CSCI366/CSCI780

Homework 1

September 23, 2024

Due Wednesday, 10/14/2024 5pm by email.
Please also email your report as a *pdf file* as a **separate attachment** and a tarball of your code to nlp.qc.cuny@gmail.com. The email should be sent before 5pm pm. If you send multiple emails, only your first (earliest) submission will be graded.
Please see further submission instructions at the end of the document.

1 LANGUAGE MODELING

PART I:
(**10 points**) Do exercise 3.4 from Chapter 3 in the textbook:
https://web.stanford.edu/~jurafsky/slp3/3.pdf
PART II:
In this assignment, you will train several language models and will evaluate them on a test corpus. You can discuss in groups, but the homework is to be completed and submitted *individually*. Two files are provided with this assignment:

1. *train.txt*
2. *test.txt*

Each file is a collection of texts, one sentence per line. *train.txt* contains about 100,000 sentences from the NewsCrawl corpus. You will use this corpus to train the language models. The test corpus *test.txt* is from the same domain and will be used to evaluate the language models that you trained.

1.1 PRE-PROCESSING

Prior to training, please complete the following pre-processing steps:

1. Pad each sentence in the training and test corpora with start and end symbols (you can use <s> and </s>, 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

Please use *train.txt* to train the following language models:

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

1.3 QUESTIONS

Please answer the questions below:

1. (**5 points**) How many word types (unique words) are there in the training corpus? Please include the end-of-sentence padding symbol </s> and the unknown token <unk>. Do not include the start of sentence padding symbol <s>.
2. (**5 points**) How many word tokens are there in the training corpus? Do not include the start of sentence padding symbol <s>.
3. (**10 points**) What percentage of word tokens and word types in the test corpus did not occur in training (before you mapped the unknown words to <unk> in training and test data)? Please include the padding symbol </s> in your calculations. Do not include the start of sentence padding symbol <s>.
4. (**15 points**) Now replace singletons in the training data with <unk> symbol and map words (in the test corpus) not observed in training to <unk>. What percentage of bigrams (bigram types and bigram tokens) in the test corpus did not occur in training (treat <unk> as a regular token that has been observed). Please include the padding symbol </s> in your calculations. Do not include the start of sentence padding symbol <s>.
5. (**15 points**) Compute the log probability of the following sentence 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? Use log base 2 in your calculations. Map words not observed in the training corpus to the <unk> token.

- I look forward to hearing your reply .

6. (**20 points**) Compute the perplexity of the sentence above under each of the models.

7. (**20 points**) Compute the perplexity of the entire test corpus under each of the models. Discuss the differences in the results you obtained.

1.4 SUBMISSION

Please include all the required files in a tarball and email those to nlp.qc.cuny@gmail.com using subject line `CSCI366/CSCI780 Homework 1`:

1. The Python code along with a README file that has instructions on how to run it in order to obtain the answers to questions in Section 1.3

2. The **report** that includes the answers to the questions in PART I and PART II in Section 1.3

Your grade will be based on the *correctness* of your answers, the *clarity* and completeness of your responses, and the *quality* of the code that you submitted.
Please refer to the course webpage on late submission policy.
Important: Submissions that do not follow the submission guidelines will not receive full credit. Submissions that do not attach the report as a separate attachment will lose 20 points. Submissions that do not have a report receive 0 credit. Submissions that do not include the code or provide a non-working link will lose 100 points. Please include your report as a separate attachment in the email.