Webcrawler Report

For the web crawler part of the project, I first started by getting a specific URL related to the actor Leonardo DiCaprio. I used the ‘https://www.biography.com/actors/leonardo-dicaprio’ as my starting URL since the website contained many different facts about the actor’s life. I then crawled through relevant URLs to get more information about the actor. The custom web crawler extracted text from each page and stored it in separate files. I then cleaned up the files using NLP techniques such as tokenizing the text, removing stop words and punctuation, and lowercasing the text. I then used the important terms from the files using tf-idf to build a knowledge base for the chatbot.

To create the knowledge base, I used Python libraries such as BeautifulSoup for web scraping, nltk for text processing, and sklearn to calculate tf-idf. The web_crawler() function takes the starting URL, maximum number of links, and the maximum depth as parameters. It then uses BeautifulSoup to parse the HTML content, get text from the website, and store it in files. This function also ensures that the links are unique and have different domains. The cleaned_files() function reads the raw text files and uses NLP techniques such as lowercasing, tokenization, and removal of stopwords and punctuation to clean the text files. The important_words() function uses the tf-idf to get important terms from the cleaned text files. I used sklearn to perform the tf-idf by using the TfidfVectorizer. I got the below top 40 terms from the text files through this function.

Important terms determined by function:
['dicaprio', 'people', 'film', 'movie', 'million', 'leonardo', 'one', 'going', 'time', 'diamonds', 'certainly', 'industry', 'actor', 'africa', 'want', 'part', 'let', 'still', 'like', 'expert', 'issue', 'playing', 'conflict', 'review', 'diamond', 'generation', 'hearing', 'issues', 'scorsese', 'fire', 'full', 'hollywood', 'think', 'get', 'name', 'oscar', 'took', 'would', 'years', 'better']

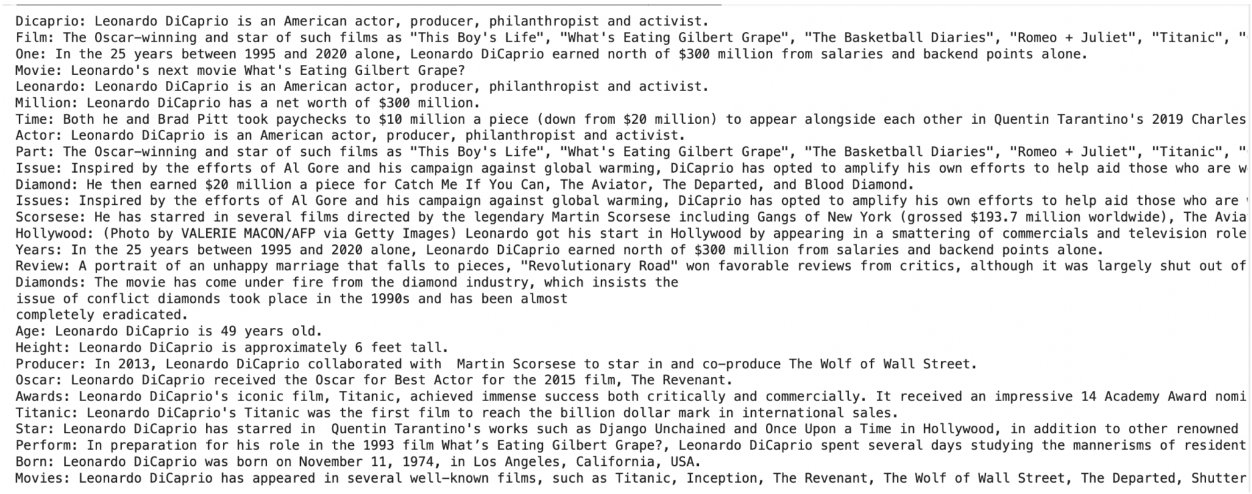
After I got the important terms, I started creating the actual knowledge base. I started by creating a get_facts() function to get relevant sentences containing the specific term from the text. The function uses sentence tokenization to break the text into sentences and then filters out sentences containing the specified term. I also created an update_knowledge_base() function to update the knowledge base with the facts based on the given URL, which is the initial starting URL. This function gets the HTML content, extracts textual content from paragraphs, and uses the get_facts() function to get the relevant sentences for each term. Only one fact per term is added to the knowledge base. I refined the selection of important terms from the initial 40 to the following terms for the construction of the knowledge base.

Narrowed down important terms:
['dicaprio','film','one','movie','leonardo','million','time','actor','part','issue','diamond','issues','scorsese','hollywood','years','review','diamonds']

I also manually added additional terms and facts to the knowledge base to make it more functional. Finally, I saved the knowledge base as a pickle file.

Manually added important terms:
['age','height','producer','oscar','awards','titanic','star','perform','born','movies']

Screenshot of the knowledge base:



```
Dicaprio: Leonardo DiCaprio is an American actor, producer, philanthropist and activist.
Film: The Oscar-winning and star of such films as "This Boy's Life", "What's Eating Gilbert Grape", "The Basketball Diaries", "Romeo + Juliet", "Titanic", "
One: In the 25 years between 1995 and 2020 alone, Leonardo DiCaprio earned north of $300 million from salaries and backend points alone.
Movie: Leonardo's next movie What's Eating Gilbert Grape?
Leonardo: Leonardo DiCaprio is an American actor, producer, philanthropist and activist.
Million: Leonardo DiCaprio has a net worth of $300 million.
Time: Both he and Brad Pitt took paychecks to $10 million a piece (down from $20 million) to appear alongside each other in Quentin Tarantino's 2019 Charles
Actor: Leonardo DiCaprio is an American actor, producer, philanthropist and activist.
Part: The Oscar-winning and star of such films as "This Boy's Life", "What's Eating Gilbert Grape", "The Basketball Diaries", "Romeo + Juliet", "Titanic", "
Issue: Inspired by the efforts of Al Gore and his campaign against global warming, DiCaprio has opted to amplify his own efforts to help aid those who are w
Diamond: He then earned $20 million a piece for Catch Me If You Can, The Aviator, The Departed, and Blood Diamond.
Issues: Inspired by the efforts of Al Gore and his campaign against global warming, DiCaprio has opted to amplify his own efforts to help aid those who are
Scorsese: He has starred in several films directed by the legendary Martin Scorsese including Gangs of New York (grossed $193.7 million worldwide), The Avia
Hollywood: (Photo by VALERIE MACON/AFP via Getty Images) Leonardo got his start in Hollywood by appearing in a smattering of commercials and television role
Years: In the 25 years between 1995 and 2020 alone, Leonardo DiCaprio earned north of $300 million from salaries and backend points alone.
Review: A portrait of an unhappy marriage that falls to pieces, "Revolutionary Road" won favorable reviews from critics, although it was largely shut out of
Diamonds: The movie has come under fire from the diamond industry, which insists the
issue of conflict diamonds took place in the 1990s and has been almost
completely eradicated.
Age: Leonardo DiCaprio is 49 years old.
Height: Leonardo DiCaprio is approximately 6 feet tall.
Producer: In 2013, Leonardo DiCaprio collaborated with  Martin Scorsese to star in and co-produce The Wolf of Wall Street.
Oscar: Leonardo DiCaprio received the Oscar for Best Actor for the 2015 film, The Revenant.
Awards: Leonardo DiCaprio's iconic film, Titanic, achieved immense success both critically and commercially. It received an impressive 14 Academy Award nomi
Titanic: Leonardo DiCaprio's Titanic was the first film to reach the billion dollar mark in international sales.
Star: Leonardo DiCaprio has starred in  Quentin Tarantino's works such as Django Unchained and Once Upon a Time in Hollywood, in addition to other renowned
Perform: In preparation for his role in the 1993 film What's Eating Gilbert Grape?, Leonardo DiCaprio spent several days studying the mannerisms of resident
Born: Leonardo DiCaprio was born on November 11, 1974, in Los Angeles, California, USA.
Movies: Leonardo DiCaprio has appeared in several well-known films, such as Titanic, Inception, The Revenant, The Wolf of Wall Street, The Departed, Shutter
```

The above screenshot shows the important terms and the facts associated with each term. This knowledge base will be used for the chatbot. For example, if the user asks the chatbot “Who is Leonardo?”, they will get a response from the chatbot saying “Leonardo DiCaprio is an American actor, producer, philanthropist, and activist.”