University of Wollongong

School of Computing and Information Technology

CSCI946 **Big Data Analytics** **Spring 2020**

Assignment 2 **(Due: 23:59, 17th November 2020, Beijing Time)** **20 marks**

Aim

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

Group work: You are to work as part of group on this assignment. Each group is to work independently from other groups on this assignment. You can form groups of your own accord. Each student can only join in one group. Each group should contain **no more than 4 members and no less than 2 members**. All group members are expected to contribute to this assignment. All your answers to this assignment must be accompanied with a justification and/or explanation. One submission per group only.

Penalties: If a group member fails to make a minimum in contributions, then the member will be awarded zero marks. Plagiarism of any part in this assignment will result in zero marks being awarded to the whole group.

Preliminaries

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

Task 1 – Prediction (Classification) (7 marks)

Cardiovascular (CV) disease is a primary cause of mortality in Australia and world-wide. Methods for preventing CV disease (CVD) are most effective when persons at risk are identified early. Identifying a target population for intervention is thus the key to preventing CVD, the principal challenge for health systems globally. The tool Australia has adopted to identify individuals at high risk is the Framingham risk-equation (FRE). The FRE is based on a logistic regression technique.

Your task is to:

- Use the dataset `heart.csv`
- Create a model based on logistic regression, and two further models by deploying two classification methods of your choice. Use N-fold cross-validation for all three classifiers.
- Analyze and compare results.

In your report, you need to

1. Describe the properties of the data set.
2. Justify and describe the data pre-processing techniques that you deployed.
3. Describe each of the three classifiers. Justify and discuss suitability of each classifier for the problem at hand.
4. Report, for each classifier, the classification accuracy, ROC, AUC, and plot the confusion matrix.
5. Compare results, draw conclusions, answer the question on whether logic regression is best for the task and why (or why not) it is best.
6. Attach your code to the ZIP file.

Task 2 – Topic Modeling (6 marks)

Perform LDA topic modeling on the Reuters-21578 corpus using R (or python). The NLTK in Python comes with the Reuters-21578 corpus. Install the nltp python package:

```
pip3 install --user -U nltk
```

To import this corpus, enter the following comand in the Python prompt:

```
import nltk
nltk.download('reuters')
reuters.readme()
reuters.categories()
reuters.raw()
```

Helpful resources:
https://www.programcreek.com/python/example/126583/nltk.corpus.reuters.words
https://www.nltk.org/book/ch02.html

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 to the ZIP file.

Task 3 – Sentiment Analysis (7 marks)

Choose a topic of your interest, such as a movie, a celebrity, or any buzz word. Then collect at least 200 tweets (from Twitter) related to this topic. Hand-tag them as positive, neutral, or negative. Next, randomly split them to 75% of tweets for the training set and the remaining 25% as the testing set. Deploy several classifiers (at least two) over these tweets to perform sentiment analysis. Report the classification accuracy, AUC, and plot the confusion matrix. Identify and deploy methods to evaluate which classifier performs better than the others.

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, AUC, and plot the confusion matrix.
6. Report which classifier performs better than the others and describe the methods you use to reach this conclusion.
7. Attach your code and datafile to the ZIP file.

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.**
4. **Specify the individual effort (in percentage) of your team members in the report.**

Make sure your report and code are correctly formatted and titled. Marks will be deducted for untidy or incorrectly formatted work. Submit your report and the source code files in a Zipped file named A2.zip via the submit link provided for Assignment 2 on the subjects' Moodle site. **Only one submission per group.**

Note: The code you submit is to assume that the required datafile is in the current working directory. Failure of your code to run may attract zero marks. Plagiarism of any part of your code or report will attract zero marks. It is the responsibility of the group to ensure that your submission does not contain plagiarized material. Each group member may be requested to demonstrate and explain the code or report. Any group member who fails to demonstrate or explain the code or report would receive a penalty in marks. Marks will be awarded for correct design, reasoning, completeness, implementation, and style. Marks will be deducted for late submissions. The deduction will be 25% for each day (or parts thereof) late. Submissions more than three days late will not be assessed.

--- END ---