University of Wollongong

School of Computing and Information Technology

CSCI446/946 **Big Data Analytics** **Spring 2019**

Assignment 2 **(Due: 9 October 2019, Wednesday)** **20 marks**

Aim

This assignment is intended to provide basic experience in conducting text analytics experiments with R. After having completed this assignment you should know how to perform text classification, topic modeling, and sentiment analysis.

Preliminaries

Read through the lecture notes and recommended readings on text analysis. Study all example programs therein so that you fully understand these techniques and know how to perform them with R.

Task 1 – Text Classification (6 marks)

The 20 Newsgroups data set is a benchmark for text classification. It consists of approximately 20,000 newsgroup documents, which have been categorised into 20 different newsgroups. Information on this dataset can be obtained from the webpage http://qwone.com/~jason/20Newsgroups/ . Download the "`20news-bydate-matlab.tgz`" from this webpage and unzip it to obtain the training and testing data sets. Train the Naïve Bayes classifier with the training data set and test it on the testing data set.

In your report, you need to

1. Describe this 20 Newsgroups data set.
2. Describe how each document is represented in your implementation.
3. Describe Naïve Bayes classifier and how you use it to classify the 20 Newsgroups data set.
4. Report the classification accuracy and plot the confusion matrix.
5. Attach your code at the end of the report.

Task 2 – Topic Modeling (6 marks)

Perform LDA topic modeling on the Reuters-21578 corpus using R (or Python) and LDA. The NLTK has already come with the Reuters-21578 corpus. To import this corpus, enter the following comment in the Python prompt:

```
from nltk.corpus import reuters
```

R comes with an `lda` package that has built-in functions. The LDA has also been implemented by several Python libraries such as `gensim`. Either use one such package/library or implement your own LDA to perform topic modeling on the Reuters-21578 corpus.

In your report, you need to

1. Describe the Reuters-21578 corpus.
2. Describe how each document is represented in your implementation.
3. Describe the whole procedure on applying LDA to this corpus to perform topic modeling.
4. Describe the parameter setting that you use in the LDA and explain their meanings.
5. Describe the output of your code and visualize the obtained topics in appropriate ways.
6. Attach your code at the end of the report.

Task 3 – Sentiment Analysis (8 marks)

Choose a topic of your interest, such as a movie, a celebrity, or any buzz word. Then collect 200 tweets related to this topic. Hand-tag them as positive, neutral, or negative. Next, randomly split them into 150 tweets as the training set and the remaining 50 as the testing set. Run one or more classifiers (such as Naïve Bayes, Maximum Entropy, or Support Vector Machines) over these tweets to perform sentiment analysis. Report the classification accuracy and

plot the confusion matrix. When you run more than one classifiers, find methods to evaluate which classifier performs better than the others. (*** It is not compulsory for the students of CSCI446 to run more than one classifier.**)

In your report, you need to

1. Describe the procedure of collecting the tweets and manually tagging them.
2. Describe the statistics of the obtained data set.
3. Describe how you represent each tweet for classification.
4. For each classifier, describe its working principle, classification procedure, and parameter setting.
5. For each classifier, report the classification accuracy and plot the confusion matrix.
6. (**CSCI946 only**) When you run more than one classifiers, report which classifier performs better than the others and describe the methods you use to reach this conclusion.
7. Attach your code at the end of the report.

Submit:

Important:

1. **The report must be in PDF format.**
2. **The report shall contain sufficient and detailed description, explanation, justification and discussion. Marks will be deducted for a BRIEF report.**
3. **Sufficient annotation shall be provided in your code to make it easy to understand.**

Neatly print your report and code (i.e. first the report then the code) on A4 pages with an appropriate cover sheet and hand it in during the lecture on the 9th of October 2019. Make sure your report and code are correctly formatted and titled. (Marks will be deducted for untidy or incorrectly formatted work.) Also, submit your report and the source code in a Zipped file named A2.zip via the submit link provided in the Moodle site.

Note: Failure of your code to run may attract zero marks. Code or reports considered to be unreasonably same due to copying will attract zero marks. You may be requested to demonstrate and explain your program when necessary. Marks will be awarded for correct design, implementation and style. Any request for an extension of the submission deadline or demonstration time limit must be made to the Subject Coordinator before the submission deadline. Supporting documentation must accompany the request for any extension. Late assignment submissions without granted extension will be marked but the mark awarded will be reduced by 25% of the assignment mark for each day (including weekends) late.

--- END ---