Information Retrieval Assignment 2

1. File structure

1. For each question there is separate folder Q1, Q2, Q3
2. Data required is saved in their folder
3. For question 1 data
   A. model are in mr folder.
      a. mr/{vector size}/{cbow | fasttext | glove | sg}
   B. And similarity data is in path "Wordsimilarity_datasets/iiith_wordsim/marathi.txt"
4. For question 2 data is in root folder (hi_dev.txt , hi_train.txt)
5. For question 3 data is saved in Q3 folder as mr.txt

2. Library needed

1. pandas
2. numpy
3. torch
4. random
5. transformers
6. joblib
7. sklearn
8. pickle
9. collections
10. gensim

To install packages run makefile

3. Q1

1. run q1.py file for question 1
2. q1.py file will genrate all result csv file.

4. Q2

1. I have uploaded python script as well as notebook for this question.
2. Running Q2.py will train the model for ner task for hindi data.
3. After training is cmopleted it will test and print f1 score. while training I got F1 score as 83.

5. Q3

1. main.py file read data line by line and process it
2. for every 25% of file it genrate file , data from this 4 files are then combined and saved in csv file in formate unigram_char.csv , unigram_word.csv, unigram_syllable.csv
3. csv file contain word ,char,syllable's frequency for top 100.
4. data required is in folder name mr.txt
5. Zipfian.py create graph of log(freq) and log(rank).
6. As graph for unigram_char and bigram_word is not inversely proportional they does not follow Zipfian distribution
7. And all other follow the distribution as they as inversely proportional