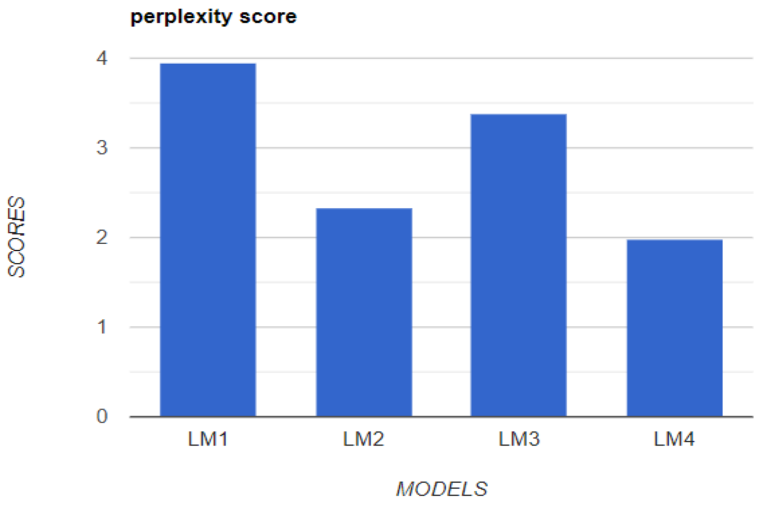
Intro To NLP

Assignment 1: Smoothing and Tokenization

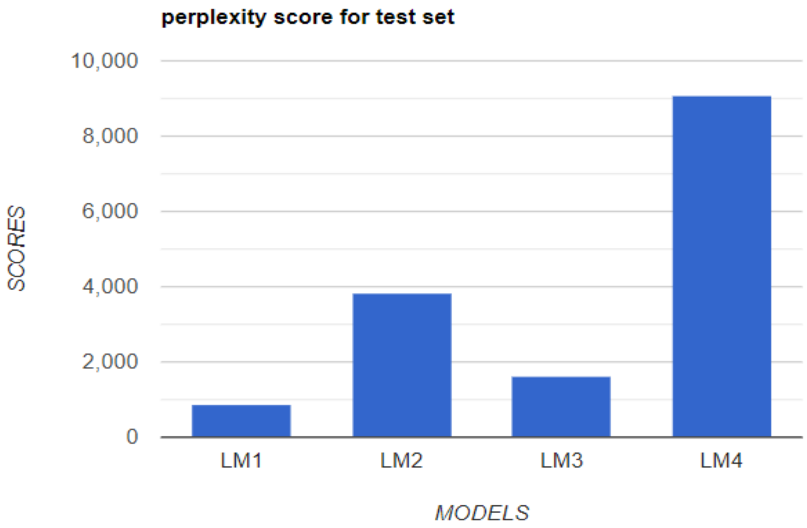
Name – Bhavya Jain(2019101095)

Results

- LM_1_test = 876.683044603444
- LM_1_train = 3.9466245082446556
- LM_2_test = 3828.0874749504433
- LM_2_train = 2.3343648826204295
- LM_3_test = 1643.6659656857248
- LM_3_train = 3.382706332961985
- LM_4_test = 9094.934809538994
- LM_4_train = 1.9902526936453713



- Above histogram plots perplexity scores vs language models.
- It is visible that Witten Bell performs better than Kneser Ney on training data.



The above histogram plots perplexity scores for the 4 language models.

LM1 and LM3 which are both Kneser Ney smoothing models and give a better perplexity avg than Witten Bell smoothing on the test data.

This means that Witten Bell has overfitted on the training vocabulary

Observation/Analysis

1. As we know earlier that the perplexity score of training set would be close to 1 (less than 4) whereas for the test set the perplexity score is much much higher (almost in order of 1000).
2. Kneser nay performs better than witten bell for the test set and performs almost similar or poorer than witten bell for training set.