SEFI Conference
Word Cloud Generator

Table of Contents

This script will generate a word cloud of the keywords from the SEFI 2014 conference in Birmingham. Words are clustered by how commonly they occur together. The word size indicates how freqently the word is mentioned in the conference proceedings.

Correlating The Words

The clustering is based on a measure of how well correlated the words are with each other. The process for measuring the clustering is:

1. Parse all paper abstracts for dictionary of unique conference keywords (3633 keywords), and their total count across all paper abstracts.

2. For most frequent N words, it counts how many times each keyword was used in each abstract.

3. This gives a histogram of keyword occurences for each abstract.

4. Then from that histogram MATLAB calculates the correlation coefficient between all N words. So we get an NxN matrix of correlation coefficients.

Forming the Clusters

Clustering is calculated using inbuilt hierarchical clustering algorithms from MATLAB statistics toolbox.

1. Correlation is used as a distance measure between the words.

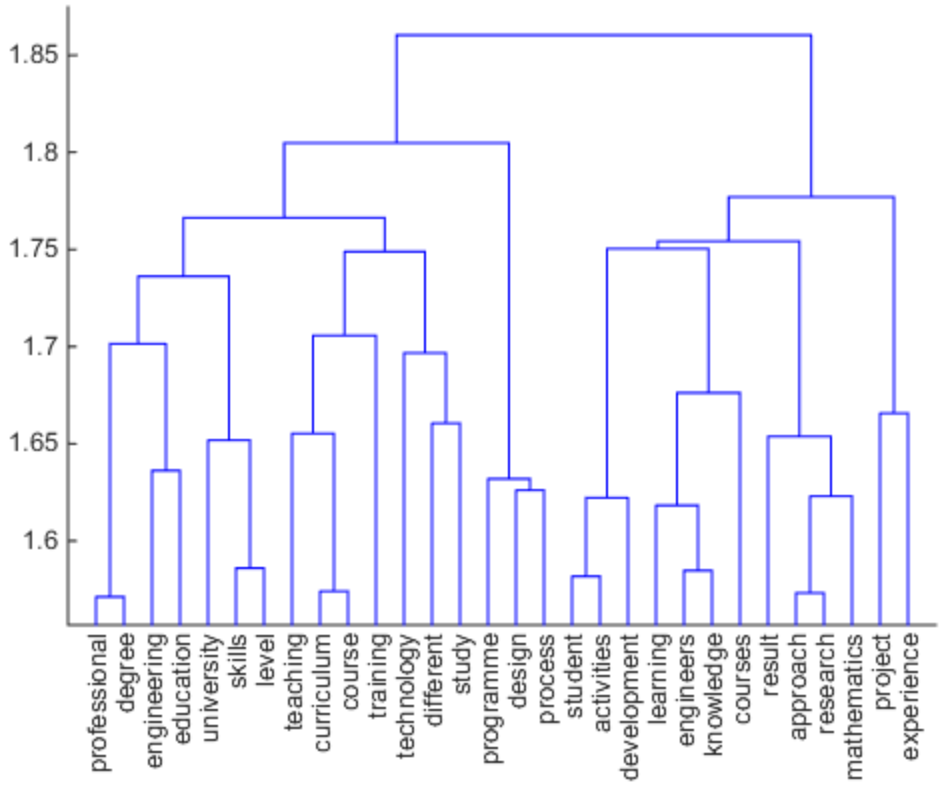
2. The algorithms works through word list, and pairs words that are most correlated, i.e. nearest.

3. The pairs can then be combined with other words or pairs which are near.

4. the process repeats until all words are grouped together into a tree.

5. This pairing is ususally shown in a diagram called a dendrogram. It shows how words are paired and the height of the link shows how close the two words being linked are.

This is the dendrogram for the SEFI conference data:



How are Clusters Displayed as Word Clouds?

The script is manually sets the cloud to display nine clusters. So the clustering algorithm will stop linking words together once there are nine groups of words.

Formatting:

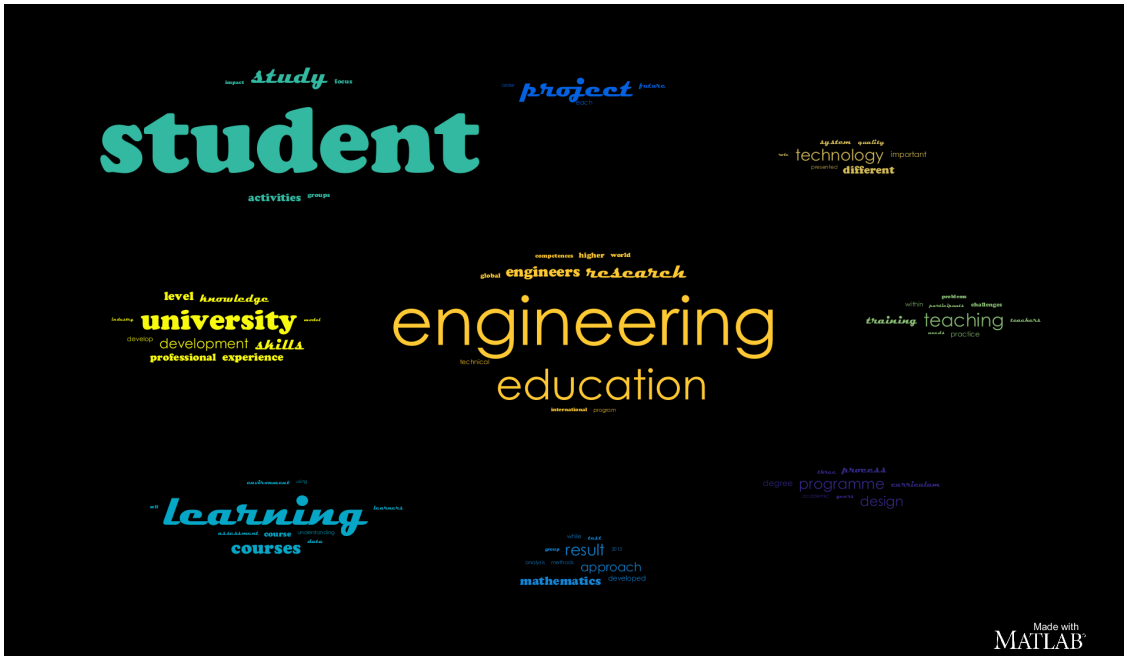
In the word cloud, each sub-cloud cluster has its own colour and the words are grouped together spatially. Word font is randomly selected, and the word size is proportional to its total count across all conference papers.

Each word cluster is formed by the following process:

- The word with the highest word count goes in the centre.
- Remaining words in the cluster are sorted by how correlated they are to the central word.
- Going from most correlated to least correlated, the words are added to the cluster in a spiral from the centre. This means the words most correlated to the central word are closer to the centre of the cluster, and the words around the outside are least correlated.

For the whole cloud:

The cluster with highest total word count (across all words in the cluster) goes in the centre of the figure. Remaining clusters are places evenly around the edge. Their distance from the centre is proportional to the mean correlation between the words in the cluster and the central cluster.



Published with MATLAB® R2014b