RMIT UNIVERSITY

School of Science

COSC2671 Social Media and Network Analytics

Assignment 1: Analysing and Tracking the Sentiment and Topics on Social Media

Assessment Type: Individual assignment; no group work. Submit online via Canvas→Assignments→Assignment 1. Clarifications/updates may be made via announcements/relevant discussion forums.
Due date: • Final report: 11:59pm, 30/Aug/2019 • Selection of entity to study: 12/Aug/2019 As this is a major assignment in which you demonstrate your understanding, a university standard late penalty of 10% of maximum total per each working day applies for up to 5 working days late, unless special consideration has been granted.
Weighting: 15 marks

1. Overview

One of the tasks a data scientists or analyst does is to answer questions from data – these questions could be business, social or even research oriented. In this assignment, you will answer some (interesting) questions using Twitter data. It will involve from data collection to communication of results.

Scenario

Imagine you are working for a hypothetical organisation, EasyAnalytics, which uses social media to help their clients do market research. EasyAnalytics has a potential client, who wishes to contact some market research about their brand. They are interested in knowing what people are saying about them online, and what their feeling are towards their brand. This client has never used social media-based analytics before, and would like a demonstration of what can be done. Your manager has asked you to provide this demonstration.

In this assignment, you'll use Twitter data to provide such a demonstration. You'll get to practice on going through the whole data science process. The assignment can be broken into 5 parts, but by no means do you have to following this order, and remember the process tends not to be a sequential, but one that has constant backtracking and refinement.

First, we examine the questions to answer.

Questions to Answer

To answer the clients' goals of understanding what people are talking about and how they perceive their brand, you should consider and answer the following questions:

- What are the trending concepts and topics associated with this person or event?
- What are the perceptions and feelings towards this person or event?

Data Collection

Next step is to collect data. Our focus is on Twitter, but your manager (and lecturer) want you to analyse something of interest to you. Hence, in this assignment and task, select a person, brand, product or event that you are interested about. Gather at least 1 weeks' worth of Tweets (remember for the REST API you can only gather about one week worth of data). For the analysis to be interesting, you'll want to select an entity that generates a fair amount of interest and want to have at least have a few thousand tweets per week in your data.

If you having issues with selecting an entity to study and analyse, consider selecting:

- Your favourite band, actress/actor, TV series, sport team
- Your favourite hobby
- Current major events
- A well known brand or company

As a suggestion, gather some initial tweets from your selected entity, and examine the number of tweets collected. If there are too few tweets to do meaningful analysis, select another topic/entity. If there are too many, e.g., millions of tweets in a week, select another one as the processing time will take an substantial amount of your time if there are too many tweets.

After you selected an entity to analyse, please email your lecturer about your choice. Please do this by no later than 12th of August, 2019.

Data Pre-processing and Exploration

The next suggested step is to pre-process and do some initial exploration of it. There might be a feedback loop between these two steps.

To help you get started, you might want to do the following. Compute basic statistics on the data, e.g., number of tweets, top K unique hashtags and words etc. Do you think the data appears adequate to answer the questions? Hint: remember you want a reasonable number of tweets, and likely to be more interesting to analyse if there are enough unique hashtags and words, indicating a diversity of topics discussed. Open some gathered tweets and read them. Do we need to do any pre-processing or cleaning? E.g., are their foreign language tweets (for this assignment we stick with English, so your lecturer can also read them!)? Are there characters that aren't useful for analysis.

After initial exploration and pre-processing of the data, consider what models or approaches you'll use to answer the questions. Do they need the data to be pre-processed? If so, what kind of pre-processing? Perform this pre-processing.

Method/Model

Run the selected models/approaches. Perform initial analysis – what do the results indicate? Does the approach selected have parameters – if so, what effect does the parameter settings have on the results obtained? Would a different approach produce different answer? This step requires some exploration and analysis, and that might lead to more data pre-processing.

To get started, examine the previous labs to get ideas on possible approaches.

Analysis

Remember we are trying to use data to answer the questions. Hence, for this component, first present outputs of your models and/or data analysis to answer the questions. E.g., present what are the topics been discussed about a user via a top-K terms, or word-cloud of the topics discovered by topic modelling. Then discuss this, e.g., what are the topics, does it correspond to recent news or other sources of information, if the results don't correspond to background knowledge, why you think that is so?

Communication

In this assignment, you'll produce a report about your answers, findings and insights to the questions. Describe your data, outline and describe your approach, your findings and insights to the questions. Use tables, plots/graphs, word clouds and other visualisations to help you communicate the results (in addition to text).

As the audience of the report is your hypothetical client and your lecturer, so you'll include things that might not typically include for a business report, such as describing the approach in some detail.

2. Assessment Criteria

This assessment will determine your ability to:

- Perform sentiment analysis and topic modelling on real data
- Perform data pre-processing to prepare for the analysis models
- Analyse the results and practice interpreting them
- Present and communicate the analysis, insights and justify your approach in a written report

3. Learning Outcomes

This assessment is relevant to the following Learning Outcomes:

1. Apply data science to analyse social media and social networks
2. Analyse social media by applying Natural Language Processing (NLP) techniques to detect sentiment and events
3. Describe the theoretical concepts behind the social media and network analytical approaches
4. Synthesise and present insights from the social media and network analysis performed

4. Assessment details

The assessment is predominantly based on your report and submitted scripts. Examine the assessment rubric at the end of the document, but you'll be assessed on how you approached in answering the questions and what findings and conclusions did you derive from your analysis. The report presentation is also important. In addition, your code will be assessed on their readability and essentially as a check to ensure your actual analysis justifies the findings and conclusions described in your report.

5. Referencing guidelines

What: This is an individual assignment and all submitted contents must be your own. If you have used sources of information other than the contents directly under Canvas→Modules, you must give acknowledge the sources and give references using IEEE referencing style.
Where: Add a code comment near the work to be referenced and include the reference in the IEEE style.
How: To generate a valid IEEE style reference, please use the citethisforme tool if unfamiliar with this style. Add the detailed reference before any relevant code (within code comments).

6. Submission format

- Your report, up to 10 A4 pages in length. Note this is a maximum, not a length you need to write.
- Your scripts used to perform your analysis. Please comment and style the code, as we will read and assess them.
- A sample of the data, extract first 1000 tweets from your collected dataset.

Submission should be made to Canvas. Closer to submission we will describe the submission process in more details.

7. Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarised, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods,
- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites.

If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.
RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviours, including:

- Failure to properly document a source
- Copyright material from the internet or databases
- Collusion between students

For further information on our policies and procedures, please refer to the University website.

8. Assessment declaration

When you submit work electronically, you agree to the assessment declaration.

9. Rubric/assessment criteria for marking

Criteria	Excellent	Good	Fair	Poor
Data Pre-processing (25%)	Sufficient data has been collected. Excellent data cleaning, initial exploration and cleaning. All steps are explained, and reasons and/or evidence provided to justify the different pre-processing operations.	Sufficient data has been collected. Reasonable data cleaning, initial exploration and cleaning, but one or two obvious steps missing. Most steps are explained, and reasons and/or evidence provided to justify the different pre-processing operations.	Data collection appears rushed or last minute, and possibly an insufficient sample has been collected. Poor data cleaning, initial exploration and cleaning, a significant number of obvious steps are missing. Few or no steps are explained, and reasons and/or evidence provided to justify the different pre-processing operations.	Data collection appears rushed or last minute, and possibly an insufficient sample has been collected. Absent data cleaning, initial exploration and cleaning. Steps are not explained, and reasons and/or evidence provided to justify the different pre-processing operations.
Analysis Methodology (25%)	Selection and parameter choices are fully justified with explanations and/or results and evidence.	Selection and parameter choices are mostly justified with explanations and/or results and evidence, but there are one or two missing or inadequate justifications.	Justification for the selection and parameter choices are few or missing, and not sufficiently backed by explanation or results and evidence.	Selection of approaches and parameter settings have no explanation or justification.
Analysis & Discussion (25%)	Excellent discussion of results, with good attempts at explaining what the results indicate. Reasonable conclusions and discussion supported by data, outputs of analysis, tables, graphs and other visualisations.	Decent discussion of results from analysis, with fair attempts at explaining what the results indicate. Adequate conclusions and discussion supported by data, outputs of analysis, tables, graphs and other visualisations.	Not much discussion of results, results from the models are just presented. Conclusions and discussion rarely supported by data, outputs of analysis, tables, graphs and other visualisations.	Almost no discussion of results from analysis, results from the models are just presented. Conclusions and discussion are not supported by data, outputs of analysis, tables, graphs and other visualisations.

Criteria	Excellent	Good	Fair	Poor
Report Presentation (15%)	Report is easy to read and flows well. It is structured well, leading the reader through the process of answering the questions. Tables, figures and other visualisation are easy to read and interpret.	Report is mostly easy to read, but there may be one or more paragraphs that are hard to parse. It is structured adequately, but there may be one or more sections that appears out of place. Tables, figures and other visualisation are generally easy to read and interpret, but there may be one or more factors that causes some readability issues.	Report is easy to read in places, but generally there are a several sections that are difficult to parse. It doesn't flow very well and structure is not well thought out. Tables, figures and other visualisation are not easy to read and interpret.	Report is difficult to read, and requires hard work to parse it. It doesn't flow at all and structure is not well thought out. Tables, figures and other visualisation are not easy to read and interpret.
Code Style and Readability (10%)	Code is styled and organised well, following general good programming practices. It is well commented and easy to follow the logic.	Code is styled and organized reasonably, following general good programming practices. Commenting could be improved, but generally possible to follow logic after some work.	Code is styled and organised poorly, not following general good programming practices. Commenting is rare.	Code is styled and organised poorly, not following general good programming practices. Commenting is absent.