COMP 598 Final Project - Data Science Project

Assigned Nov 10, 2021
Due Dec 13, 2021 @ 11:59 PM

This is a GROUP assignment. This document contains two project descriptions and a fine print section. Your group will select and complete ONE of these projects – with the end and graded result being a project report (to be clear, we will not be reviewing your code or data – only the final report document).

Project 1: COVID in Canada

Overview

Your team has been hired by a not-for-profit that wants to understand the discussions currently happening around COVID in Canadian social media. They have indicated that they are especially concerned with vaccine hesitancy. Specifically, they want to know:

1. The salient topics discussed around COVID and what each topic primarily concerns
2. Relative engagement with those topics
3. How positive/negative the response to the pandemic/vaccination has been.

You will conduct this analysis and submit a report discussing your findings.

Analysis Details

Your analysis will draw on Twitter posts (tweets). To inform your analysis, you should collect 1,000 tweets within a 3 day window. You should set filters such that all 1,000 posts mention either COVID, vaccination, or a name-brand COVID vaccine **AND all are in English** (ensure that the language field is set to "en"… this isn't exact, but it gets close). You can filter by hashtags or words when collecting Twitter data. You can choose the exact words, as long as they are related to the context that we mentioned before.

Each post in your collection should be unique – meaning that you shouldn't include an identical tweet or retweet twice.

To develop your topics, conduct an open coding on 200 tweets (approach the exercise requiring each tweet to belong to exactly one topic). You should aim for between 3-8 topics in total.

Once your topics have been designed, manually annotate the rest of the 1,000 tweets in your dataset. During this annotation, also code each post for positive/neutral/negative sentiment. While double annotation would usually be used, for this project (given time constraints), use single annotation. While double annotation would usually be used, for this project (given time constraints), use single annotation.

Characterize your topics by computing the **10 words in each category with the highest tf-idf** scores (to compute inverse document frequency, use all 1,000 posts that you originally collected.

Project 2: Movie Release

Overview

Your team has been hired by a media company that wants to understand the discussions currently happening around the film "(insert a recently-released movie that your team selected here)". They have indicated that they are especially concerned with the favorability of the audience response. Specifically, they want to know:

1. The salient topics discussed around their film and what each topic primarily concerns
2. Relative engagement with those topics
3. How positive/negative the response to the movie has been

You will conduct this analysis and submit a report discussing your findings.

Analysis Details

Your analysis will draw on Twitter posts (tweets). To inform your analysis, you should collect 1,000 tweets within a 3 day window. You should set filters such that all 1,000 posts have a very high likelihood of being related to the movie AND all are in English (ensure that the language field is set to "en"… this isn't exact, but it gets close). You can filter by hashtags or words when collecting Twitter data. You can choose the exact words, as long as they are related to the context that we mentioned before.

Each tweet in your collection should be unique – meaning that you shouldn't include an identical tweet or retweet twice.

To develop your topics, conduct an open coding on 200 tweets (approach the exercise requiring each tweet to belong to exactly one topic). You should aim for between 3-8 topics in total.

Once your topics have been designed, manually annotate the entire set of the 1,000 tweets in your dataset. During this annotation, also code them for positive/neutral/negative sentiment. While double annotation would usually be used, for this project (given time constraints), use single annotation.

Characterize your topics by computing the **10 words in each category with the highest tf-idf** scores (to compute inverse document frequency, use all 1,000 posts that you originally collected.

Final Report Details

Your report should be written using the AAAI format found here (you may use either Word or Latex). The template formatting (e.g., font, font size, spacing, citation style) should be followed strictly. The report structure should consist of the following sections (the lengths are suggestions):

1. Introduction (0.5 page) – general overview and key findings
2. Data (0.5 page) – describe your dataset. This should include statistics relevant to the project – the number of posts you originally started with, the number of Trump and Biden posts you had post filtering, and any design decisions you had to make around the filtering of this content.
3. Methods (0.5 page) – explanation and justification for what you did. Focus on the design decisions you made NOT listed in this document that impacted your results.

4. Results (1 pages) – share all your findings including the topics selected (and their definitions), topic characterization, and topic engagement.
5. Discussion (1 pages) – interpret your results in terms of what they reveal about the way each candidate was being discussed and perceived. Make extensive use of your results to justify your interpretations.
6. Group Member contributions (0.25 page) – a description of the contributions each group member made to this project.
7. References (< 1 page) – this is an optional section should you reference other works in your report.

The report must be between 5 and 7 pages in length, not including references. Figures are encouraged – but should be used to maximum effect (fluffy or otherwise unnecessary images that do not make strong contributions to the report will lead to point deductions).

Fine Print

- Each group will submit one report which will receive one grade that all members of the group will share. The one exception to this is in the case of strong evidence of delinquent group members. In this case, each member's grades may be adjusted up or down as appropriate.

- While there are no rules about how work should be divided up, good team participation and fair sharing of the workload are absolutely expected in this project.

Evaluation Rubric

Criteria	**Points (total 100)**	**Details**
Style	10	Is the text written in a clear, concise way? Is good grammar and spelling employed throughout?
Data collection correctness	10	Was the dataset prepared correctly? Did it have baseline characteristics that would allow this study to deliver meaningful insights?
Topic design validity	15	Was a process followed that would produce valid topics? Insufficient details should be treated the same as if something was not done.
Topic validity	15	Are the topics appropriate to the task? Are they well-defined? Are they defined to minimize subjectivity?
Annotation quality	10	Does the annotation process give us confidence in the quality of the annotations?

Results	20	Are all results requested present? Do the results make sense? Are outliers or unusual trends appropriately explained?
Findings	20	Are insightful candidate-level interpretations provided? Are these grounded in results? Do the findings integrate results and prior knowledge in a sound, well-reasoned way?

FAQ

- You can use https://www.overleaf.com/ if you decide to work with LaTeX
- There are multiple libraries for fetching Tweets from Twitter API. You can use anything you like, in any language you like. TAs will support you with Python specific tools.