

Natural Language Processing

@Fionn McCarthy

@G00414386

@version Java 11

Description

This is a console-based application that creates a document containing words with the most similar meanings to a word or phrase, and their relative distance from each other, using a word embeddings file specified by the user.

How to Run

While in the directory of the jar file on Windows Command Line (or OS alternative), enter:

- `java -cp ./dsa.jar ie.atu.sw.Runner`

to run the application.

```
C:\Users\Admin\OneDrive - Atlantic TU\Desktop\G00414386>java -cp ./dsa.jar ie.atu.sw.Runner
*****
*   ATU - Dept. of Computer Science & Applied Physics   *
*                                                         *
*                   Fionn McCarthy                       *
*                   G00414386                             *
*                                                         *
*           Similarity Search with Word Embeddings       *
*                                                         *
*****

(1) Specify Embedding File (current: ./word-embeddings.txt)
(2) Specify an Output File (current: ./out.txt)
(3) Calculate and print a word or phrase
(4) Enter number of outputs (current: 10)
(?) Quit

Select Option [1-?]>
```

Then select (1) to edit the word embeddings file path to the desired path. By default, the application uses the path './word-embeddings.txt'.

Select (2) to edit the output file path to the desired path. By default, the application uses the path './out.txt'.

Select (4) to change the number of similar words outputted to the file. By default, the application displays the top 10 words.

Select (3) to enter a word or string of words for calculation. The inputted words must be valid. The program will then calculate and display the most similar words on screen and write to the specified output file.

Features

- Specify the word embeddings file to read from
- Specify the output file to write to
- Specify number of words outputted to file
- Specify word or multiple words to compare to
- Specify formula to calculate distance (Euclidean distance formula or Cosine distance formula)
- Type mismatches are will not cause crashes
- Non-existent file directories will not cause crashes