Coursework 2 Specifications

Your second coursework for COM61332 will be based on a group project focussed on Social Media Analytics (SMA) enabled by Text Mining.

Intended Learning Outcomes

- to develop and employ a social media analytics (SMA) text mining pipeline on a student-chosen problem
- to analyse the results obtained by employing the SMA pipeline on the chosen problem
- to explain how the SMA pipeline was applied to answer questions central to the problem, and summarise results
- to act as a responsible member of a team, communicate with team mates, and contribute to the team's self-organisation, planning and conflict resolution for the duration of the group work

Instructions

Given the chosen topic for your group, you will answer the research questions that you set out for that topic by developing a Social Media Analytics (SMA) pipeline.

Before starting to write any program code, however, your first step should be to identify, together with your group mates, the text mining tasks that will help you answer your research questions. For example:

1) What is the dominant sentiment towards self-driving cars? [Sentiment analysis]

2) Which brands/manufacturers of self-driving cars have been most talked about recently? Do people prefer/trust any particular brands? [Named entity recognition + Sentiment analysis]

3) Have people been discussing the pros and cons of self-driving cars? If yes, what are these? [Topic modelling + sentiment analysis]

A Jupyter notebook with boilerplate code will be made available to the class. This includes code for the following tasks, that you can build upon:

A. Data pre-processing
B. Topic modelling
C. Sentiment analysis
D. Named entity recognition
E. Named entity linking

As part of your group project, you are expected to **develop your own enhancements/novelties** to any of the code provided to you, in order to ensure that you can address your research questions adequately. For example, you could explore:

- additional, custom steps for data cleaning;
- types of models which have not been applied to your problem before;
- incorporating your own features (where applicable);
- developing/training your own models; or
- developing post-processing techniques.

The list above is only indicative and not exhaustive.

You have the choice to implement any of the above NLP tasks from scratch. For example, if you found a data set relevant to your topic that is already labelled with sentiments, then you are free to

train your own sentiment analysis classification model, as long as you can integrate it into the notebook.

Your final SMA pipeline should be in the form of a Jupyter notebook---similar to what you are provided with---which will allow you to document every step in your pipeline, as well as to more easily visualise results. While the notebook provided to you does not include any code for visualisation, you are encouraged to add this, as you might find the generated diagrams useful in preparing your report (see Deliverables section below).

Note that during the Week 5 Laboratory Session, you will be working on data set retrieval and cleaning (specifications to be given as a Lab Exercise, separately). This should allow you to focus on the NLP tasks thereafter.

Deliverables

There are two deliverables for this coursework:

1. Social media analytics pipeline. This should come in the form of a Jupyter notebook containing well-documented code for each step in the pipeline, as well as a folder containing any resources required by the code (e.g., data sets that will allow the markers to reproduce your results).
2. Short paper reporting results. This should be in the form of a research paper (rather than a technical report/user manual), something that can potentially be published (e.g., on ResearchGate, if not in a workshop). The content of a sample paper[1] published in The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS-2018) can be used as a guide. Avoid plagiarism and discuss ideas in your own words.

   Recommended template: ACL Rolling Review Template[2]

   Recommended length: 3-4 pages (in two-column format), excluding references.

Timeline

27th Feb	Submission of group topic proposals (by 18:00, via Blackboard)
28th Feb	Feedback on topic proposals given to students Decision on group project topic Lab exercise for data set retrieval/cleaning
17th Mar	Submission of deliverables (by 12:00 midnight, via Blackboard)

Marking scheme

This coursework accounts for 25% of your final mark for COMP61332, and is worth 100 points. The following rubric will be used in marking your group project, where the first column specifies the various criteria and the second column indicates the maximum number marks your group can possibly be given. The raw score (out of 40) will be multiplied by a weight (2.5) to obtain a final mark out of 100.

SMA Pipeline		
Functionality	0	The SMA pipeline is not functional.
	2	The SMA pipeline is functional but was not carefully designed to ensure that all of

[1] https://bit.ly/2tuCUXy
[2] https://www.overleaf.com/latex/templates/acl-rolling-review-template/jxbhdzhmcpdm

		the research questions set out by the group during the proposal stage are answered.
	4	The SMA pipeline is functional and was designed to answer all of the research questions set out by the group during the proposal stage.
Adaptations	0	There was no attempt to incorporate any enhancements or novelties into the code provided to the class.
	2	Some adaptations and enhancements were added to the code provided to the class, although more creativity could have been put into these.
	4	Novel adaptations and enhancements were added to the code provided to the class.
Documentation	0	The notebook does not contain any form of documentation (e.g., in-line comments, descriptions).
	2	The notebook contains some documentation but not enough to clearly explain any adaptations/enhancements incorporated into the pipeline (e.g., how these were implemented or how they improve the pipeline).
	4	The code contains documentation that clearly explains any adaptations/enhancements to the pipeline (e.g., how these were implemented or how they improve the pipeline).
Short paper		
Academic writing	0	The short paper looks like a technical report/user manual rather than a research paper.
	2	The short paper was written for an academic audience. However there are some ideas which were not clearly presented, or it seemed like the discussion lacked originality/argumentation.
	4	The short paper was written for an academic audience and can potentially be published in a research workshop or symposium. Ideas were presented in a clear and well-argued manner.
Background and introduction	0	The paper does not provide an introduction to the proposed topic (e.g., why it is interesting) and does not clearly present the analytical questions that the SMA project seeks to answer.
	2	The paper provides an introduction to the proposed topic and presents the analytical questions that the SMA project seeks to answer. However, the motivation for the choice of topic/questions is not argued well enough nor very convincing.
	4	The paper provides an introduction to the proposed topic and presents the analytical questions that the SMA project seeks to answer. The motivation for the choice of topic/questions is well-argued and convincing.
Review of related work	0	The paper does not provide any review of related work.
	2	The paper provides a review of related work although there are some shortcomings, e.g., it is not clear how these relate to the group's own work, or some of the mentioned work is outdated/more recent work could have been reviewed.
	4	The paper provides a good review of related work, showing awareness of recent

		relevant efforts. How these relate to the group's own work was clearly presented.
Methodology	0	The paper does not provide sufficient details on the group's methodology (including steps for data collection).
	2	The paper provides details on the group's methodology (including steps for data collection), although some parts need further elaboration.
	4	The paper provides sufficient and clear details on the group's methodology (including steps for data collection). Any adaptations (to the code) were explained well.
References	0	The references included in the paper are not sufficient to support the group's arguments. These were cited and added without following the recommended style/format.
	2	Sufficient references were included in the paper. However they were cited and provided in a slightly inconsistent style.
	4	Sufficient references were included in the paper, cited and provided in a consistent style.
Interpretation of results (should be included as a Results or Discussion section in the paper)		
Analysis and interpretation	0	Quantitative results were obtained by the group's SMA pipeline but were not analysed and interpreted in order to answer the research questions set out during the topic proposal stage.
	2	Quantitative results obtained by the group's SMA pipeline were analysed in order to answer the research questions set out during the proposal stage, but in some parts the interpretation seems exaggerated (or are not aligned with the results).
	4	Quantitative results obtained by the group's SMA pipeline were adequately interpreted, allowing the group to answer the research questions they set out during the proposal stage.
Exemplification/ visualisation	0	The group did not provide any examples nor make use of visualisation to evidence their analysis and interpretation of quantitative results.
	2	The group provided examples and made use of visualisation, however some of these do not support or are not aligned with their findings/interpretation.
	4	The group explained their findings and interpretation by providing supporting examples or making use of suitable visualisation.

Your deliverables will be assessed based on the marking scheme above, which will lead to one overall mark (out of 100). Everyone in your group will get the same mark: one of the Learning Outcomes of this coursework is focussed on acting as a responsible team member (see Intended Learning Outcomes), hence it is everyone's duty to ensure that tasks are delegated fairly, that there are equal contributions, and that integration goes smoothly.

In the exceptional case where one or more group members have not put in an acceptable contribution, despite the team's effort to bring them back into the team and suitable discussions of the issues arising, we implement a grievance procedure. A group can bring forward a "case of grievance" to the COMP61332 teaching staff by providing the following pieces of evidence:

- minutes of team meetings

- a written description of the events that led to the break-down of the team work, including a description of the actions that were taken to get the team back on track

The case should be brought forward to the teaching staff no later than one working week after the coursework deadline. The COMP61332 teaching staff will then decide whether the situation is indeed exceptional and warrants a mark re-distribution.