COMP 370 Final Project - Data Science Project

Assigned Nov 5, 2024
Due Dec 4, 2024 @ 11:59 PM

This is a GROUP assignment. This document contains three projects 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, except in the case of suspected cheating, we will not be reviewing your code or data – only the final report document).

Project 1: Politician in the Media

Overview

Your team has been hired by a media company that wants to understand how "(insert a specific political figure your group selected)" is being covered in the media. They have indicated that they are especially concerned with North American coverage and

(1) whether coverage is positive or negative

(2) what topics the coverage focuses on.

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

Analysis Details

Your analysis will draw on news articles drawn from the NewsAPI.org. To inform your analysis, you should collect at least 500 articles on this political figure from North American news outlets (ensure that you thoughtfully select these news sources).

To develop your topics, conduct an open coding on 200 articles (approach the exercise requiring each article to belong to exactly one topic). For your open and later codings, just use the title and opening of the article (i.e., you don't need to read the entire article). You should aim for between 3-8 topics in total.

Once your topics have been designed, manually annotate the rest of the articles in your dataset. 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 500 articles that you originally collected.

Conduct a second coding of the posts by assessing whether they are positive, negative, or neutral about the politician. Bear in mind that you will need to develop a defensible way in which to interpret these categories.

Project 2: Movie Release

Overview

Your team has been hired by a media company that wants to understand the news reporting currently happening around the film "(insert a recently-released movie that your team selected here)". They have indicated that they are especially concerned with visibility and reception relative to other movies that have come out at a similar time. Specifically, they want to know:

1. What aspect of the movie was the focus (topic) of the article
2. How much coverage the movie received relative to other movies that came out at a similar time

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

Analysis Details

Your analysis will draw on news articles drawn from NewsAPI.org. To inform your analysis, you should collect 500+ articles (total) on the movies you consider in your analysis. Ensure to collect articles in a way that does not bias towards or against coverage volume of your selected movies. You should set filters such that all 500 posts have a very high likelihood of being related to one of the movies AND all are in English.

To develop your topics, conduct an open coding on 200 articles (approach the exercise requiring each article to belong to exactly one topic). For your open and later codings, just use the title and opening of the article (i.e., you don't need to read the entire article). You should aim for between 3-8 topics in total.

Once your topics have been designed, manually annotate the entire set of the 500 articles in your dataset.

Characterize your topics by computing the 10 words in each category with the highest tf-idf scores (to compute inverse document frequency, use all 500 posts that you originally collected.

Project 3: Character Comparison

Overview

Your team has been hired by a production company that wants to understand the way the side characters in "(insert a movie/TV series that your team selected here)" talk. Specifically, they want to know:

1. What topics each character tends to talk about
2. How much attention the characters give each topic

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

Analysis Details

Your analysis will draw on scripts from the show selected. To inform your analysis, you should focus on 3-5 side characters. Collect 300+ lines of non-trivial (not-just-banter) dialog from EACH character selected. Ensure to collect dialog in a way that does not bias towards/against certain topics.

To develop your topics, conduct an open coding on 100 lines from each character (approach the exercise requiring each speech act 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 speech acts in your dataset.

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

Final Report Details

Your report should be written using the Camera Ready AAAI format (you may use either Word or Latex): https://aaai.org/authorkit24-2/.

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 articles you collected, filtering you did, and any design decisions you had to make around the collection 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 (including but not limited to details in the Group Member contributions section). 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 the workload are absolutely expected in this project.

Evaluation Rubric

Criteria	**Points (100 in total)**	**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? Was sampling done to avoid problematic biases? 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 interpretations provided? Are these grounded in results? Do the findings integrate results and prior knowledge in a sound, well-reasoned way?