Description

Team Members: ***Pranjul Dubey (ED21S011)*** *and* ***Bhumika Gupta (EE21S060)***

This directory contains:

1. Assignment solution pdf - “Assignment_1_answers.pdf”
2. Code and dataset folder - “Assignment_code_and_dataset”
3. Assignment output folder - “Output”

To run the code:

The directory ‘Assignment_code_and_dataset’ contains cranfield dataset in folder cranfield and the python code files. An empty ‘output’ folder is present, which is used to store the output files generated by the code.

Initialise terminal to the directory ‘‘Assignment_code_and_dataset’’ and run the following code:

```
python main.py
```

To generate output for the entire dataset with defaults being punkt sentence tokenizer and penn treebank word tokenizer.

Otherwise,
python main.py [-custom] [-dataset DATASET FOLDER] [-out_folder OUTPUT FOLDER]
[-segmenter SEGMENTER TYPE (naive|punkt)] [-tokenizer TOKENIZER TYPE (naive|ptb)]

When the -custom flag is passed, the system will take a query from the user as input. When the flag is not passed, all the queries in the Cranfield dataset are considered, for example:
> python main.py -custom
> Enter query below
> Papers on Aerodynamics

Output:

The directory ‘Output’ contains the output for custom given query using punkt and ptb tokenizers and for the complete cranfield dataset using punkt and ptb.

References:

[1] https://www.nltk.org/_modules/nltk/tokenize/treebank.html
[2] https://www.nltk.org/api/nltk.tokenize.TreebankWordTokenizer.html
[3] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
[4] https://www.nltk.org/howto/stem.html
[5] https://www.holisticseo.digital/python-seo/nltk/lemmatize