Dataset

1.1 PRE - PROCESSING:

- 1. Pad each sentence in the training and test corpora with start and end symbols (you can use "~~ and ~~", respectively).
- 2. Lowercase all words in the training and test corpora. Note that the data already has been tokenized (i.e. the punctuation has been split off words).
- 3. Replace all words occurring in the training data once with the token `<unk>`. Every word in the test data not seen in training should be treated as `<unk>`.

1.2 TRAINING THE MODELS:

- A unigram maximum likelihood model
- A bigram maximum likelihood model.
- A bigram model with Add-One smoothing.

- **A bigram model with discounting and Katz backoff. Please use a discount constant of 0.5.

1.3 QUESTIONS:

- How many word types (unique words) are there in the training corpus? Please include the padding symbols and the unknown token.
- How many word tokens are there in the training corpus?
- What percentage of word tokens and word types in each of the test corpora did not occur in training (before you mapped the unknown words to in training and test data)?
- Compute the log probabilities of the following sentences under the three models (ignore capitalization and pad each sentence as described above). Please list all of the parameters required to compute the probabilities and show the complete calculation. Which of the parameters have

zero values under each model? • He was laughed off the screen . • There was no compulsion behind them . • I look forward to hearing your reply .

- Compute the perplexities of each of the sentences above under each of the models.
- Compute the perplexities of the entire test corpora, separately for the brown-test.txt and learner-test.txt under each of the models. Discuss the differences in the results you obtained.