F21AA Applied Text Analytics: Coursework 1

Handed Out: Sunday 26th January 2020.
Work organisation: group work, in groups of 3-4 students.
What must be submitted: A report of 5-10 pages of A4 and accompanying software.
Submission deadline: 11:55pm Sunday 1st March 2020 -- via Vision
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)
- Experimenting with different classification models
- Topic modelling and text clustering

Problem Formulation:

Given a data set of text reviews and the corresponding ratings provided by the user. The ratings score different food 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:

This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plaintext review.

Dataset source: https://snap.stanford.edu/data/web-FineFoods.html

You will be working on *a subset of this data* set which follows the same description as the original data. You can download the .csv file of this data set at train.zip

Implementation and Requirements:

You should use Python (or any other programming language) on the data set to conduct the following steps:

1. Data Exploration and Visualization: (10%)
   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: (20%)
   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. 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.

6. Discussion and conclusion from experiments in steps 1-5: (15%)
   Summarize the insights you gained from the experiments conducted in step 1-5. Draw conclusions and provide your findings in a well-structured report. You might find it useful to compare your findings with results reported on the same data set in the literature.

What to Submit

You will submit:
(a) All evidence of conducted experiments: data sets, scripts, tables comparing the accuracy, screenshots, etc. Supply a link to your HW web space, GitHub or Google drive.
(b) A report of 5-10 pages (11 pt font, margins 2cm on all sides) documenting and discussing your findings
(c) Task distribution per group member

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

In order to get an A grade (70 points and higher), you will need to complete steps 1-6. 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.

Grade B (up to 69 points) will be awarded for implementing steps 2,3,4 and 6 and providing a well-documented report. 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.
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

Interviews will be scheduled during class on Monday 2nd March. Each group will present the highlights and the contribution of the conducted work. All members of the group must be present for the interview. You will receive your marks and a feedback on the interviews and your submitted report by March 10th, 2020