This readme is divided into 4 sections.

- The first section walks through how to run the code for 4 different algorithms, Vector Space, LSA, ESA and Query Expansion with WordNet.
- The second section talks about how to run the code for Word2Vec
- The third section talks about how to run the code for the Vector Space Model improved with spellcheck
- The fourth section details how to run hypothesis tests to compare models.

**NOTE: The main file for each of the sections is different and cannot be replaced**

**The data files to be downloaded are present in the following link:**
https://drive.google.com/drive/folders/1THVWfs051h2ix8yQeXTjLBDYIVDbN6x3?usp=sharing

## Section 1 (Running ESA, LSA and WordNet) :

### Files Required

Ensure that the folder consists of the following files:

| | | | |
|---|---|---|---|
| __pycache__ | 03-06-2021 14:47 | File folder | |
| cranfield | 02-06-2021 23:19 | File folder | |
| output | 02-06-2021 23:19 | File folder | |
| .DS_Store | 07-03-2020 13:13 | DS_STORE File | 7 KB |
| doc_words_final | 28-05-2021 16:15 | Microsoft Excel Co... | 45,467 KB |
| docs_wiki_final | 28-05-2021 16:15 | Microsoft Excel Co... | 1,09,632 KB |
| evaluation | 06-04-2021 20:18 | Python File | 12 KB |
| inflectionReduction | 06-04-2021 20:50 | Python File | 2 KB |
| informationRetrieval | 28-05-2021 23:27 | Python File | 3 KB |
| informationRetrieval_ESA | 28-05-2021 17:47 | Python File | 3 KB |
| informationRetrieval_LSA | 02-06-2021 06:08 | Python File | 4 KB |
| informationRetrieval_W2V | 02-06-2021 22:49 | Python File | 6 KB |
| informationRetrieval_WN | 02-06-2021 16:37 | Python File | 3 KB |
| main | 03-06-2021 15:01 | Python File | 19 KB |
| README | 07-03-2020 13:12 | Text Document | 2 KB |
| sentenceSegmentation | 03-06-2021 14:46 | Python File | 2 KB |
| stopwordRemoval | 02-06-2021 23:46 | Python File | 2 KB |
| tokenization | 03-06-2021 00:13 | Python File | 3 KB |
| util | 06-04-2021 20:43 | Python File | 1 KB |
| wiki_words_intersect | 28-05-2021 16:14 | Microsoft Excel Co... | 1,31,065 KB |

- The following .py files should be present in the folder in addition to the previous ones:
    - informationRetrieval_ESA.py
    - informationRetrieval_LSA.py
    - informationRetrieval_W2V.py
    - informationRetrieval_WN.py
- The following .csv files should be present in the folder (these are pre-calculated TF-IDF matrices for documents and wikipedia articles)
    - docs_words_final.csv
    - docs_wiki_final.csv
    - Wiki_words_intersect.csv

**Running:**
Run the following code in terminal:

python main.py -dataset cranfield/ -out output/ -segmenter naive -tokenizer naive -method [VS|LSA|ESA|WN]

The highlighted part is to select the method. The codes correspond to methods as follows:

- VS: Vector Space
- LSA: Latent Semantic Analysis
- ESA: Explicit Semantic Analysis
- WN: Query Expansion with Wordnet

The Sentence Segmenter and Tokenizer can be selected as desired between the naive model or otherwise.

## Section 2: Word2Vec

### Files Required

| | | | |
|---|---|---|---|
| __pycache__ | 03-06-2021 17:59 | File folder | |
| cranfield | 27-05-2021 21:13 | File folder | |
| out | 27-05-2021 22:11 | File folder | |
| .DS_Store | 07-03-2020 13:13 | DS_STORE File | 7 KB |
| evaluation | 06-04-2021 20:18 | Python File | 12 KB |
| inflectionReduction | 06-04-2021 20:50 | Python File | 2 KB |
| informationRetrieval | 03-06-2021 17:55 | Python File | 7 KB |
| main | 03-06-2021 17:58 | Python File | 9 KB |
| README | 07-03-2020 13:12 | Text Document | 2 KB |
| sentenceSegmentation | 07-04-2021 01:58 | Python File | 2 KB |
| stopwordRemoval | 15-03-2021 21:01 | Python File | 1 KB |
| tokenization | 15-03-2021 20:26 | Python File | 2 KB |
| util | 06-04-2021 20:43 | Python File | 1 KB |

**Running:**
Run the following code in terminal:
python main.py -dataset cranfield/ -out out/

There is no option present for the  segmenter and tokenizer as naive was fixed for both

## Section 3: Spell Check

### Files Required

| | | | |
|---|---|---|---|
| __pycache__ | 03-06-2021 15:12 | File folder | |
| cranfield | 03-06-2021 13:56 | File folder | |
| out | 03-06-2021 17:02 | File folder | |
| .DS_Store | 07-03-2020 13:13 | DS_STORE File | 7 KB |
| doc_words_final | 28-05-2021 16:15 | Microsoft Excel Co... | 45,467 KB |
| docs_wiki_final | 28-05-2021 16:15 | Microsoft Excel Co... | 1,09,632 KB |
| evaluation | 06-04-2021 20:18 | Python File | 12 KB |
| inflectionReduction | 06-04-2021 20:50 | Python File | 2 KB |
| informationRetrieval | 06-04-2021 22:58 | Python File | 4 KB |
| main | 07-03-2020 14:27 | Python File | 9 KB |
| README | 07-03-2020 13:12 | Text Document | 2 KB |
| sentenceSegmentation | 03-06-2021 14:00 | Python File | 2 KB |
| stopwordRemoval | 02-06-2021 23:46 | Python File | 2 KB |
| tokenization | 03-06-2021 15:09 | Python File | 3 KB |
| util | 06-04-2021 20:43 | Python File | 1 KB |
| wiki_words_intersect | 28-05-2021 16:14 | Microsoft Excel Co... | 1,31,065 KB |

**Running:**

Run the following code in terminal:

python main.py -dataset cranfield/ -out out/ -segmenter naive -tokenizer naive

## <u>Section 4: Hypothesis Testing</u>

### Files Required

| | | | |
|---|---|---|---|
| __pycache__ | 03-06-2021 18:43 | File folder | |
| cranfield | 03-06-2021 01:23 | File folder | |
| output | 03-06-2021 20:11 | File folder | |
| .DS_Store | 07-03-2020 13:13 | DS_STORE File | 7 KB |
| doc_words_final | 28-05-2021 16:15 | Microsoft Excel Co... | 45,467 KB |
| docs_wiki_final | 28-05-2021 16:15 | Microsoft Excel Co... | 1,09,632 KB |
| evaluation | 03-06-2021 03:20 | Python File | 12 KB |
| inflectionReduction | 06-04-2021 20:50 | Python File | 2 KB |
| informationRetrieval | 28-05-2021 23:27 | Python File | 3 KB |
| informationRetrieval_ESA | 28-05-2021 17:47 | Python File | 3 KB |
| informationRetrieval_LSA | 02-06-2021 06:08 | Python File | 4 KB |
| informationRetrieval_W2V | 02-06-2021 22:49 | Python File | 6 KB |
| informationRetrieval_WN | 02-06-2021 16:37 | Python File | 3 KB |
| main | 03-06-2021 22:40 | Python File | 23 KB |
| README | 07-03-2020 13:12 | Text Document | 2 KB |
| sentenceSegmentation | 03-06-2021 14:46 | Python File | 2 KB |
| stopwordRemoval | 02-06-2021 23:46 | Python File | 2 KB |
| tokenization | 03-06-2021 00:13 | Python File | 3 KB |
| util | 06-04-2021 20:43 | Python File | 1 KB |
| wiki_words_intersect | 28-05-2021 16:14 | Microsoft Excel Co... | 1,31,065 KB |

- The following .csv files should be present in the folder (these are pre-calculated TF-IDF matrices for documents and wikipedia articles)
    - docs_words_final.csv

- docs_wiki_final.csv
- Wiki_words_intersect.csv

**Running:**

Run the following code in terminal:

The baseline for all methods is considered as the vector Space Model

```
python main.py -dataset cranfield/ -out output/ -segmenter naive -tokenizer naive -method [LSA|ESA|WN]
```

The highlighted part is to select the method. The codes correspond to methods as follows:

- LSA: Latent Semantic Analysis
- ESA: Explicit Semantic Analysis
- WN: Query Expansion with Wordnet

Running the program will give the following results:

- The p value of the hypothesis test along with a comment about whether the null hypothesis can be accepted or rejected
- It outputs a plot which includes two subplots:
  - The first is a scatterplot of the nDCG/MAP values of one method vs the other
  - The second is a histogram showing the nDCG/MAP of each method