Assignment 1

The corpus used for this assignment consists of tweets tweeted during the 2016 U.S Presidential Elections. The corpus consists of **6,72,893 words** from tokenized from **40,992 tweets.** The vocabulary (unique words) consists of **21,369 words**. We find that an average sentence is **16** words long.

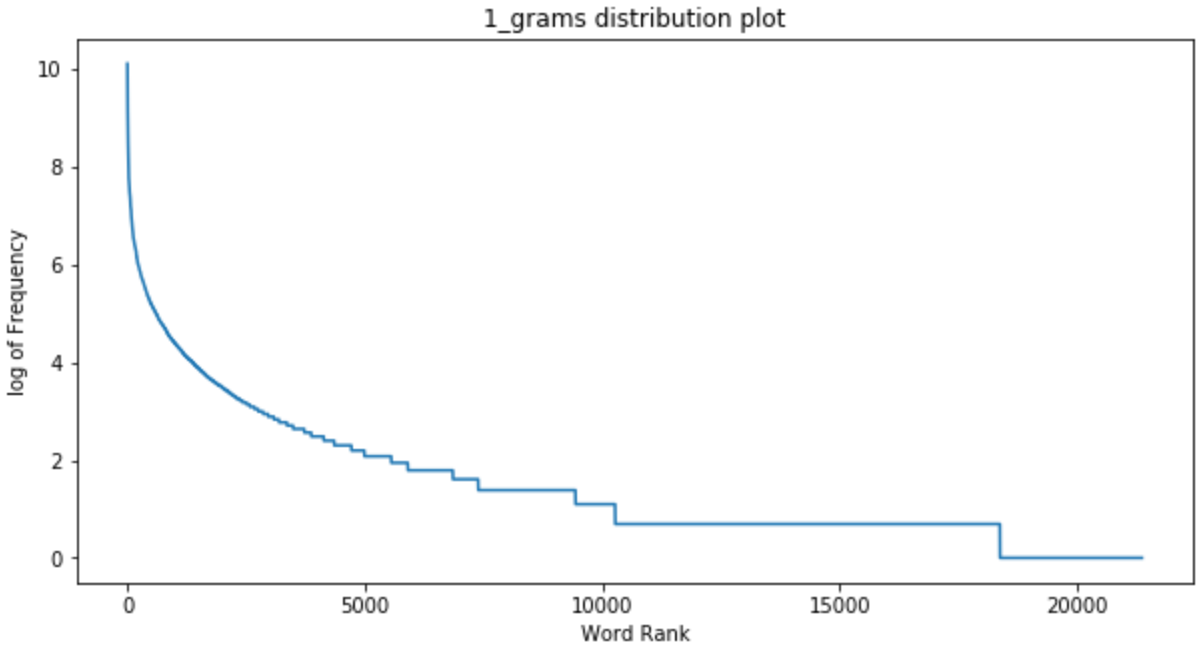

Fig 1: Unigram Distribution Plot

Top 5 Bigrams	Counts
of the	1918
in the	1840
to the	1445
on the	1387
we need	1244

Table 1: Most Common bigrams in the corpus

The figure below shows the probable words given by a trigram model for the context word **‘the world’.**

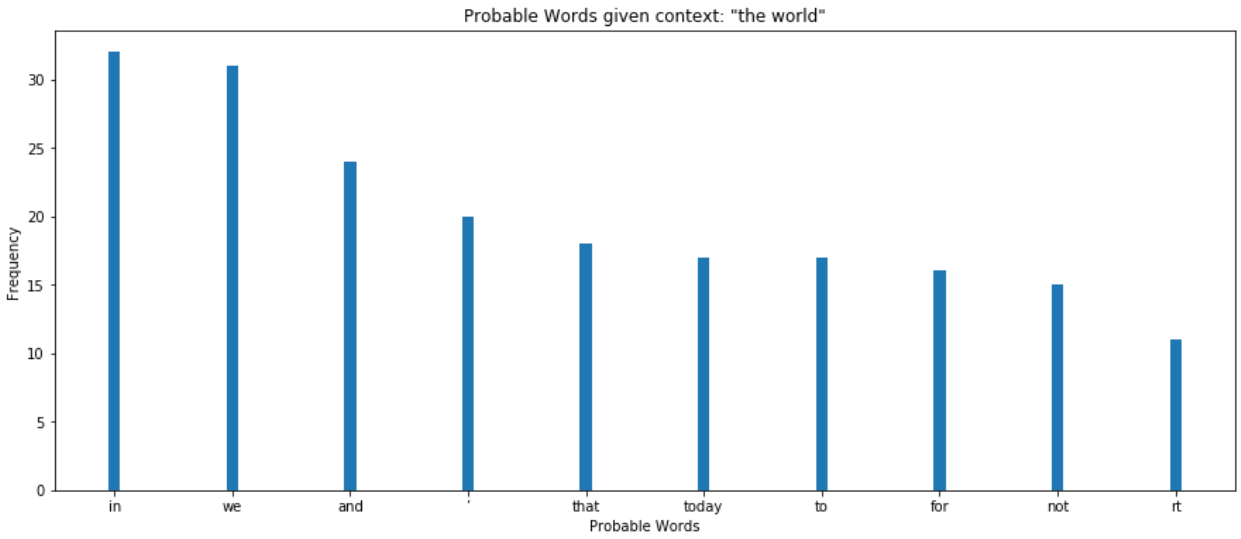

Fig 2: Probable words given the context ‘the world’

Model Evaluation

We seperate the corpus into training and test sets. After training the model on the training set, we evaluate the model on the test set.

Here is the evaluation of the model without any smoothing,

	Entropy	**Perplexity**
Unigram	0.0345	1.0246
Bigram	0.1312	1.0952
Trigram	0.1046	1.0751

Table 2: Results of a language model without smoothing

Next we apply Add-one smoothing to the model and here are the results.

	Entropy	Perplexity
Unigram	0.0337	1.0236
Bigram	4.35e-06	1.00003
Trigram	3.41e-11	1.0000

Table 3: Results of model after add-one smoothing

The next smoothing technique is the good turing smoothing technique. We see the following results after applying the above method.

	Entropy	Perplexity
Unigram	0.03544	1.0248
Bigram	0.12911	1.0936
Trigram	0.10178	1.0731

Table 4: Results of model after good-turing smoothing

The final smoothing technique applied is Kneser-Ney smoothing technique. Here are the results

	Entropy	Perplexity
Trigram	0.1282	1.0929

Table 5: Results of model after Kneser-Ney smoothing