CS 122 Assignment 4

Due: Oct 21 , 11:59 PM
Topics: Text Mining, Data Cleaning, Sentiment Analysis, Visualization
Total Points: 100

Overview

In this assignment, you will analyze a collection of social-media posts to uncover insights about popular topics, engagement levels, and sentiment. You will clean the text data, extract hashtags and mentions, calculate engagement metrics, perform sentiment analysis, and visualize the results.

Homework Setup and Submission

1. File Structure

You must create the following directory structure, replacing *SJSUID* with your 9-digit San José State University ID number:

```
hw4_SJSUID/
├── social_media_analysis.py
├── test_homework.py
├── social_media_posts.csv
├── wordcloud.png
├── monthly_posts.png
├── likes_vs_shares.png
├── avg_sentiment_per_user.png
└── README.md
```

2. Submission

Compress the entire hw4_SJSUID directory into a single file named:

hw4_SJSUID.zip

Submit this .zip file through Canvas before the deadline.

3. Testing Your Work

A file named *test_homework.py* is provided. Place it in the root of your hw4_SJSUID directory as shown above.
This script will help you check if your implementation is correct.

4. Grading Rubric

Category	Description	Points
A. Data Wrangling	Regex cleaning, datetime conversion, missing value handling	20
B. Data Analysis	Top users and top hashtags identified and printed	20
C. Sentiment Analysis	Sentiment distribution and average per user computed and printed	20
D. Unit Testing	All provided tests pass successfully	20
E. Visualization	Four plots saved (wordcloud.png, monthly_posts.png, likes_vs_shares.png, avg_sentiment_per_user.png) with proper labels and titles	20

Task Breakdown by File

1. social_media_analysis.py

Implement all data-processing and analysis tasks here.

A) Data Wrangling (20 points)

- Clean post_content using regular expressions to remove URLs, mentions (@), hashtags (#), punctuation, and emojis.
- Create a new column cleaned_post_content.
- Convert post_date to datetime format.
- Handle missing values: fill likes with median, shares with mean, and post_content with “No Text”.
- Extract hashtags and store them in a new column Hashtags.

B) Data Analysis (20 points)

- Compute average likes and shares for each user.
- Identify the Top 3 users by total engagement (average likes + average shares).
- Determine the Top 5 hashtags based on frequency of appearance.

C) Sentiment Analysis (20 points)

- Use TextBlob to calculate a sentiment score for each post.
- Classify each post as positive, neutral, or negative based on polarity.
- Print the overall sentiment distribution and the average sentiment score per user.

D) Visualization (20 points)

Generate and save the following plots in PNG format:

1. Word Cloud of hashtags → *wordcloud.png*
2. Monthly Posts (line plot) → *monthly_posts.png*
3. Likes vs Shares (scatter plot) → *likes_vs_shares.png*
4. Average Sentiment per User (bar chart) → *avg_sentiment_per_user.png*

Each plot must have clear axis labels and titles.

Required Function Definitions

The following functions must be defined exactly with these names in your *social_media_analysis.py* file.
Your submission will be tested using these function names.

extract_hashtags(text): Returns a list of all hashtags (including #) found in the given text.
Example: "I love #AI and #Python" → ['#AI', '#Python']

extract_mentions(text): Returns a list of all mentions (including `@`) found in the text.
Example: "Thanks @Alice and @bob" → ['@Alice', '@bob']

clean_post_content(text): Cleans the text by removing URLs, mentions, hashtags, punctuation, emojis, and extra spaces.
Returns a plain string containing only the cleaned text.

get_sentiment(text): Uses `TextBlob` to compute the sentiment polarity (float).
Returns a value > 0 for positive sentiment, < 0 for negative, and 0 for neutral.

get_avg_engagement(df): Accepts a pandas DataFrame containing `user_id`, `likes`, and `shares`.
Returns a new DataFrame showing each user's average likes and average shares.

You may include other helper functions as needed, but **these five functions must exist** for your code to pass the provided tests

2. test_homework.py

This file contains unit tests for the main functions in *social_media_analysis.py.*
Tests include:

- Hashtag and mention extraction using regular expressions.
- Sentiment polarity verification (positive vs negative text).
- Average user engagement computation with pandas.

Your goal is to ensure that all these tests pass successfully.

Deliverables

Students must submit:

- All code files (*social_media_analysis.py, test_homework.py*)
- The dataset (*social_media_posts.csv*)
- All generated plots (4 PNG files)
- A short README.md summarizing insights (Top users, Top hashtags, Sentiment trends)