ASSIGNMENT 2

Helper Functions:

1. For text pre-processing
2. Pickle for storing the Bigram Smoothing Model

Pre-processing Steps:

1. Removal of white spaces using Regex
2. Tokenization using NLTK
3. Removal of stop words using NLTK
4. Removal of Punctuations using python function
5. Removal of URL using Regex
6. Spelling correction using Speller
7. Lemmatization
8. Defining the starting and ending boundaries with <s> and </s>

Methodology:

- **Calculation of Beta:**
  - First, we created two models that is one for Positive Bigram Model and Negative Bigram Model.
  - In order to generate the sentiment component of sentences with greater extent (or increasingly larger)we willingly increased the number of positive(positive '1' tweets) bigrams counts for the Positive Bigram Language Model and similarly the number of negative(negative '0' tweets) bigram counts for the Negative Bigram Language Model with their frequency of occurrences in both of the Positive and Negative Corpus.
  - After doing this, we applied the Laplace Smoothing so that the zero probability can be removed and hence we don't get infinite perplexity.
- Below is the formula for generating the "beta" value:

$$(\mathrm{w(i)} \mid \mathrm{w(i-1)}) = \frac{Count(\mathrm{w(i-1)w(i)}) \ + \ 1 \ + \ k \times Count * (\mathrm{w(i-1)w(i)})}{Count(\mathrm{w(i-1)}) + \ V \ + \ \mathrm{k\,x}\ Count * (\mathrm{w(i-1)})}$$

Where

- $(\mathrm{w(i)} \mid \mathrm{w(i-1)})$ is Bigram Probability of w(i-1)w(i)

- $Count(\mathrm{w(i-1)w(i)})$ is the count of the Bigrams $\mathrm{w(i-1)w(i)}$ in the actual corpus.
- $Count * (\mathrm{w(i-1)w(i)}$ is the count of the Bigrams $\mathrm{w(i-1)w(i)}$ in the positive and negative corpus.
- $Count(\mathrm{w(i-1)})$ is the count of the unigram w(i-1) in the actual corpus.
- $Count * (\mathrm{w(i-1)}$ is the count of the unigram w(i-1) in the positive and negative corpus.
- v is the vocabulary of unique words in the actual corpus.

- We first separated the actual corpus into two copra. One with the Positive labels and the other one with the Negative labels.
- We created dictionaries for the bigrams, bigram counts, unigrams and unigram counts required for the calculations in the above bigram probability formulas.

1) Report the Top-4 bigrams and their score after smoothing.

Top 4 Bigrams
(('http', </s>), 0.014)
(('day', </s>), 0.0089)
(('lol', </s>), 0.0082)
(('work', </s>), 0.0064)

2) Report the accuracy of test set using dataset A for training.

accuracy a=0.7857142857142857

3) Report the average perplexity of the generated 500 sentences.

average perplexity score of 500 generated sentences : 3391.67

4) Report 10 generated samples: 5 positives + 5 negatives

a. Without Beta

['<s> joeymcintyre rfargnoli studios chicks durham perezhilton pictureeee tedmurphy by chess bakerzin di garage nude problems coke iilovejbxox eat sharenwatson shell </s>',
'<s> split wb goodness jennluvs2sing partially jimmycostello everybody yard gps greater emmys20 gooooooooooooood swim you iris mnr barney iris plain sal </s>',
'<s> back bacon pirrofina potency bendaubney oo online foreverivy bathroom lace shame business calibre passion unlike tinyurl stub mi88s mmmmmmm broadband </s>',
'<s> michaels create dizzee phoooone configurable stardust channel foundation mygoldenchild17 2getting hahahahah facebook york ff curious soil purple tho caitlinlynn nose </s>',
'<s> bananabby lucyfurleaps cinetrip 09 devonmarie78 fill subset chesterfield fly boleyn explorations 100 melissa officer plague paloooos case quieter assets 3rd </s>']

b. With Beta Positive

['<s> bnp nicole unblock conference tight ahahahhahaahhahh range91 bball comformfortable itschristablack fancy reign bella vis stevebrunton carcassonne brothers teemwilliams receive survive </s>',
'<s> poster quit piratesswoop cartoon mynameisforge rd kudos africa airlines advantage brothers god4movers thunderstorm algebra iranelection kris daviddjfrancis twice scratch macabroso </s>',
'<s> thelarssan 45sinyoureyes hairbrained nuggets straight gotta menu19 arent soo kirstiealley stay joey stuff woofwednesday cant decline gv odds handy jaffacakes </s>',
'<s> get lucas mrsmcflygrimmy redemption teenhearts worth katie freshman form nm animal chloe border lay greatness food broadcast creamcheese mwuahmwuah nanny </s>',
'<s> goodnight boil downstairs u ataxia pink doingwork xiaomantous moscow mate recognition marshal app mitchelmusso lonely psp ten latitude euphemism faire </s>']
1

5) **Report the accuracy of the test set using dataset B for training.**

accuracy b=0.796583850931677

Contributions:

1. Ayush Agarwal: Coding
2. Jahnvi Kadia: Report