CS 4063 - Natural Language Processing

Due Date: Thu, Nov 25th by 11:55pm.

Assignments are to be done individually. No late assignments will be accepted.
Submissions that do not comply with the specifications given in this document will not be marked, and a zero grade will be assigned.
Write your name and e-mail id in a comment line on top of each source file. You are required to submit a single zip file containing an archive of your documentation and ipython notebook on Google Classroom. You should name your notebook as l18-XXXX.ipynb where l18-XXXX represents your student id.

Language Modeling

1 Introduction

After first assignment, you should have good idea about the stucture of language modeling. In this assignment, you will use n-gram language modeling for the given training set using the **spaCy** library for text processing. For example, following is an altered stanza from Corpus.

To this urn let those repair
That are either true or fair
For these dead birds sigh a prayer.
Sir Charles into my chamber coming in,

When I was writing of my Fairy Queen
I praysaid hewhen Queen Mab you do see
Present my service to her Majesty
And tell her I have heard Fames loud report
Both of her beauty and her stately court.
When I Queen Mab within my fancy viewed,

The task is to genrate 3 sentences of different lengths by using different n-gram models and calculate the perplexity of each model on these sentences and select the best model. Moreover, you have to justify your selected model by providing some facts. You will train unigram, bigram and trigram models using this corpus. These models will be used to generate senetence. Several online solutions are available that use the NLTK library. However, we will be using the spaCy library to accomplish this task!

2 Assignment Task

The task is to calculate perplexity using different models. The perplexity calculation problem can be solved using the following steps:

- Load the corpus
- Tokenize the corpus in order to split it into a list of words
- Generate ngram models
- Calculate perplexity of each model
- Select the best model

3 Model Selection

You will calculate the perplexity of self-generated text and provided test set using each ngram model. Then you have to select the model according to perplexity of each model.

3.1 Challenges in Implementing a Model

You have to compile a report on your assignment in which you will explain the implementation, results, and challenges you encountered while working on it.

3.2 Bonus Acitivity

Extra points will be given to your model if it generates a sentence with the least amount of perplexity. You can increase the model's performance by adding some additional constraints.

Honor Policy

This assignment is a learning opportunity that will be evaluated based on your ability to think in a group setting. Work through the problem in a logical manner, and write a research report on your own. You may however discuss verbally or via email the assignment with your classmates or the course instructor, but you are to write the actual report for this assignment without copying or plagiarizing the work of others. You may use the Internet to do your research, but the written work should be your own. **Plagiarized reports or code will get a zero**. If in doubt, ask the course instructor.