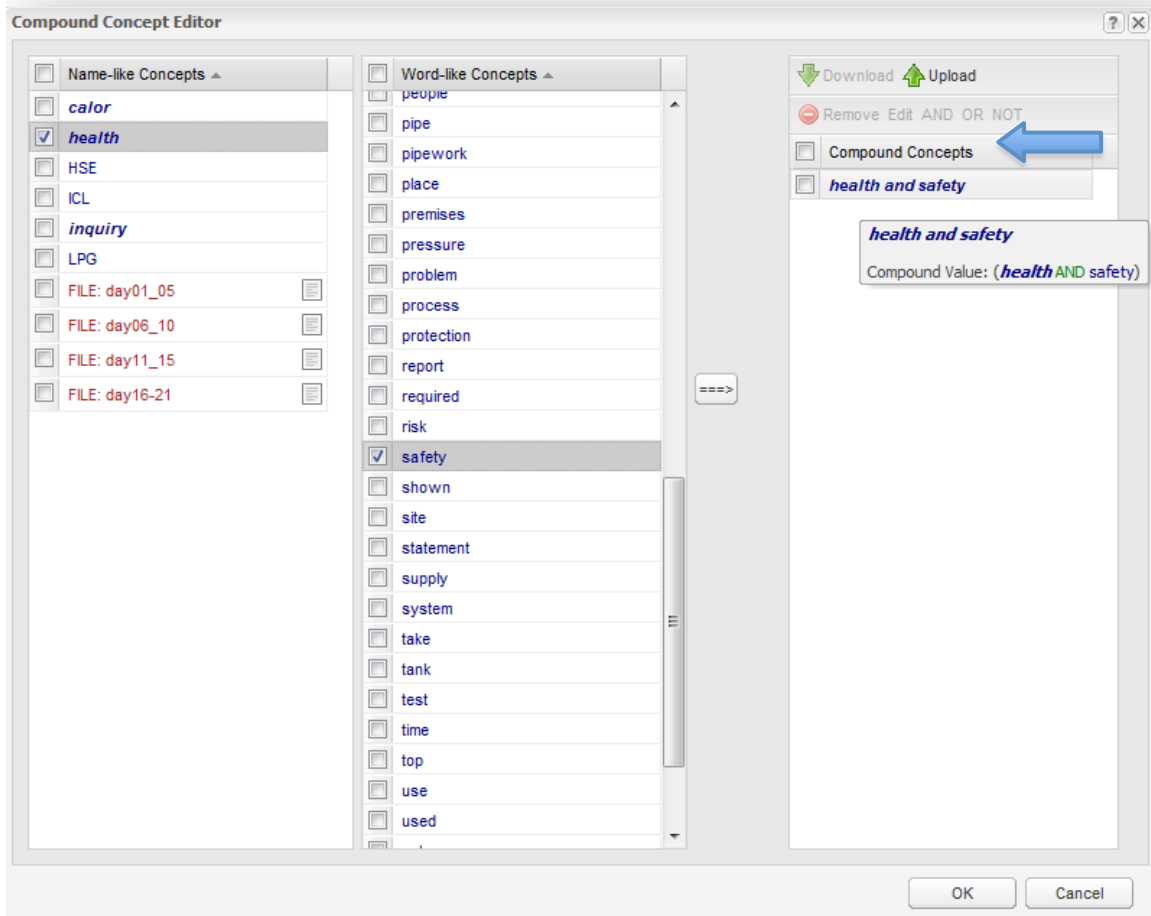


Select the boxes for any concepts you wish to combine into compound concepts. There are lists of all tags, name concepts and word concepts in the left and centre columns. Move them across to the compounding workbench using the right-hand arrow:



Combing the concepts using the AND operator requires both concepts to appear in the same (2-sentence) piece of text for the compound to be coded.

After moving the concepts you wish to combine into the right column, tick both of them again and click the 'AND' button in the header:

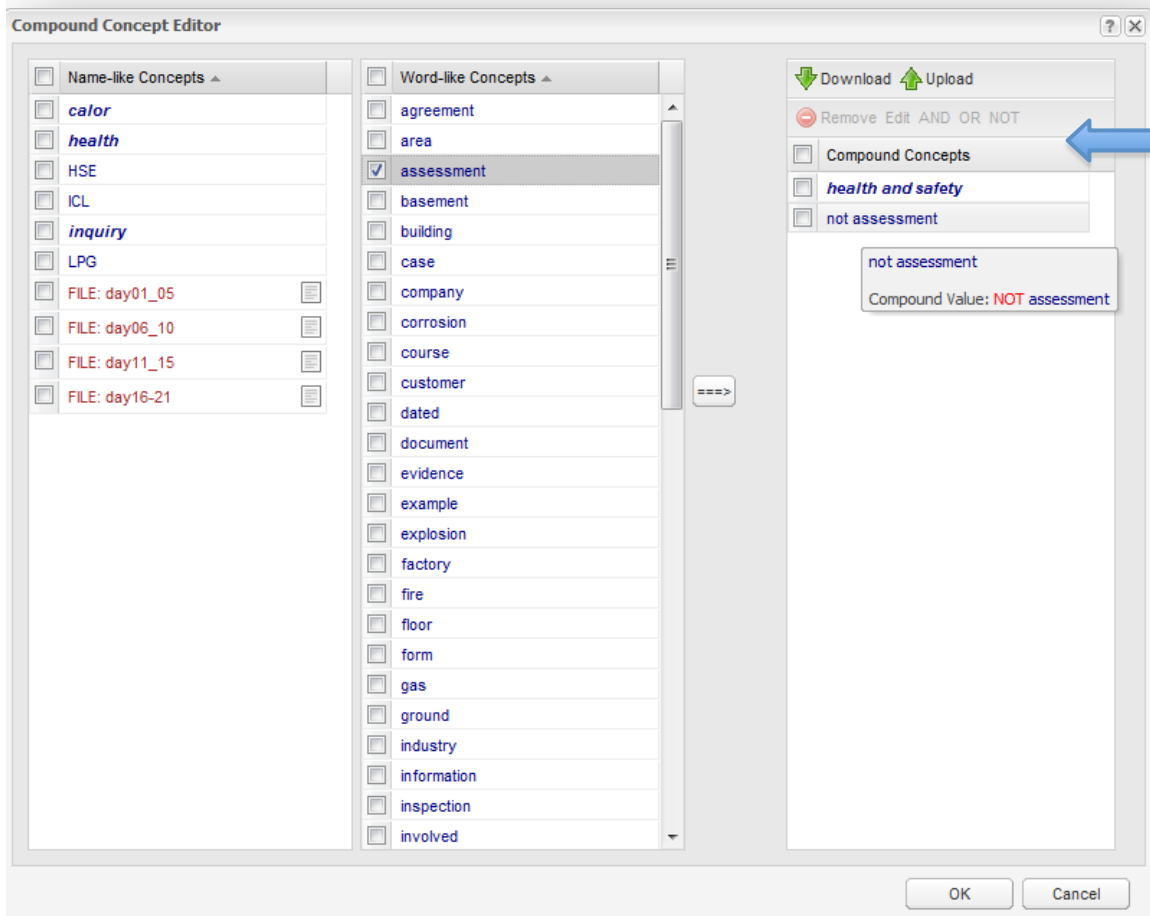


Hover your mouse over the compound concept to see the equation defining it.

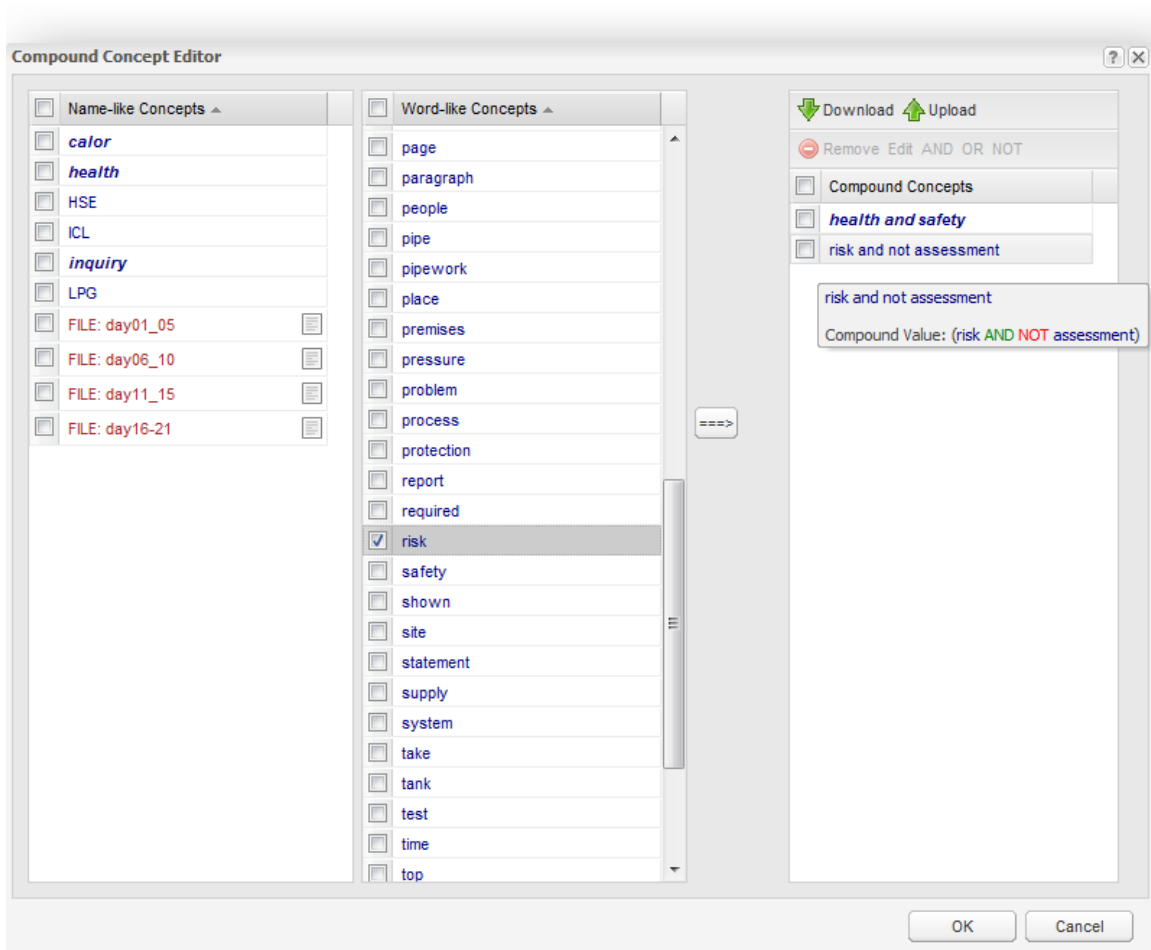
Note: Concepts used to build a compound remain available as singular concepts in the project results as well (unless you exclude the constituent concepts from the Mapping Concepts list in the Concept Coding Settings).

Use the same method to combine concepts with the Boolean operator 'OR'. Using the 'OR' conjunction means that evidence for your compound concept will be calculated to include evidence for either of your concepts. Compounding concepts using the OR operator is similar to Merging concepts in the Edit Concept Seeds interface.

To make a compound concept using the 'NOT' operator, first select the concept you wish to negate. The 'NOT' button at the top of the column will become available. Clicking the 'NOT' button will negate the concept you have selected:

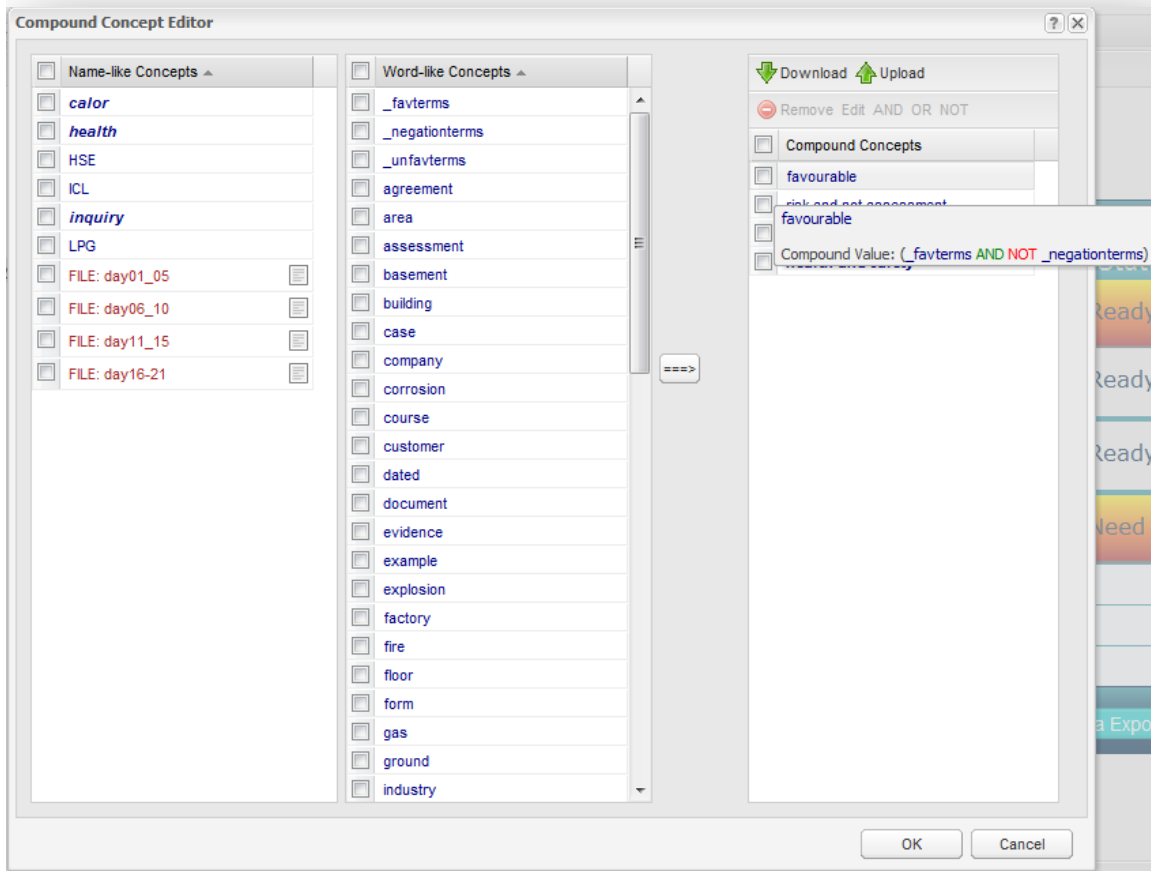


You can now combine the negated concept with another concept by following the steps for combining concepts with 'AND' as outlined above. Doing so will include instances of the positive concept, and exclude instances of the negated concept:



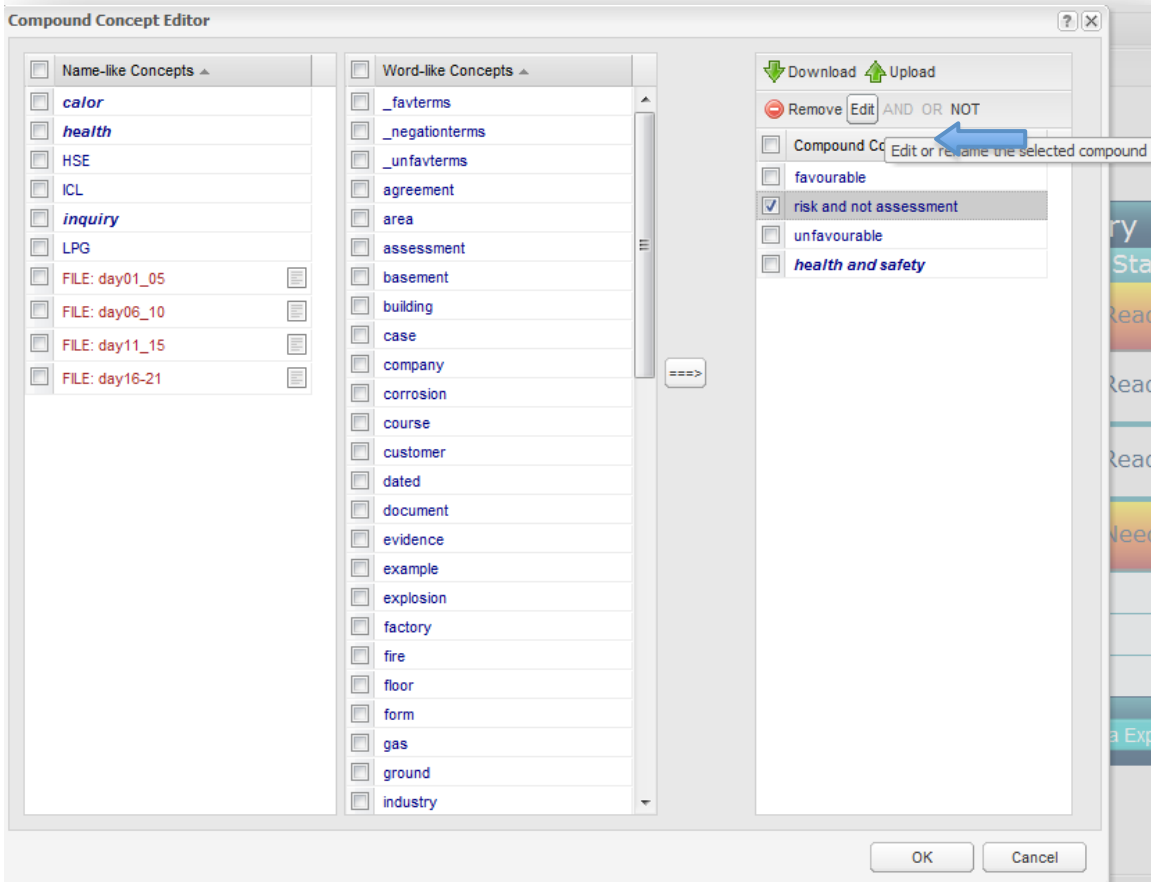
The same procedures outlined above can also be used to build more complex Sentiment concepts.

This is done automatically when you click **Sentiment Lens** in the Concept Seeds Editing interface. For example, the Sentiment Lens creates a compound concept for Positive Sentiment that includes favourable terms and excludes negation terms:

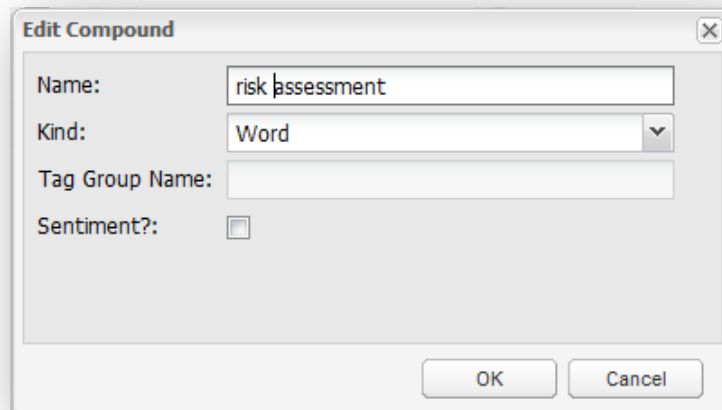


You can specify different rules for coding sentiment by combining the concepts used as building blocks for the Sentiment Lens ('favterms', 'unfavterms', 'negationterms') in different ways if you wish.

Finally, once you have created a compound concept you may edit it. Do so by selecting a compound, and then clicking 'Edit':



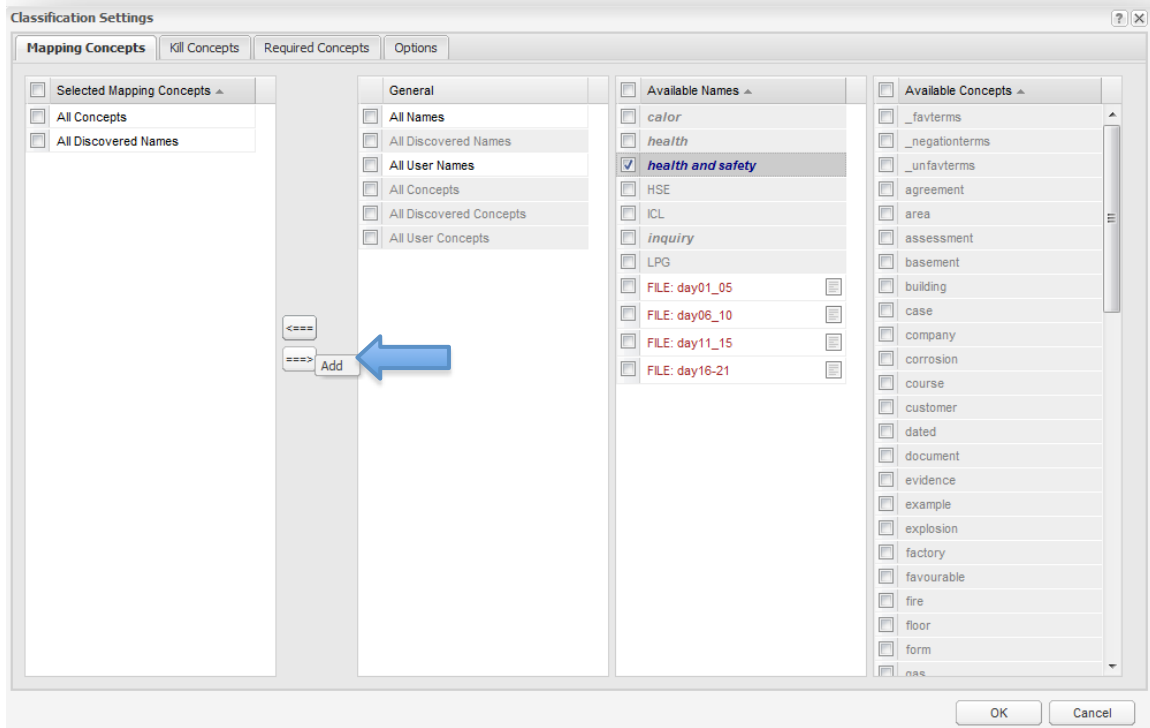
This allows you to rename your new compound concept, specify whether it is a name or a regular word, and whether or not it should be treated as sentiment:



(NOTE: Sentiment terms will not appear on the map, but will be present in the report tabs to the right of the map).

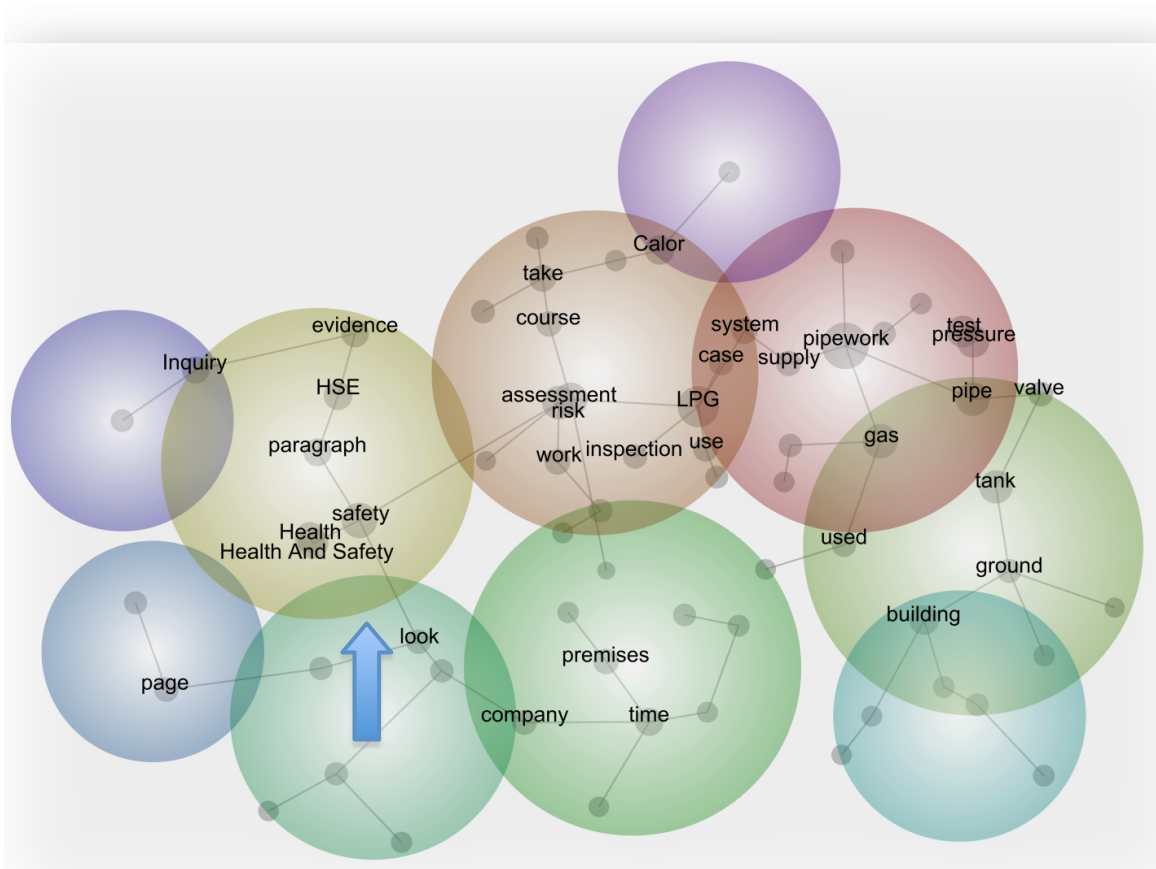
Compound concepts will automatically appear on the map if they are regular word concepts. Name-like compounds need to be specifically added to the Mapping Concepts list in the Concept Coding Settings.

After Clicking Ok to leave the Compound Concepts interface, click the Code Concepts Settings Edit button to open this interface:



Select the compound concept you wish to add to the map, and move it to the left hand column using the arrow.

Now when you click Ok, Run the Project, then open the concept map, your newly created compound concept will appear:

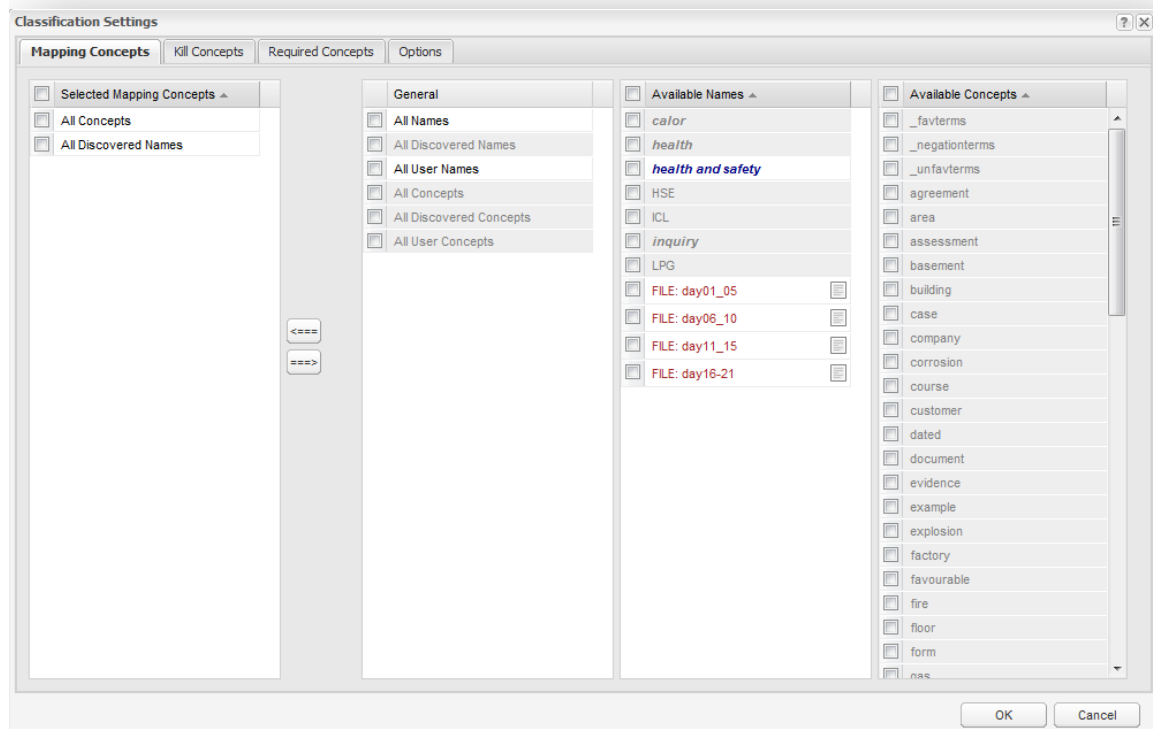


4b. Concept Coding

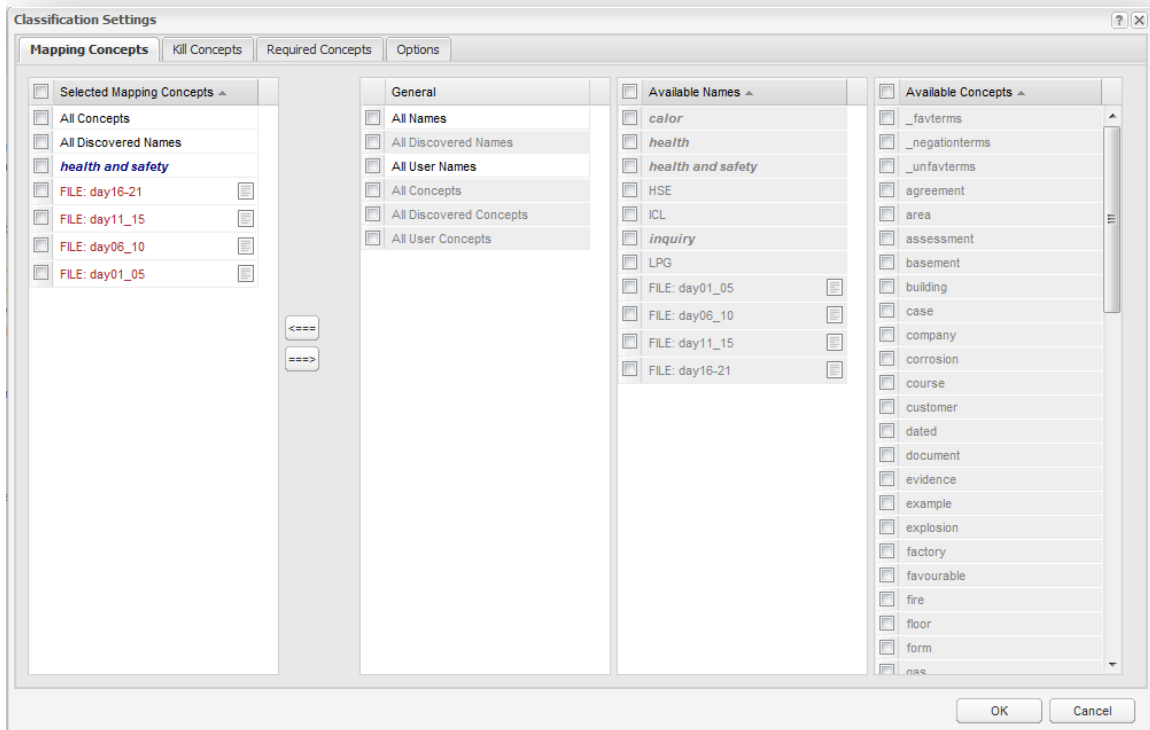
These are the concepts that appear on the conceptual map, and represent the top-level of classification of the text. Generally all concepts can be used as mapping concepts, but there are some cases in which you may want to map only a subset of the concepts. For example, you could map only the concepts discovered automatically by Leximancer in one instance, and then only the names seeded by you the user in another. You can also choose to map specific names and concepts from the lists provided if you wish.

Configuring the Code Concepts Stage

The first part of configuring the Concept Coding phase is to specify which of the Available Names and Concepts you wish to include on your map. Clicking on Edit for the Concept Coding phase opens the following interface:



By default, All Concepts and All Discovered Names appear in the Mapping Concepts tab. This means that all word-like and name-like concepts discovered by Leximancer will appear on the concept map. Tags, compound concepts and name-like user-defined concepts must be added to the list manually. Using the arrows to replace these wildcards with others from the General list allows you map other groups of concepts:



Discovered names and concepts refer to those automatically-identified by the program, and User names and concepts refer to those created by you the user. Tags are treated as User-defined.

Instead of using the option in the General list, you can choose to map particular names (including tags) and concepts using the lists on the right. Simply use the arrow buttons to add or delete concepts from each of the lists.

This interface also allows you to filter records in and out of the analysis by specifying Kill Concepts and Required Concepts. Simply click on the appropriate tab and move the desired concept(s) or tag(s) across using the arrow buttons.

Kill Concepts and Required Concepts

Kill concepts are concepts that if found in a classified block of text, cause all other classifications of that block to be suppressed. For example, you could use this option to suppress the processing of questions asked by an interviewer to focus on the responses of the interviewee. If you identified the dialogue spoken by the interviewer using the correct form (speaker name starting with a capital letter on a new line, and followed with a colon and a space, eg: Alan:), Leximancer has a setting in the Preprocessing Options called Apply Dialogue Tags which will automatically identify the dialogue tags. These will be presented to you in the Auto Tags tab in the Concept Seed Editor. You can then set the interviewer dialogue tag as a Kill class to suppress the processing text spoken by the interviewer.

Required concepts by contrast, are classifications that must be found in blocks of text, or else the blocks are ignored. That is, at least one of the required classifications must be found in any context block for it to be included in the concept map.

Options (Classification Settings)

Once the concept definitions have been learned, each block of text is tagged with the names of the concepts that it contains. This process is

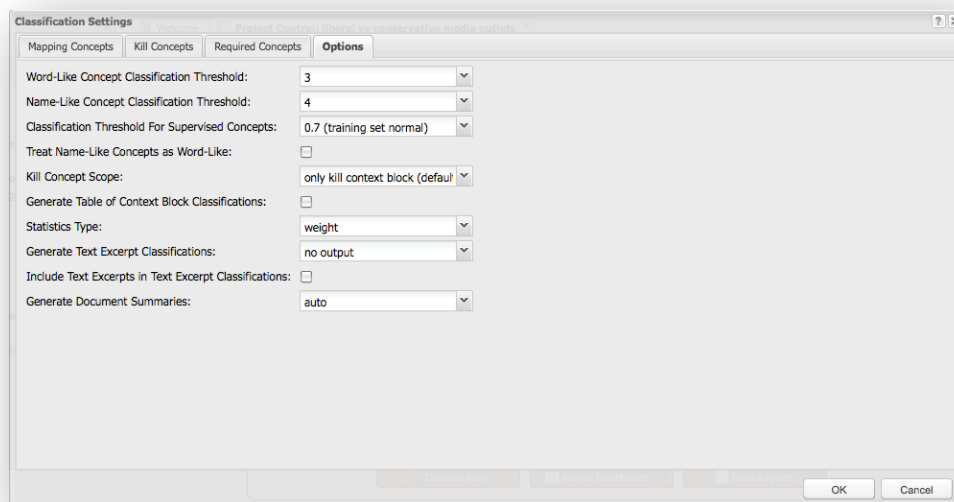
similar to manual coding used in ‘content analysis’. However, the benefit of using Leximancer for this task is that it is fast (compared to the time taken for human coders), and is more objective (as opposed to humans, where there is much variability in coding performance).

In Classifying the text document, the following steps are taken:

- The text is broken into context blocks of n sentences
- For name-like concepts, all the associated terms that are present in the block are noted. The block is said to contain the concept if the word with the highest relevancy to the concept is above a set threshold.
- For word-like concepts, the relevancies of all the associated keywords that are present in the block are summed. The block is said to contain the concept if this sum is greater than a predefined threshold.

Classification Settings

Clicking on the Options tab opens the following dialogue:



Word Classification Threshold (2–4.9): For word-like concepts, the relevancies of all the associated keywords that are present in the block

are summed. The block is said to contain the concept if this sum is greater than a predefined threshold. This threshold specifies how much cumulative evidence per sentence is needed for a word concept classification to be assigned to a context block.

Commentary: Note that the actual threshold used is this value multiplied by the average number of sentences per context block found in the data, not the number of sentences found in any particular segment of text. This means that a fixed threshold is applied for all context blocks. It may seem that the actual threshold should be calculated from the number of sentences found, but this would mean that less evidence would be required to trigger a classification in some places than others. After all, one sentence is less evidence than three. The units of this threshold are the same as the relevancy standard deviation values shown for terms in the thesaurus, so you can get a feeling for how much cumulative evidence you need by looking at the learned thesaurus which can be viewed through the map interface.

Name Classification Threshold (2.6–5): For name-like concepts, all the associated terms that are present in the block are noted. The block is said to contain the concept if the word with the highest relevancy to the concept is above the threshold value specified here. That is, this value gives the minimum strength of the maximally weighted piece of evidence to trigger a name classification.

Commentary: The idea behind this threshold is that one strong piece of evidence is enough to indicate the presence of a named thing such as company. For example, the name of the CEO or the

stock exchange code would be enough to indicate the company is involved. However, lots of weak evidence for the company is not sufficient, as it could be describing the same area of the market but actually a competitor company. This is related to the notion that a named thing is an instance of a class, rather than a class. If you want cumulative tagging of named things based on a similar lexical context, use the Treat Names as Words option. Like the Word classification Threshold, the units of this threshold are relevancy standard deviation units as shown in the thesaurus.

The next step in processing is to measure the co-occurrence of the identified concepts in the text (indexing) for the generation of the conceptual map. The positions of the discovered concepts and groups of important co-occurring concepts are also noted for use with the query browser.

Classification Threshold for Supervised Classifiers (0.7-1.4):

Supervised Concepts are used to find instances of a particular concept in the text. Generally, such concepts are 'trained' by tagging exemplars with a code word (such as 'violence'). They are trained to be sensitive to the vocabulary surrounding such tags, without including the tag itself within the definition. Thus, generally there is weaker evidence compared to normal concepts such as 'dog' in which the key word is present in the script. For this reason, supervised concepts require a lower classification threshold. This threshold specifies how much cumulative evidence per sentence is needed for a supervised classification to be assigned to a context block.

Treat Name-Like Concepts as Word-Like (Yes|No): This setting forces name-like concepts to be classified using the same system as word-like concepts, allowing more intuitive coding. This option allows tagging of named things based on similar lexical context, rather than similar identity.

Kill Concept Scope (Kill Whole Document|Only Kill Context Block): This option lets you choose whether to suppress the classification of the entire document should a killed class be present, or just the context block in which the kill class is located.

Generate Table of Text Segment Classifications (Yes|No): If this option is enabled, a delimited text output file listing the concepts tagged in each section of text through the data file is created. This output file, called Table of Text Segment Classifications, is accessible under the Export tab once processing is complete.

Statistics Type (Count|Weight): This setting affects the type of statistics produced in the High Resolution Data Output. The output file lists the concepts tagged in each section of text through the data. If the Generate Table of Text Segment Classifications option is enabled, this file is created and accessible under the Export tab. Normally an assigned tag is simply counted. You can change this setting so that tags are assigned a confidence weight. This results in weighted sums rather than count statistics.

Generate Document Section Classification (No Output|Document as Vector|Document as Matrix): Some advanced applications require classification metadata for each document.

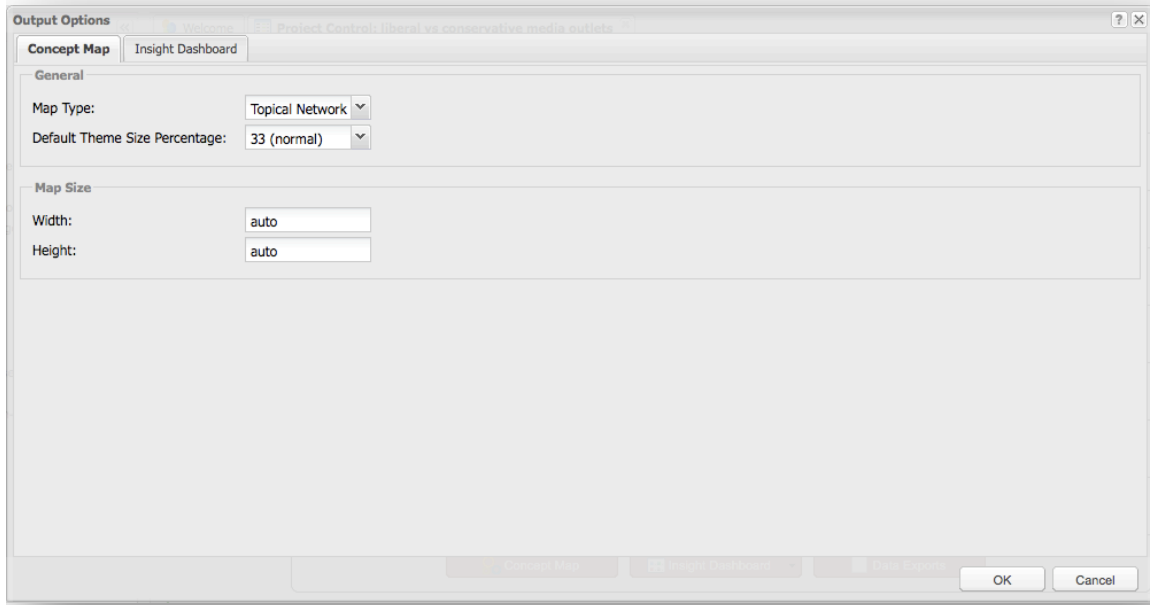
If Document Metadata is set to Document as Vector, an XML file is created that lists the concepts tagged in each document or document section. This output enables analysis of the conceptual content of the data by document section.

If Document Metadata is set to Document as Matrix, an XML file is produced that reports the matrix of concept co-occurrences within each document or document section. Matrix classification enables more accurate document comparison.

If either of these outputs is created, it is called Document Section Classification, and is accessible under the Data Exports button below the main Project Control Panel, or via the Exports tab in the concept map interface.

4c. Project Outputs

The last settings that can be edited through the main interface is Project Outputs. In this phase, the map displaying the relationship between variables is constructed. Clicking 'Edit' at the 'Project Output Settings' stage opens the following interface:



Generating the Concept Map

One of the principal aims of Leximancer is to quantify the relationships between concepts (i.e. the co-occurrence of concepts), and to represent this information in a useful manner (in a concept map) that can be used for exploring the content of the documents. The concept map can be thought of as a bird's eye view of the data, illustrating the main features (i.e. concepts) and how they interrelate.

The mapping phase generates a two dimensional projection of the original high dimensional co-occurrence matrix between the concepts. It must be emphasised that the process of generating this map is stochastic. Concepts on the map may settle in different positions with each generation of a new map.

In understanding this, consider that concepts are initially scattered randomly throughout the map space. If you imagine the space of possible map arrangements as a hilly table top, and you throw a marble from a

random place on the edge, the marble could settle in different valleys depending on where it starts. There may be multiple 'shallow valleys' (local minima) in the map terrain if words are used ambiguously and the data is semantically confused. In this case the data should not form a stable pattern anyway. Another possibility is that some concepts in the data should in fact be stop words, but aren't in the list. An example of this is the emergence of the concept 'think' in interview transcripts. This concept is often bleached of semantic meaning and used by convention only. The technical result of the presence of highly-connected and indiscriminate concept nodes is that the map to loses differentiation and stability. The over-connected concept resembles a mountain which negates the existence of all the valleys in the terrain. To fix this, remove the over-connected concept.

The practical implication is that for a strict interpretation of the cluster map, the clustering should be run several times from scratch and the map inspected on each occasion. If the relative positioning of the concepts is similar between runs, then the cluster map is likely to be representative. Note that rotations and reflections are permitted variations. If the map changes in gross structure, then revision of some of the parameters is required. The concept map display supplied with Leximancer allows easy animated re-clustering and map comparison using the buttons above the map.

Topical versus Social Mapping

In concept mapping, there is one setting to choose, namely whether to use a Topical or Social map. The **Social map** has a more circular symmetry and emphasises the similarity between the conceptual contexts in which the words appear. A Social map is best when entities tend to be

related to fewer other entities, such as a map made up of many name concepts.

The **Topical map**, by comparison, is more spread out, emphasising the co-occurrence between items. It tends to emphasise differences and direct relationships, and is best for discriminant analysis. The Topical map is also much more stable for highly connected entities, such as topics. The most common reason for cluster instability is that the concepts on the map are too highly connected, and no strong pattern can be found. The Topical variant of the clustering algorithm produces more stability in maps of this kind, so switching to this setting will often stabilise the map. However, the most important settings which govern the connectedness of the map are the classification thresholds and the size of the coded context block, which are located in the Classification Settings in the Locate Concepts node. If the coded context block is too large, or the classification threshold is too low, then each concept will tend to be related to every other concept. If you have some highly-connected concepts which are effectively bleached of meaning in your data, removing from the concept lists in the Concept Seeds Editor will often stabilise the map. Words such as 'sort', 'think', and 'kind' often appear in spoken transcripts and may be used as filler words which are essentially stopwords. Inspect the actual text locations to check the way words like these are being used before removing them.

In summary, the Topical clustering algorithm is more stable than the Social, but will discover fewer indirect relationships. The cluster map should be considered as indicative and should be used for generating hypotheses for confirmation in the text data. It is not a quantitative statement of fact.

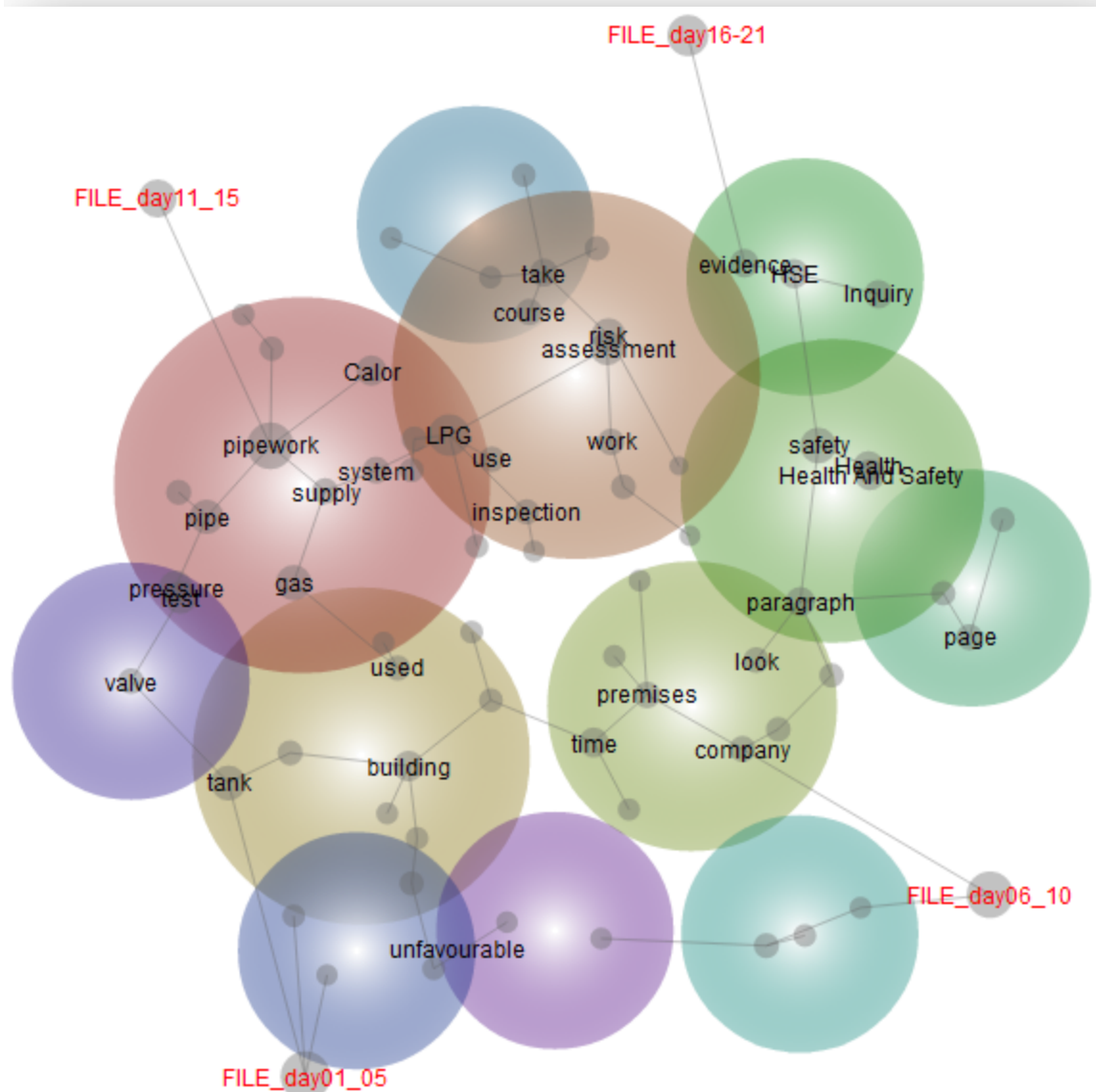
Default Theme Size Percentage (10–65): This option sets the size of themes (concept groupings) visible when the map interface opens initially. The theme size slider beneath the concept map also allows this parameter to be adjusted through the map interface.

Create Theme Hierarchy (Yes|No): If this option is enabled, an xml file is created that contains a nested hierarchy of the discovered concepts. The hierarchy is induced from the discovered hierarchy of concepts shown in the Leximancer concept map. This output file, called Table of Text Segment Classifications, is accessible under the Export tab.

You can also specify the Size (Width and Height) of the concept map to be produced in this interface.

Final Outputs

The Concept Map



See page 8 for an explanation of the Concept Map.

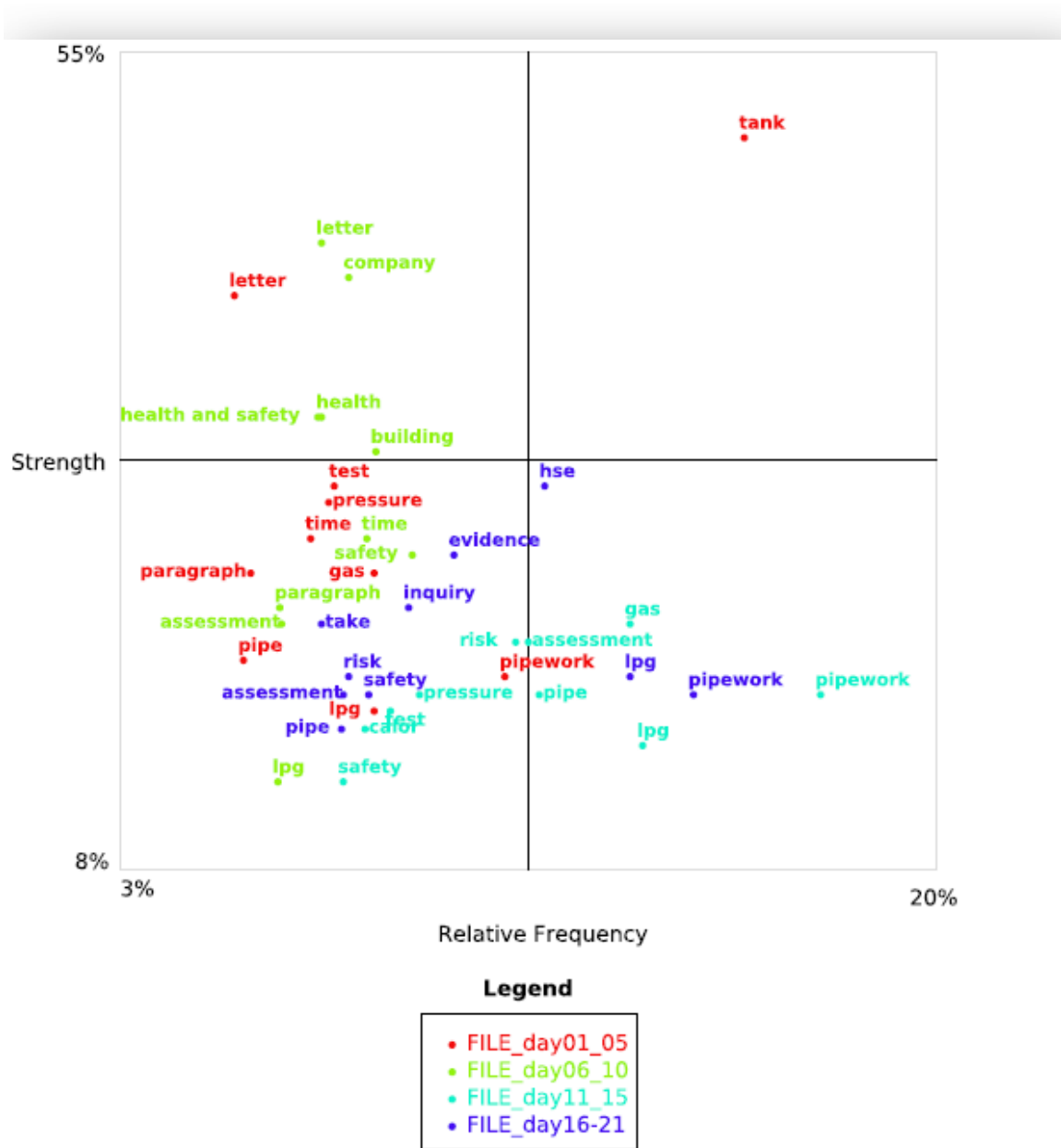
The Concept Cloud

The Insight Dashboard Report

The Insight Dashboard is a standalone pdf or html report that can be produced within a Leximancer project. The Dashboard report adopts a more quantitative focus than the concept map, and is designed to provide a quick understanding of the project results.

The report is not set up to be created automatically when you run a project, as not all projects are suitable for the dashboard report format. The Dashboard is designed for some sort of comparison, or difference analysis. Usually you must have created some (source document, tabular, speaker or folder) tags to use as categories for comparison in the Dashboard report. You can then use the Report to investigate the Attributes (independent variables) associated with certain Categories (dependent variables) in the data.

In this example, File tags were aggregated to create a category for each week of hearing transcripts after an explosion at a plastics factory. The Dashboard quadrant below shows the concepts most relevant to the discussion in each week (trends). The concepts towards the upper, right-hand (magic) quadrant were talked about most frequently in each week, and are most characteristic of the discourse during that time:

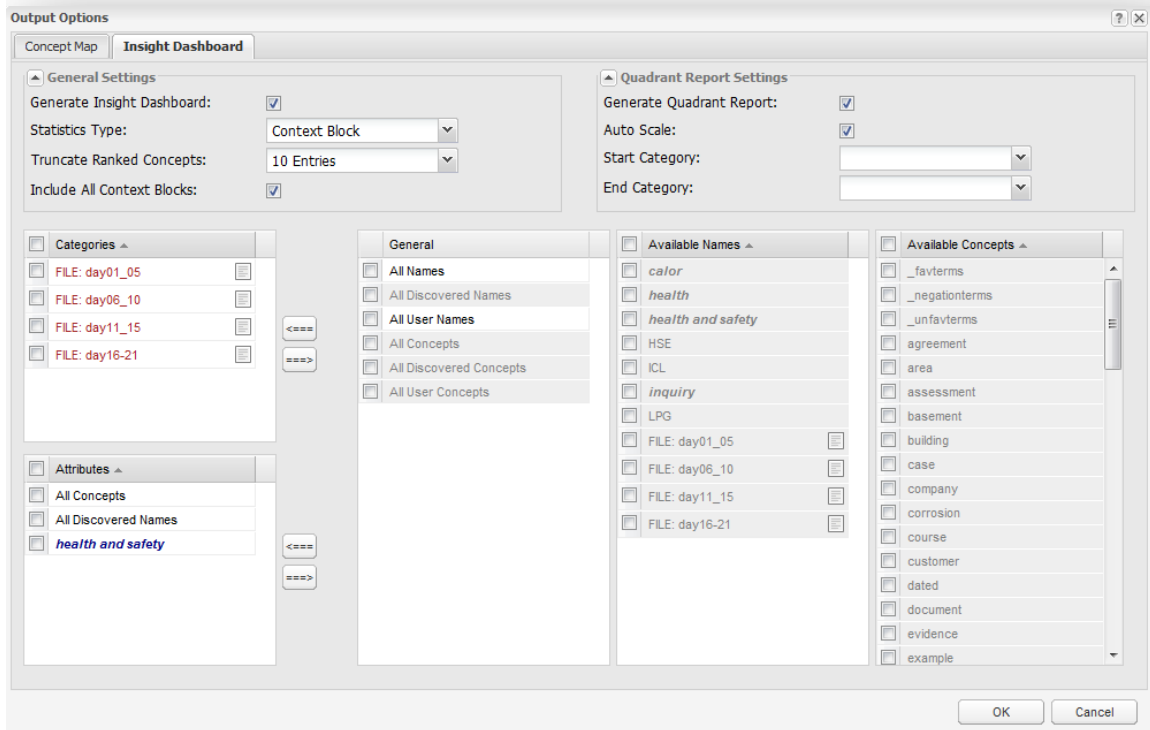


The Dashboard is produced as a standalone document so that it can be shared with others. Once created, it can be Downloaded from the Project Control header, or from the Data Exports beneath the main Project Control Panel. It can then be printed or sent to colleagues, collaborators and clients. The report does not require any knowledge of Leximancer to be understood.

The Dashboard is intended as a general purpose reporting tool, and the user can configure the options to suit their particular application or research question.

Configuring the Insight Dashboard Report

To configure a Dashboard Report, expand the Run Project Settings, and Edit the Project Output Settings. Click on the Insight Dashboard tab to reveal the following interface:



- ▶ Tick the **Generate Insight Dashboard** option.
- ▶ Tick the **Generate Quadrant Report** option. This includes a 'magic quadrant' graphic in the Report for identification of key issues.

► Set the **Statistics Type setting (Segment|Segment)**

Using the Segment Statistics setting, concepts are coded and classified at the level of the text segment. You can define a text segment (the coding resolution) using the Sentences per Block setting in the Pre-processing Options.

Using the Section Statistics setting, concept codes are applied to sections of the data. The definition of a document 'section' depends on the type of data:

- If you are processing Microsoft Word or pdf documents, then a document section is usually an individual file.
 - If you are processing delimited spreadsheet data, however, then a section is a single free text cell. This setting can be used to report the number of responses (whole comments) coded with particular concepts.
- The **Truncate Ranked Concepts** setting allows you to specify the number of Attributes reported for each Category. You can control the level of detail in the Report using this setting. The Do Not Truncate default does not apply any cut-off, but presents all the Attributes associated with each Category.
- The **Auto Scale** setting scales the axes in the Quadrant graphic. This spreads the Attributes in the Quadrant space for improved visibility.

- If this option is enabled, the Prominence statistics (described in the preamble to the Report) are expressed as relative probabilities.
- ▶ The **Include All Text Segments** setting introduces an optional final section to the Report that presents all text segments matching each concept query.
- ▶ Specify the **Categories** (or dependent variables) of interest. You can specify any number of Categories, though using upwards of 10 can clutter the Quadrant graphic.

Often the Categories will be (auto- or user-defined) Tags identifying different levels of variables, or groups for comparison, within the data. Gender (male or female) and tone (favourable or unfavourable) are examples of possible Categories.

- Select the Tags (or concepts) of interest from the Available Names (or Concepts) list(s), and use the appropriate left arrow to add these to the Categories list.

Note: If you wish to use Tags as Categories, you must add these to the Mapping Concepts list by hand in the Select Concept to Locate phase. This codes the data with the Tags so that they can be used in the Dashboard Report. It will also cause them to be clustered on the map among the topical concepts.

- ▶ Specify some **Attributes** (or independent variables) of interest.

You might use the emergent concepts as Attributes in the Dashboard Report. In this case, the Report will compare the concepts associated with each of your Categories

- Select the All Concepts wild card from the General list, then click the Attributes left arrow to use all the word-like concepts as Attributes. This wildcard includes all the entries in the Available Concepts list on the right. Alternatively, you can select individual concepts from the list and add them as Attributes by hand.
 - Select the All Discovered Names wild card from the General list, then click the Attributes left arrow to use all the word-like concepts as Attributes. This wildcard includes all the entries in the Available Names list in the centre. Alternatively, you can select individual names from the list and add them as Attributes by hand.
- ▶ When the settings have been configured, click Ok and run the final stages of processing to produce the Dashboard Report.

After clicking the Run Project button, you can download the Dashboard Report from the button beneath the main Project Control Panel.

The Dashboard can be downloaded in pdf or html format.

- The pdf version can be viewed in Adobe Acrobat Reader. The Quadrant graphic requires at least Acrobat version 9 to be displayed.

- The html version is useful if you wish to make edits to the Report. This version can be saved as a zipped folder (or archive) on your local machine. You may need to rename the folder (changing the extension from insight-dashboard-zip to insight-dashboard.zip) to allow it to be extracted or opened. Click on the insight-dashboard.html file to view the Report in a browser tab, or right click and select Open With to view the Report in another application (such as Microsoft Word).

The Dashboard is named after the project in which it was created. The header provides counts of the Total text Segments or Sections coded in the Report. It also presents counts for the number of Concepts and Categories specified.

Note: The preamble explains the various sections of the Report, and is included as part of Dashboard to allow others to understand its contents.

The Dashboard is designed to be interactive, and the Table of Contents includes clickable links, as do most many other sections of the Report. In the html version, you may have to hold down the <control> key and click to navigate the Report links.

Explanation of the Statistics in the Insight Dashboard Report

The Frequency axis on the Quadrant graphic represents a conditional probability. Given that a text extract comes from a particular Category, it gives the chance that the Attribute is coded in this text extract. This measures frequency of mention in the data, and is affected by the

distribution of comments across the Categories. The frequency score is in fact a log scale (so that it can be mapped on the quadrant).

The Strength score is the reciprocal conditional probability. Given that the Attribute is present in a section of text, it gives the probability that this text comes from that Category. Strong concepts distinguish the Category from others, whether or not the Attribute is mentioned frequently.

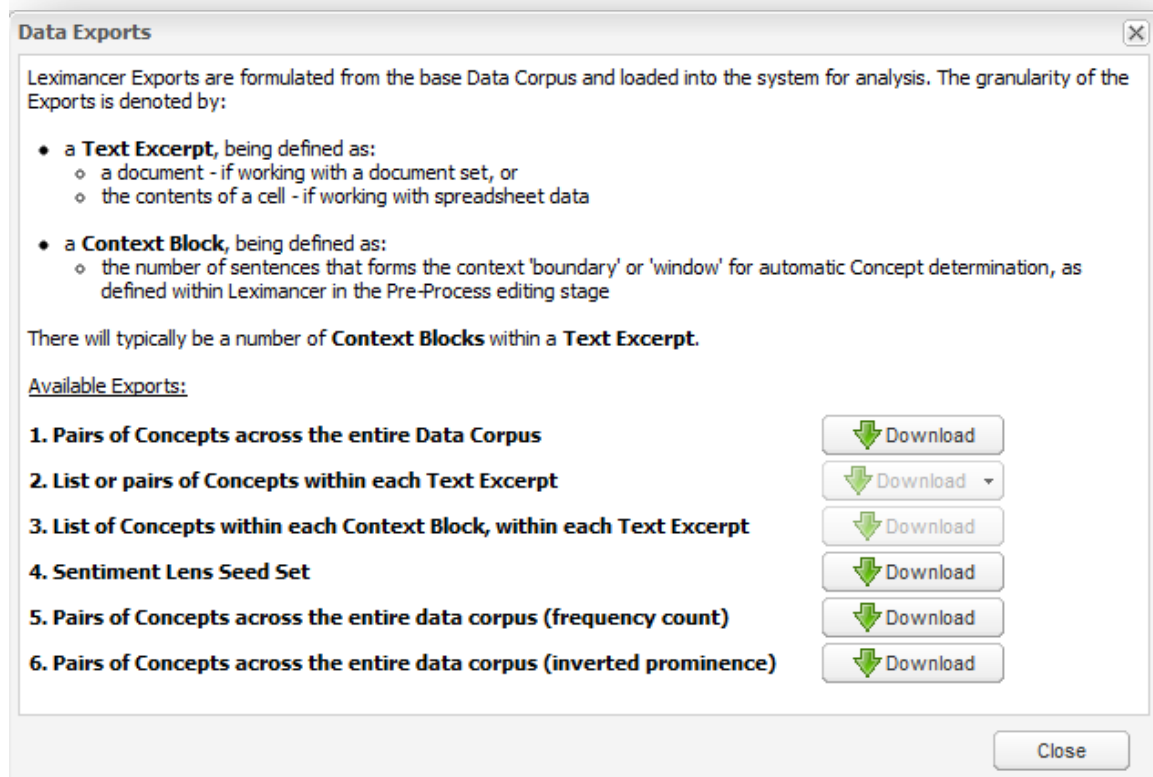
The percentages in the Ranked Concept for Categories lists match the quadrant coordinates. They reflect the same Strength and Frequency conditional probabilities.

The Prominence score combines the Strength and Frequency scores using Bayesian statistics. Prominence scores are absolute measures of correlation between category and attribute, and can be used to make comparisons over time.

Data Exports

Among the projects results, Leximancer produces several statistical reports. These can be exported for reporting or to allow further analysis in other applications.

Several reports are available for Download from the Data Exports button beneath the Project Control Panel. These include: (1) the Pairs of Concepts across the Entire Corpus (the co-occurrence matrix); (2) the List or Pairs of Concepts within each Text Excerpt; (3) the List of Concepts within each Context Block, within each Text Excerpt; and (4) The Sentiment Lens Seeds Set:



Hover your mouse over the Download button for a description of the level of detail in the each of the reports.

Leximancer Exports are formulated from the base Data Corpus and loaded into the system for analysis. The granularity of the Exports is denoted by:

- a **Text Excerpt**, being defined as:
 - a document – if working with a document set, or
 - the contents of a text cell – if working with spreadsheet data
- a **Context Block**, being defined as:
 - the number of sentences that forms the context 'boundary' or 'window' for automatic Concept determination, as defined within Leximancer in the Pre-Process editing stage

There will typically be a number of **Context Blocks** within a **Text Excerpt.**”

Available exports include:

- 1. Pairs of Concepts across the entire Data Corpus (the co-occurrence matrix)**

Downloads a comma delimited file displaying the matrix of co-occurrences between concepts. This file will open a spreadsheet program, including recent versions of Excel. It contains co-occurrence counts, listed for every concept pair combination, as well as x,y coordinates for each concept on the map, and the weight for each concept, which is the sum of its co-occurrences with all the other concepts:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	concept	x	y	weight	FILE_day0	pipework	FILE_day0	LPG	FILE_day1	FILE_day1	safety	gas	tank	pipe	assessment	risk	pressure	test
2	FILE_day01_05	0.58475	0.751729	13045	6522	557	0	429	0	0	205	428	900	329	95	87	391	396
3	pipework	0.397195	-0.3131	11772	557	2931	257	604	537	547	128	362	389	596	294	293	375	364
4	FILE_day06_10	-0.62752	0.716412	11721	0	257	5863	317	0	0	416	278	294	201	320	311	141	134
5	LPG	0.09467	-0.2616	9586	429	604	317	2409	473	382	211	406	319	193	197	196	138	136
6	FILE_day16-21	-0.45908	-0.83443	9413	0	537	0	473	4298	0	279	160	121	264	265	268	171	164
7	FILE_day11_15	0.488835	-0.76946	8177	0	547	0	382	0	3390	209	373	199	310	304	296	244	230
8	safety	-0.51635	-0.0999	7244	205	128	416	211	279	209	1546	106	69	25	186	171	35	37
9	gas	0.411764	-0.07581	7030	428	362	278	406	160	373	106	1687	239	220	150	146	193	182
10	tank	0.598727	0.23291	6869	900	389	294	319	121	199	69	239	1798	186	106	105	203	203
11	pipe	0.531192	-0.21937	6593	329	596	201	193	264	310	25	220	186	1633	129	123	295	282
12	assessment	-0.20232	-0.34456	6528	95	294	320	197	265	304	186	150	106	129	1428	1384	89	77
13	risk	-0.20993	-0.36558	6422	87	293	311	196	268	296	171	146	105	123	1384	1405	84	72
14	pressure	0.610098	-0.11463	5948	391	375	141	138	171	244	35	193	203	295	89	84	1326	1247
15	test	0.606003	-0.09522	5799	396	364	134	136	164	230	37	182	203	282	77	72	1247	1285
16	Health	-0.59202	-0.04492	5216	137	55	346	94	151	126	983	54	28	17	141	130	11	10
17	Health And Safety	-0.59118	-0.031	5185	135	55	344	94	150	122	983	53	28	16	139	129	11	10
18	building	0.291737	0.274931	5040	300	246	387	187	124	163	80	165	168	183	65	65	69	60
19	Calor	0.199357	-0.39914	4886	288	376	162	179	233	219	97	98	175	129	90	88	82	82
20	time	-0.02617	0.307556	4637	377	171	379	131	185	130	78	91	109	128	85	84	94	92
21	paragraph	-0.42566	0.153561	4484	334	238	318	160	241	112	86	71	37	113	45	45	122	121
22	HSE	-0.54317	-0.39587	4473	127	160	180	226	398	101	179	70	27	32	87	93	38	38
23	evidence	-0.46439	-0.43264	4153	148	176	169	128	331	152	101	60	56	79	62	60	48	49
24	Inquiry	-0.67567	-0.32863	4116	172	110	162	156	302	181	92	52	33	22	56	56	31	31
25	take	-0.12647	-0.49505	3948	151	190	147	123	254	158	98	100	78	96	96	93	68	69
26	unfavourable	-0.39444	-1.01994	3439	203	177	152	124	165	136	71	101	65	92	76	76	53	45

2. List or pairs of Concepts within each Text Excerpt

This CSV or XML file contains classifications of text excerpts. Each block (row) indicates:

- the file and section number for the text excerpt;
- the surrogate id for viewing this context block;
- the tabular input data, a field for each quantitative column in the input data;
- for tabular input, a field whose value indicates which text column in the row is indicated. This is for situations where there are multiple text cells in a row of input data;
- a nested block containing classifications assigned to this text excerpt, with occurrence counts and cumulative weights

This output is designed for import of document or text excerpt classifications into a database. This output file is not sparse.

3. List of Concepts within each Context Block, within each Text Excerpt

This tab delimited text file contains one row for each context block. Each row indicates:

- the file, text excerpt, and starting sequence number of the context block;
- the html surrogate link for viewing this context block in a browser;
- the presence or absence of a concept or tag in the context block.

There is one column for each concept or tag class. As a result, this table is very sparse.

This import is specifically designed for input into statistics or data mining packages for building models such as: decision trees, rule sets, logistic regression, or market basket analysis. There is a setting in the Classifications Settings tab to generate either real valued or binary values.

4. Sentiment Lens Seed Set

This export will open a list of all the Sentiment seeds remaining after they have gone through the tuning process of Sentiment Lens.

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<seeds>
  ▼<concept value="_unfavterms" kind="WORD,USER">
    <term value="abuse" kind="WORD" positive="true"/>
    <term value="angry" kind="WORD" positive="true"/>
    <term value="bad" kind="WORD" positive="true"/>
    <term value="blame" kind="WORD" positive="true"/>
    <term value="castigated" kind="WORD" positive="true"/>
    <term value="complicated" kind="WORD" positive="true"/>
    <term value="condemn" kind="WORD" positive="true"/>
    <term value="cutting" kind="WORD" positive="true"/>
    <term value="delay" kind="WORD" positive="true"/>
    <term value="delays" kind="WORD" positive="true"/>
    <term value="detract" kind="WORD" positive="true"/>
    <term value="difficult" kind="WORD" positive="true"/>
    <term value="disappointed" kind="WORD" positive="true"/>
    <term value="disappointing" kind="WORD" positive="true"/>
    <term value="disapproval" kind="WORD" positive="true"/>
    <term value="dissatisfied" kind="WORD" positive="true"/>
    <term value="dubious" kind="WORD" positive="true"/>
    <term value="fail" kind="WORD" positive="true"/>
    <term value="failed" kind="WORD" positive="true"/>
    <term value="failing" kind="WORD" positive="true"/>
    <term value="fails" kind="WORD" positive="true"/>
    <term value="failure" kind="WORD" positive="true"/>
    <term value="fault" kind="WORD" positive="true"/>
    <term value="fear" kind="WORD" positive="true"/>
    <term value="fearful" kind="WORD" positive="true"/>
    <term value="frustrated" kind="WORD" positive="true"/>
```

This file can be Uploaded directly into other projects via the Concept Seeds Settings Edit interface.

Logbook Exports


In the map interface, Leximancer allows complex records of queries to be stored and exported. Find a particular query of interest, and its example text, and add it to the logbook:

← Themes Concepts Thesaurus Pathway **Query** Summary Log

WORD:gas AND WORD:explosion Search

Export Page Export All Log All

Result

[/ICL Explosion Inquiry/Day01_02_July~7.html 1_2098](#) Add to Log | Full Text 

"I did not notice the smell of gas prior to the explosion.

[/ICL Explosion Inquiry/Day01_02_July~7.html 1_2102](#) Add to Log | Full Text
[Tags](#)

THE CHAIRMAN: Just before we go on to *Mr Moir*, I am not quite clear what this last witness really is saying about the gas smell. First of all, he says he smelt gas in the car park which I think was before the explosion.

[/ICL Explosion Inquiry/Day01_02_July~7.html 1_2145](#) Add to Log | Full Text
[Tags](#)

"While waiting on the police to inform us of the escape route, I smelt a very strong smell of *Calor* or propane gas. I was worried about the smell as a spark could cause another explosion.

[/ICL Explosion Inquiry/Day05_09_July~1.html 1_138](#) Add to Log | Full Text
[Tags](#)

If I may, *Mr Ives*, a BLEVE is where not only is the gas igniting but it is also because it is being heated up turning from liquid into vapour at a significant rate which increases the ferocity of the burning or explosion; is that right?

[/ICL Explosion Inquiry/Day08_15_July~2.html 1_285](#) Add to Log | Full Text
[Tags](#)

This meant that when the flame had gone out the gas would have continued to escape and as LPG gas is heavier than air, then this would have built up creating a pocket of gas and would have caused an explosion when lit.

Go to the 'Logbook' tab to review all the logged text:

← :s Concepts Thesaurus Pathway Query Summary **Logbook**

Logbook

.

Matching Log Entries: 3 [export page](#) [export all](#)

1. /ICL Explosion Inquiry/Day01_02_July~7.html/1/1_2098

I did not notice the smell of gas prior to the explosion.

Author: julia **Date:** Oct 23, 2011 5:56:47 AM **Query:** [Edit](#) | [Delete](#)

2. /ICL Explosion Inquiry/Day01_02_July~7.html/1/1_2145

"While waiting on the police to inform us of the escape route, I smelt a very strong smell of *Calor* propane gas. I was worried about the smell as a spark could cause another explosion.

Author: julia **Date:** Oct 23, 2011 5:56:50 AM **Query:** [Edit](#) | [Delete](#)

3. /ICL Explosion Inquiry/Day08_15_July~2.html/1/1_285

This meant that when the flame had gone out the gas would have continued to escape and as LPG gas is heavier than air, then this would have built up creating a pocket of gas and would have caused an explosion when lit.

Author: julia **Date:** Oct 23, 2011 5:56:52 AM **Query:** [Edit](#) | [Delete](#)

From here, you can export either just the current page, or every entry in the logbook:

← :s Concepts Thesaurus Pathway Query Summary **Logbook**

Logbook

.

Matching Log Entries: 3 [export page](#) [export all](#)

1. /ICL Explosion Inquiry/Day01_02_July~7.html/1/1_2098

"I did not notice the smell of gas prior to the explosion.

Author: julia **Date:** Oct 23, 2011 5:56:47 AM **Query:** [Edit](#) [Delete](#)

2. /ICL Explosion Inquiry/Day01_02_July~7.html/1/1_2145

"While waiting on the police to inform us of the escape route, I smelt a very strong smell of *Calor* or propane gas. I was worried about the smell as a spark could cause another explosion.

Author: julia **Date:** Oct 23, 2011 5:56:50 AM **Query:** [Edit](#) [Delete](#)

3. /ICL Explosion Inquiry/Day08_15_July~2.html/1/1_285

This meant that when the flame had gone out the gas would have continued to escape and as LPG gas is heavier than air, then this would have built up creating a pocket of gas and would have caused an explosion when lit.

Author: julia **Date:** Oct 23, 2011 5:56:52 AM **Query:** [Edit](#) [Delete](#)

Your exported logbook entries will pop up in a new window of your browser. Pop-ups need to be enabled for this to occur. Leximancer will have logged the Document ID, Folder/s, Author, Date and Time, Query Terms, and the example text:

Logbook

Matching Log Entries Found: 3

1. /ICL Explosion Inquiry/Day01_02_July~7.html/1/1_2098
"I did not notice the smell of gas prior to the explosion.

Author: julia
Date: Oct 23, 2011 5:56:47 AM

2. /ICL Explosion Inquiry/Day01_02_July~7.html/1/1_2145
"While waiting on the police to inform us of the escape route, I smelt a very strong smell of Calor or propane gas. I was worried about the smell as a spark could cause another explosion.

Author: julia
Date: Oct 23, 2011 5:56:50 AM

3. /ICL Explosion Inquiry/Day08_15_July~2.html/1/1_285
This meant that when the flame had gone out the gas would have continued to escape and as LPG gas is heavier than air, then this would have built up creating a pocket of gas and would have caused an explosion when lit.

Author: julia
Date: Oct 23, 2011 5:56:52 AM

Section 5: Example Advanced Techniques

This section is intended to be a guide for some of the styles of analysis that are possible with the Leximancer system. However, the system is designed to be a general-purpose tool, and the confident user is encouraged to experiment.

Specific tutorials include: Manual Concept Seeding; Profiling; Profiling Using Tag Categories; Extracting a Social Network; Analysing Transcripts; Analysing Spreadsheet Data

1. Manual Concept Seeding

You can seed your own concepts in the User-Defined Concepts tab if you Show the Generate Concepts Settings, and click Edit Concept Seeds Settings. Move to the User-Defined Concepts tab. If you create new manual seeds, thesaurus definitions will be extracted for these and any automatically-identified concepts during the Generate Thesaurus phase. You can also create your own Manual Tags. These act like lists of keywords, or fixed dictionaries – they are not modified by the learning process.

In many cases, an automatically generated map may contain concepts that are irrelevant to your interests or concepts that are similar (such as thought and think), or the map may be lacking concepts that you wish to locate in the text. You can ‘seed’ your own concepts prior to running the Thesaurus Learning phase by clicking on Edit Concept Seeds in the Project Control Panel. There you can add, edit, merge or delete concepts in order to produce cleaner or more tailored concept maps. If you add (User-Defined) concept seeds manually, thesaurus definitions for these will be extracted from the text along with any automatically-extracted concept seeds. You can also create your own User-Defined Tags and

these act like lists of keywords, or fixed dictionaries – they are not modified by the learning process.

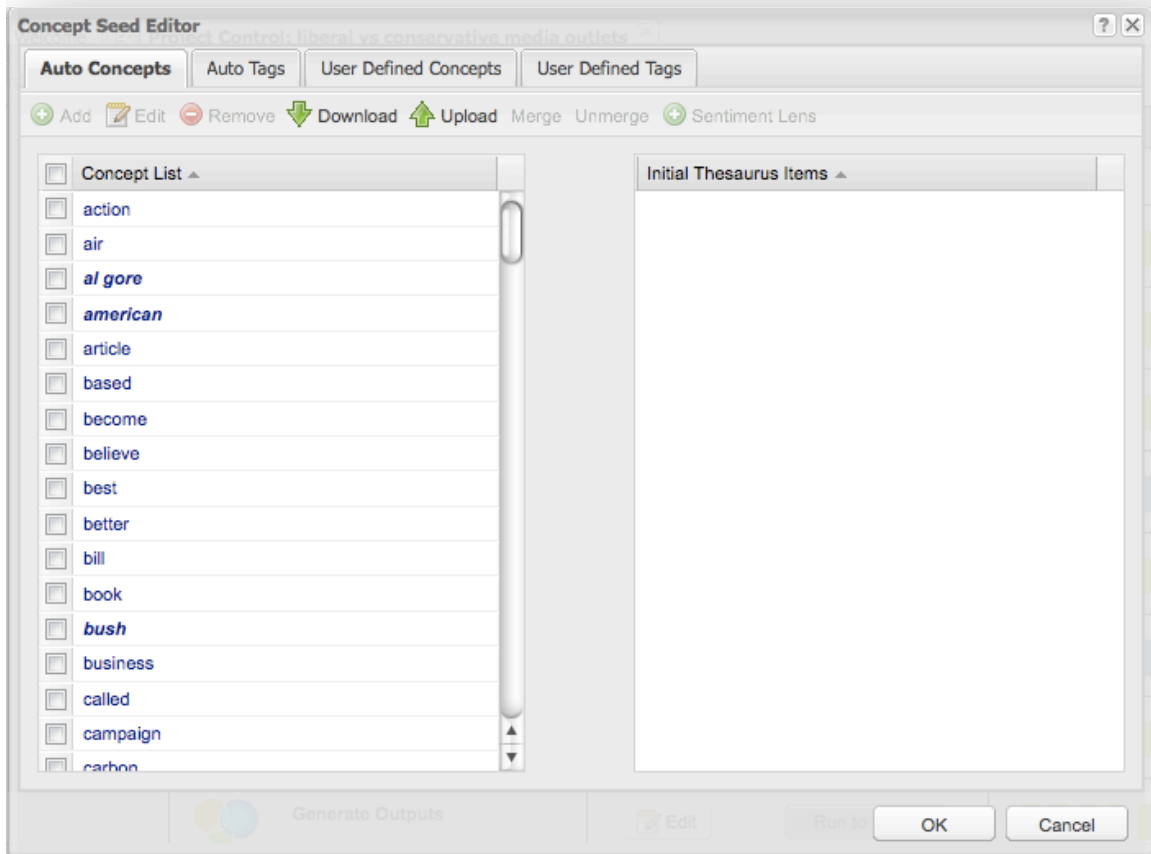
Configuring Manual Concept Seeding

You can Add, Edit, Remove, Merge or Unmerge concepts in Generate Thesaurus, then Edit Concept Seeds. This is important for a number of reasons:

- Among the Auto Concepts (those automatically–extracted by the software), there may be concepts seeds (such as think and thought) that are similar, or seeds that are not of interest to you. You can Merge similar concept seeds into a single concept, or Remove concepts that you do not wish to see on the map
- in the User Defined Concepts tab, you may wish to Add your own seeds (such as violence) to search for concepts that you are interested in exploring, or create categories (such as dogs) containing the specific instances found in your text (such as hound and puppy).

In order to modify concepts automatically extracted by Leximancer, Generate Concept Seeds (the node prior to Generate Thesaurus) needs to have been run. If this node has been run (i.e. if you have previously generated a map), it will be green and it's status will say Ready. If not, click on the Generate Thesaurus button to run this stage.

After Generate Concepts has been run, showing the Generate Concepts Settings, and then clicking on Edit Concept Seeds Settings reveals the following interface:

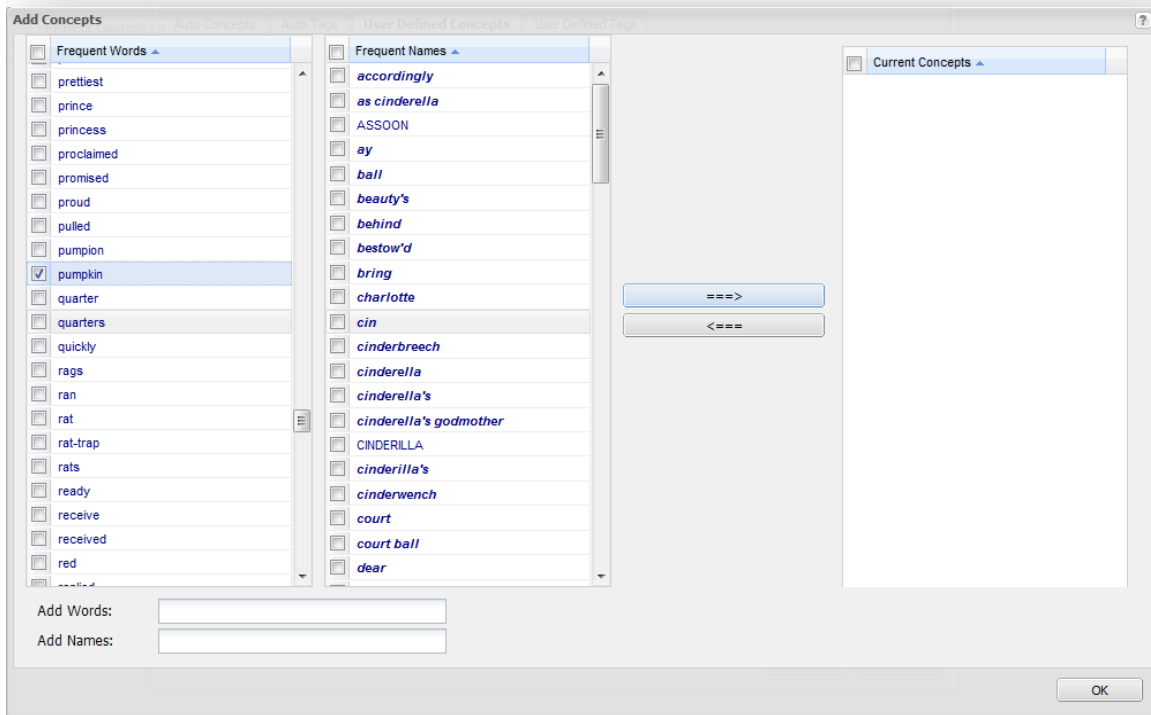


There are tabs for editing the Automatic Concepts (concepts identified by Leximancer) and User Defined concepts (concepts that you wish to define yourself).

At this stage, only the central key word for each concept has been identified. The learning of associated terms and their weightings occurs in the following Thesaurus Learning phase.

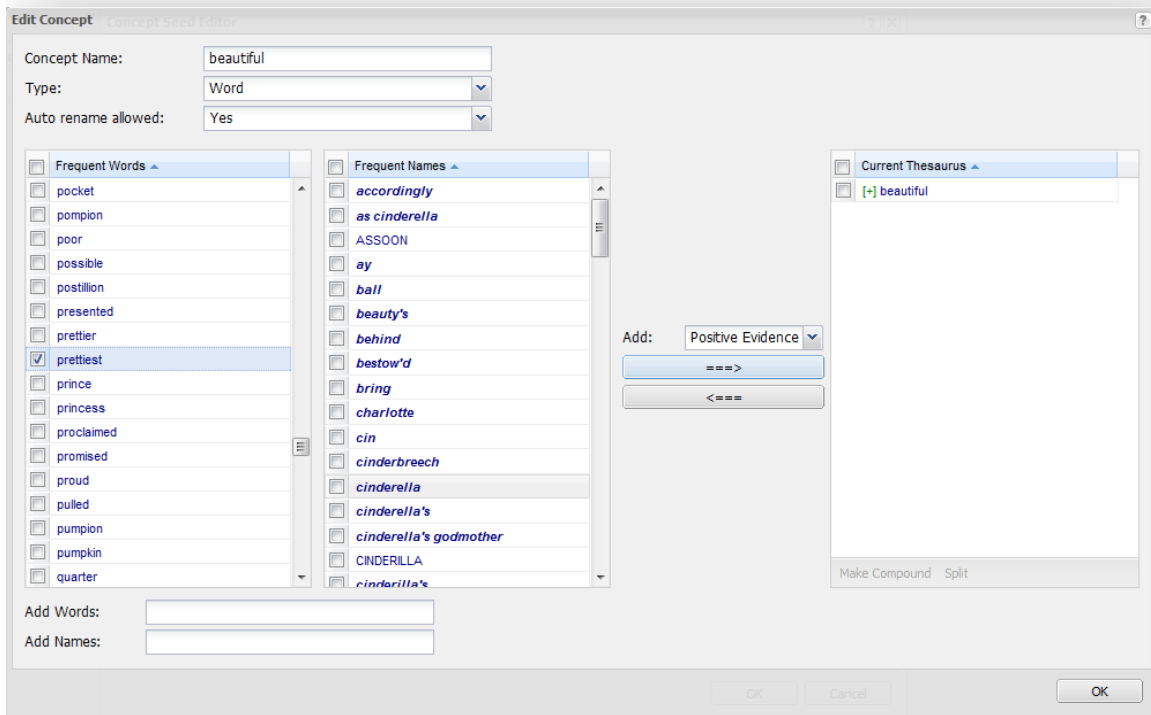
In the interface above, you can select and Merge or Delete concept seeds (note: holding down <ctrl> while clicking allows you to select multiple items).

Open the User-Defined Concepts tab and click on Add to open the Add Concepts interface:



You can select terms from the lists of Frequently-occurring Words or Names on the left, or type in your own name for the new concept using the text boxes beneath the lists. Use the arrow buttons to move the term to the right to create your concept. Click on Ok to return to the Concept Seed Editing interface.

To add thesaurus terms to the concept's early definition, select the concept in the Auto- or User-Defined concepts tab, and click on Edit. The following interface will open:



Here you can add terms strongly-related to your concept. For example, if you are interested in finding sections in your text containing violence, create the concept 'violence' and add any terms from the Frequent Words or Names list above that you think indicate a violent act. Only use words that fairly unambiguously indicate this concept in the text. Leximancer will automatically find additional terms from the text during the Thesaurus Learning phase, so you don't have to know all the relevant words in advance. Click on Ok to save your changes and exit.

When you return to the Concept Seed Editing interface, you can also create and edit Tag categories. These are concepts for which no associated terms will be learned by Leximancer. This useful if you want to compare groups in the data (using file or folder tags) or perform a simple keyword search for terms.

Click on OK to close the Concept Seed Editor and return to the main Project Control Panel. Click on Run Project to run the remaining phases

on default settings. Click on Concept Map to view the map containing the new concepts.

2. Profiling

This function is **not** the same as automatic concept discovery. The aim here is to discover new concepts during learning which are relevant to the concepts defined in advance, either in the Automatic- or in the User-Defined Concepts tabs. For example, this setting would allow you to extract the main concepts related to stem cell research from a broader set of documents. Concept profiling settings can be found under Edit for the Thesaurus Settings in the Generate Thesaurus stage. Note that Tags do not take part in this process automatically. If you have Tag categories, folder tags for example, which you wish to profile, you must turn on the Learn From Tags option in the Thesaurus Settings. It is important to understand that although these discovered concepts are seeded from words that are relevant to the prior concepts, they are then learned as fully-fledged independent concepts. As a result, the map will contain some peripheral areas of meaning that do not directly contain the prior concepts. Contrast this with the Required Concepts function described below.

The profiling function has three alternative behaviours: ALL, ANY, and EACH. You can ask for the related concepts to be relevant to many of the prior concepts, and thus follow a theme encompassed by all the prior concepts - this is the ALL option, and resembles set intersection. Alternatively, the discovered concepts need only be related to at least one of the prior concepts - this is the ANY option, which is similar to set union. The EACH option discovers equal fractions of profile concepts for each predefined concept, and these concepts show what is peculiar to

each predefined concept. The EACH option is very useful for enhanced discrimination of prior concepts.

If you wanted to extract the main concepts related to stem cell research from a broader set of documents, for example, you could disable Automatic Concept Identification in Concept Seeds Edit in the Generate Concepts stage. Instead you would create user-defined seeds for multiple simple concepts that encompass the scenario. You might seed concepts such as 'research', 'ethics', 'debate' and so on. Keep the seed words for each concept simple, and don't try too hard to restrict the seeds of each concept to just the topic you are after. In this instance we will be considering the intersection of all these elements. Expand the Generate Thesaurus Settings, and Click on Edit for Thesaurus Settings. Specify a quota of concepts to be discovered in the Concept Profiling options. Choose to discover several profiled concepts per prior concept. Select Concepts in ALL as the operator.

If you are attempting to discover the network of associated names around a name or a scenario, you can choose to only discover name-like concepts during profiling. You should try the Social mapping algorithm first for this style of map.

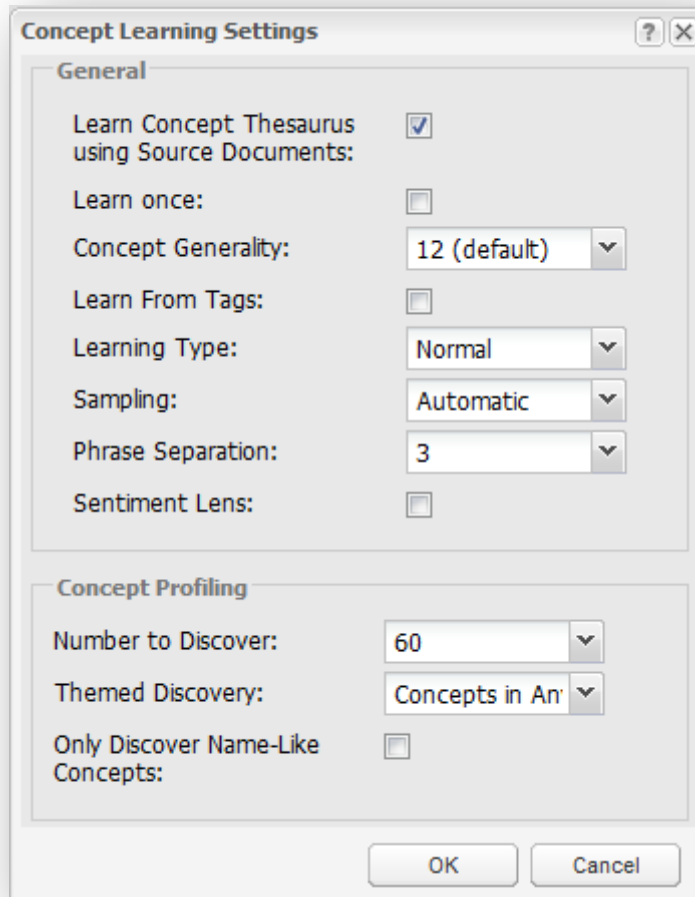
It is important to understand that although the profiled concepts are seeded from words that are relevant to the prior concepts, they are then learned as fully-fledged, independent concepts. As a result, the map will contain some peripheral areas of meaning that do not directly contain the prior concepts. Contrast this with the Required Concepts function described below.

Configuring Concept Profiling

Leximancer will generally extract the main concepts that occur within your documents. However, in some instances you may be interested in inspecting certain aspects of your text in more detail. For example, given 1000 newspaper articles, you may only be interested in the events containing references to violence. In such cases, you can use profiling to extract concepts that are specific to violence, rather than those that occur in all the articles that you have been given.

To profile an issue:

- Select your data files as usual, expand the Generate Concepts Settings.
- Click on Edit Concept Seeds Settings to untick the Automatically Identify Concepts box, because you don't want any concepts present other than the ones you are interested in profiling.
- Run the Generate Concepts stage.
- Expand the Generate Thesaurus Settings, and click on Edit Concept Seeds (as discussed above– Configuring Concept Editing) to create one or more concepts to profile. In this case, a concept called pipework is of interest for profiling.
- Then click on Thesaurus Settings, and enter the number of concepts that you would like to see on the map related to this theme (60 in this example) in the Concept Profiling section:



- Click OK to exit the Thesaurus Settings. Click on Run Project in the main Project Control Panel to run all remaining stages of processing, then click Concept Map to reveal the profile map.

The concept map will contain profiled concepts specific to your area of interest. The map below is a profile of the Pipework concept in the Explosion Enquiry (hearing transcript) data set:

Folder and Filename Tag Generation

Filename or folder tagging converts categories of text (or variables) into explicit tags in the text, and hence into concepts on the map. The names of all parent folders of a file, and optionally the name of the file itself, are embedded as separate tags on each sentence within a file. For example, a file called “patient1” inside the folder “control1” below the data folder would have separate tags [control1] and [patient1] inserted in each sentence (if folder and filename tags are applied). These tags are imported into the Automatic Tags Tab so long as Automatic Concept Identification is enabled. Note that the hierarchical order of the folders is not important, and is not carried into the Leximancer map. If you have a folder called ‘Control’ under several different branches, then this becomes one concept called ‘Control’ on the map, and this can be a powerful feature for freeing up the exploration of category co-variances.

Comparing Names or Categories based on Shared Semantics

The strategy here is to compare categories based on shared concept usage. The categories are usually Tags, often generated using the folder tags setting. You might have several submissions on an issue from different people and organizations, and you want to know how they focus on the major topics. Enabled the Apply Folder Tags setting will cause each file to generate a tag. This is also very useful for looking at trends over time. Simply put documents into folders by month or by year etc. Use automatic concept identification to generate a set of concepts that characterise the whole text collection, then complete the map, making sure that you add the tags of interest to the Mapping Concepts list in the Project Output Settings in the Run Project stage. Make sure that you are using the Topical network algorithm to cluster the concept map. The map

should then show the tag categories distributed around the concepts. If the locations of the tags relative to the concept field show little repeatability, then you should conclude that the tag categories are difficult to differentiate based on the global concept selection. Essentially, they all address most of the same global concepts to similar degrees. This is a result in itself, but if you actually wish to discriminate between tag categories, see the section on Discrimination of Concepts, Names, or Categories based on Semantics.

Configuring Folder and Filename Tags

If you are going to analyse a set of multiple text files using Leximancer, you should consider making use of the Apply Folder Tags function. Essentially, if the name of each file is a category of interest, or you can group the files into folders describing categories of interest, this feature is a simple way to add a lot of power to your analysis.

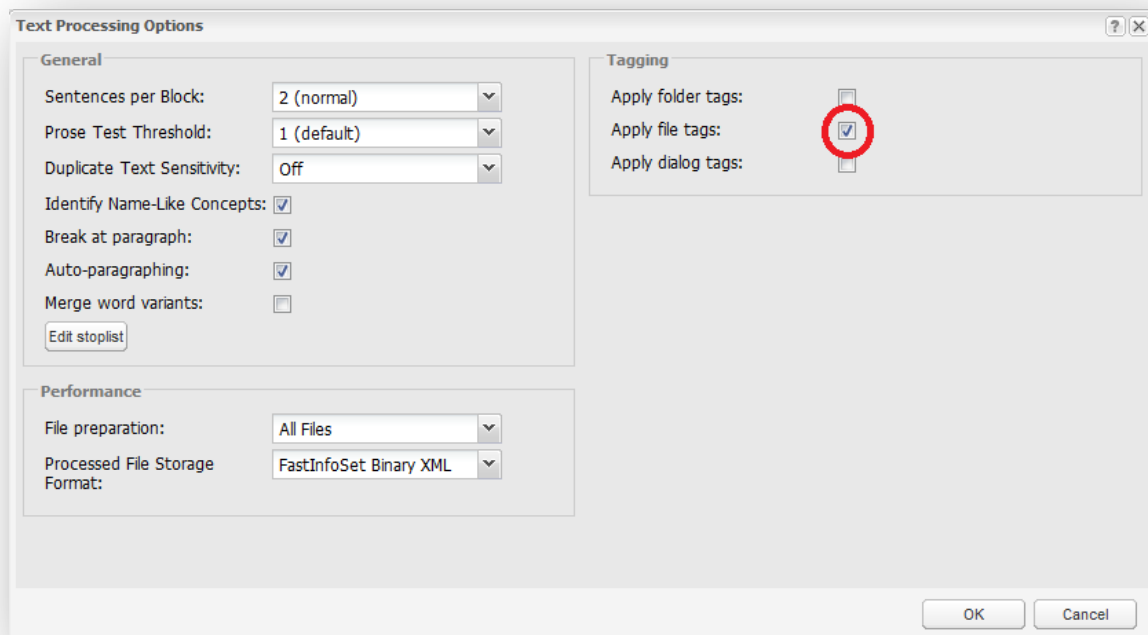
For example:

- break up a book into various chapters (one file per chapter) to allow you to explore what topics or characters appear in each chapter,
- naming letters or reports – make each file name indicate the person or organisation who wrote the document to let you see each of these bodies on the map,
- group newspaper articles into folders by newspaper.

Leximancer can create a category called a Tag for each folder and / or file name in your data documents. You can create multiple levels of folders under your parent (top-level) data folder. For example, you can create a folder for each newspaper, and under each of those a folder for each

month, and under each of those a folder for each journalist. You would place the text file for each article in the appropriate folder at the bottom of this tree. Leximancer can then create a Tag for each folder at each level of the tree.

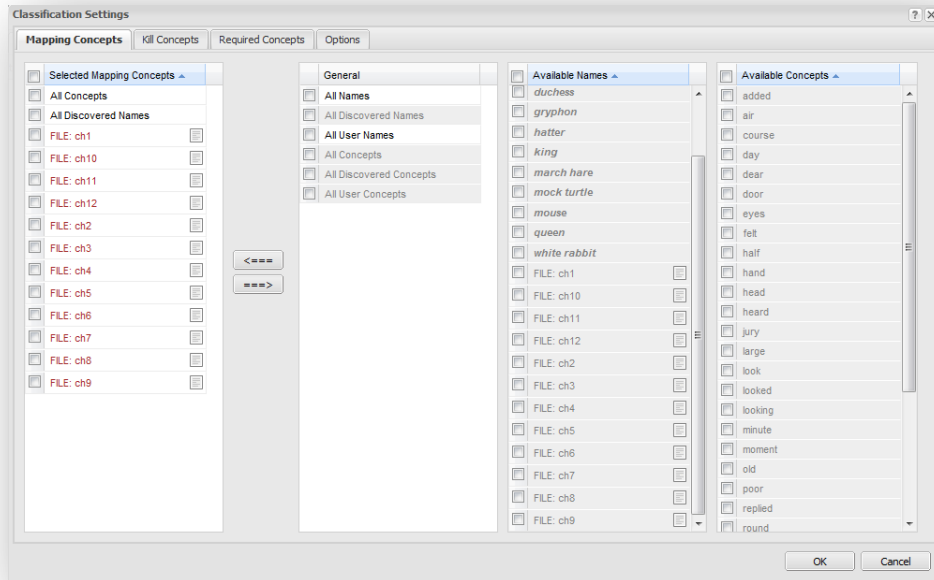
To enable this function, expand the Generate Concepts Settings and click on Edit Text Preprocessing Settings to bring up the interface below. In the Tagging Options, select Apply folder tags or Apply file tags:



In this example, the chapters comprising the Alice in Wonderland story were loaded into Leximancer as separate documents. When the Apply File Tags option is enabled therefore, a tag is automatically created to represent each the chapter in the story.

Click OK to exit the Preprocessing Settings.

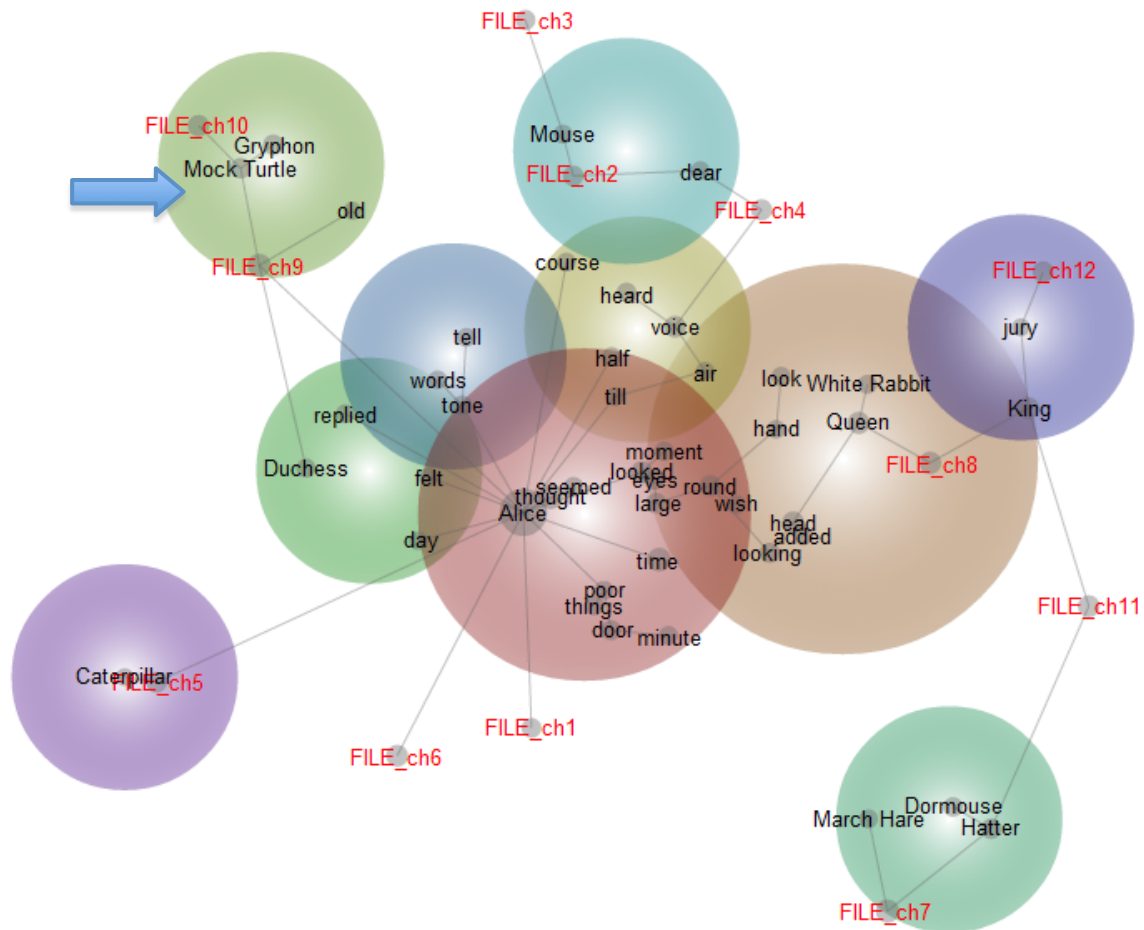
Run the Generate Concepts and Generate Thesaurus stages of processing. Expand the Run Project Settings, and Edit the Concept Coding Settings. Use the left-hand arrow button to add the file tags to the list of Mapping Concepts:



Click Ok, and then click on Run Project in main Control Panel to complete the final phase of processing. Click on Concept Map when the project is complete.

The concept map now includes the chapter file tags, and the concepts are clustered around these according to their relationships. Concepts coming from the content of a particular chapter will tend to settle near that chapter's file tag in the map space.

You can explore the topics characteristic of a chapter by clicking on a file tag. A ranked list of related topics is revealed in the panel on the right. These are the concepts that are coded into the chapter frequently:



In the example above, the Mock Turtle and Gryphon are clustered near Chapter 10 file tag, indicating that they are fairly specific to this chapter.

Filename and folder tags are useful if you wish to explore similarities or differences between various conditions. For example, placing speeches by various politicians into folders according to the party to which they belong will allow you to explore the views of the party, whereas placing files into folders corresponding to different periods of time will allow you to explore temporal differences.

Discrimination of Categories based on Semantics

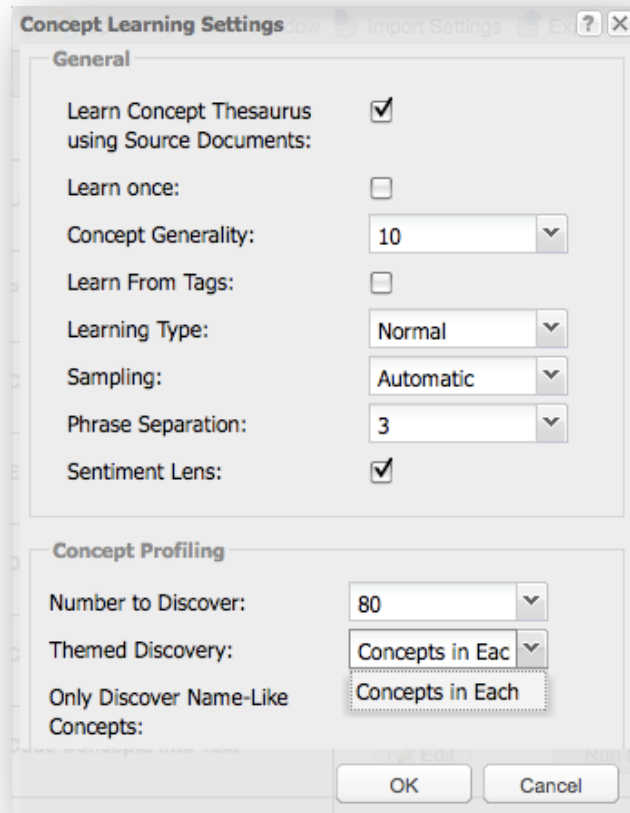
The goal here is to generate concepts that distinguish each element of a set of prior categories. For example, we might want to know what attributes discriminate failing students from successful students. First, ensure that the categories are defined either in the manual- or automatic- tags tabs within the Concept Seeds Editor in the Generate Concepts Settings. Be sure to delete any other concepts that are not required for profiling. Endorse the Learn From Tags option in the Thesaurus Settings in the Generate Thesaurus stage, so that the categories can be used in the learning phase to generate related concepts. Nominate how many concepts you wish to discover to profile your categories, and use the Concepts in EACH setting to discover concepts that distinguish your categories from one another. To create the map, make sure the tag categories and the discovered concepts are in the Mapping Concepts list in the Concept Coding Settings in the Run Project stage. Also make sure that you use the Topical network mapping setting.

The final map will show the tag categories in proximity to their discriminating discovered concepts. Be aware that the discovered concepts in this case do not characterise the whole text. In fact they are quite specific to the tag categories, and do not necessarily cover the major themes of the whole data set at all.

In order to profile tags:

- Select your data files as usual. In this example, the data consists of 4 folders containing speeches by members of Australian political parties concerning the stem cell debate.
- Expand the Generate Concepts Settings, and Edit the Text Processing Settings. Use the Apply Folder Tags option to create a tag for the speeches of each political party.

- Edit on the Concept Seeds Settings and disable the radio button near the top. Here we are interested in profiling the tags, and don't want any automatic- or user-defined concepts present.
- Run the Generate Concepts stage.
- Expand the Generate Thesaurus Settings, and Edit the Concept Seeds Settings open. In the Auto Tags tab you will see that a tag has been created for each data sub-folder. Delete the Nationals and Independents tags, so that you are only left with tags for the Labour and Liberal parties. Click Ok.
- Edit the Thesaurus Settings, and tick the Learn From Tags option. In the Concept Profiling section, enter the number of concepts that you would like to see on the map related to the two tags (60 in this example). Since we want to discover concepts that distinguish the party tags, use the Concepts in EACH operator this time:



- Click Ok, then expand the Run Project Settings, and Edit the Concept Coding Settings. Move the Folder tags of interest into the Mapping Concepts list on the left (to allow them to be shown on the map).
- Click Ok, then click the Run Project button in the main Project Control Panel to complete the project, and then click Concept Map to see the party profile map.

The concept map will contain concepts that distinguish the tags from one another. The map below shows the clusters of concepts that contrast the arguments made by the Australian Labour and Liberal parties on the subject of the stem cell debate:

After running Generate Concepts, tidy-up the automatic concepts in the Edit Concept Seeds Settings in Generate Thesaurus. For instance, you might remove any unwanted proper nouns, and add any missing names that you want to watch.

Next, go into the Thesaurus Settings and enable the Concept Profiling function. Select the Concepts in ANY operator, and choose to discover one concept per prior name. You can increase the number of discovered concepts if you want a richer map.

Run the Generate Thesaurus phase, then expand the Run Project Settings and go into the Concept Coding Settings. Make sure the tags, names and concepts of interest are included in the Mapping Concepts list. If not, use the left-hand arrow to add them to the list.

The resulting map will show a network of names intertwined with concepts that describe and mediate the relationships.

For a more traditional and constrained social network which uses structural variables:

- Edit the Concept Seeds Settings to select only names as described above, then run this phase.

- In the Edit Concept Seeds, manually seed some concepts as structural variables.
- Run the Thesaurus Learning phase.
- Open the Concept Coding Settings, and put only the name-like concepts in the Mapping Concepts list. Place the desired structural concepts in the list of Required Concepts, then run the remaining stages.

The result will be that only text segments that contain one of the Required Concepts will be mapped. Consequently, the map will show a network of names based on relationships that involve at least one of the required concepts

5. Analysing Transcripts

Transcripts of meetings, interviews and focus groups can be analysed in Leximancer as normal text, and if you group the interviews into files and folders, you can use Folder Tagging (Chapter 14) to enhance your analysis. Moreover, if your transcripts are in plain text or Microsoft Word and are suitably formatted, Leximancer allows you to select, ignore, or compare all the utterances of each distinct speaker. To allow the program to identify the speaker of any text segment, they must be identified in a certain way, and a new speaker label must be inserted whenever any new speaker begins. The format requires dialogue markers which are at the start of a paragraph; use upper case first letters for each constituent term; are made up of a maximum of three terms; and end in a colon followed by a space. For example:

Interviewer: So what does your son like to do for leisure, Susan?

Susan: Every Friday he plays uh ten pin bowling with the oldies. He's not bad either.

Alan: Oh yes, he excels at ten pin bowling, he's one of the better players there.

Interviewer: And do you have any plans for travel coming up?

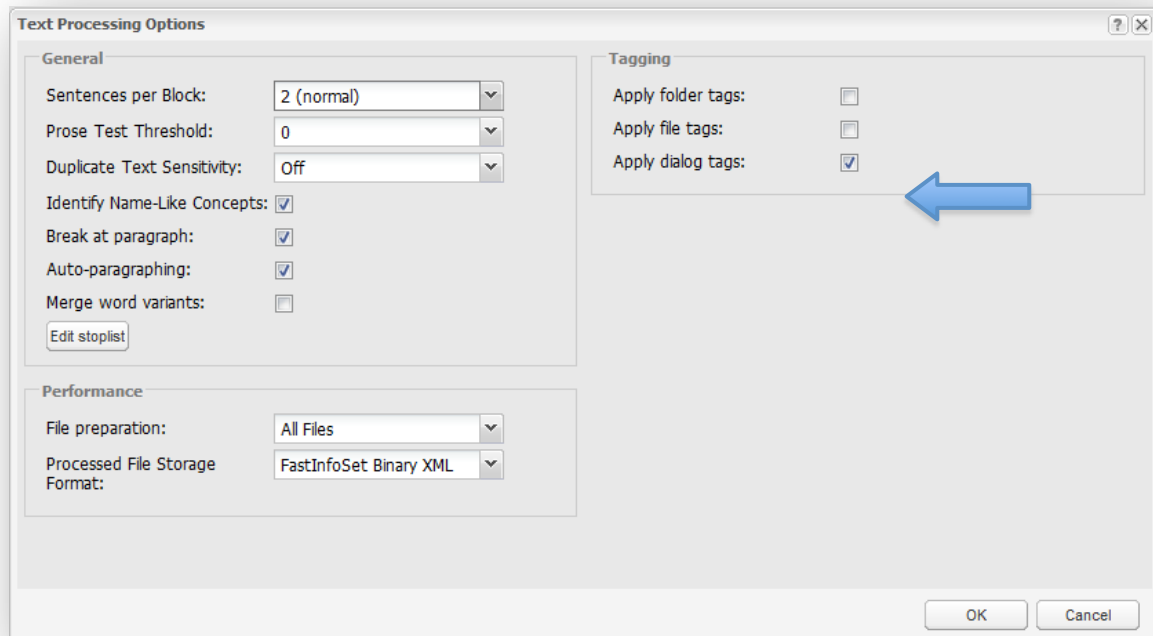
Susan: Yes actually we're going up to Thailand on the 13th of October, (B)'s coming with us for 14 days. My daughter in law is there, and they've got a little boy.

Alan: Yeah, so we'll show you a photo of that, she's very cute. Four boys, four grandsons, and one granddaughter.

Given text data in this format, Leximancer can extract the dialogue markers as tags and identify the speaker of every subsequent sentence until the next dialogue marker.

Configuring Transcript Analysis

Select your text data as usual. To create dialogue tags, expand the Generate Concepts Settings and Edit the Text Processing Settings. Tick Apply Dialogue Tags:

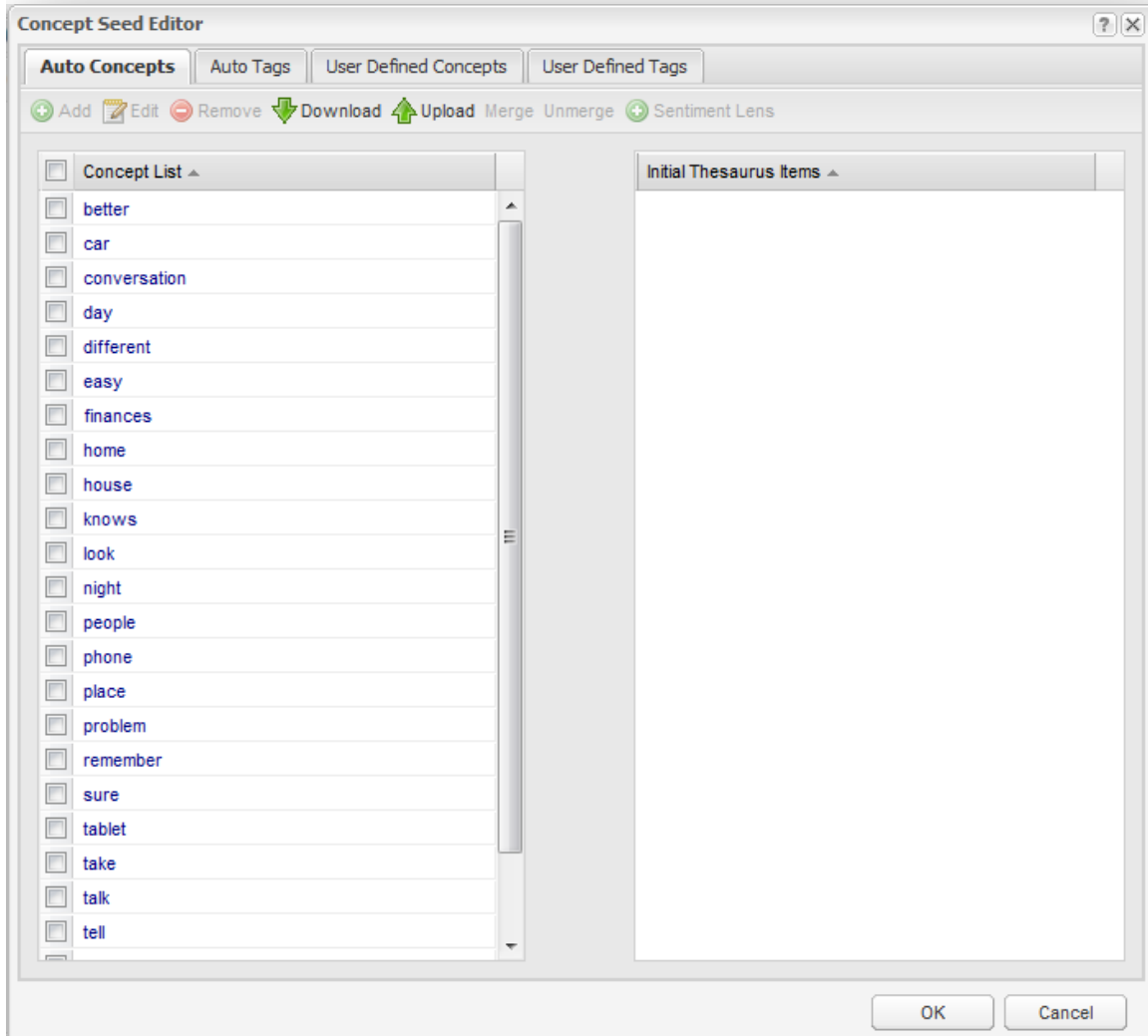


This transforms each dialogue marker in the text into a tag, which is then inserted into each relevant sentence and displayed for you under Autotags tab in the Edit Concept Seeds interface.

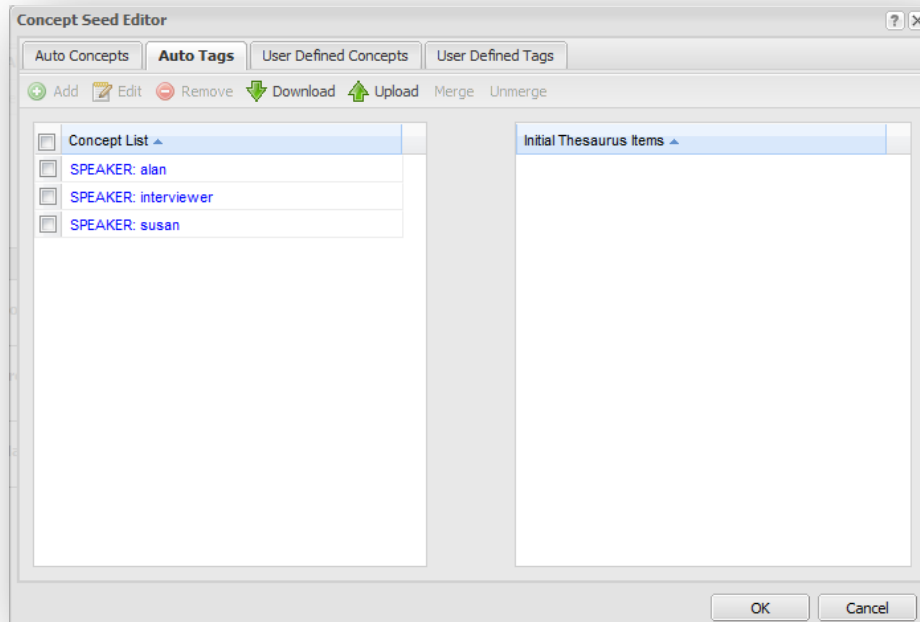
There is another setting in the Pre-process stage called the Prose Test Threshold. If your interview text is quite colloquial and does not conform to standard stop-word usage, set the Prose Test Threshold to 0. This filter is most useful for prose interspersed with non-textual material, such as web pages.

Run the Generate Concepts stage.

When this stage is complete, expand the Generate Thesaurus Settings and Edit the Concept Seeds node to inspect the extracted textual concept seeds in the Auto Concepts tab:

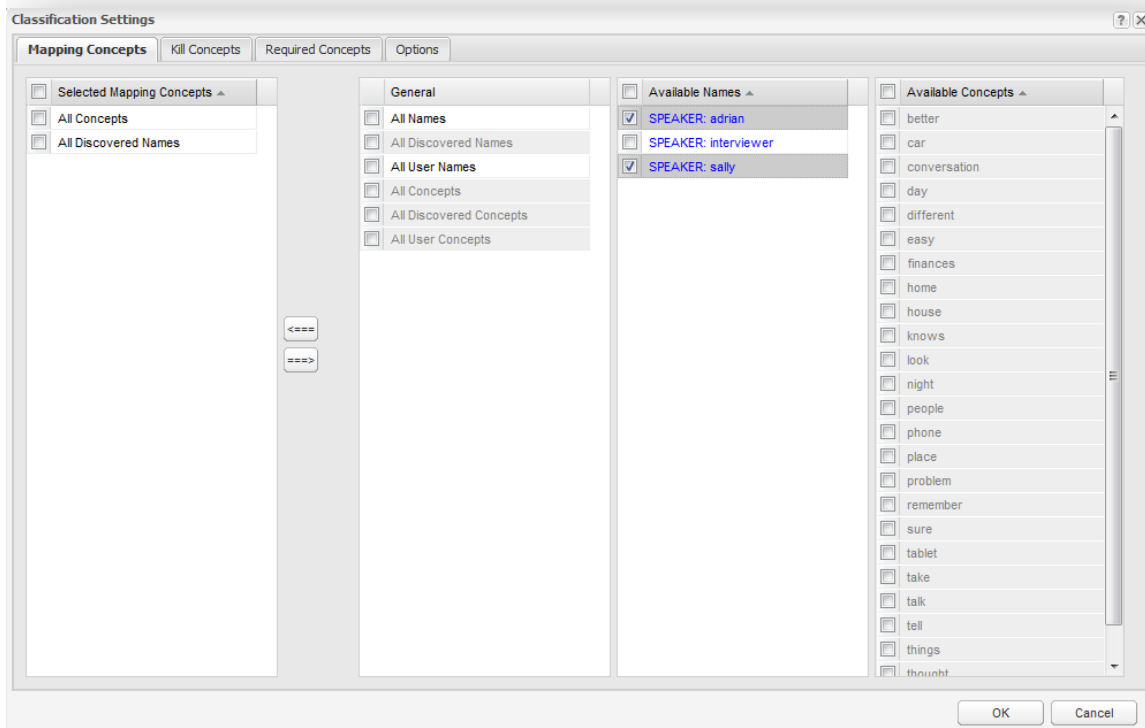


Then click on the Auto Tags tab to see the list of speakers identified by Leximancer:



Now run the Generate Thesaurus phase.

When this is done, you can choose whose utterances you wish to analyse, and whose you wish to leave out. You can also choose which items you wish to see on the map. These settings can be changed by expanding the Run Project Settings, and Editing the clicking Concept Coding Settings:

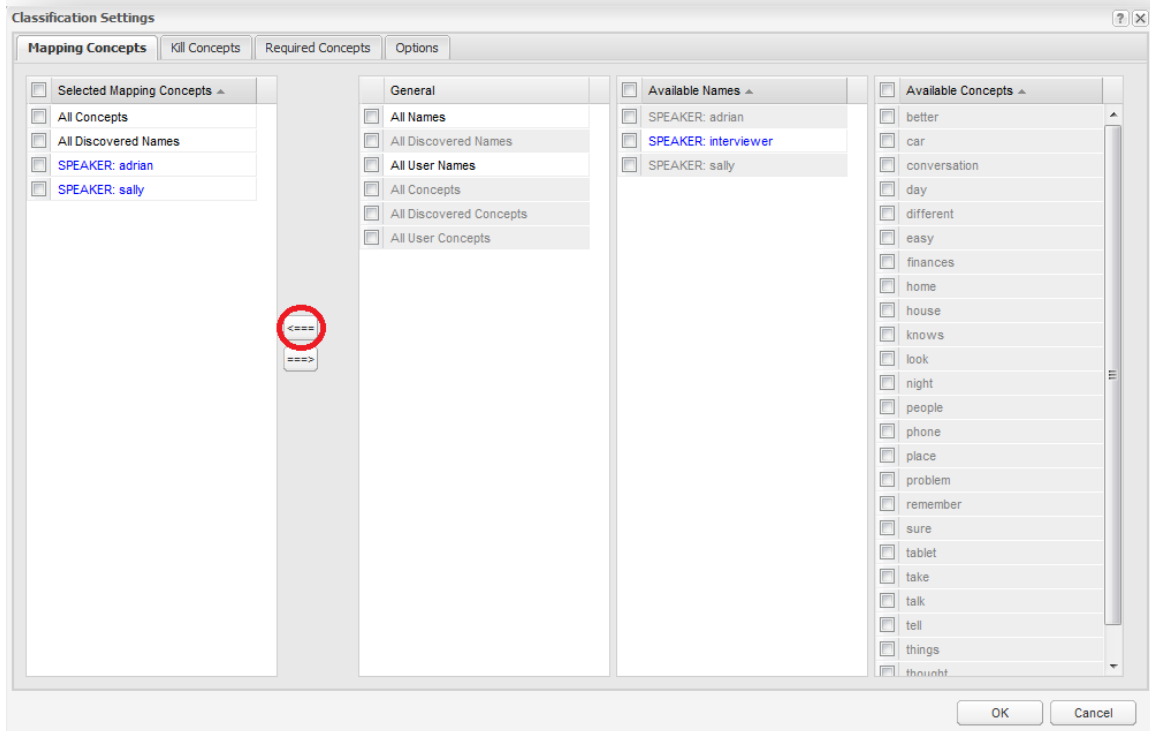


The Mapping Concepts list lets you choose which tags and concepts you wish to appear on the concept map. By default, there are two items in the list, All Concepts and All Discovered Names.

- The All Concepts wildcard represents all of the concepts identified for this project, be they automatically-discovered or user-defined.
- The All Discovered Names wildcard represents only those name-like concepts discovered automatically by the software.

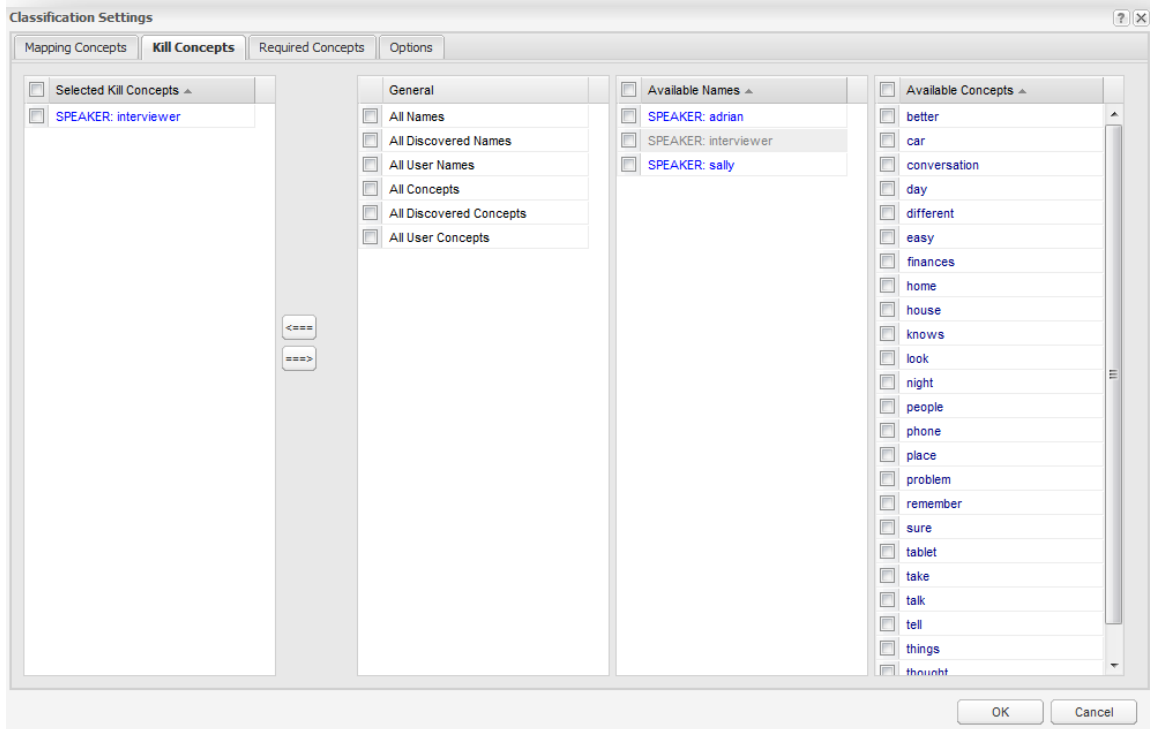
Full lists of the possible name-like and word-like concepts appear in the right-hand panels so that you can choose which entries you would like to see on the map.

If you wish to inspect the relative ownership of the textual concepts between your speakers, select the desired speaker tags and use the left arrow to add them to the Mapping Concepts list:

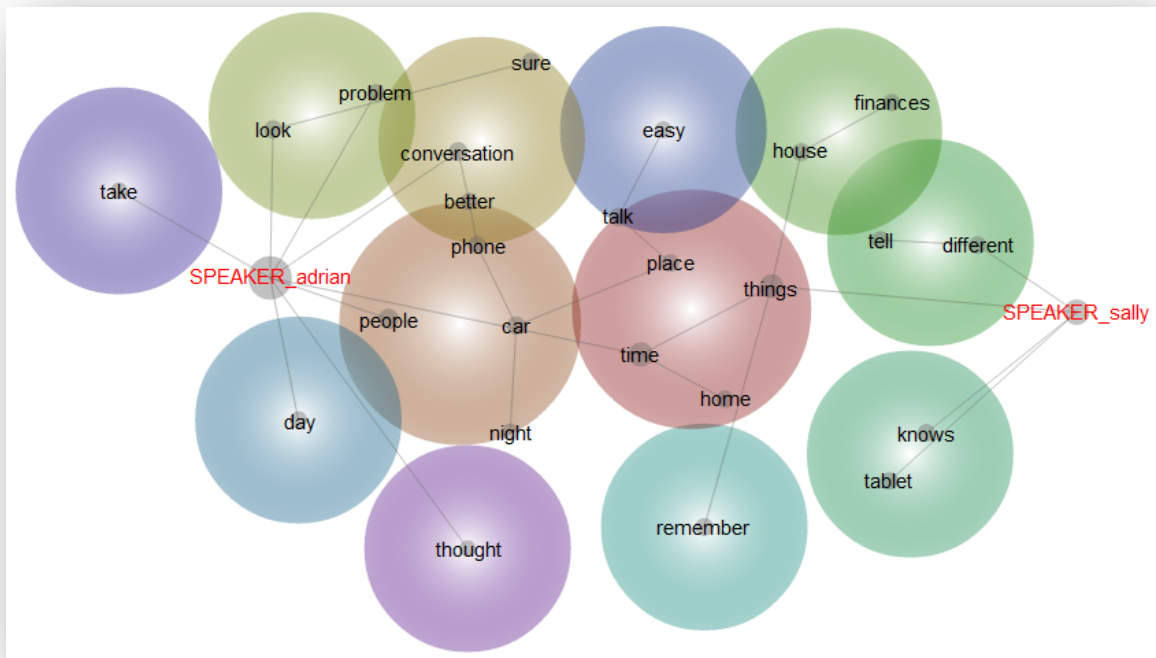


Since the full list of word-based concepts is of interest in this case, we can leave the All Concepts wildcard in the Mapping Concepts list.

The Required and Kill Concepts tabs allow you to select whose utterances you wish to analyse, and whose you wish to leave out from the analysis. For example, if you wanted to suppress all the utterances of the Interviewer, you would move the Interviewer speaker tag into the Kill Concepts list. This causes all the concepts coded into questions asked by the Interviewer to be removed from the analysis:



With these changes made, you can click Run Project to complete the final phase of processing and produce a concept map:



6. Analysing Spreadsheet Data

Leximancer can effectively analyse spreadsheets containing text fields and category fields (data in tabular format). You can analyse multiple text fields in each record, and also include the categorical fields as variables in your analysis. This enables very powerful text mining.

Practical: Analysing Spreadsheet Data

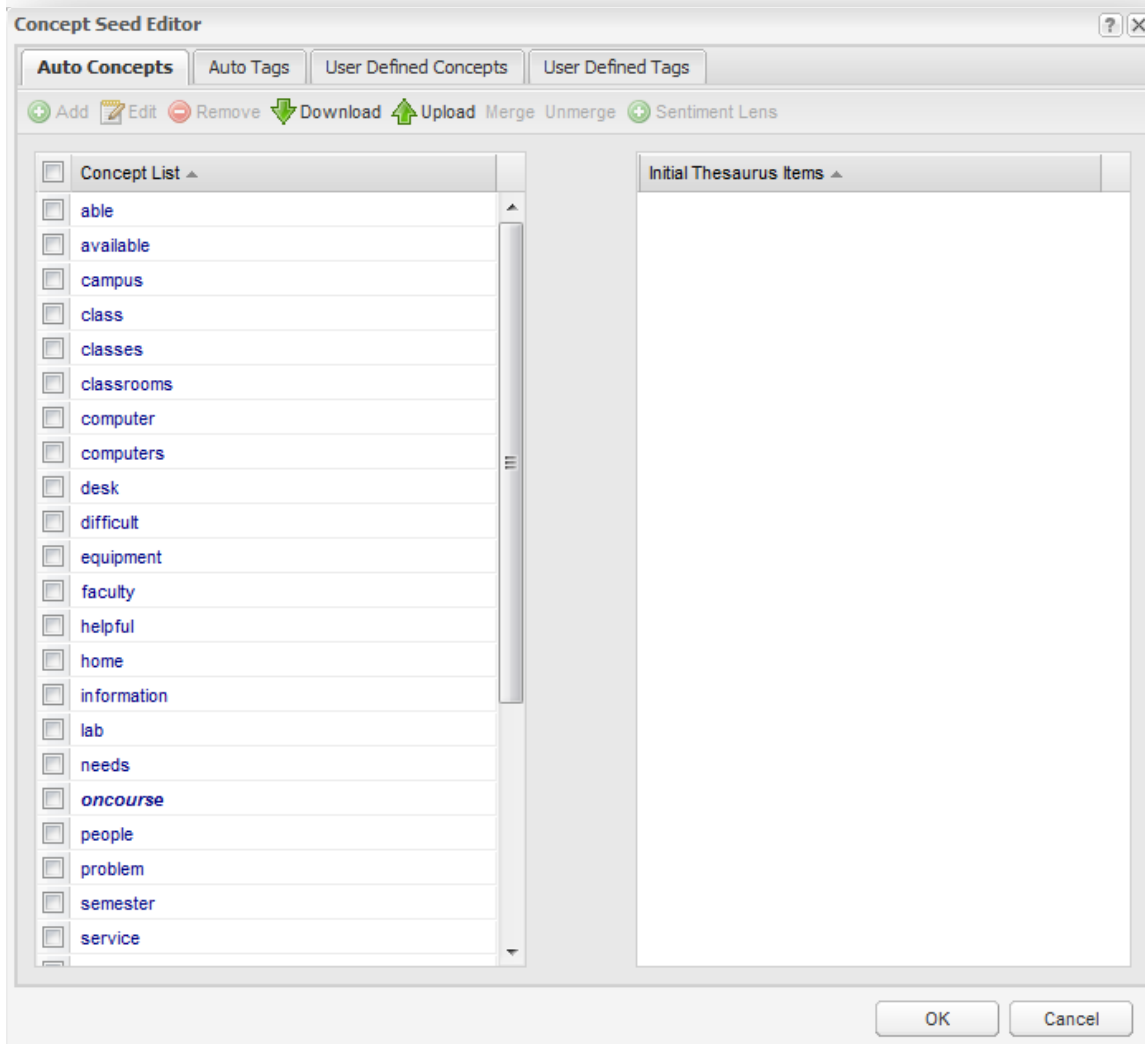
Your data may be in a spreadsheet application such as Microsoft Excel or in a database application such as Access. The best spreadsheet arrangement or layout for analysis is to have one separate response item per column (the first line should contain the column headers), and one respondent per row.

Save the spreadsheet in delimited text format (for example, in tab delimited (.tsv) or comma separated variable (.csv) format).

	A	B	C	D	E	F
1	Respondant #	Gender	Age	Position	Ratings of satisfaction	IT Feedback Comments
2		1 Female		23 Faculty		The main problem that I have with the computer labs is that there is almost always one computer that is not functioning. This is a waste of time. If the stations are full, that means at least one student is just sitting around or looking over another student's shoulder.
3		2 Male		45 Faculty		Most faculty have no clue how to use the smart classrooms. I think they would be utilized far more completely if there were several opportunities during the summer and during the early part of each semester to learn the basics of the equipment. Often though the carts are not working properly for those that do know how to use them, and getting someone to take care of the problem often does not happen for weeks. There is no "emergency" help, especially for evening classes to troubleshoot these
4		3 Male		65 Faculty		Sometimes decisions and changes are made and we just have to "adjust". This takes additional time and energy which must be diverted from other activities.
5		4 Female		43 Faculty		Should never have migrated to Oncourse CL--also a problem to have students on both the Angel and Oncourse systems in the same semester. Help desk people are of no assistance for online courses. If after hours, hard to get help from someone knowledgeable about Angel--the Bloomington people aren't helpful when the IUE people are not around. [IRD] has been very responsive to help with hardware problems--some of other staff not as knowledgeable as [IRD]. [IRD] is good with Angel
6		5 Male		42 Faculty		questions--but should have someone as a back-up when she is gone. Response time is generally very
7		6 Female		34 Faculty		3 TLC location is out of the way. This inhibits faculty use.
8		7 Male		37 Faculty		2 Too many unused computers--better to spend funds for workspace & software
9		8 Female		48 Faculty		3 Computer desks in lab are generally too small or cluttered to spread out materials for writing projects.
10		9 Male		57 Faculty		1 Cable routing of computer stations & especially the tech carts is typically messy & tangled.
11		10 Female		29 Faculty		Smart equipment is "too smart." There are simpler (and cheaper) switching products to manage which component is put up on screen. Some days I fumble with the remotes and buttons for five minutes in front of my waiting class before I can get PowerPoint or a video up on-screen.
12		11 Female		38 Faculty		3 ML 127 equipment is very slow and difficult to use...
						4 We need more support for Oncourse
						Hopefully IT will provide training sessions this summer when I have time to "get educated." I am naive

Create a Leximancer project as usual, and click on Load Data. Browse and select the spreadsheet file (the .tsv or .csv file). Drag and drop it into the Document Set area, then click Ok.

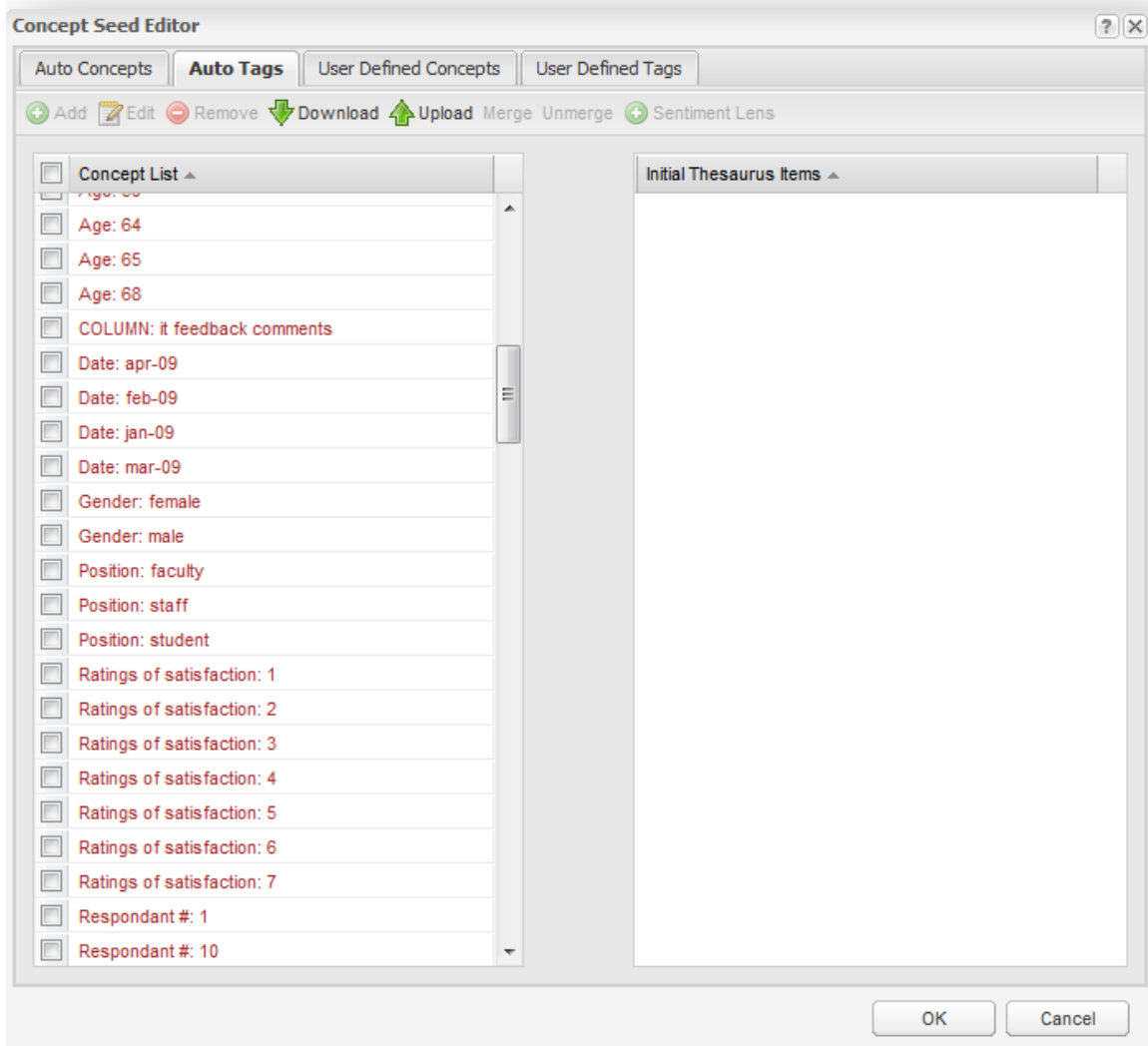
Run the first phase of processing (Generate Concepts), then expand the Generate Thesaurus Settings and Edit the Concept Seed Settings. You will see the concept seeds automatically extracted by the software in the Auto-Concepts tab:



You can seed your own concepts in the User Defined Concepts tab. If you want more (or fewer) automatic concepts, just go back one node to the Concept Seeds Settings and change the Total Number of Concepts to be suggested by Leximancer.

If you process a spreadsheet that uses the layout described above, Leximancer will automatically create a tag to represent each free-text column, and a tag for each of the levels (or possible responses) to the categorical variables in the data (to a limit of 500 unique responses). Entries that read '@none' represent null entries or empty response cells.

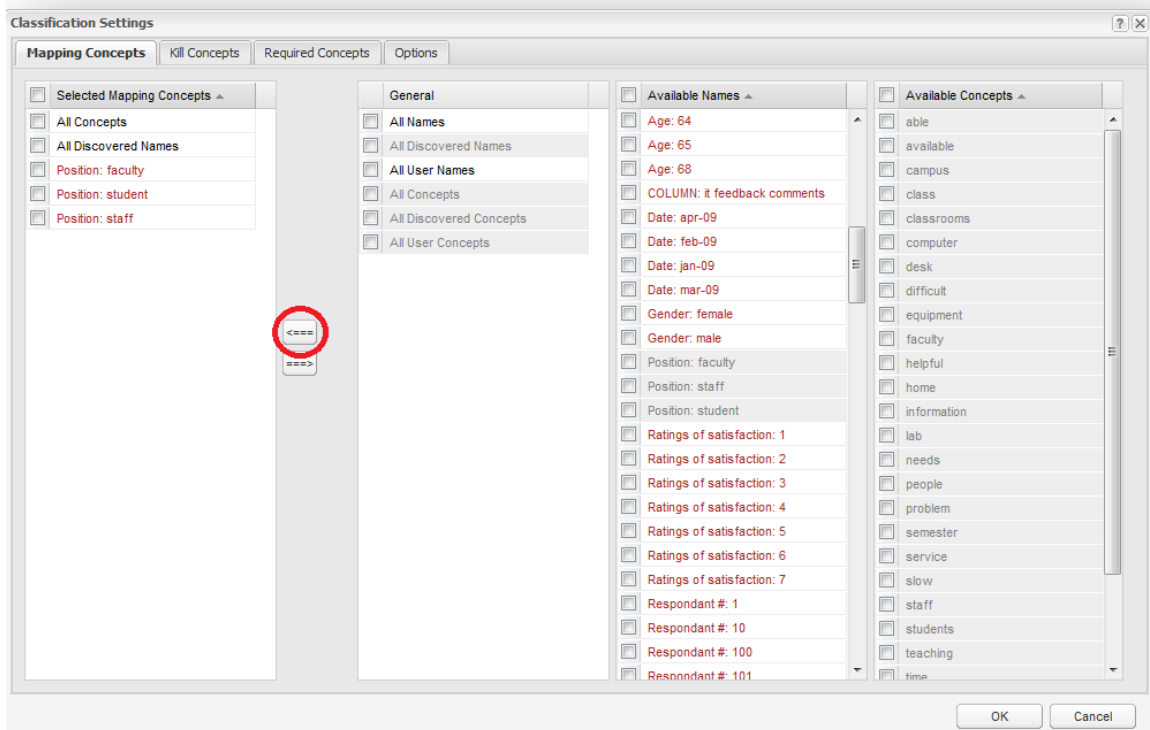
You can review what tags Leximancer has extracted by clicking on the Auto Tags tab. The tags take their names from the column headers and categorical responses in the data. There is no hard limit to the number of free-text and categorical variables that can be analysed in Leximancer. The Auto Tags can be used for data mining correlations with textual concepts, and for selecting which text column(s) you wish to map at any time:



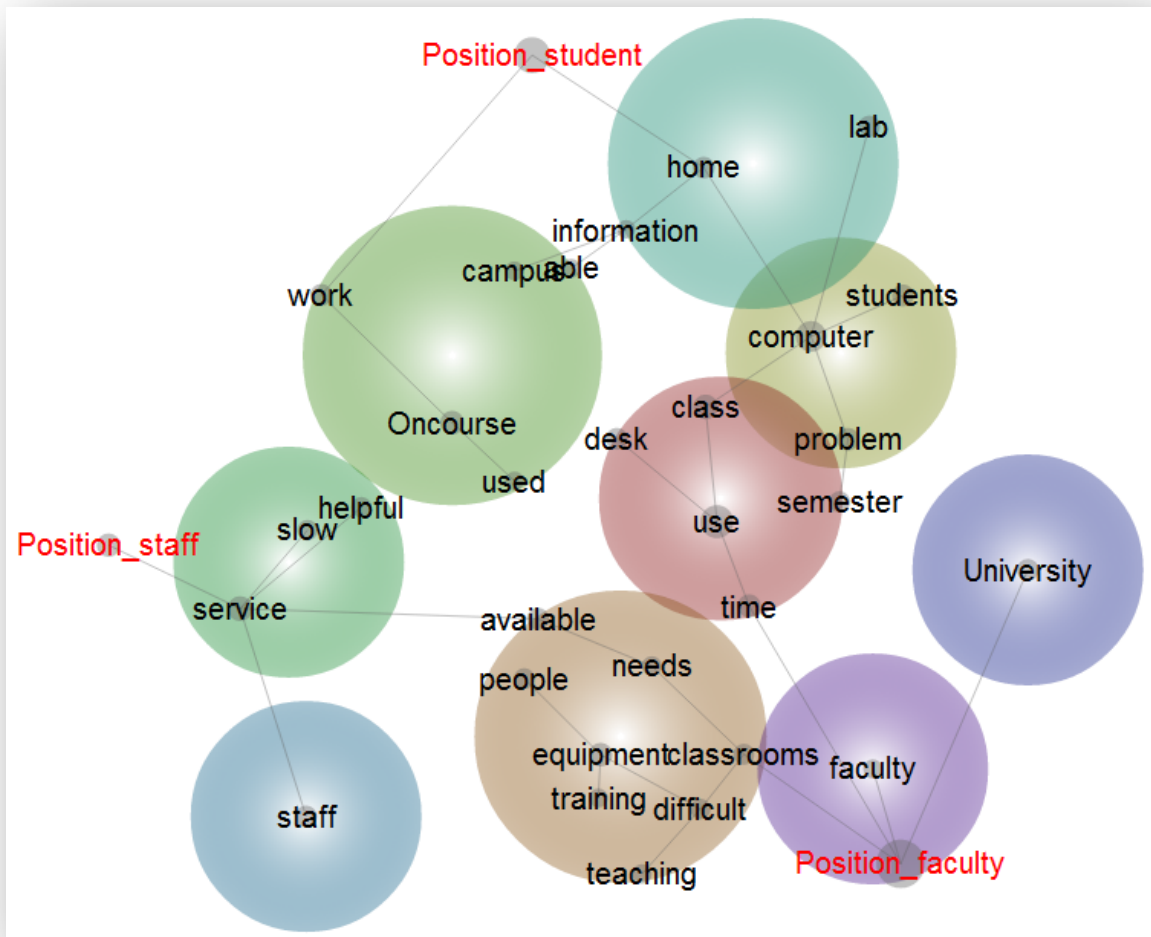
Run the Generate Thesaurus phase to extract a thesaurus from the data describing each concept seed.

Expand the Run Project Settings, and Edit the Concept Coding Settings to access the data mining options.

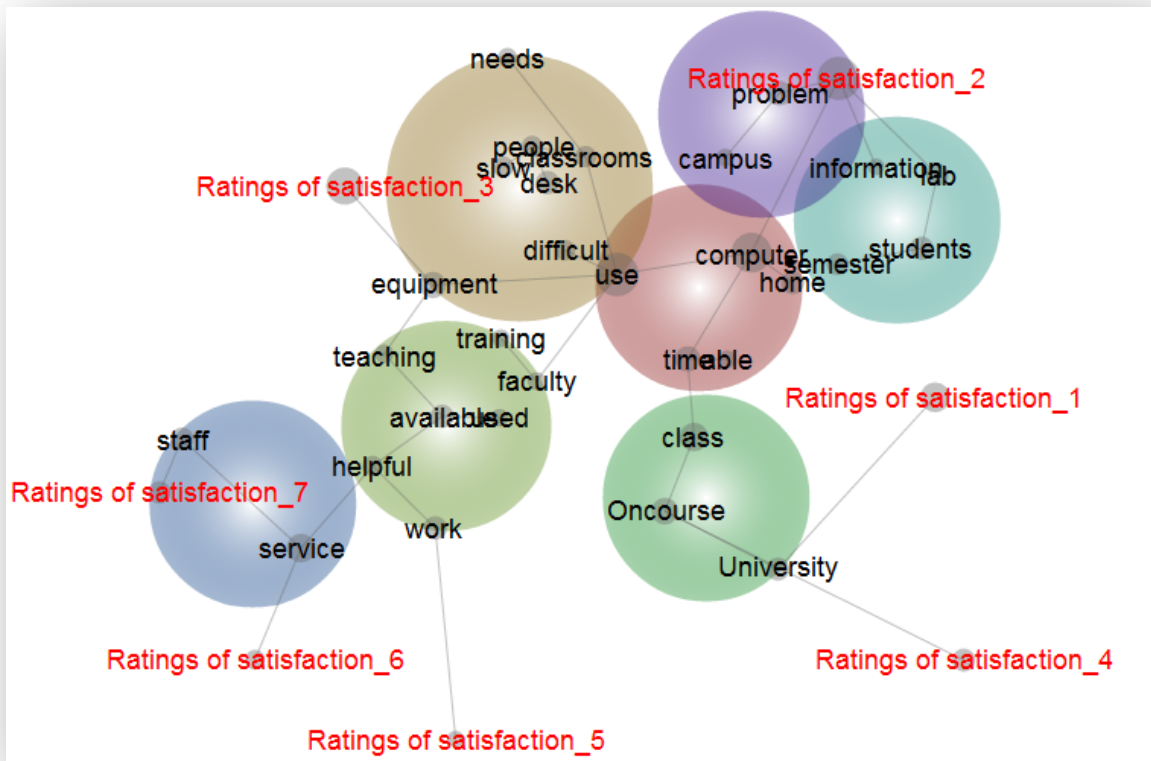
The Mapping Concepts list lets you select what variables you want on the concept map, like choosing the columns you want in a database query. For example, let's say you wish to examine the correlations of the three position types (faculty members, staff and students) with the comments from the text column called 'IT feedback comments' in the data. You would like all the textual concepts on the map, so retain the All Concepts wildcard in the Mapping Concepts list. You would also need to select and add the three respondent type tags (Position: Faculty, Position: Staff and Position: Student) to the Mapping Concepts tab using the left arrow:



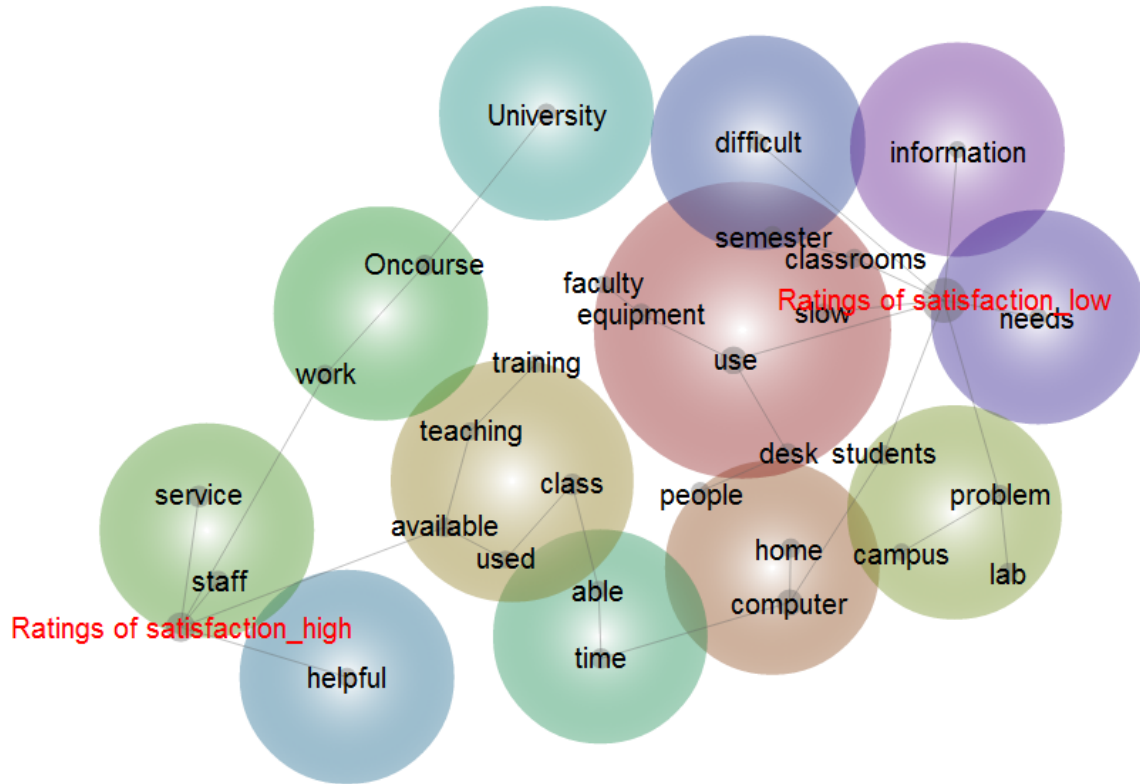
Once you have configured the data mining settings, run the last phase of processing and inspect the resulting concept map:



You could choose to correlate the textual concepts with the satisfaction ratings instead of the position categories. To do so, return to the data mining settings in the Concept Coding Settings, and remove the position tags from the Mapping Concepts list. Replace them with the satisfaction score tags. Rerun the final phase to produce this new view of the data very quickly:



You could also aggregate the satisfaction tags in the Edit Concept Seeds stage to produce 'Low' and 'High' satisfaction tags to place on the concept map. For instance, you could Merge the satisfaction scores of 1–3 and Edit this to rename it as tag as 'Low'. Then Merge the 4–7 tags and Edit to rename them as 'High'. Adding the 'Low' and 'High' tags to the Mapping Concepts list in the Concept Coding Settings produces the following map:

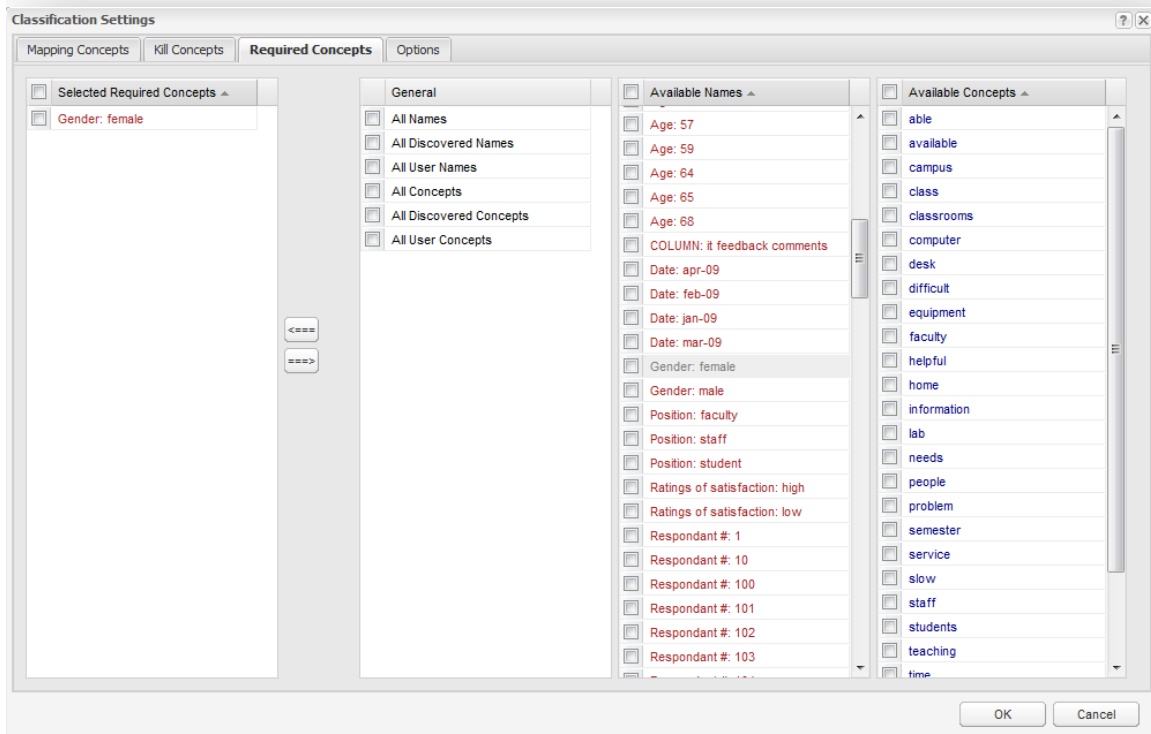


You can also use the Required Concepts and Kill Concepts tabs in the Concept Coding Settings to filter records in or out of your analysis. For instance, if you had more than one text column in your spreadsheet, you could choose to examine the concepts associated with a particular text response. You would do this by moving the tag denoting the text column of interest into the Required Concepts tab. In this case, if a data cell does not come from that text column, it will not be coded for concepts and mapped.

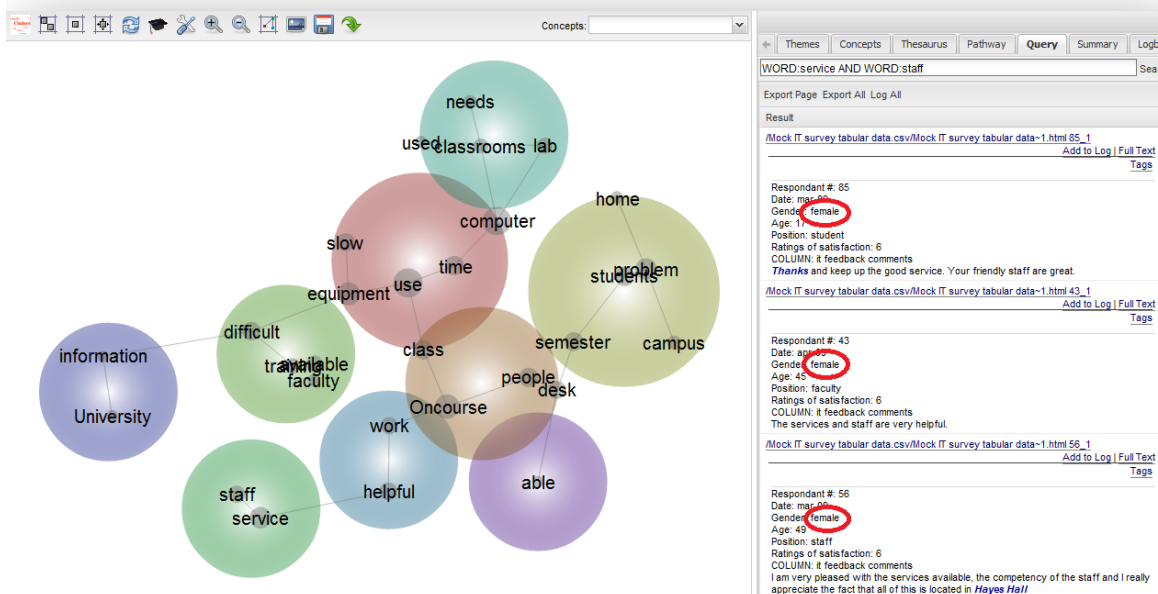
A Kill Concept is almost the opposite of a Required Concept. If a text segment matches a Kill Concept tag or concept, it will not be coded with any other classifier. For example, you could suppress the analysis of all text segments which match the concept 'available' by identifying 'available' as a Kill Concept.

By way of example, if we wished to map only the comments made by women in this spreadsheet, we could either add the Gender: female tag to

the Required Concepts tab, or by add the Gender: male tag to the Kill Concepts tab:



If we remove all the tags from the Mapping Concepts tab (so that only the default All Concepts and All Discovered Names wildcards remain), the resulting map reflects all the responses made by women in the data:



This concludes the Leximancer 4 Manual. Visit our website at www.leximancer.com

This documentation is Copyright 2011 Leximancer Pty Ltd,

<http://www.leximancer.com/>.

All rights reserved.
