Hons. Degree in Computing H4016 Text Mining & Information Retrieval

Unit 2 – Preparing text data for data mining

Part 3: Text visualisations

Ref for this section: *Konchady, Text Mining Application Programming, Thomson 2006, chapter 3*

Recap on last week:
Text Mining Process

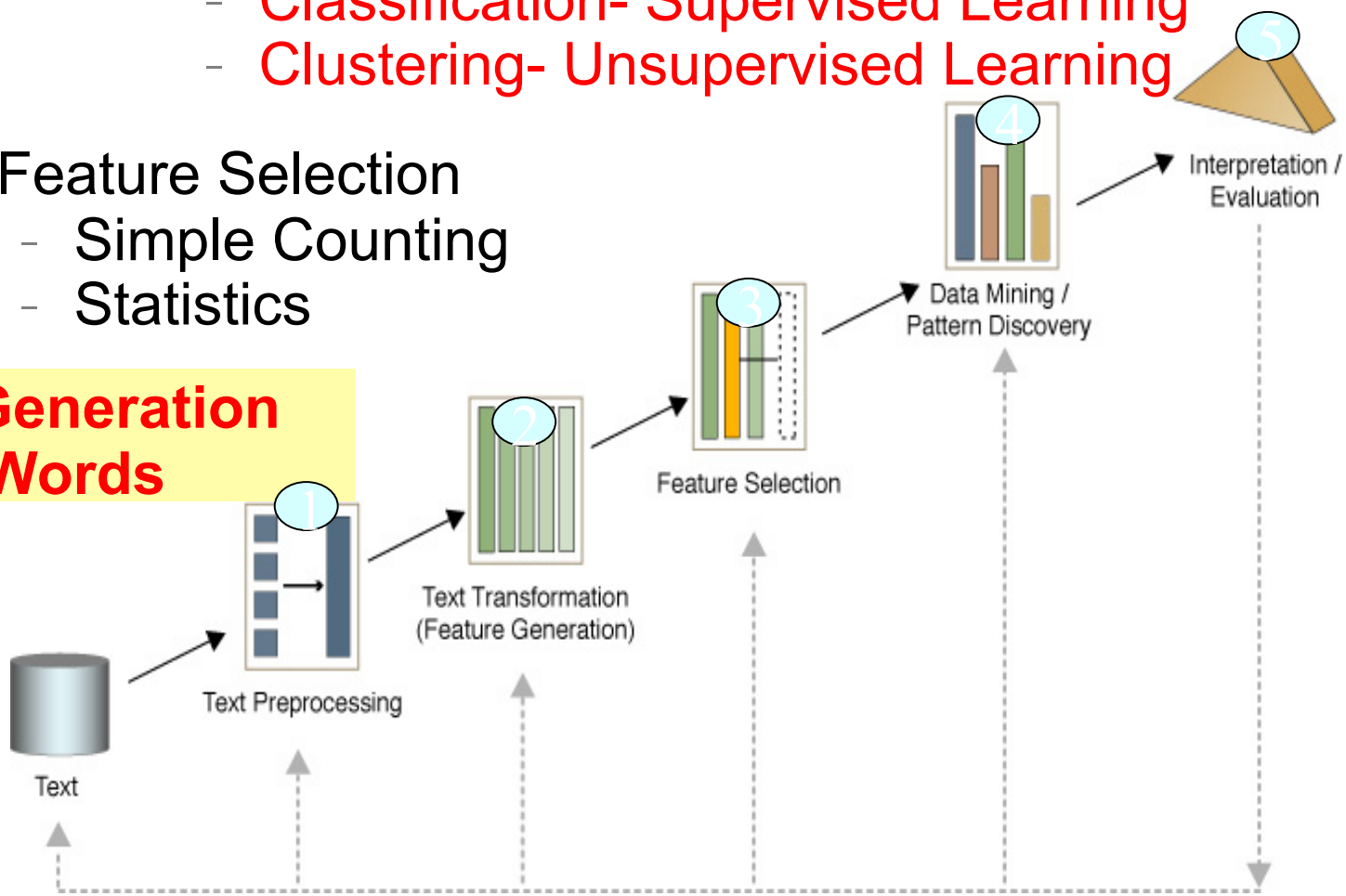
5. Analyzing Results

4. Text/Data Mining
   - Classification- Supervised Learning
   - Clustering- Unsupervised Learning

3. Feature Selection
   - Simple Counting
   - Statistics

2. **Features Generation**
   - **Bag of Words**

1. Collect text and identify business objective



G.Gray

Objectives

- Experiment with visualisation techniques for unstructured text:
  - Term frequencies:
    - Word clouds
    - Tree maps
  - Phrases:
  - Word trees
  - Phrase nets

1: Word clouds

Word clouds

A word cloud is a visual representation of the frequency of terms in a text.
There are many tools available to do this, for example http://www.wordclouds.com or http://www.wordle.net

Download the following three files from moodle: allHealthCare.txt; allCrime.txt and allKenya.txt. Each is a single file with all five texts from that category.

Go to http://www.wordclouds.com ; click **word list**; and then click on the **Paste/Text dialog** hyperlink. Paste in one of the three texts above.

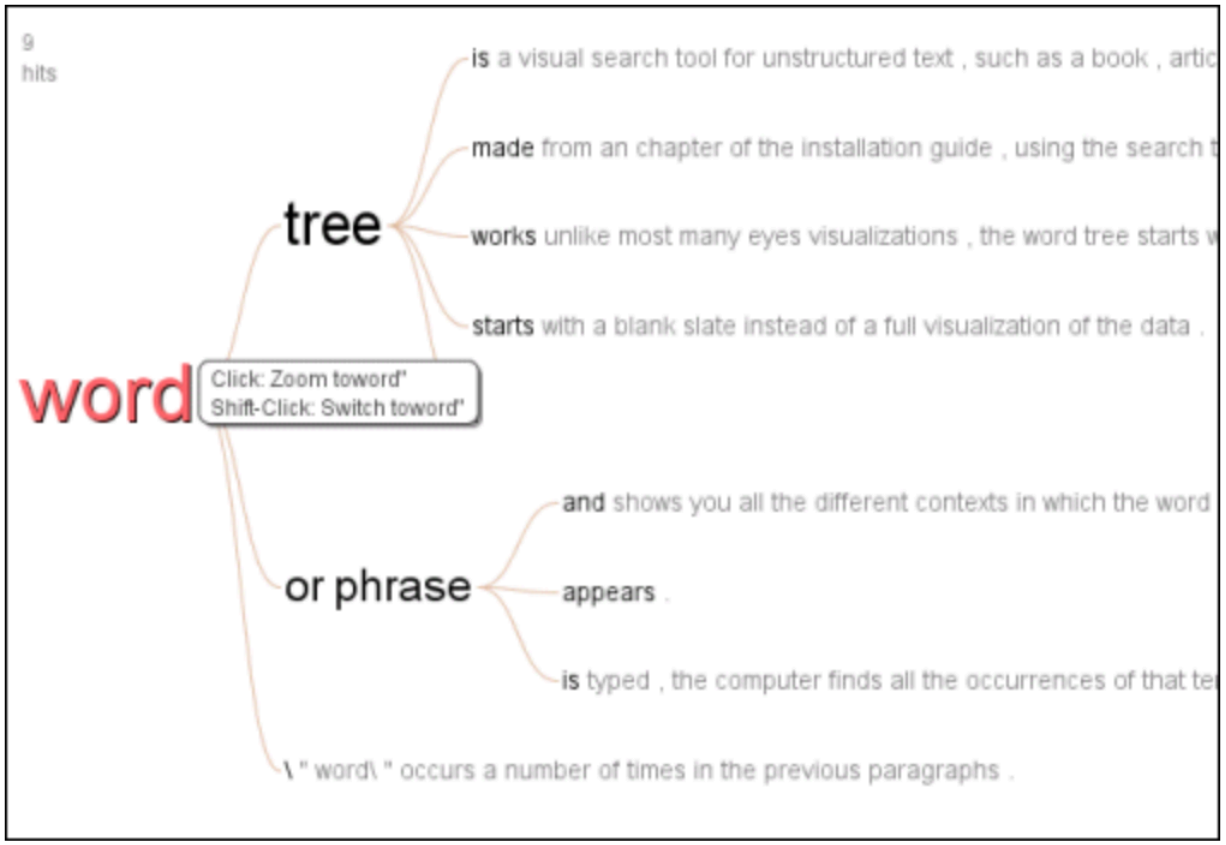
Word cloud is generated from the text. Colours and shape can be configured. Clicking on **word list** again shows the counts for each term.
Which are the common terms in each group of document?.


year
legislation
murders
crime
manslaughters
murder
last
rights
victims
Court
number
wounds
Crime
Kenya
Healthcare

G.Gray

2. Word tree

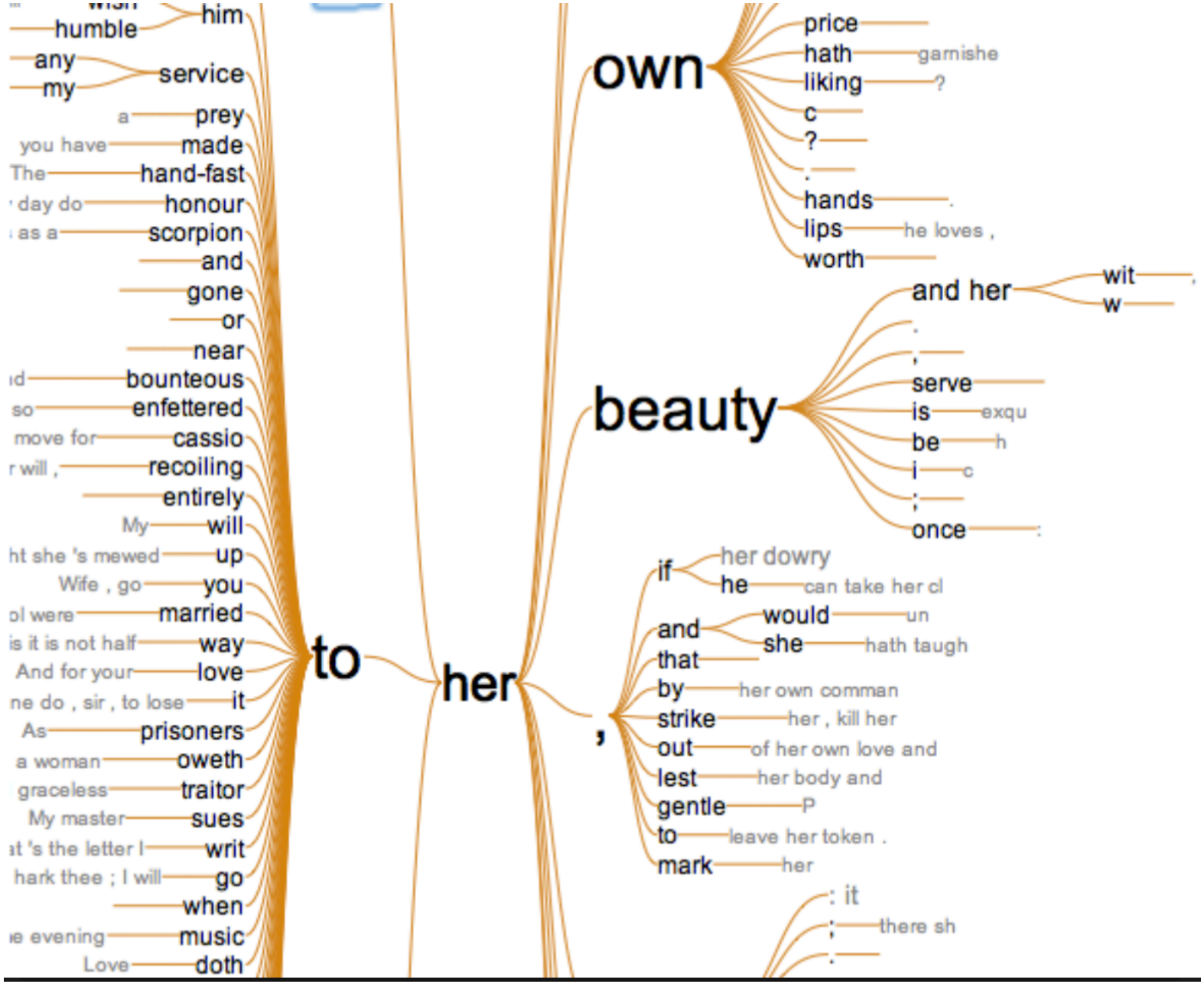
A word tree visualises all the occurrences of a selected term, along with the phrases that appear before or after it.



G.Gray

3. Word tree

Word tree for ‘her’ in Shakespear’s texts



G.Gray

Try it:

Take a look at:
https://www.jasondavies.com/wordtree/

Select one of the examples on the site (e.g. Obamas speech on Iraq, or the Cat in the Hat.)

To change the term, select any term in the text of the right using ‘shift-click’

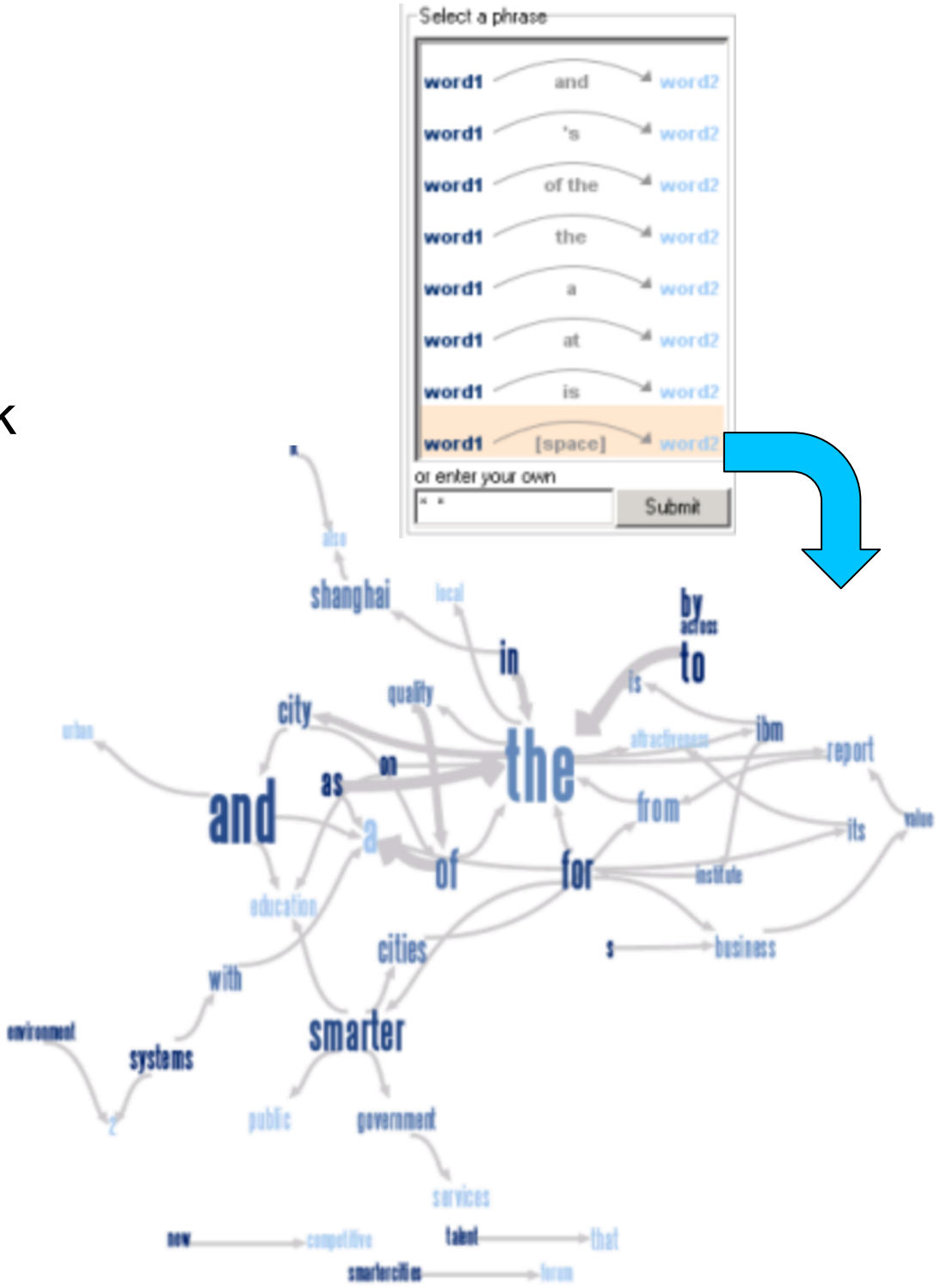
You can paste in your own text as well. Generate a word tree for one of the more common words according to the word cloud, e.g. ‘valley’ from the Kenya texts.

3. Phrase net

Used to search for phrases in a text.

Looks for terms that are linked using common link terms.

For example, if you select the link term ‘of the’, the network will return all terms that match the pattern: ‘*term1* of the *term2*’.



Phrase Net

Below is a Phrase Net form ‘Pride and Prejudice’ of terms joined by ‘and’