Flickpedia

Digital Archaeology Final Project Proposal

Ann McClure, Isha Bhandari, Jason Choi, Justin Henley, Nayana Patil, Pooja Masani, Shashank Bandaru

Abstract—This project creates a web-based tool that allows users to search for episodes by entering quotes from a television show. By processing transcript data and episode features, the platform identifies and displays the corresponding season and episode, making it easier for fans to locate their favorite scenes.

1 Objective

The primary objective of our project is to develop a web-based platform that enables users to easily identify specific episodes and seasons of television series by entering memorable quotes. Our goal is to provide a seamless and efficient user experience, making it easy for users to search for episodes without needing to sift through entire transcripts manually. The platform will feature an intuitive search interface that will quickly match user input to the corresponding episode, providing episode titles, summaries, and relevant details.

This project is designed with accessibility and ease of use in mind, ensuring that users of all technical skill levels can interact with the platform effortlessly. By hosting this tool on a website, we aim to make the service widely available and highly responsive across different devices and browsers.

Stretch Goal: Beyond direct quote searches, our stretch goal is to implement an advanced search feature that allows users to input more general descriptions or summaries of scenes. This feature would leverage natural language processing or refined search algorithms, enabling users to find episodes even when they can't recall the exact wording of a quote. This enhancement would significantly improve the flexibility and power of the platform, catering to a broader range of user inputs.

Future Enhancements: As the project evolves, we also plan to explore potential integration's with machine learning models to improve search accuracy and responsiveness. Additionally, incorporating external APIs for episode metadata and further refining the UI for a more dynamic and engaging experience are part of our long-term vision.

2 Motivation

The inspiration for this project stems from the availability of tools for finding songs based on lyrics. For example, searching a portion of song lyrics on Google usually leads to the exact song you're looking for, and services like Shazam can even identify music based on a short audio clip. These tools have made finding songs incredibly easy and intuitive for users.

However, when it comes to TV shows, the process of finding specific episodes based on a quote is far more challenging. There aren't many accessible options for users who remember a memorable line but not the season or episode it came from. This gap in functionality motivated us to create a platform that replicates the ease of song-finding tools but for TV shows like Friends. Our goal is to enable users to search quotes and quickly locate the corresponding episode and season.

3 Data Collection and Usage

Data collection for this project involves acquiring two key datasets from Kaggle: the Friends Series Dataset and the FRIENDS TV Series - Screenplay Script dataset. The Friends Series Dataset provides structured information about each episode, including features like the episode title, season, and a summary, indexed from IMDb. The Screenplay Script Dataset contains the full text of the screenplay and dialogue for all episodes in the series.

To facilitate accurate searching, both datasets will undergo pre-processing to normalize the data. This includes converting all text to lowercase, removing special characters, and performing other text-cleaning operations to ensure consistent formatting for our search functionality. The cleaned scripts will serve as the core data for matching user-entered quotes to specific episodes. We may also store the cleaned scripts in a specialized JSON document to allow for easier recognition of episodes, reducing reliance on episode names within the datasets.

Once a relevant quote is identified, the Friends Series Dataset will be used to retrieve additional contextual information about the episode, such as the episode name and summary, to enhance the user experience.

Stretch Goals for this project may include the implementation of machine learning models or advanced search algorithms to improve the flexibility and accuracy of searches, allowing users to enter partial or approximate quotes. Additionally, there may be integration with APIs to fetch more detailed episode summaries or other supplementary information.

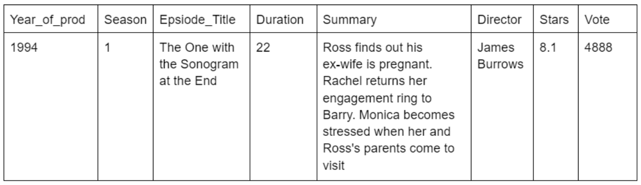
3.1 Example of Data

3.1.1 Transcript Data

"The One Where Monica Gets a New Roommate (The Pilot-The Uncut Version)
Written by: Marta Kauffman and David Crane
[Scene: Central Perk, Chandler, Joey, Phoebe, and Monica are there.]
Monica: There's nothing to tell! He's just some guy I work with!
Joey: C'mon, you're going out with the guy! There's gotta be something wrong with him!
Chandler: All right Joey, be nice. So does he have a hump? A hump and a hairpiece?
...

The transcript data represents the dialogue from Friends episodes, capturing the interactions between characters in each scene. This example, taken from the pilot episode titled "The One Where Monica Gets a New Roommate," showcases the format of the screenplay, including scene descriptions and character dialogues. This data is essential for our project, as it allows users to search for specific quotes and find the corresponding episode. The dataset consists of the full text of the scripts, enabling detailed searches and episode identification based on the user's input. [1]

3.1.2 Feature Data



Year_of_prod	Season	Epsiode_Title	Duration	Summary	Director	Stars	Vote
1994	1	The One with the Sonogram at the End	22	Ross finds out his ex-wife is pregnant. Rachel returns her engagement ring to Barry. Monica becomes stressed when her and Ross's parents come to visit	James Burrows	8.1	4888

Fig. 1. The first row of our features dataset

The feature data is a structured dataset containing key details about each episode, such as the episode title, season, IMDb ratings, air date, and summary. This visual example highlights the first row of the dataset, which includes relevant metadata for one episode. The feature dataset complements the transcript data by providing additional context and episode details, making it easier for users to understand the episode information beyond just the quote. [2]

4 Member Responsibilities

Our group has divided responsibilities to ensure efficient progress and collaboration, focusing on three primary components: User Interface (UI), backend development including web scraping, and the logic for connecting quotes to episodes. Each member is assigned based on their strengths and expertise, fostering a balanced workload and allowing us to leverage individual skills effectively.

- User Interface (UI): *Ann, Pooja, and Isha* This team will design and develop the frontend, ensuring a user-friendly interface for seamless searching and displaying episode results. They will focus on layout, aesthetics, and interaction design to provide an intuitive experience for users.
- Web Scraping and Backend Development: *Nayana and Shashank* Nayana and Shashank will be responsible for building the backend infrastructure, including web scraping to collect episode transcripts. They will ensure the data is processed, cleaned, and stored efficiently to support the search functionality.
- Connecting Quote to Episode: *Justin and Jason* Justin and Jason will focus on developing the core logic that links user-entered quotes to the corresponding episodes. This will involve searching the processed scripts and integrating the results with the UI, as well as refining search algorithms for accuracy.

5 Project Timeline

- Oct 3: Complete project proposal, outlining objectives, datasets, and methodologies.
- Oct 15: Finalize data collection and begin preprocessing, including text normalization and data cleaning.
- Oct 31: Deliver a working prototype, including a functional search feature that identifies relevant episodes based on user-entered quotes.
- Nov 22: Expand dataset with additional features, refine search algorithms, and test user interface for improved usability.
- Nov 30: Finalize all core features and polish UI/UX; complete user testing.
- Dec 5: Prepare and finalize the presentation, including analysis of project outcomes, insights, and future developments.
- Dec 5 - Finals Week: Deliver the project presentation and submit all required materials.

6 Expected Outcome

The final product we aim to produce is a website that acts like search engine for television shows. The back-end will store a database of scripts from multiple shows and will handle search logic that will narrow down the user's search to a set of episodes or one specific episode. The identified episodes will also have a concise summary attached to it to give users additional context about the plot and setting. The search logic will use advanced algorithms to analyze user input and match it with the most relevant scripts in the database.

References

[1] B. Densil, *Friends tv series - screenplay script*, https://www.kaggle.com/datasets/blessondensil294/friends-tv-series-screenplay-script?select=S01E01+Monica+Gets+A+Roommate.txt, Accessed: 2024-10-01.

[2] M. R. Ghari and M. Dhade, *Friends series dataset*, https://www.kaggle.com/datasets/rezaghari/friends-series-dataset?resource=download, Accessed: 2024-10-01.