Group Project: Analyzing IMDb Reviews for Disney's "Snow White (2025)"

Problem Statement:

Disney has recently released its live-action adaptation of the classic tale, *Snow White (2025)*. Despite the anticipation surrounding this film, early audience reactions on platforms like IMDb suggest mixed-to-negative feedback. As a team of data scientists employed by Disney, your task is to analyze the reviews and sentiment of this movie to provide actionable insights that can help Disney better understand audience preferences, improve marketing strategies, and guide future movie production decisions.

The goal of this project is to:

1. **Scrape and Collect Data**: Extract all user reviews from IMDb for *Snow White (2025)*.
2. **Preprocess and Clean Data**: Use Natural Language Processing (NLP) techniques to clean and preprocess the review text.
3. **Perform Sentiment Analysis**: Analyze the sentiment of the reviews to determine overall audience reception and identify key themes in positive and negative feedback.
4. **Generate Insights and Recommendations**: Provide Disney with data-driven recommendations to improve their storytelling, casting, marketing, and production strategies for future projects.

Key Questions to Answer:

1. What is the overall sentiment of the audience toward *Snow White (2025)*?
2. What are the most common positive and negative themes in the reviews?
3. How do audience expectations for live-action remakes differ from those for original animated films?
4. What specific aspects of the movie (e.g., casting, visuals, storyline) received the most criticism?
5. What recommendations can be made to improve future Disney productions?

Project Instructions: IMDb Reviews Scraping and Sentiment Analysis

Web Scraping Steps:

Tools and Libraries:

- **Selenium**, **Pandas**

Step 1: Set Up WebDriver

- Install WebDriver (e.g., ChromeDriver)
- Install Selenium (`pip install selenium`)

Step 2: Navigate to IMDb Page

- URL: `https://www.imdb.com/title/tt6208148/reviews/?ref_=tt_urv`

Step 3: Scroll and Load Reviews

- Scroll using JavaScript

```
 driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
time.sleep(2)
```

- Click "Load More" repeatedly

```
 try:
     load_more_button = WebDriverWait(driver, 5).until(
         EC.element_to_be_clickable((By.XPATH, "(//span[@class='ipc-btn__text'])[16]"))
     )
     load_more_button.click()
     time.sleep(2)
     return True
except Exception:
     return False
```

Step 4: Extract Data

- Extract review text and ratings

Step 5: Save Data

- Save data to CSV (`imdb_reviews.csv`) with columns `Review Number` , `Review` , `Rating`

Expected CSV Output Example:

Review Number	Review	Rating
1	"As always, whenever there's a negative hype around a movie..."	1/10
2	"This movie isn't just bad—it's a grotesque, steaming pile of corporate excrement..."	1/10
3	"I like some of the new characters..."	7/10

Sentiment Analysis Steps:

Tools and Libraries:

- **NLTK**, **Pandas**

Step 1: Preprocessing

- Load CSV data into Pandas
- Tokenize reviews
- Remove stopwords and retain alphabetic tokens
- Lemmatize tokens
- Rejoin tokens into cleaned reviews

Step 2: Sentiment Analysis

- Initialize VADER Sentiment Analyzer
- Calculate Compound, Positive, Negative, Neutral scores

Step 3: Aggregate Results

- Calculate average compound score to determine overall sentiment
- Classify sentiment (Positive, Negative, Neutral)
- Visualize sentiment distribution using Matplotlib or Seaborn

Expected Sentiment Analysis Output Example:

Review Number	Review	Rating	Cleaned_Review	VADER_Compound	VADER_Positive	VADER_Negative	VADER_Neutral
1	"As always, whenever there's a negative hype around a movie..."	1/10	"always negative hype prove wrong"	-0.3	0.1	0.8	0.1
2	"This movie isn't just bad—it's a grotesque, steaming pile of corporate excrement..."	1/10	"movie bad grotesque pile corporate"	-0.9	0.0	0.9	0.1
3	"I like some of the new characters..."	7/10	"like character support snow white"	0.4	0.8	0.1	0.1

Deliverables:

- **GitHub Repository**:
  - Create a GitHub repo named `IMDb-Snow-White-2025-Analysis`.
  - Include contributors with clear documentation of roles.
  - Each submission of code must be properly commented, and version control must follow best practices (e.g., merging and pull requests).
- **CSV File**: Include codes used for web scraping and sentiment analysis.
- **Written Report**: Answers to all key questions listed above, based on your analysis.
- **Visualizations**: Clearly presented graphs demonstrating the distribution of sentiment scores.

By following these instructions, your team will provide Disney with essential insights into audience sentiment and actionable recommendations for improving future productions.