CS 6320.001 Natural Language Processing

Project 1 – Chatbot Report (Part 1)

Part 1 - Web Crawler

Summary:

The web-crawler part of the project starts with a single link provided by the user and gathers several other links related to that topic and scrapes information from those websites. The scraped data is then stored into several text files which are used to create the knowledge base.

Program Flow:

- Start URL: The program begins with a start URL of Keanu Reeves' Wikipedia page.
    - It parses through the website and collects all URL's present in it.
    - It does not include certain URL's containing phrases which are hardcoded such as '%', 'wiki' (to avoid same Wikipedia page in different languages, and 'donate'.
- URL Collection: It collects related URLs from the start URL and subsequent URLs.
    - All the URL's are added to the local list variable.
- URL Filtering: Filters URLs to include only those related to Keanu Reeves.
    - All the unnecessary URL's are filtered by length if length is lesser than 20.
    - Additionally, they should contain either 'keanu' or 'reeves' in it.
- Data Scraping: Scrapes data from the final list of URLs.
    - Using requests and BeautifulSoup libraries, data which is visible is collected from the websites.
    - All this data are stored in the text files.
    - The first corpus contains the table data of movies and tv shows from the Wikipedia page.
    - The second corpus contains the Wikipedia page data of the subject.
    - Rest of the corpuses are non- Wikipedia page data- articles, newsletters and such.
- Data Cleaning: Cleans the scraped data.
    - The clean_text() function does a basic clean on the text files to remove unnecessary spaces, non-alphanumeric characters, references ('[29]'), and tag names.

- Data Filtering: Further filters the cleaned data to remove irrelevant information.
    - The filter_text() function does the second level of filtering to the text files based on the user's ban list. To eliminate lines like 'contact us for more', 'read more' etc.

CS 6320.001 Natural Language Processing

Project 1 – Chatbot Report (Part 1)

- Data Storage: For every link parsed, the data scraped is stored in a text file.
    - The filenames are in incremental fashion, so counters are used to traverse through.
- **Knowledge base creation:**
    - The first corpus file – corpus0.txt contains the movies and tv shows tabled data. This file is used to create pickle dictionaries in the knowledge_base.py. The key is the year number, and the values is the list of movies or tv shows from that year.
    - From the second corpus files, all the text is stored into a dictionary with the
    - doc numbers as key and list of sentence tokenized strings as the values.
      Ex: { 'doc1' : ['sentence 1', 'sentence 2' , 'sentence 3' ……. 'Sentence n'],
      'doc2' : ['sentence 1', 'sentence 2' , 'sentence 3' ……. 'Sentence n'],
      .
      .
      .
      'doc20' : ['sentence 1', 'sentence 2' , 'sentence 3' ……. 'Sentence n']}
    - The tf_idf function calculates top important words from all corpus files and returns 50 most common words.

Function Descriptions:

- **calc_similarity():**

Calculates the similarity between two strings using the difflib function. Used to reduce redundancy in storing strings by comparing the similarity of URLs.

- **get_urls(starter_url):**

Takes a URL as input and returns a list of related URLs. Ignores irrelevant URLs based on user specified criteria (e.g., not containing 'wiki', 'donate', etc.).

- **visible(element):**

Determines if an HTML element is visible on a webpage. Used to filter out invisible elements during data scraping.

- **datascraper(my_url, counter):**

Scrapes data from a given URL and saves it to a text file. Special cases are hardcoded – filmography table for corpus0 and Wikipedia page for corpus1. All the other corpuses are scraped with common code.

- **clean_text(text_string):**

Cleans the text in each file by removing references, non-alphanumeric characters, and extra whitespaces.

- **should_append_line(i, flag):**

Determines whether a line of text should be appended to the final text based on certain criteria (e.g., containing specific strings, not matching regex patterns).

- **filter_data(i):**

Filters the data in a text file to remove irrelevant information. Uses should_append_line to decide which lines to keep.

- **main():**

The main function that orchestrates the entire flow of the program. Collects URLs, scrapes data, cleans and filters the data, and saves the final URLs to a file.

NLP Tools Used:

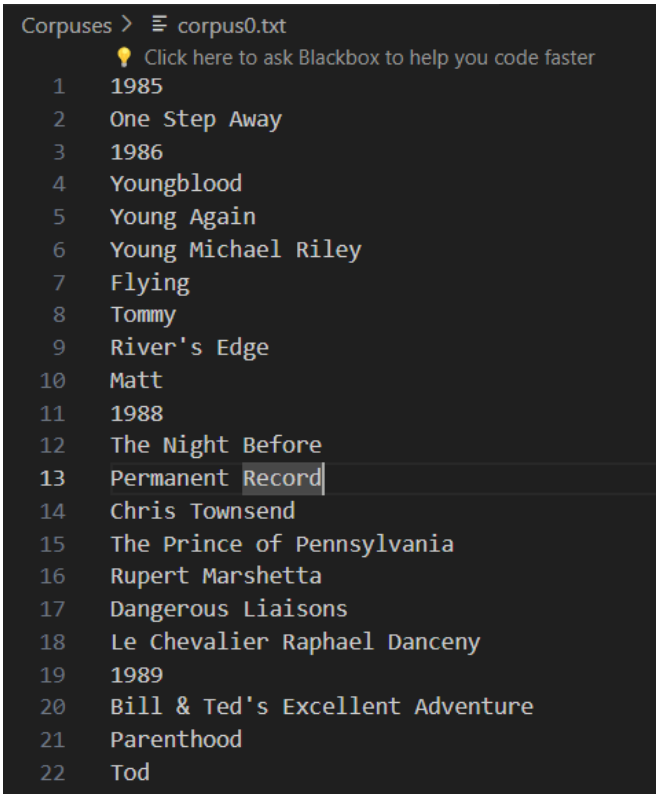
- **Regular Expressions (regex):** Used extensively for text cleaning and filtering, such as removing references, non-alphanumeric characters, and irrelevant lines of text.
- **Cosine Similarity:** Used to measure the similarity between two URLs to reduce redundancy in URL collection.
- **Information Extraction:** Utilized in functions like datascraper and filter_data to extract relevant information from the scraped web pages, such as details from Keanu Reeves' filmography and Wikipedia page.
- **Beautiful Soup:** A Python library for pulling data out of HTML and XML files. It is used for parsing the HTML content of web pages to facilitate information extraction and data cleaning.
- **URLlibrary:** Library used to parse through websites and access data.
- **Pickle:** Used to store and retrieve pre-processed data for efficient access during runtime.

CS 6320.001 Natural Language Processing

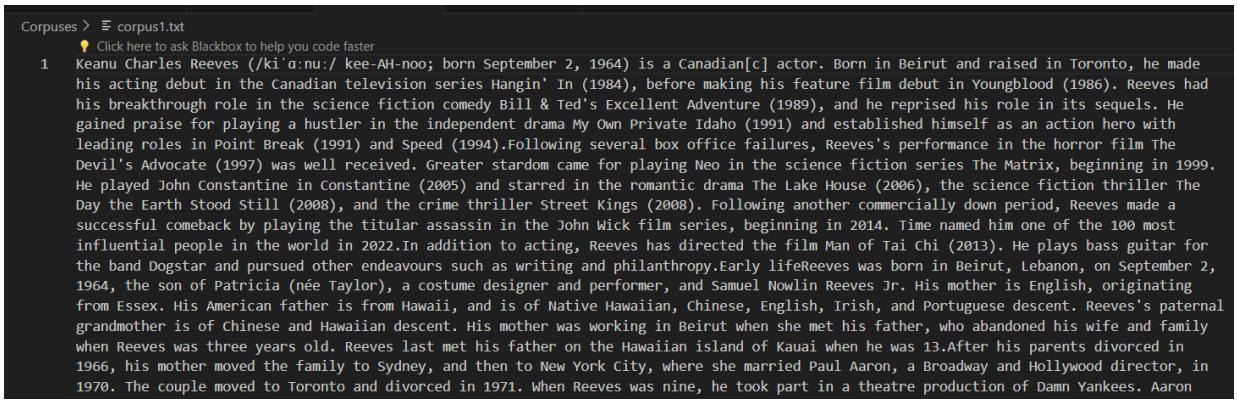
Project 1 – Chatbot Report (Part 1)

Screenshots of Knowledge base:

- corpus1.txt:



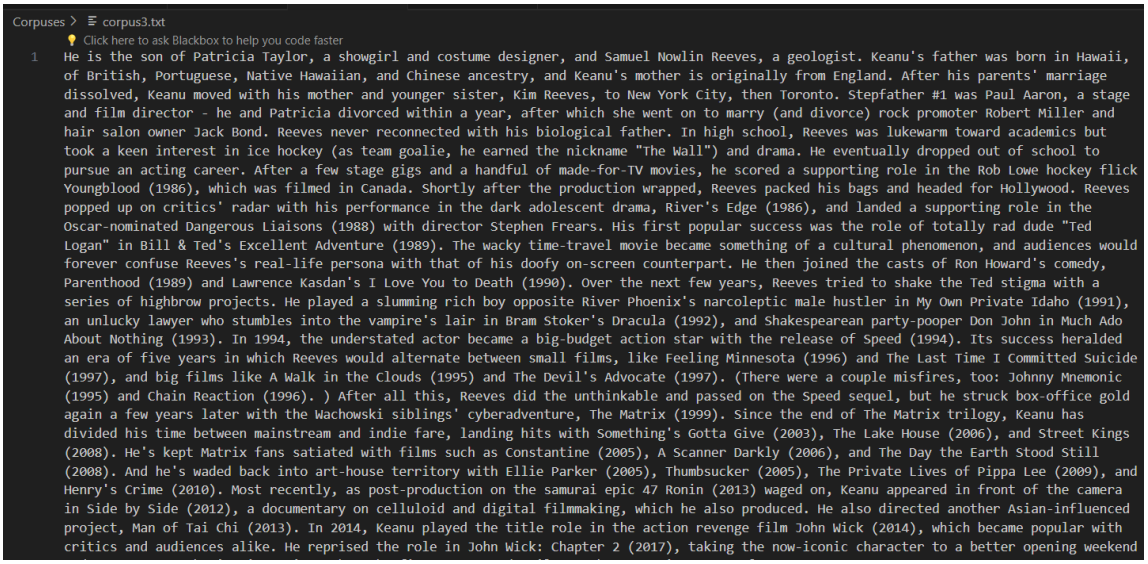
- corpus2.txt



CS 6320.001 Natural Language Processing

Project 1 – Chatbot Report (Part 1)

- corpus3.txt



Corpuses > corpus3.txt

Click here to ask Blackbox to help you code faster

```
He is the son of Patricia Taylor, a showgirl and costume designer, and Samuel Nowlin Reeves, a geologist. Keanu's father was born in Hawaii,
of British, Portuguese, Native Hawaiian, and Chinese ancestry, and Keanu's mother is originally from England. After his parents' marriage
dissolved, Keanu moved with his mother and younger sister, Kim Reeves, to New York City, then Toronto. Stepfather #1 was Paul Aaron, a stage
and film director - he and Patricia divorced within a year, after which she went on to marry (and divorce) rock promoter Robert Miller and
hair salon owner Jack Bond. Reeves never reconnected with his biological father. In high school, Reeves was lukewarm toward academics but
took a keen interest in ice hockey (as team goalie, he earned the nickname "The Wall") and drama. He eventually dropped out of school to
pursue an acting career. After a few stage gigs and a handful of made-for-TV movies, he scored a supporting role in the Rob Lowe hockey flick
Youngblood (1986), which was filmed in Canada. Shortly after the production wrapped, Reeves packed his bags and headed for Hollywood. Reeves
popped up on critics' radar with his performance in the dark adolescent drama, River's Edge (1986), and landed a supporting role in the
Oscar-nominated Dangerous Liaisons (1988) with director Stephen Frears. His first popular success was the role of totally rad dude "Ted
Logan" in Bill & Ted's Excellent Adventure (1989). The wacky time-travel movie became something of a cultural phenomenon, and audiences would
forever confuse Reeves's real-life persona with that of his doofy on-screen counterpart. He then joined the casts of Ron Howard's comedy,
Parenthood (1989) and Lawrence Kasdan's I Love You to Death (1990). Over the next few years, Reeves tried to shake the Ted stigma with a
series of highbrow projects. He played a slumming rich boy opposite River Phoenix's narcoleptic male hustler in My Own Private Idaho (1991),
an unlucky lawyer who stumbles into the vampire's lair in Bram Stoker's Dracula (1992), and Shakespearean party-pooper Don John in Much Ado
About Nothing (1993). In 1994, the understated actor became a big-budget action star with the release of Speed (1994). Its success heralded
an era of five years in which Reeves would alternate between small films, like Feeling Minnesota (1996) and The Last Time I Committed Suicide
(1997), and big films like A Walk in the Clouds (1995) and The Devil's Advocate (1997). (There were a couple misfires, too: Johnny Mnemonic
(1995) and Chain Reaction (1996). ) After all this, Reeves did the unthinkable and passed on the Speed sequel, but he struck box-office gold
again a few years later with the Wachowski siblings' cyberadventure, The Matrix (1999). Since the end of The Matrix trilogy, Keanu has
divided his time between mainstream and indie fare, landing hits with Something's Gotta Give (2003), The Lake House (2006), and Street Kings
(2008). He's kept Matrix fans satiated with films such as Constantine (2005), A Scanner Darkly (2006), and The Day the Earth Stood Still
(2008). And he's waded back into art-house territory with Ellie Parker (2005), Thumbsucker (2005), The Private Lives of Pippa Lee (2009), and
Henry's Crime (2010). Most recently, as post-production on the samurai epic 47 Ronin (2013) waged on, Keanu appeared in front of the camera
in Side by Side (2012), a documentary on celluloid and digital filmmaking, which he also produced. He also directed another Asian-influenced
project, Man of Tai Chi (2013). In 2014, Keanu played the title role in the action revenge film John Wick (2014), which became popular with
critics and audiences alike. He reprised the role in John Wick: Chapter 2 (2017), taking the now-iconic character to a better opening weekend
```

- movies.pickle

```
Movies:
1985 : ['One Step Away']
1986 : ['Youngblood', 'Young Again', 'Flying', "River's Edge"]
1988 : ['The Night Before', 'Permanent Record', 'The Prince of Pennsylvania', 'Dangerous Liaisons']
1989 : ["Bill & Ted's Excellent Adventure", 'Parenthood']
1990 : ['I Love You to Death', 'Tune in Tomorrow']
1991 : ['Point Break', "Bill & Ted's Bogus Journey", 'My Own Private Idaho']
1992 : ["Bram Stoker's Dracula"]
1993 : ['Much Ado About Nothing', 'Even Cowgirls Get the Blues', 'Freaked', 'Little Buddha']
1994 : ['Speed']
1995 : ['Johnny Mnemonic', 'A Walk in the Clouds']
1996 : ['Chain Reaction', 'Feeling Minnesota']
1997 : ['The Last Time I Committed Suicide', "The Devil's Advocate"]
1999 : ['The Matrix', 'Me and Will']
2000 : ['The Replacements', 'The Watcher', 'The Gift']
```

- tvshows.pickle

```
TV Shows:
1984 : ["Hangin' In"]
1985 : ['Letting Go', 'Night Heat', 'Fast Food']
1986 : ['The Disney Sunday Movie', 'Babes in Toyland', 'Act of Vengeance', 'Brotherhood of Justice', 'Under the Influence']
1987 : ['Trying Times']
1989 : ['Life Under Water', 'The Tracey Ullman Show']
1990 : ["Bill & Ted's Excellent Adventures"]
2009 : ['Bollywood Hero']
2016-2018 : ['Swedish Dicks']
2020 : ['A World of Calm']
2023 : ['Ride with Norman Reedus', 'Brawn: The Impossible Formula 1 Story', '']
```

CS 6320.001 Natural Language Processing

Project 1 – Chatbot Report (Part 1)

- Important words list:

```
TF-IDF list of words:
 (['reeves', 'keanu', 'film', 'grant', 'first', 'said', 'name', 'like', 'years', 'people', 'would', 'john', 'book', 'one', 'two', 'matrix', 'life', 'ted', 'wi
ck', 'interview', 'movie', 'also', 'time', 'played', 'together', 'star', 'sad', 'role', 'think', 'bill', 'much', 'actor', 'couple', 'toronto', 'films', 'way',
 'school', 'happy', 'speed', 'run'], {'parker', 'test', 'together', 'catches', 'shallow', 'others', 'ago', 'u', 'constant', 'year', 'remember', 'know', 'insid
e', 'played', 'fiercely', 'committed', 'volodymyr', 'trinity', 'unkind', 'blige', 'us', 'handsome', 'stressful', 'reprieve', 'interactions', 'eyes', 'successf
ul', 'day', 'well', 'supported', 'bigger', 'come', 'character', 'kind', 'man', 'something', 'home', 'screen', 'always', 'zelensky', 'characters', 'edit', 'liv
es', 'could', 'need', 'training', 'still', 'count', 'keanu', 'rather', 'become', 'years', 'friendship', 'protective', 'neo', 'listens', 'actor', 'checks', 'in
spired', 'shows', 'intense', 'often', 'hope', 'friend', 'resurrections', 'last', 'essence', 'glimpses', 'kindly', 'felt', 'plays', 'gives', 'met', 'humanity',
 'pedestal', 'ease', 'put', 'world', 'thoughtful', 'time', 'first', 'everyday', 'talented', 'disappointment', 'matrix', 'every', 'way', 'already', 'get', 'ill
uminates', 'post', 'generous', 'actions', 'candace'})
```

- List of files stored: