F20AA Applied Text Analytics: Coursework 2

Handed Out: Friday 9th February 2024.
Work organisation: group work, in groups of 3-4 students. (*For Malaysia, groups of 2-3 students)
What must be submitted: A report of 5-10 pages of A4 and accompanying software.
Submission deadline: 11:55pm Monday 1 April 2024 -- via canvas/ group space
Worth: 30% of the marks for the module**.**

Objectives:

In this coursework you will practice using essential text processing, representation, analysis and categorization tools.

In particular you will gain experience in:

- Text processing techniques: tokenization, stemming, normalization and stop-word removal, ..etc
- Exploring the effect of using n-gram features vs. unigram features
- Vector space representation (binary, frequency count & tf-idf) and word-embeddings
- Experimenting with different classification models
- Presenting and comparing results to Sequence Models, deep learning models or pre-trained models
- Topic modelling and text clustering
- Participating in a Kaggle competition

Problem Formulation:

You are given a data set of text reviews and the corresponding ratings provided by the user. The ratings score different 'Arts and Crafts' related products to 1 (low rating) up till 5 (maximum rating).

This can be considered a multi-class classification problem where your classes are the scores provided by the user {1,2,3,4,5}. The features are the text reviews.

You are requested to use your 'analytics' skills to provide insights on the data provided and to propose a categorization system that can automatically rate similar text reviews.

The Data Set:

The original dataset consists of reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014. The original data includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

Dataset source: http://jmcauley.ucsd.edu/data/amazon/links.html

You will be working on *a modified subset of the* ***Arts and Crafts*** *related product purchases.* This data set will only include the 'Review Text" and the corresponding rating {1,2,3,4,5} provided by the user. You will be provided with a train.csv file and a test.csv file to use as your training/test data sets. You will use this train.csv data set to train your models and optimize your model parameters.
You will receive detailed instructions how to participate in private class organized Kaggle competition for building a text classifier. The competition will open at least 6 weeks before submission deadline.

Implementation and Requirements:

You will use Python on the data set to conduct the following steps:

1. Data Exploration and Visualization: (5%)
   Provide an initial step to inspect, visualize and analyse the different attributes in your data set. Document your findings and make conclusions for your next steps.

2. Text Processing and Normalization: (15%)
   Thoroughly experiment with different text processing and normalization alternatives. Explain the trade-off and benefits of using each and justify their effectiveness for the current data set.

3. Vector space Model and feature representation: (20%)
   Experiment with different representation techniques. Document your findings and make conclusions. Show how choosing n-gram features can influence your results

4. Model training, selection and hyperparameter tuning and evaluation:(20%)
   You should at least experiment with 3 models and show how you can optimize model parameters using cross validation. For each model discuss your choices of text processing, representation and features from steps 1-3.

5. Modelling text as a Sequence[1]: (25%)
   So far you have investigated the document as a 'bag of word model'. Bag of word models fail to take advantage of the semantic meaning present in word ordering. Experiment with a least one model that looks at the text as a sequence (RNN, LSTM, Transformers…). Document your results and compare to the bag of word representation models in steps 2-4.

6. Topic Modelling of high and low ratings: (15%)
   Examine the five-star reviews and the one-star reviews separately. Categorize each review into a set of topics (10-20 topics). Can you infer any particular observations regarding the topics discussed in the high rated reviews vs. the low rated reviews? Document any other observations you have gained with this analysis. You may use a smaller subset of the reviews to better demonstrate your findings.

Your Report: Your report should include discussions and conclusion from experiments in steps 1-6. Summarize the insights you gained from the experiments conducted. Draw conclusions and provide your findings in a well-structured document. Justify any choices and attempts for your submissions in the Kaggle competition.

[1] For higher marks you are expected to implement multiple models that consider the text as a sequence, discuss effects of word-embeddings and compare to pre-trained models.

What to Submit (Canvas CW2 link)

(a) Well documented python notebook that includes all relevant experiments. Additionally, please upload a link to a google collab version of the notebook.
(b) A report of 5-10 pages (11 pt font, margins 2cm on all sides) documenting and discussing your findings.
(c) Multiple successful entries in the class organized Kaggle competition. Choose to apply classification models based on your investigation in steps 1-5 of the coursework. Make sure to justify your attempt and choices in your report (b). Your results should be reproducible. The competition will open at least 6 weeks before the submission deadline. 10% penalty will apply to the overall CW mark if groups do not have a Kaggle entry. *5% penalty will apply for not improving over the baseline.*
(d) Task distribution per group member

Marking: See detailed marking Rubric on Canvas. Maximum points possible: 100.

Higher marks will be assigned for work that shows original thinking, thorough discussions and critical analysis in each step. You are also required to show what you have learned in the course in addition to your independent research and findings.

Provide well-documented report and code. You are expected to present thorough experiments, be able to draw conclusions and discuss findings.

Plagiarism and Collusion

This project is assessed as **group work**. You must work within your group and not share work with other groups. Please Register your group through vision.
Students must never give hard or soft copies of their coursework reports or code to students in another group. Students must always refuse any request from another student not in their group for a copy of their report and/or code. It is expected that all group members will have read and write access to the report and code for their group.
Sharing a coursework report and/or code with another group is collusion, and if detected, this will be reported to the School's Discipline Committee. If found guilty of collusion, the penalty could involve voiding the course.

Readings, web sources and any other material that you use from sources other than lecture material must be appropriately acknowledged and referenced. Plagiarism in any part of your report will result in referral to the disciplinary committee, which may lead to you losing all marks for this coursework and may have further implications on your degree. https://www.hw.ac.uk/students/studies/examinations/plagiarism.htm

Lateness penalties

Standard university rules and penalties for late coursework submission will apply to all coursework submissions. See the student handbook.

Feedback and Interviews

A 10 min presentation time will be scheduled for each group in week 12. During this time, you will present your findings and explain your code. All group members need to attend. You will receive your marks and detailed feedback on Canvas within 14 working days from your submission deadline.