Evaluation Program Execution Instructions

***** Make sure you have the following packages installed: Numpy, Scipy, NLTK, Summa, and Gensim.

The evaluate.py requires four mandatory files.

1. Make sure you have pre-trained Google News word2vec file. If not, download it from https://github.com/3Top/word2vec-api. The vectors will be in .bin file titled GoogleNews-vectors-negative300.bin
2. The pre-trained TF-IDF values dictionary stored in the file named tf_idf_dict.txt
3. The stop words list with more than 550 words named stopwords.txt.
4. The ground truth is available in a comma separated file named Keyword_selection.csv.

The sample documents are in the **"docs"** folder.

Evaluation Options

The program takes two arguments: First argument is the method number, and the second argument is the distance metric.

The output for all the methods is filename along with F-score.

Method 1

Applies clustering keyword extraction procedure using one of the three distance metrics (avg/chain/skip)

Example: python evaluate.py 1 chain

Method 2

Applies skip-agglomeration keyword extraction procedure using one of the three distance metrics (avg/chain/skip)

Example: python evaluate.py 2 avg

Method 3

Applies traditional TextRank algorithm to extract keywords from the documents.

Example: python evaluate.py 3

Method 4

Uses pre-defined TF-IDF scores to naively rank all the words in the document.

Example: python evaluate.py 4

Method 5

Applies the best method available (clustering with cluster selected using chain length, sorted with pre-trained TF-IDF scores)

Example: python evaluate.py best