Text Source: Movie/TV Show Subtitles - "Friends" TV Show

ExactSubtitles (https://exactsubtitles.com/friends-season-1-10-english-subtitles-complete/):

This is one of the largest databases of subtitles for movies and TV shows. We can search for "Friends" and download the subtitles for each episode in multiple languages (I have used English language).

In [9]:
```
from google.colab import files

# Upload the PDF file
uploaded = files.upload()
```

```
Choose Files No file chosen    Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving friends_s01e01_720p_bluray_x264-sujaidr.srt to friends_s01e01_720p_bluray_x264-sujaidr.srt
```

In [12]:
```
!pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk) (2023.12.25)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.4)
```

In [13]:
```
import nltk
from nltk.tokenize import word_tokenize
import os

# Download necessary NLTK data
nltk.download('punkt')

# Initialize a variable to store all text
all_text = ""

# Loop through each uploaded subtitle file
for filename in os.listdir('/content'):
    if filename.endswith('.srt') or filename.endswith('.txt'):
        with open(os.path.join('/content', filename), 'r', encoding='utf-8') as file:
            all_text += file.read() + " "

# Tokenize the text
tokens = word_tokenize(all_text)

# Display the number of tokens
print("Number of tokens in the text:", len(tokens))
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Number of tokens in the text: 35071
```

What is your interest in analyzing this text?

Analyzing the subtitles from the "Friends" TV show is interesting for many reasons. "Friends" is a very popular and influential TV show with a wide range of viewers. By looking at the subtitles, I can understand the way characters talk, the jokes they make, how they interact with each other, and the cultural references they use. It also helps us learn about everyday language and communication. This makes it a great source for studying language, social interactions and media.

What questions might you answer given this dataset?

1. What are the most common themes and topics discussed in the show?

This question will help identify recurring subjects and motifs, such as relationships, work, and social life, which are central to the show's narrative.

2. What are the linguistic characteristics of each main character?

Analyzing the dialogue can reveal unique speech patterns, vocabulary that define each character's personality.

3. How is language used to convey emotions and humor?

Analyzing the dialogue can uncover techniques for expressing emotions, creating humor and engaging the audience.