MI2 – Establish your Data & Register your Analysis Plan

Group 4: Dylan Beyerlein, Kelly Kohout, Logan Lee, Shreeja Tangutur (leader)
DS 4002: DS Project Course - Jan. 28th 2026

Research Question: What content words appear most frequently in Cavalier Daily News article titles in recent years, and how do these frequencies differ between News and Opinion sections?

Model Approach:

- Collect article titles from The Cavalier Daily published from 2024-2025.
- Filter the dataset to include only articles titles from the News and Opinions section.
- Preprocess the title text by lowercasing, tokenizing, and removing stopwords to focus only on content-driven words.
- Construct a unigram Bag-of-Words representation for all News article titles.
- Compute the frequency counts for each content word.
- Rank words from most to least frequently occurring within the News section.
- Rank words from most to least frequently occurring within the Opinion section.
- Create visualizations (graphs and plots) to examine vocabulary patterns
- Compare vocabulary patterns and assess whether the two sections emphasize similar or different themes.

Executive Summary: This document outlines our dataset construction and analysis plan for a study examining the most frequently used content words in Cavalier Daily article titles published after 2021. Using a Bag-of-Words frequency model, we will identify the dominant terms appearing in News and Opinion titles and compare their word-frequency distributions to evaluate whether the two sections differ in the vocabulary they emphasize.

Data set establishment details:

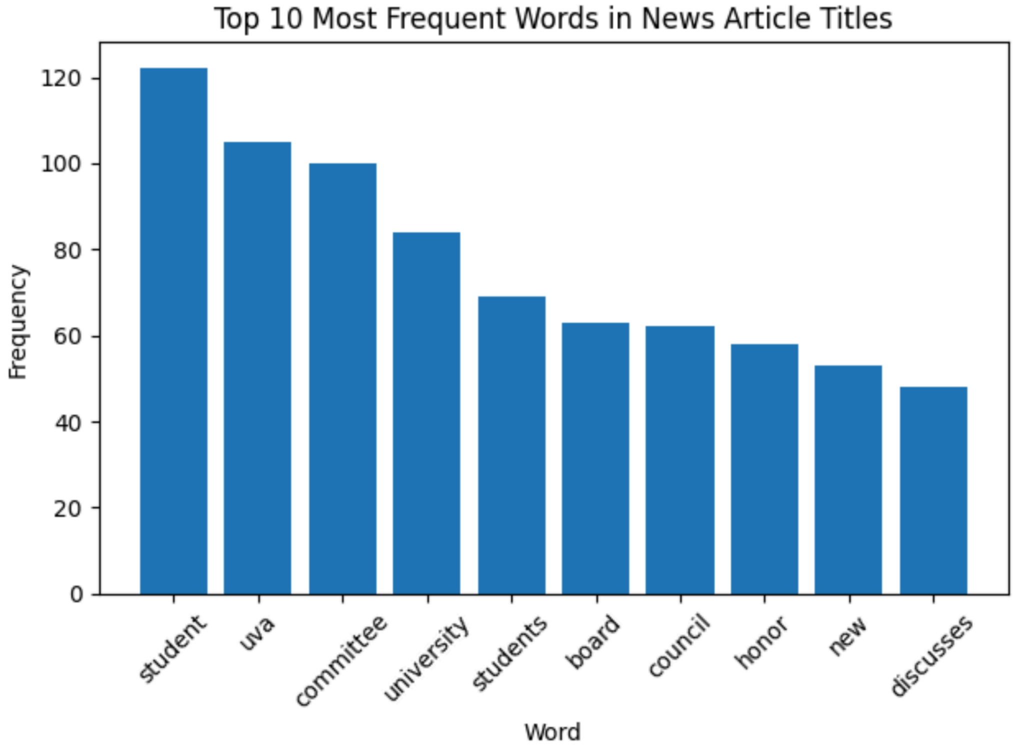
- Goal
  - Our dataset will be gathered from News and Opinion article titles published by The Cavalier Daily from 2024-2025. The data set will include the abstract, authors, content, image author, image caption, image publication date, section, slug, subhead, and title. The dataset was created by collecting articles from The Cavalier Daily website (cavalierdaily.com) and organizing it into a Firebase database using Python scripts. From this database, only the necessary fields including document ID, author, publication date, section, and title will be taken and stored in a CSV file using additional Python scripts. The CSV file will then be loaded into Python, where a Bag-of-Words model will compute the frequencies of words across News and Opinions article titles. This process allows us to observe the most popular words that represent commonly covered topics in The Cavalier Daily's sections.
- Summary of the established data set

    - Number of rows:
        - The data set contains 1003 rows, with each row representing 1 News/Opinion article.
            - News articles: 735
            - Opinion articles: 268
    - Number of columns
        - The data set contains 5 columns: doc_id, authors, published_date, section, title.
    - Time Range
        - Articles span from 2024-2025.
    - Cleaning
        - Titles were lowercased, stripped of punctuation, and stopwords were removed.
        - Publication dates were standardized to the format 01-Jan-25.
- Data Dictionary:

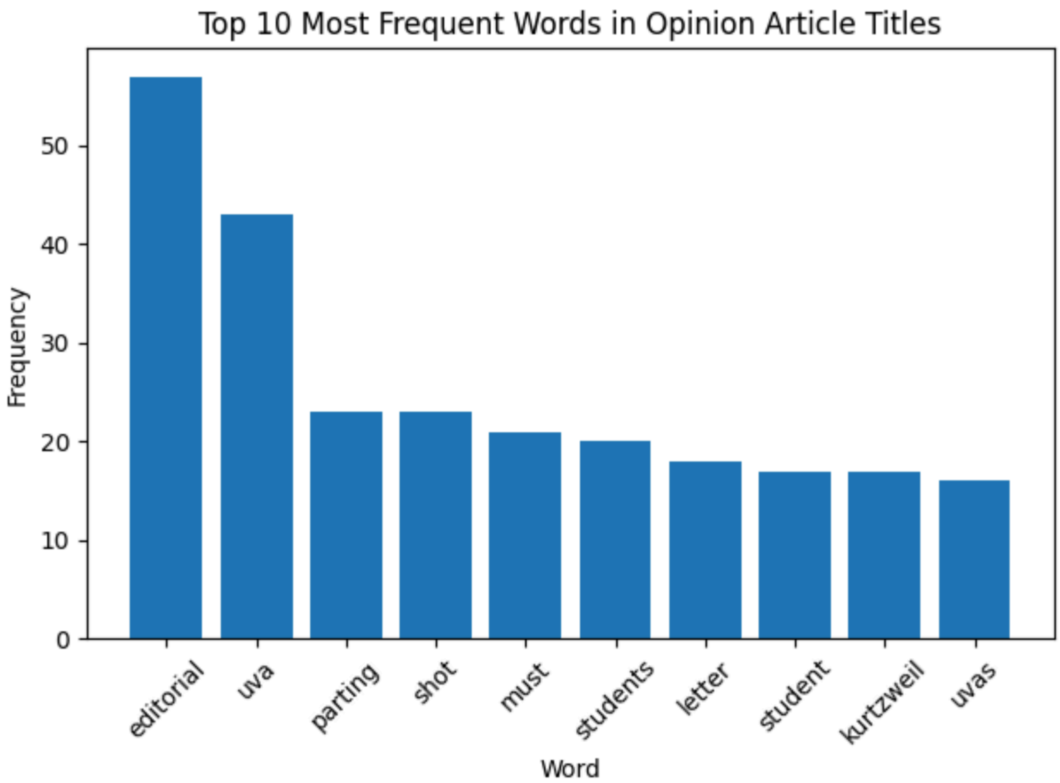
Column Name	Description	Data Type	Example Value
doc_id	The unique identifier for the article.	String	“00337a3a-5eed-4ff4-9a8e-663c59de852d”
authors	The writer(s) of the article.	String	“Ann Brown; Chris Ford”
published_date	The date the article was published.	Timestamp	17-Sep-24
section	The newspaper genre label.	String	“News”
title	The headline of the article.	String	“ncaa enacts unprecedented changes national letter intent program”

- Link to data set:
    - News/Opinion Articles
- State all the questions you explored and answered about the data set so far
    - What are the most frequently occurring content words in The Cavalier Daily News and Opinion article titles from 2024-2025?
    - How many News articles are published per year from 2024-2025?
    - How many Opinion articles are published per year from 2024-2025?
    - What columns in the dataset are most in need of cleaning?
    - How many unique words appear in News and Opinion titles after cleaning?
    - What proportion of words appear only once versus multiple times?
    - Are there noticeable differences in common words between the News and Opinion sections?

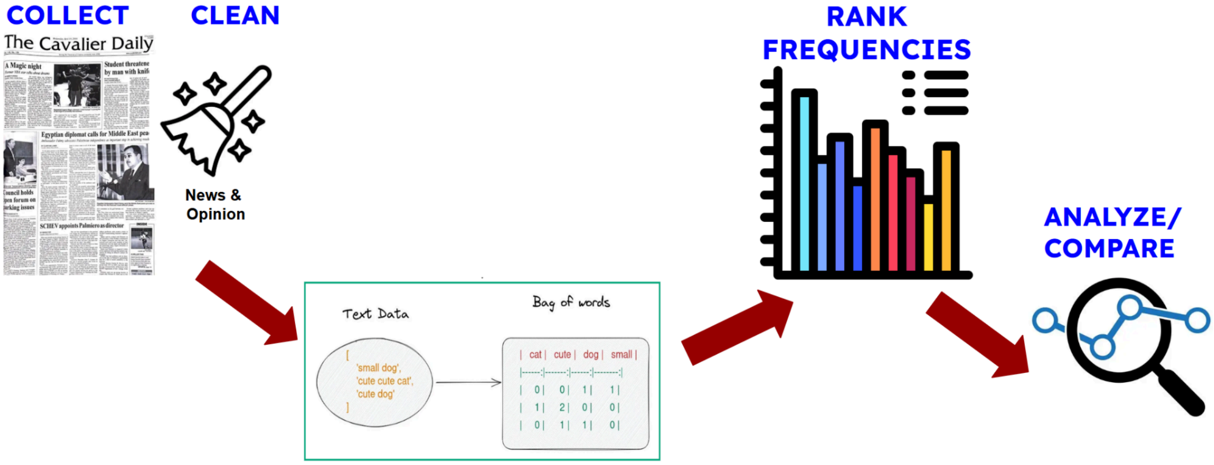
    - Do particular words dominate headlines year round or do they vary over time?
- State current unknowns or questions and where they are resolved or mitigated in your analysis plan
    - When making our plan, our main point of confusion was our evaluation method. After revising our goal and what we want to determine from our questions, we believe that a proportion test will be the best model. However, we have also entertained the idea of a t-test, z-test but they do not work out as we are not comparing one proportion at a time. We will move forward with the chi-squared, but during our analysis phase, we may try the additional models to see what results are revealed and verification. In addition, our visualizations below don't include the cleaned data as we just wanted to get a high-level understanding of the frequencies, but this will be resolved throughout the analysis process.
- State any refinement of the goal/research question/model plan that arose
    - Initially, our analysis focused solely on identifying the most frequent content words within the News section of *The Cavalier Daily*. As we progressed, we recognized the value in broadening the scope to compare vocabulary patterns across different sections of the newspaper. To deepen our analysis, we refined our research question to examine whether word-frequency distributions differ between two distinct types of articles. As a result, we expanded our model plan to include both the News and Opinion sections, allowing us to compare their Bag-of-Words frequency distributions and evaluate whether the two sections emphasize similar or different themes.
- Show any exploratory plots you created in the establishment of the data set



    -



Analysis Plan:



```
[
  'small dog',
  'cute cute cat',
  'cute dog'
]
```

cat	cute	dog	small
0	0	1	1
1	2	0	0
0	1	1	0

[2]

To perform our analysis, we will divide our responsibilities into three main sections, preprocessing involving cleaning and formatting the data, methodology where carry out the Bag-of-Words model and perform the analysis with respective visualizations, followed by an evaluation where we provide concrete metrics to understand and quantity patterns in the data.

To ensure that the data is in its most usable form, we want to start by preprocessing the data. This requires several steps. The first step is filtration: because our analysis focuses on News and Opinion articles, we keep only rows where the section equals “News” or “Opinion.” Afterwards, we address missing values by removing any rows without a valid title. While this data should be all the published articles in the News Section in the Cavalier Daily, it follows the assumption that all articles are named. However, in the case that there isn’t a title name, we would want to remove this value from the dataset. Alongside this, we would want to lowercase all the text to ensure consistency. This way, words such as “Students” and “students” would be treated as one token. Following this, we would want to tokenize the titles. Tokenization involves splitting the title into individual words. This is a critical step as it prevents frequency counts from being misleading and allows us to apply the Bag-of-Words model to the data. We would also want to remove stopwords. For instance, we would remove the words, “the,” “and,” “of,” “to,” etc, as these words don’t have any meaning and would dominate frequency distribution. To determine the correct stopwords, we can use the NLTK stopword list [1]. Finally, we remove punctuation to ensure that only clean tokens remain.

After preprocessing the titles, we will use a descriptive natural language processing approach based on a unigram Bag-of-Words model. A Bag-of-Words model builds a vocabulary of all the unique words in the dataset, counts how often each word appears, and then represents each title as a vector of these counts. We will aggregate these counts separately for News and Opinion articles to generate two frequency distributions. Words will then be ranked from most to least common within each section. To evaluate our hypothesis, we will compare the two distributions to determine whether the most frequent words in News titles differ from those in Opinion titles. This descriptive NLP approach provides a transparent framework for identifying section‑specific vocabulary patterns and assessing whether the two sections emphasize different themes.

To evaluate our analysis, we will examine whether the word‑frequency distributions produced by the Bag‑of‑Words model meaningfully differ between the News and Opinion sections. Because our project is descriptive rather than predictive, evaluation focuses on the stability and interpretability of the frequency patterns rather than metrics such as accuracy or R‑squared. We will compare the ranked word lists for each section and assess whether the most frequent content words overlap or diverge. If inferential testing is desired, we can use a chi‑square test of independence on a subset of high‑frequency words to determine whether differences in word usage between sections are statistically significant. A p‑value below 0.05 would indicate evidence against the null hypothesis that the two sections share the same word‑frequency distribution.

Specific Goal/ Hypothesis: To identify the most frequently used content words in Cavalier Daily article titles and determine whether the News and Opinion sections use similar or different vocabulary patterns.

- Null (H_0): The distribution of content words in News article titles does not differ from the distribution of content words in Opinion article titles.
- Alternative (H_1): The distribution of content words in News article titles differs from the distribution of content words in Opinion article titles.

References:

[1] 262588213843476. “NLTK’s List of English Stopwords.” Gist, 27 Aug. 2010, gist.github.com/sebleier/554280.

[2] All images Free-Use from Google Images