**Report**

**Name: Sharath Chandra Bagur Suryanarayanaprasad**

**Email:** [**sbagursu@uncc.edu**](mailto:sbagursu@uncc.edu)

**ID: 800974802**

**Probability calculation:**

The probability calculations for the Unigram, Bigram and Trigram models were calculated as follows:\

Unigram Model:

Consider a sentence “This is a sentence” as an example. The probability calculation for this is done as follows-

P(this) = Count(this)/ total number of words

P(is) = Count(is)/ total number of words

P(a) = Count(a)/ total number of words

P(sentence) = Count(sentence)/ total number of words

Bigram Model:

Consider a sentence “This is a sentence” as an example. The probability calculation for this is done as follows-

P(is|this) = Count(is|this)/ Count(this)

P(a|is) = Count(a|is)/ Count(is)

P(sentence|a) = Count(sentence|a)/ Count(a)

Trigram Model:

Consider a sentence “This is a sentence” as an example. The probability calculation for this is done as follows-

P(a|is|this) = Count(a|is|this)/ Count(is|this)

P(sentence|a|is) = Count(sentence|a)/ Count(a|is)

The above model is based on occurrences of words. However I have extended the same logic to be applicable to characters so that I can predict unseen characters thereby predicting unseen words.

**Challenges Faced:**

* Handling zero occurrences of a character i.e a character that has not occurred in the training set. Resolved by considering only the characters that occur in the training set.
* Zero occurrences of a particular bigram sequence which is not present in the training set. Created the model on the basis of only the sequences that have occurred in the training set.
* Numerical Underflow due to multiplication of probability values. Used data type that handles storing large value in memory.

**Accuracy Results:**

Unigram accuracy of english model is 73.9

Unigram accuracy of french model is 26.1

Bigram accuracy of english model is 39.300000000000004

Bigram accuracy of french model is 49.8

Trigram accuracy of english model is 60.4

Trigram accuracy of french model is 10.0

Unigram accuracy of spanish model is 49.6

Unigram accuracy of italian model is 50.4

Bigram accuracy of spanish model is 49.4

Bigram accuracy of italian model is 31.8

Trigram accuracy of spanish model is 45.1

Trigram accuracy of italian model is 15.1