Keyword Extraction – Simple Instructions

This program executes an unsupervised keyword extraction model that works on individual documents.

Things to have before running the code.

1) Pre-trained Google News Word2Vec word vectors. These can be downloaded from here. (https://github.com/3Top/word2vec-api)

2) Pre-trained Wikipedia TF-IDF scores. These are in the file pretrained_tfidf.txt.

3) A stop word list. These words are available in stopwords_en.txt

Packages to have before running.

1) Numpy and Scipy – for fast numerical computing

3) NLTK – for word lemmatization and frequency distribution calculation

4) Gensim – for loading and working with Word2Vec word embeddings.

All the installations can be done using pip.

The model has two clustering mechanisms and three different cluster-selection mechanisms. The formal syntax for running this model on command line is as follows.

>>> python KEX.py filename mode number-of-keywords[optional] topic-word[optional] distance-measure[optional]

Execution Instructions:

1 – python KEX.py clouds.txt 1 (Recommended)
Applies clustering method and returns all the keywords that are relative to the document.

2 – python KEX.py clouds.txt 2
Applies skip-agglomeration method and returns all the keywords that are relative to the document.

3 – python KEX.py clouds.txt 3 10
Applies clustering method and returns the top 10 keywords in the document.

4 – python KEX.py clouds.txt 4 10
Applies skip-agglomeration method and returns the top 10 keywords in the document.

5 – python KEX.py clouds.txt 5 10 stratus

Applies clustering method and returns the top 10 words of the cluster whose center is the word specified (“stratus” in this case)

6 – python KEX.py clouds.txt 6 10 stratus

Applies skip-agglomeration method and returns the top 10 words of the cluster whose center is the word specified (“stratus” in this case)

For all the modes, the last argument is used to specify the type of distance measure to use. By default, they are set of the best. If you want to change, just add [avg/w2w/skip] one of those as the last argument. [avg – average diameter, w2w – word-to-word distance, skip- skip-agglomerative distance]