The University of Adelaide, School of Computer Science

Applied Natural Language Processing

Semester 1, 2022 Assignment 1: Building a sentiment analysis system with Naïve Bayes

DUE: 10 April. 2022, Sunday 11:55 PM

Submission

Instructions and submission guidelines:

• You must sign an assessment declaration coversheet to submit with your assignment.

• Submit your assignment via the Canvas MyUni.

Task

You are required to write code for building a sentiment analysis system based on the Naïve Bayes classifier. Specifically, the tasks you need to complete are:

- 1. Write code to read data from the dataset file and perform text pre-processing. You have the flexibility to choose the text-processing method. However, you need use at least two text-processing steps, e.g., stop words removal and stemming or Byte pair encoding. (15%)
- 2. Write code to build the Naïve Bayes classifier from the training set and evaluate its performance on the test set with F1 measure. (20%)
- 3. Investigate on the impact of different factors in the pipeline of building the classification system. For example, you can conduct experiment to examine the impact of using Byte-pair-encoding for text normalization or the impact of using add-k smoothing for building the Naïve Bayes classifier. At least two factors need to be investigated. Analysing more factors are encouraged and you may earn up to 5% bonus points from that (the total mark is capped to 100%). (15%)

You will use Python to write the program. Third party packages are allowed in the following cases:

1. Read text file or load data
2. Tokenization and stop words removal
3. Stemming or Byte pair encoding or other related text normalization
4. Calculating the occurrence frequency of tokens
5. Use of scientific calculation packages, e.g., numpy

The construction of Naïve Bayes classifier and F1 measure calculation need to be implemented without using third party libraries.

Note that libraries that perform elementary matrix calculation such as numpy are not counted as third part library.

You are also required to submit a report (<10 pages in PDF format), which should have the following sections (report contributes 50% to the mark; code 50%):

• 1. A description of the text pre-processing technique you used. (15%)

• 2. A description of how to train the Naïve Bayes and how to evaluate the performance. (15%)

• 3. Analysis of various factors in your implementation. You are encouraged to use tables or figures to visualize the results. (20%)

In summary, you need to submit (1) the code and (2) a report in PDF.

Data

You will use IMDB movie review dataset which is provided at MyUni

This dataset is a csv files which has Five columns

Data_ID, Type, Review Text, label, file_name.

The type indicates if the review is in the training set or testing set. **Please build your model by using information from the training set only and evaluate your model on the test set.**

Label is the class label. In this dataset, there are only two classes, pos or neg. 50% of samples are labelled “unsup” which means that they are unlabelled samples. You can simply ignore those samples.