REPORT

Roll Number: 2018101052

1. Preprocessing for Tweets

a) Converted text to lowercase
b) Replaced URLs with <URL> tag.
c) Replaced HASHTAGs with <HASHTAG> tag.
d) Replaced MENTIONs with <MENTION> tag.
e) Removed repeated punctuations of the same type with single punctuation.
f) Tokenized the tweets.

2. Language Model

There were 2 types of Language Models that we built: Kneyser Ney and Witten Bell.

a) Kneyser-Ney Smoothing:
Kneyser Ney is an absolute discounting method which gives more weightage to higher order n-gram if it has seen that before and gives less weightage to lower order n-grams in that case. In the other case, where the highest order n-gram did not appear previously in the training set, it gives more weightage to the lower order n-grams and less weightage to the higher order n-gram.

b) Witten Bell Smoothing:
The idea relies on a recursive interpolation method where we backoff to the lower order n-grams with a certain probability depending on the lambda, which is calculated based on the

context and the remaining weightage is given to the decided n-gram value.

c) Observation:

i) Perplexity of LM1_train_set: 1.1847685265071486
ii) Perplexity of LM2_train_set: 0.7550009539954785
iii) Perplexity of LM3_train_set: 1.0074428281018581
iv) Perplexity of LM4_train_set: 0.5755490379392041
v) Perplexity of LM1_test_set: 4.700360929363577
vi) Perplexity of LM2_test_set: 5.587184640523427
vii) Perplexity of LM3_test_set: 4.999124170499596
viii) Perplexity of LM4_test_set: 6.19456277203167

NOTE: The values reported are log(perplexity) values.

From the above values, we observe that the log(perplexity) for the test set is much higher than that for the training set. This is because of the fact that the n-grams dictionary was constructed using the training set only and hence all the n-grams of the training set would be present in the highest order n-gram itself and would have been given more weight and hence the probability would naturally be more for the training examples. Since probability and perplexity are inversely proportional the perplexity for the training examples is much lower than those for the testing set.

d) Analysis:

By tweaking the values of n, it was analyzed that Witten Bell smoothing gives better results on lower order n-grams whereas Kneyser-Ney smoothing outperforms it for higher order n-grams.