ASSIGNMENT 1

Subject : Language Models for E-mails
Handed Out : 28.02.2018
Due date : 13.03.2018

Please submit your solution (code and a README file) by 17:00 pm on the due date. Please describe your code in detail in README file.

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 according to the training set and use your language models to check e-mails given in the test set are grammatically correct or not and generate e-mails.

0.1 Language Models

Implement unigram, bigram and trigram language models trained on the training set.

0.2 Add-One (Laplace) Smoothing

In the previous section, we do not handle unknown words. If a word is unseen in the test set, the probability is going to be zero. To handle unseen words, we use add-one (Laplace) smoothing.

Estimate the probability of the each e-mail in the test set according to the **smoothed trigram** model.

0.3 Generating E-Mails

Use your smoothed and unsmoothed unigram, bigram and trigram models to generate 10 e-mails in each models and compute probabilities of these e-mails. Compare the probabilities of these e-mails. Stop criteria of e-mails are getting end of the sentence punctuation or reaching number of words in e-mail up to 30.

0.4 Evaluation

Evaluate the smoothed bigram and trigram models' performance using perplexity of the each e-mail in the test set.

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

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

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

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

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

Dataset The Enron email data set contains approximately 500,000 (517402) emails generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse. We use %60 of the data set as the training set and %40 of the data set as the test set. We need to use the body of the mail. You may use regular expressions to extract body of the e-mail (That is after X-FileName: part). You also need to separate punctuation marks by white spaces in order to take those as also tokens. Please use punctuation used end of the sentences to separate e-mails into sentences and add sentence boundaries to each sentence.

To download the data set, please click the link :
Dataset : `https://www.kaggle.com/wcukierski/enron-email-dataset/downloads/emails.csv`

Notes

- Do not miss the submission deadline.
- Compile your code on *dev.cs.hacettepe.edu.tr* before submitting your work to make sure it compiles without any problems on our server.
- Save all your work until the assignment is graded.
- The assignment must be original, individual work. Duplicate, very similar assignments or code from Internet are going to be considered as cheating.
- You can ask your questions via Piazza and you are supposed to be aware of everything discussed on Piazza. You cannot share algorithms or source code. All work must be individual! Assignments will be checked for similarity, and there will be serious consequences if plagiarism is detected.
- You need to implement either in **Java** or **Python** (Python 3). Please submit your source codes and README file in the following submission format.
- You will be graded not only for the output, but also readability, comment lines and README.md.

- I will run your programs from the command line as following. Any other command line format will not be accepted!
  Python

  ```
  python3 assignment1.py emails.csv results.txt
  ```

  Java

  ```
  Java Main emails.csv results.txt
  ```

  → <student id>
  → code.zip
  → README.md