README FILE

IBM Model 1 and 2:

Files used in this Project :

1. IBM_Model.py
2. utils.py
3. sims.py
4. CLI.py
5. d2e_wordmap.pkl
6. e2d_wordmap.pkl
7. trainFile_English.txt
8. trainFile_Dutch.txt
9. testFile_English.txt
10. testFile_Dutch.txt

To use training and test Files, download the dataset from this link :-

https://drive.google.com/drive/folders/1zYq0FPzwoD_8nIY441PeukOGQ1CsTt-E?usp=sharing

Provide the path for these files as and when required for training or testing.

Open the Command Line Interface(CLI) by executing:
python CLI.py

The Command Line Interface(CLI) provides User with 5 options :

1. Training a Corpus
2. Translation of Documents
3. Computing Cosine Similarity for 2 documents
4. Computing Jaccard Coefficient for 2 documents
5. Exit the Console

This CLI will be executing in a loop unless the user wants to exit.

[Note: The average cosine similarity and Jaccard coefficient for all the test cases is computed during the Translation phase and printed on User's Demand.]

On **Input = 1** (Training) :

The user will have to give following inputs:

- Which Model to train the data on? - IBM model 1 or IBM model 2 .
- Path of English Document
- Path of Dutch Document

(Length of both the Documents must be same)

The Training will be done for both cases :

Dutch to English and English to Dutch

After Training ,
User will be asked if this model is to be used for translation of documents.
If Yes, it performs the same steps as when input = 2,with this trained model.
If No, the loop continues.

On **Input = 2** (Translation) :

Input needs to be provided for the following:

1. How does the document need to be translated, from Dutch to English(Enter 1) or from English to Dutch(Enter 2).
2. Number of Test Documents on which translation to be performed.
3. Whether Cosine similarities(average) and Jaccard coefficient(average) needs to be printed for all the test documents(Y/N).
4. Path of Test document to be translated.

5. If input in step3 was "Y", path of expected translated document needs to provided.

Translated Output Document is saved as "Document (Doc Number).txt"

Cosine similarities and Jaccard coefficient is printed for all documents along with their average if it was asked to.

On **Input = 3** (Cosine Similarity) :

Give input for Path of document1 and Path of document2.

Cosine Similarity is printed for this pair of documents.

On **Input = 4** (Jaccard Coefficient) :

Give input for Path of document1 and Path of document2.

Jaccard coefficient is printed for this pair of documents.

On **Input = 0** (Exit) :

The Console exits.

Note : If any required input is not correct, CLI either prompts user for valid input or may result in an error.