# Data – pre-processing

## Data sets

As mentioned, we used a data set for native and non-native English speakers.

We converted both datasets to lower case and we only took samples that have 45 tokens in them.

Since Toefl was the smaller dataset, we started with it and then took the same size of samples from Reddit.

For the non-native English speakers, we took 11044 samples. For the native English speakers, we had 6 different files from different countries. So, we took 1840 from each and had 11040 samples on hand. To avoid taking similar samples (sentences from the same discussion) we shuffled the files before picking the 1840 samples. We took all the 22084 samples and wrote them to a file shuffled after prepending the class label.

## Function words

We used several sources to gather a dictionary of function words. We achieved a dictionary of size 311. Every sample was represented as a vector in the size of 311 and each entry i was the count the ith function word. Below we show the top function words used by native and non-native English speakers:

|  |  |  |  |
| --- | --- | --- | --- |
| Native (11040 samples) | | Non-native (11044 samples) | |
| Function word(s) | Count | Function word(s) | Count |
| The | 29792 | The | 30795 |
| To | 16770 | To | 23452 |
| And | 15245 | And | 17640 |
| Of | 14777 | Of | 14846 |
| A | 12432 | In | 14372 |
| In | 10226 | A | 13726 |
| That | 9192 | That | 12590 |
| It | 7575 | Is | 10852 |
| Is | 7381 | I | 7780 |
| I | 6888 | It | 7280 |
| total | 130278 | total | 153333 |

We can see that the top 10 words are the same words, only the count and some of the words order are different.

Another thing that could be interesting is to see the entropy of the usage. The Larger the entropy will be, the diversion of words is bigger and that might point to conclusion that the speaker is a native speaker.

We used Minimum squared error. We calculated the average usage by summing all the counters of all the words and dividing by the number of the words. Then we summed the distance of each word counter from the average in the power of 2 and when we finished we divided by the number of words. The results:

* Native
  + Average usage=909.54
  + MSE =7,066,042
* Non-native
  + Average usage=1089
  + MSE =9,702,525

It seems the non-natives are repeating specific function words, while the natives are using more diverse function words.