CSCI 3450: Natural Language Processing
Fall 2022
Assignment 01
Performance Analysis of N-gram Language Models
Assigned: Sep 15, 2022
Due: Sep 26, 2022
Total points: 100

****

This is a group activity.

If you prefer to work with a particular individual, inform (email) your professor by Sep 18, 2022 (10:00 PM). Keep the partner in the cc list. If you don't do so, your professor will assign you a partner.

****

Problem Description:

Language modeling is a fundamental task in NLP. It has applications in many related fields: sentence generation, next-word prediction, spelling correction, speech recognition, and so on. In class, we have studied a simple probabilistic language modeling technique called n-gram language modeling. We have discussed how to create a vocabulary of words and form a knowledge base from a set of training documents. We also have discussed issues like normalization, smoothing, unknown token handling, and evaluation. In this assignment, we will apply all the techniques to test the effectiveness of n-gram models (where n= 2, 3, 4). Finally, we will write a technical report.

Our learning goals are the followings:

a) to conduct scientific studies
b) to write an analytical report

The overview of the activities:

I. Preliminary Work/Implementation:

Given the training corpus,

a) Conduct preprocessing (cleaning, tokenization, lemmatization, punctuation removal, etc.)
b) Create unigram, bigram, trigram, and 4-gram dictionary.
c) Record the statistics with and without smoothing (you may use Laplace-k smoothing).
d) Write necessary methods (with comments if needed) to calculate the probability (likelihood) of a given sentence(s)/phrase.
e) Test your code.

II. Analytical Study:

Given the test corpus,

For each document:

For each sentence, find the likelihood value.

You may use log probabilities for n = 2, 3, 4-gram models.
Find the average likelihood value for each model.
Report the analysis (like the perplexity chart).

III. Report Writing:

Once the analysis is done, write a short (3~5 page, template link is provided later) report that contains at least the following sections:

a) **Title:** Performance Analysis of N-gram Language Models
b) **Author:** Your and partner's names
c) **Abstract:** In two/three sentences, outline the goal and the findings of your experiment.
d) **Introduction:** Write one/two paragraphs about Language Modeling, the n-gram model, and their applications.
e) **Dataset:** Introduce the dataset. You may add something like the following in this section, along with some narratives.

	Statistical Analysis of the Corpus
Training Corpus	#total file #unigrams #bigrams #trigrams #4-grams #unknown words #sentences
Test Corpus	

Also, talk about the preprocessing (format of the files, necessary steps to clean, etc.)

f) Result Analysis:

Take some sample sentences from the training set (three is enough) and show the likelihood calculation for each n-gram (n=2,3,4) model.

Now, report the average log-likelihood and the Perplexity for each model.
Sample:

	2-gram		3-gram		4-gram	
	Without smoothing	With Laplace-k smoothing	Without smoothing	With Laplace-k smoothing	Without smoothing	With Laplace-k smoothing
Average log-likelihood						
PP						

Before showing this table, report some charts like frequency distribution of top k unigrams (or more) before and after smoothing.

g) **Discussion/Conclusion:** Add your comments/remarks about this study.

h) **Reference:**

You **must** add your textbook and DUC 2005 (https://duc.nist.gov/duc2005/tasks.html) as references. The provided corpus (training and test) is taken from the DUC 2005 conference with their permission.

You may add any other resources that you use while completing the assignment.

IV. Submit:

◊ Code (has to be well documented)
◊ README.txt (has to contain at least some instruction about how to compile and test different methods of your implementation)
◊ Report (a pdf file)

V. Report Template:

Download the template from ACM:
https://www.acm.org/publications/proceedings-template

You may use the latex or word template. The final product (report) should be a pdf file. Name your report as **LASTNAME.pdf**. You may use either or both of the partners' last name(s) while naming the file.

VI. Point Distribution:

◊ Code: 35 points
◊ README: 10 points
◊ Report: 55 points

Final Remarks:

I will be reading your reports carefully. Put the information and data that you have discovered. Do not make any false claims.

If any of my instructions are ambiguous, feel free to email. I will get back to you at my earliest convenience.

Based on your progress and need, I may make some minor changes to the instruction set/expectations.