BT5152 Decision Making Technology for Business

Assignment 4 (Due 30th October 2018 5:59 PM)

Problem Statement

In this assignment, we want to use text mining techniques, such as sentiment analysis, text classification and topic modelling to understand content in a text dataset.

Tasks

1. Conduct all relevant data-preprocessing of your textual data (TED Talk Transcripts). You need to describe your steps and processes taken. You may pre-process your dataset once and used it for all the following 3 tasks. [3 marks]
2. Using the dictionary approach, determine the emotions that are associated with the talks by counting the occurrences of the emotions terms. [3 marks]
3. Text Classification: using an appropriate algorithm you deem appropriate, perform a multi-class classification on the ratings on the talks to predict how the talks will likely to get what rating from the viewers. [3 marks]
4. Using topic modelling, find the top 10 related talks given a specific talk. You can think of this as sort of related articles feature that is common in many media sites. [3 marks]

Task 2:

The final deliverable for this task is a heatmap with one axis as emotions and the other as the talks. The colour of the heatmap will show the count of the emotions.

You can expect it to be very dense and you might not want to show the label for the talks. This visualisation basically helps us to understand what kind of emotions are in the talks.

Task 3:

For this task, you should evaluate your model based on metrics such as micro and macro F1 and describe what are the common ratings that are misclassified in your model.

Notes:

1. You will need to replace single quote with double for ratings column before using `jsonlite` to parse it.
2. Use the rating with the highest count as the label for each talk. In other words, you are predicting a categorical variable, NOT a numerical variable such as average rating.

Task 4:

You are to show a list of 10 articles for any specified article; you can show a few that have similarity above 0.5. Besides that you should also show your methodology (quantitatively or qualitatively) to derive the optimal topic model and justify your judgement.

Common notes for all tasks

- For consistency purpose, at the beginning of your script set the random seed to 5152.

Dataset Description

In order to do the assignment, you will need to use the following datasets:

1. General Inquirer Category Listings: http://www.wjh.harvard.edu/~inquirer/homecat.htm
2. Ted Talks: https://www.kaggle.com/rounakbanik/ted-talks

General Inquirer Category Listings

You should read up on what are the coverage of this dataset and determine which part of the dataset will help you to achieve the requirements of the tasks.

Ted Talks Dataset

For this dataset, it comes with 2 tables in forms of files: `ted_main.csv` and `transcripts.csv`.

You should treat `url` as the primary key to link the 2 files, `ratings` and `transcript` should be the 2 main columns you will use and they provide the count of the rating that viewers think and the content of the talk respectively.

You may use any additional fields in the dataset to help you achieve the tasks.

Submissions and Grading

- You need to submit two files :
    a. .R (or .Rmd) file
    b. .PDF (or .html generated from .Rmd) file
- Zip up all your files and name it using your matric number, e.g. A0123456X_A4.zip
    a. Name your main code as `main.R/main.Rmd` and the report as `report.PDF/report.html`
- If you are using parallel processing, do stop the cluster at the end of the script.
- You may use any packages, but make sure your R file is runnable and has all the dependency packages imported e.g. `library(C50)`.

- The page limit of the pdf file is maximum 3 pages including everything. The formatting is A4, default margin, 12 font size, single-spacing. There is no need to try to fill 3 pages. Correct answers are much more important than the length of your answers for grading.
- You may revise and submit as many times before deadline. Make sure to remove any old version that you don't wish to be graded.
- If you have questions about the assignment, feel free to email TA and cc me. Later, if you have questions about grading of assignment, then you can email TA and cc me because TA (not me) will grade your assignment by following my grading rules listed below.

Grading Policy

- Zero tolerance for plagiarism.
- Every day of late submission will result in 3 marks deducted, i.e. 4 days late = 0 mark.
- Submissions without a runnable R/Rmd file will receive a failing grade. Make sure all the dependency packages are imported e.g. `library(C50)`
- TA can judge the quality of your code and deduct up to 2 marks. For example, if you included quite a number of unnecessary codes (which shows you do not really know which line of R command is the real one that helps you conduct analysis). You can provide comments into your codes to show your understanding.
- TA can give you up to +2 bonus if you did an excellent job. The maximum marks of A4 is still 12 marks.