Professional Assignment Document

From: Data Science Lead, Broadcast Analytics Division
To: Christos Zogas
Subject: Assignment Briefing - Sentiment and Entity Analysis of Movie Reviews
Project: Rotten Tomatoes - Movie & Review Intelligence Extraction

---

Objective:
To extract, structure, and analyze insights from Rotten Tomatoes data using advanced data science methodologies, natural language processing (NLP), and sentiment analysis techniques. Your task was to derive actionable intelligence regarding public and critic perception of movies for strategic content placement and audience profiling.

Assigned Tasks:

1. **Data Acquisition & Preparation:**
    - Source data from the Kaggle dataset titled **"Rotten Tomatoes movies and critic reviews dataset"**.
    - Merge the movie metadata and review datasets into a unified structure.
    - Clean and preprocess the data to ensure consistency, removing nulls and unwanted characters.
2. **Data Structuring:**
    - Split the unified dataset into two dedicated DataFrames:
        - **Critics reviews from Rotten Tomatoes (df_rotten)**
        - **User-submitted reviews from general audience (df_reviewers)**
3. **Text Preprocessing:**
    - Implement **stemming** and **lemmatization** techniques to normalize textual data in both DataFrames.
    - Ensure all reviews are lowercased and tokenized appropriately.
4. **Descriptive Analytics & Visualization:**
    - Utilize **Matplotlib** and **Seaborn** to:

        - Display movie frequency per score level.
        - Rank and visualize **movie content ratings** distribution.
        - Identify and plot the **Top 10 Action Movies** by score.
5. **Named Entity Recognition (NER):**
    - Apply **SpaCy's NER pipeline** on critic reviews to extract entities.
    - Identify and visualize the most frequently mentioned **organization names** within critic reviews.
6. **Sentiment Analysis:**
    - Use **TextBlob** to classify reviews into **Positive**, **Negative**, or **Neutral**.
    - Apply sentiment classification to both DataFrames.
    - Further segment audience reviews into:
        - Recognized experts
        - General public
7. **Insights Extraction & Visualization:**
    - Identify movies receiving the **highest number of positive reviews** from each segment (critics, recognized, and general).
    - Display all insights using **professional bar charts** with annotations and styling for clarity.
8. **Development Environment:**
    - All analysis performed using **Python**, **VSCode**, and essential data science libraries: Pandas, NumPy, Matplotlib, Seaborn, NLTK, SpaCy, and TextBlob.

Deliverables:

- Cleaned and well-documented Jupyter Notebook / .py script.
- High-quality bar chart visualizations.
- Insight-driven commentary for each chart.
- GitHub repository with organized structure and professional documentation.

Please ensure all code and visual output are reproducible and follow best practices for readability and modularity.

For any enhancements or production-grade integration, contact the analytics engineering team.

Best regards,
Head of Data Science
Broadcast Analytics Division