NLP Assignment-1 Report

201501207

I. Tokenization

The main challenges faced were URLs and contractions. The special character were removed from URLs as they add no extra information. Contractions such as “n't” and “'d” were also extracted as separate tokens as they could be continuations to multiple words.

II. Language Modeling

Question 2: The Zipf's curve follows an inverse relationship between frequencies and ranks whereas the log-log curve is linear.

Question 3: As the value of V increases, the probability masses quickly drop.

Question 6: From the graph it is evident that Laplace smoothing takes away too much probability mass from already seen n-grams. The conditional probabilities are of similar order for KN and WB smoothing. KN estimates are slightly lower.

Question 7: When WB counts are used, the KN estimates are similar to absolute discounted ones. However, when Laplace estimates are used, the counts are drastically reduced. This is explained the fact that Laplace smoothing takes away too much of the probability mass.

Question 8: The unigram text is completely meaningless. The bigram generated text shows some patterns of natural language and the trigram generated text even more.

III. Naive Bayes

Question 1: The 3 Zipf curves are very similar at the lower ranks and start to differ a little more as the ranks increase. The lower rank words are usually stop words which have similar frequencies in different texts.

Question 2: In the training corpus, we tag each character as 'I' if it lies within a token and tag it as 'O' if it lies outside all tokens.
Then, given a new sentence, we can tag each character as 'I' or 'O' using Naive Bayes Classifier

We can formulate it as:

tag_of_char = argmax_(tags) P(tag | character)
where, P(tag | character) = P(tag, character)/N

Question 3: It performs well for cases where all tokens are continuous characters in the original corpus. Since our corpora only have words and URLs, the classifier performs pretty well.