CSCI381/CSCI780

Homework 1

Due 02/28/2019 11:59 pm

1 LANGUAGE MODELING

In this assignment, you will train several language models and will evaluate them on two test corpora. You can discuss in groups, but the homework is to be completed and submitted *individually*. Three files are provided with this assignment:

1. *brown-train.txt*
2. *brown-test.txt*
3. *learner-test.txt*

Each file is a collection of texts, one sentence per line. *Brown-train.txt* contains 26,000 sentences from the Brown corpus.[1] You will use this corpus to train the language models. The test corpora (*brown-test.txt* and *learner-test.txt*) will be used to evaluate the language models that you trained. *brown-test.txt* is a collection of sentences from the Brown corpus, different from the training data, and *learner-test.txt* are essays written by non-native writers of English that are part of the FCE corpus.[2]

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).

[1] http://clu.uni.no/icame/brown/bcm.html
[2] http://ilexir.co.uk/applications/clc-fce-dataset/

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 *brown-train.txt* to train the following language models (see lectures 2 and 3):

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 padding symbols and the unknown token.

2. (**5 points**) How many word tokens are there in the training corpus?

3. (**10 points**) 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 <unk> in training and test data)?

4. (**20 points**) What percentage of bigrams (bigram types and bigram tokens) in each of the test corpora that did not occur in training (treat <unk> as a token that has been observed).

5. (**20 points**) 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? Use log base 2 in your calculations. Map words not observed in the training corpus to the <unk> token.
   - He was laughed off the screen .
   - There was no compulsion behind them .
   - I look forward to hearing your reply .

6. (**20 points**) Compute the perplexities of each of the sentences above under each of the models.

7. (**20 points**) 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.

1.4 SUBMISSION

Please place the following on the **server venus.cs.qc.edu** and email me the path to the directory (in addition, please include all the required files in a tarball and email those to arozovskaya@qc.cuny.edu using subject line `CSCI381/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 writeup that includes the answers to the questions 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: Please make sure that you have uploaded the required homework files on the server. Your assignment will not be graded if the homework solution files are not accessible on the server.