Alejo Vinluan

Abv210001

CS 4395

Web Crawler Assignment

This assignment breaks down how to scrape sentences from various websites stemming from one root website. After scraping sentences, words are tokenized, cleaned, and filtered until a Corpus is created. Finally, a knowledge base is created from the Corpus to determine important terms associated with the root website. In this project, I decided to set the root website as the Wikipedia page for the National Basketball Association.

The program found 15 websites that were listed on the Wikipedia page, scraped all 15 websites, cleaned the raw data from the website, and utilized each website to build an internal Corpus. The websites are listed as:

```
Websites:
https://en.wikipedia.org/wiki/1978_NBA_Finals
https://en.wikipedia.org/wiki/Fort_Wayne_Pistons
https://en.wikipedia.org/wiki/1998_NBA_draft
https://en.wikipedia.org/wiki/NBA_Finals_Most_Valuable_Player_Award

https://en.wikipedia.org/wiki/Fiserv_Forum
https://en.wikipedia.org/wiki/2010_NBA_All-Star_Game
https://en.wikipedia.org/wiki/1969_NBA_draft
https://en.wikipedia.org/wiki/Major_League_Baseball
https://en.wikipedia.org/wiki/Turkish_Basketball_Second_League
https://en.wikipedia.org/wiki/Women%27s_National_Basketball_Associa
tion
https://www.nba.com/history/top-moments/golden-state-warriors-win-7
3-games
https://www.nba.com/news/history-all-defensive-team
https://chicago.suntimes.com/sports/25-facts-to-celebrate-the-dream
-team-25-years-later/
https://sportslawexpert.com/2022/05/08/team-market-report-publishes
-nba-fan-cost-index-slight-increase-reported/
https://sports.yahoo.com/game-1-nba-finals-lakers-heat-ratings-low-
abc-225636681.html
```

Sentences from the website were filtered by first tokenizing every sentence within the website's text. After tokenization, punctuation was removed, all words were lower cased, and stop words from the English dictionary were removed. Finally, the Top 40 Most Frequent words are printed, alongside their rank and count. Here is an example of the first 5 words given:

```
Rank: 1
Term: retrieved
Occurrences: 846

Rank: 2
Term: team
Occurrences: 614

Rank: 3
Term: league
Occurrences: 565

Rank: 4
Term: nba
Occurrences: 551

Rank: 5
Term: pistons
Occurrences: 464
```

Finally, I handpicked 10 words from the Top 40 words and inputted their definitions as a Python dictionary, with the key being the term and the value being the term's definition. The dictionary is then dumped as a pickle file within the same directory as the program.

```
# Create a knowledge base from the top 10 terms by hand
knowledge_base = {
    "team": "a group of players forming one side in a competitive game or sport",
    "league": "a collection of people, countries, or groups that combine for a particular purpose, typically mutual protection or cooperation",
    "nba": "the National Basketball Assosciation",
    "season": "a fixed time in the year when a particular sport is played",
    "wnba": "the Womens National Basketball Assosciation",
    "players": "a person taking part in a sport or game",
    "basketball": "a game played between two teams of five players in which goals are scored by throwing a ball through a netted hoop fixed above each end of the court",
    "sport": "an activity involving physical exertion and skill in which an individual or team competes against another or others for entertainment",
    "game": "a form of play or sport, especially a competitive one played according to rules and decided by skill, strength, or luck",
    "may": "The month where the NBA Playoffs and NBA Finals generally occur"
}
```

```
# Create a pickle of the dictionary for the knowledge base
pickle.dump(knowledge_base, open('knowledge_base.p', 'wb'))
```

I would like to have the chatbot eventually answer specific questions about NBA games or specific NBA trivia that is scraped directly from the NBA Wikipedia page and other associated websites. Some examples of chat interactions should include:

"How many points did Lebron James score on February 26, 2023?"
"Lebron James scored 26 points."

"What year was the NBA founded?"
"The NBA was found in 1946."