Visualisation of textual data through collocate clouds

David Beavan

Corpus of Modern Scottish Writing
Department of English Language
University of Glasgow
Scottish Corpus of Texts & Speech

Features

• Scots and Scottish English
• 4 million words, 20% spoken
• 1945 – present day
• Diverse range of genres e.g. conversations to legal documents
• Wide geographic spread
• Free and publicly available at www.scottishcorpus.ac.uk

Challenges

• Variant spellings e.g. how do we link home = hame = haim
  potatoes = tatties
• How to place a document on a continuum from Scots to Standard Scottish English?
• Can regional or dialectal words be reliably found?
blood

Small = low frequency, big = high frequency. Dim = low MI, bright = high MI
Click on word above or visualize new words: Outward | Go

activation

Small = low frequency, big = high frequency. Dim = low MI, bright = high MI

glycogenin-2

Small = low frequency, big = high frequency. Dim = low MI, bright = high MI
Collocate Clouds visualisation

Benefits

• Clouds familiar to many users e.g. Flickr shows shared photograph keywords as a cloud
• Good introduction to analysis for new users and beginners
• Acts as a window into what may be complex documents or data sets
• See how lexical terms interact
• Visualise data subsets and vocabularies
• Promotes data exploration and browsing by allowing collocate words to trigger new clouds using selected word as node
• Could be used as a gateway into more specific statistics in tabular form

Method

1. Choose node word e.g. ‘blood’
2. Search entire document for node word
3. Collect collocates (surrounding words) five words before node, five words after
4. Total up occurrences of each type
5. Keep 100 most frequent collocates
6. For each collocate calculate MI (mutual information). The measure of how likely the node and the collocate occur together.
7. Display the 100 collocates alphabetically
8. Scale font size to frequency of occurrence
9. Scale brightness to MI score