Assignment 1

Due on 21, 2019 (23:59:59)

Click here to accept your Assignment 1

Introduction

Language Modeling is one of the fundamental concepts of Natural Language Processing (NLP). In this assignment, you will build some of the basic language models and use your language models to perform authorship-determination on a list of documents.

0.1 Authorship Detection

The main aim behind this task is to practice language models. Language models define the characteristics of word order in a language. The language models not only change from one language to another, but also it changes from one author to another, which defines the writing style of an author. Therefore, we can distinguish a given text whether it is written by a specific author.

In this task, the goal is to determine the author of a given list of essays, which are written by Alexander Hamilton or James Madison. Those essays are known as The Federalist Papers (see Figure 1) and were written in 1787 to promote the ratification of the United States Constitution[1]. We leave John Jay since he contributed to the Federalist Papers with only a few essays. In this assignment, the goal is to build a language model to classify a list of essays that were written by Hamilton or Madison.

0.2 Task 1: Building Language Models

You will build unigram, bigram and trigram language models using the Federalist Papers as your training data. You may use various preprocesssing steps on the given dataset (removing punctuation, tokenizing punctuation, lowercasing the tokens). You are free to try out various preprocessing steps to observe the performance of your model under those operations.

Please use **add-one (Laplace)** smoothing for the out-of-vocabulary (OOV) problem.

Your language models will be used separately for Task 2 and Task 3.

[1] `https://en.wikipedia.org/wiki/The_Federalist_Papers`

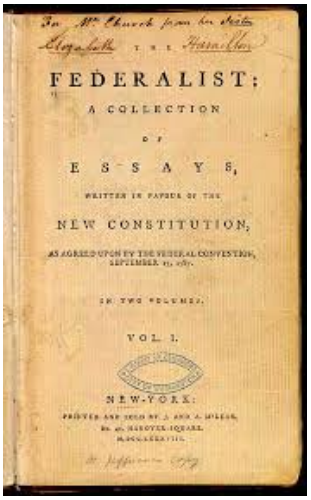

Figure 1: The Federalist Papers

0.3 Task 2: Automatically Generating Essays

Once you build your language models separately for unigram, bigram and trigram language models, you will generate 2 essays for each author using each model (unigram, bigram, trigram). Therefore, you will generate 6 essays in total for each author. Then you will estimate the probability of the automatically generated essays and compare them. What are the probabilities like in each model? Compare the probabilities of each model (unigrma, bigram and trigram) and discuss in your report.

While you generate an essay, the stopping criteria will be either getting to the end of the sentence punctuation or reaching a number of words in the essay up to 30. Therefore, each essay will have 30 words maximum.

0.4 Task 3: Classification and Evaluation

In this task you will use your language models to perform authorship detection. Once you perform authorship detection on a given test set (which are not a part of your training set), you will evaluate how your models are good at predicting the author of a given essay.

Using your language models that are built in Task 1, you will compute the perplexity of a given list of essays which are used for only testing purposes. Those essays are held-out essays and they are not a part of the training essays that you used in Task 1. Moreover, the author of the held-out essays are not known. Once you estimate the perplexity of each test essay using your language models, you will decide which essay was written by which author (Madison or Hamilton). For this task, you will use the essays 49, 50, 51, 52, 53, 54, 55, 56, 57, 62, 63.

You will use your bigram and trigram language models to estimate the perplexity separetely for two models. Are the results coherent in both models? Do the predicted authors and the actual authors match? How good are your models? What percent of the test essays are detected correctly? Discuss in your report.

What is Perplexity?
Perplexity is the inverse probability of the set, normalized by the number of words.

$$PP(W) = P(w_1 w_2 \cdots w_N)^{-\frac{1}{N}} \quad (1)$$

$$= \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_1 w_2 \cdots w_N)}} \quad (2)$$

When we use the log probabilities for the calculation, perplexity is calculated as follows:

$$PP(W) = 2^{-\frac{1}{N} \sum_{i=1}^{N} log_2 P(w_1 w_2 \cdots w_N)} \quad (3)$$

Regarding the perplexity, you may look at this blog.

Dataset

The Federalist Papers are a set of essays published in 1788, authored by Alexander Hamilton, John Jay, and James Madison.

- Download the federalist papers. Each file consists of a single essay. The first line of the file is the name of the author.
- For this assignment, you will use only some of the essays of Hamilton (1, 6, 7, 8, 9, 11, 12, 13, 15, 16, 17, 21, 22, 23, 24, 25, 26, 27, 28, 29), and all of the essays of Madison (10, 14, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 58). You will hold out three essays (9, 11, 12) of Hamilton and three essays (47, 48, 58) of Madison for testing in Task 3. The rest of the essays (17 Hamilton essays and 12 Madison essays) will be used for training in Task 1.

Submit

You are required to submit all your code. You will implement the assignment in **Python** (Python 3.5). You will submit a report in latex format template). The codes you will submit should be well commented. Your report should be self-contained and should

contain a brief overview of the problem and the details of your implemented solution. Give the answers of all questions raised in the definition of the assignment above. You can include pseudocode or figures to highlight or clarify certain aspects of your solution.

- report.pdf
- code/ (directory containing all your codes as Python file .py)

Grading

- Code (85 points): Task 1: 25, Task 2: 30 , Task 3: 30
- Report 15

Note: Preparing a good report is important as well as the correctness of your solutions! You should write your results for each part of the task and answer the questions raised in the related sections.

Academic Integrity

All work on assignments must be done individually unless stated otherwise. You are encouraged to discuss with your classmates about the given assignments, but these discussions should be carried out in an abstract way. That is, discussions related to a particular solution to a specific problem (either in actual code or in the pseudocode) will not be tolerated. In short, turning in someone else's work, in whole or in part, as your own will be considered as a violation of academic integrity. Please note that the former condition also holds for the material found on the web as everything on the web has been written by someone else.