CSCI 59000 NLP Spring 2024 Homework 2

January 26, 2024

Instructions

We will be using Canvas to collect your assignments. Please read the following instructions to prepare your submission.

1. Submit your solution in a pdf file and a zip file (<yourLastName_FirstName>.pdf/zip). Your write-up must be in pdf. Your code must be in the zip file.
2. In your pdf file, the solution to each problem should start on a new page.
3. Latex is strongly encouraged to write your solutions, e.g., using Overleaf (https://www.overleaf.com/). However, scanned handwritten copies are also acceptable. Hard copies will not be accepted.
4. You may discuss the problems and potential directions for solving them with another student. However, you need to write your own solutions and code separately, and not as a group activity. Please list the students you collaborated with on your submission.

Problem 1 (30 points)

We are given the following corpus:

```
<s> I am Tom </s>
<s> Tom I am </s>
<s> I am Tom </s>
<s> I am Tom and she is Jane </s>
<s> I like pink eggs and Tom </s>
<s> I like Tom </s>
<s> Tom like Jane </s>
```

(a) Using a bigram language model without smoothing, what is $P(Tom|am)$? Include <s> and </s> in your counts just like any other token. Please show your work – how you get your answer.

(b) Using a bigram language model with add-one smoothing, what is $P(Tom|am)$? Include <s> and </s> in your counts just like any other token. Please show your work – how you get your answer.

(c) If we use linear interpolation smoothing between a maximum-likelihood bigram model and a miximum-likelihood unigram model with $\lambda_1 = 1/2$ and $\lambda_2 = 1/2$, what is $P(Tom|am)$? Include <s> and </s> in your counts just like any other token. Please show work – how you get your answer.

(d) Using the above model from (a), what will be the probability of "<s> I like Jane </s>"? Please show work – how you get your answer.

(e) Using the above model from (b), what will be the probability of "<s> I like Jane </s>"? Please show work – how you get your answer.

(f) Using the above model from (c), what will be the probability of "<s> I like Jane </s>"? Please show work – how you get your answer.

Problem 2 (Programming, 70 points)

For this problem, we will use Python 3.

First, we are going to download the Brown corpus using NLTK.

```
from nltk.corpus import brown

news_data = brown.sents(categories='news')
romance_data = brown.sents(categories='romance')
```

Note that the texts are already split into sentences and also are tokenized.

Write a program to compute unsmoothed unigram and bigram models.

Before doing that, you need to lowercase all the words.

In addition, remove tokens that only consist of punctuation. For example, change ['Phil', 'nodded', '.'] into ['Phil', 'nodded'] before adding <s> and </s>. Therefore, ['Phil', 'nodded', '.'] will become ['<s>', 'Phil', 'nodded', '</s>']

Don't remove tokens that consist of both alphanumeric characters and punctuation. e.g., Richard's, 30$, Inc.

Finally, you need to include <s> and </s> before and after each sentence.

You can use the `nltk.util` package, but don't use the `nltk.lm` package, which means you need write your own function to create vocabularies, calculate MLE, etc.

Run your program on the news data and the romance data. Now compare the statistics of the two corpora.

(a) How many non-zero unigrams (in terms of counts) did you get for each corpus?

(b) How many non-zero bigrams (in terms of counts) did you get for each corpus?

(c) List the 10 most common unigrams (in terms of counts) from each dataset with their probabilities $P(w_t)$ (using MLE). You can create a table to show the numbers. Any interesting differences between the two?

(d) List the 10 most common bigrams (in terms of counts) from each dataset with their probabilities $P(w_t|w_{t-1})$ (using MLE). You can create a table to show the numbers. Any interesting differences between the two?

Write a function to compute a probability of a given sentence using a n-gram model you built above.

(e) What is the probability for "<s> I loved her when she laughed </s>" when using the bigram model from the news data?

(f) What is the probability for "<s> I loved her when she laughed </s>" when using the bigram model from the romance data?

Add an option to your program to do add-one smoothing.

(g) After applying add-one smoothing to your bigram models, what are the probabilities for "<s> I loved her when she laughed </s>" when using the model from the news data and the model from the romance data, respectively?

Bonus 2 points

If our TA and instructor can understand your code without difficulty (e.g., proper documentation, clear logic, no hard-coding, clean and modular code), you will get extra 2 points.