Ho Vinh Thinh (2562630)
Prabal Agarwal (2561419)
Nurzat Rakhmanberdieva (2561148)

Problem 1 | Short and Long Range Dependencies

In this exercise you will experience analyzing the Short/Long Range Dependencies by Correlation Function.

- Use the document "poem.txt" provided in the materials. Start with text normalization and change all different versions of the word "you" like "your", "you'll", "you've" into "you".
- Use the correlation function provided in page 47 of Slide "Chapter 4" and compute it for the word "you" with different distances of 1 to 50.

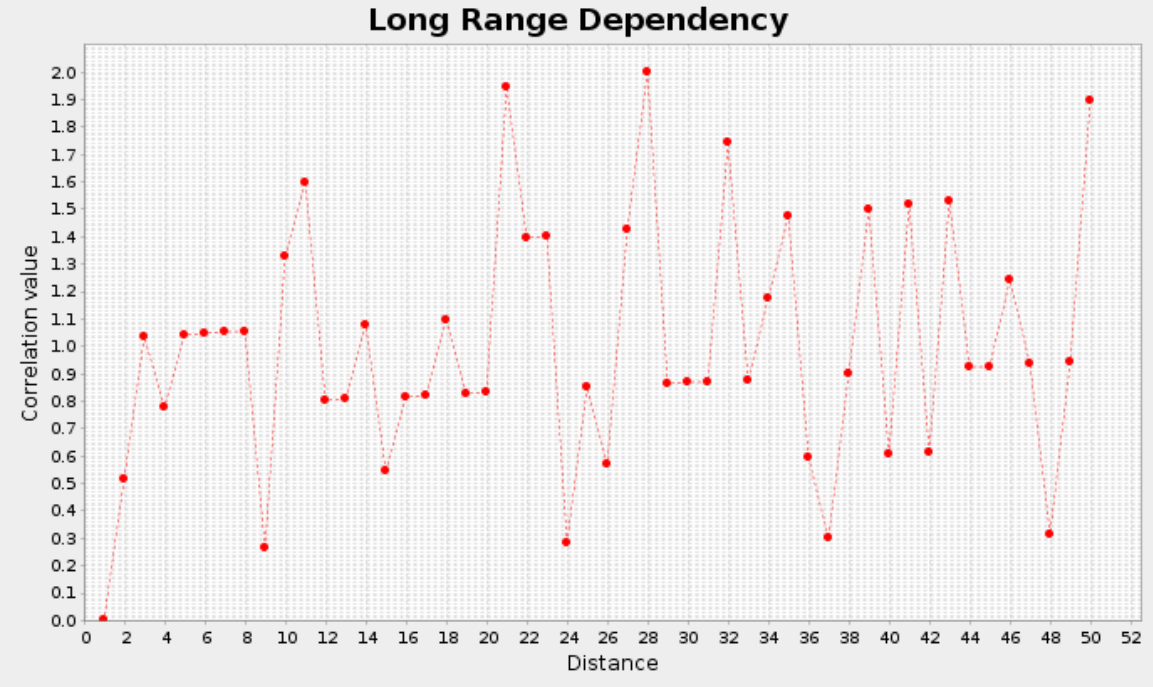
  We got the following results:

With d = 1: 0.00000000	With d = 26: 1.49792552
With d = 2: 0.54321659	With d = 27: 0.60175282
With d = 3: 1.09067702	With d = 28: 2.11525249
With d = 4: 0.82121563	With d = 29: 1.21397090
With d = 5: 0.82444876	With d = 30: 0.30481803
With d = 6: 0.82770747	With d = 31: 1.83692968
With d = 7: 1.10798931	With d = 32: 1.53751826
With d = 8: 1.11240363	With d = 33: 0.61772859
With d = 9: 0.00000000	With d = 34: 1.24094808
With d = 10: 1.40167332	With d = 35: 0.93486601
With d = 11: 1.40732515	With d = 36: 0.93905824
With d = 12: 0.84781373	With d = 37: 0.00000000
With d = 13: 0.85126013	With d = 38: 1.57926083
With d = 14: 0.85473466	With d = 39: 1.26915145
With d = 15: 0.85823768	With d = 40: 0.95621002
With d = 16: 0.57451302	With d = 41: 0.64039755
With d = 17: 0.86533052	With d = 42: 1.28669727
With d = 18: 1.73784220	With d = 43: 1.29265428
With d = 19: 0.29084721	With d = 44: 0.64933330
With d = 20: 0.87619245	With d = 45: 0.97855133
With d = 21: 1.75974786	With d = 46: 1.63857579
With d = 22: 0.88358647	With d = 47: 0.65852195
With d = 23: 1.18310726	With d = 48: 0.33082145
With d = 24: 0.59407091	With d = 49: 1.66198409
With d = 25: 0.59660965	With d = 50: 1.00196171

Ho Vinh Thinh (2562630)
Prabal Agarwal (2561419)
Nurzat Rakhmanberdieva (2561148)

- Plot the correlation vs. distance with the values obtained in the previous section. How do you explain the plot?

Figure 1:



We can see from the graph that there are fluctuations in the correlation value of Long Range Dependency. This is because of variants of word 'you' are not statistically independent. A poem has a property of following special pattern. In general, the length of each sentence in the poem does not vary much, plus it has sounding rhyme. To create such effect, the positions of some words are usually fixed. That is why, words "you" and "your" words follow a repeated pattern. Also to mention, the difference in any two spikes in the graph is the multiple of the sentence length.

Ho Vinh Thinh (2562630)
Prabal Agarwal (2561419)
Nurzat Rakhmanberdieva (2561148)

Problem 2 | OOV

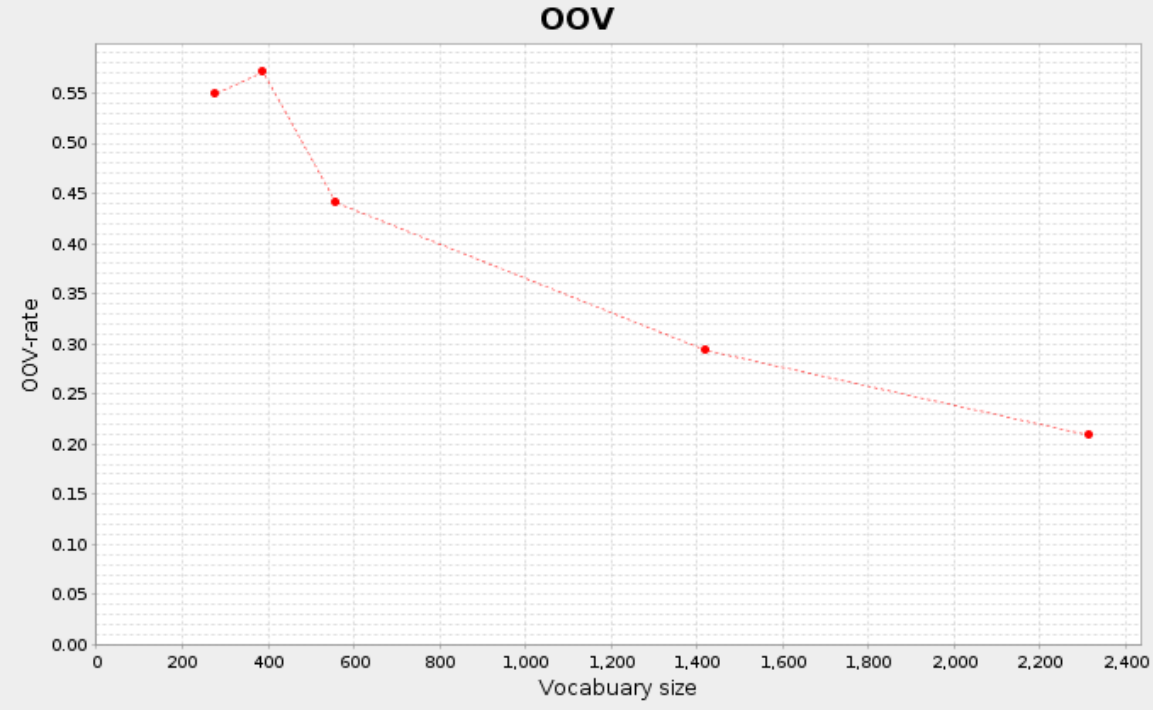
In this exercise you will experience the relationship between Out Of Vocabulary rate and Size of Vocabulary.

- Use each of the documents provided in train folder to construct 5 different vocabularies.
- Use the "test.txt" document provided in the materials which is about Physics as a test to compute OOV using each of the vocabularies.

  We got the OOV-rate of "test.txt" with 5 training documents:
  "train1.txt": 0.54828172
  "train2.txt": 0.56978101
  "train3.txt": 0.44041865
  "train4.txt": 0.29309023
  "train5.txt": 0.20752642

- Plot OOV vs. Size of the Vocabulary. How do you explain the plot?

Figure 2:



$$OOV - rate = \frac{count(unseen\ words\ in\ test\ corpus)}{count(of\ tokens\ in\ test\ corpus)} \tag{1}$$

From the graph, we can notice that as the vocabulary size increases, the $OOV - rate$ decreases. There is negative correlation between them. Because as we increase vocabulary size, we decrease the chance of seeing unseen word in test set.

- How do you explain the importance of OOV rate? Do you think out of vocabulary words will cause issues in SNLP like computing probability of a sequence of words? If yes, how do you offer to solve this issue?

Ho Vinh Thinh (2562630)
Prabal Agarwal (2561419)
Nurzat Rakhmanberdieva (2561148)

$OOV - rate$ indicates the error rate when using certain size of vocabulary. By knowing $OOV - rate$ value, we will be able to choose the size of vocabulary with certain percentage of error which can be tolerable. The out of vocabulary words will cause the problem of zero probability, which can be solved using smoothing. Smoothing will add some constant to the probability of the unseen word.

- How many different bigrams does the training data contain? How many bigrams in the test file do not occur in the training data. Give the same numbers for trigrams.

  Detail about 5 training documents:

  Document "train1.txt":
  Number of different Bigrams: 514
  Number of missing Bigrams in "test.txt": 28580/30000
  Number of different Trigrams: 548
  Number of missing Trigrams in "test.txt": 29968/29999

  Document "train2.txt":
  Number of different Bigrams: 712
  Number of missing Bigrams in "test.txt": 28419/30000
  Number of different Trigrams: 804
  Number of missing Trigrams in "test.txt": 29971/29999

  Document "train3.txt":
  Number of different Bigrams: 1423
  Number of missing Bigrams in "test.txt": 27119/30000
  Number of different Trigrams: 1672
  Number of missing Trigrams in "test.txt": 29892/29999

  Document "train4.txt":
  Number of different Bigrams: 5295
  Number of missing Bigrams in "test.txt": 22989/30000
  Number of different Trigrams: 7109
  Number of missing Trigrams in "test.txt": 27079/29999

  Document "train5.txt":
  Number of different Bigrams: 12543
  Number of missing Bigrams in "test.txt": 20113/30000
  Number of different Trigrams: 18472
  Number of missing Trigrams in "test.txt": 26472/29999

Instruction to run the the code (for Linux):

```
tar -xzvf code.tar.gz
cd code
bash build
bash run [folder]
```

bash run takes 1 parameter:
[folder]: Materials folder.
To run the code with given materials:

```
bash run ./Materials
```

The code will print out the result for both questions 1 and 2 as well as plotting the charts.