Homework 5: Web Scraping

**Important Terms**

The way that I created my important terms was using TF-IDF. I calculated the TF-IDF for every word in every document. Then I sorted the TF-IDF weights for each document. Then I took the 5th root of the length of the set of tokens and round it to determine how many words to retrieve for that document. After, I simply add the N top terms of that document to the whole important terms list. I do this for every document. Which ultimately gets me my important terms list.

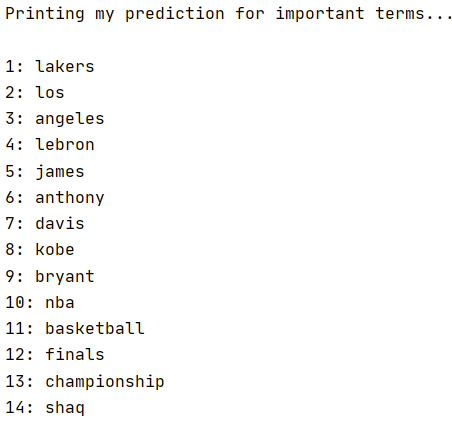
My prediction of the output of top terms include: lakers, los, angeles, lebron, james, anthony, davis, kobe, bryant, nba, basketball, finals, championship, shaq.

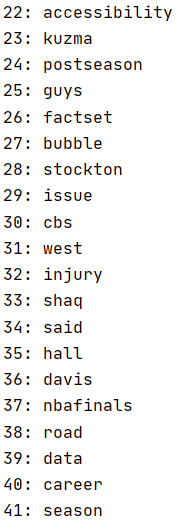
The following is the output from my program. The program is concerned with URLs about the Lakers and has a limit of anything greater than or equal to 25 URLs. Finally, keep in mind that not all URLs will be processed to text because not URLs will contain relevant info.

Output:



Note: The initial link is not included as it was a link to a google search for ‘lakers’ and it does not have relevant info we can extract besides the links.





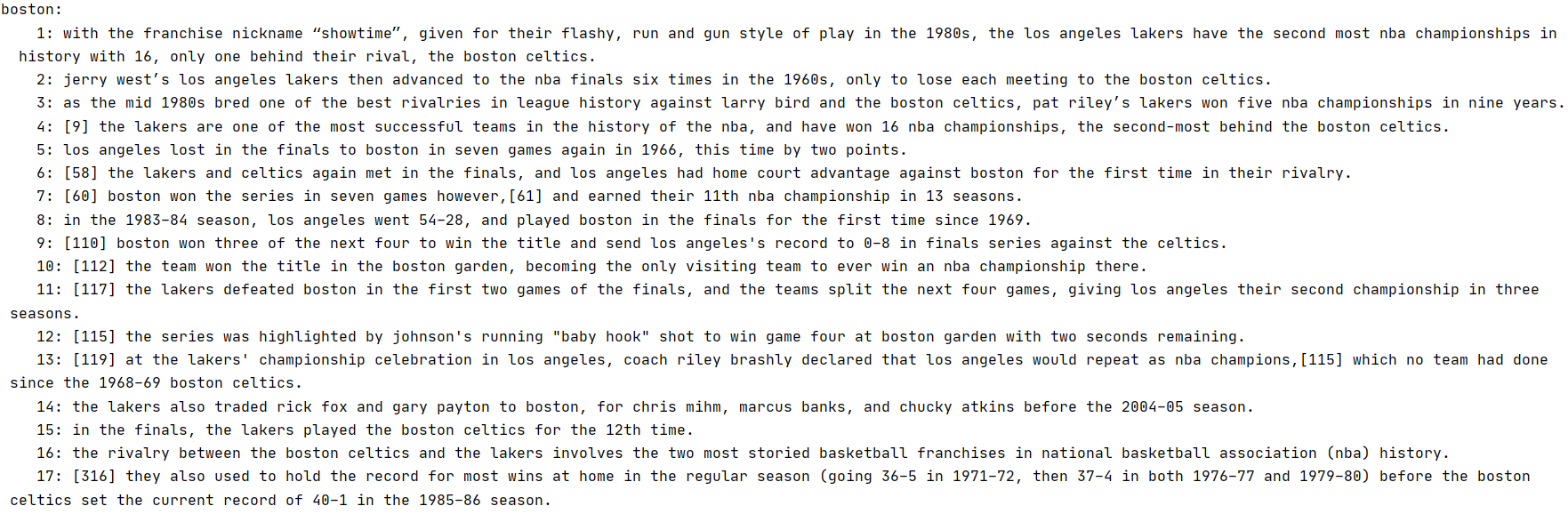
Note: A lot of these words makes sense. For example, ad is often used to refer to Anthony Davis (current player of the Lakers), the Boston Celtics are the rivals of the Lakers, and many other words are names of former/current players of the Lakers or another team in the nba.

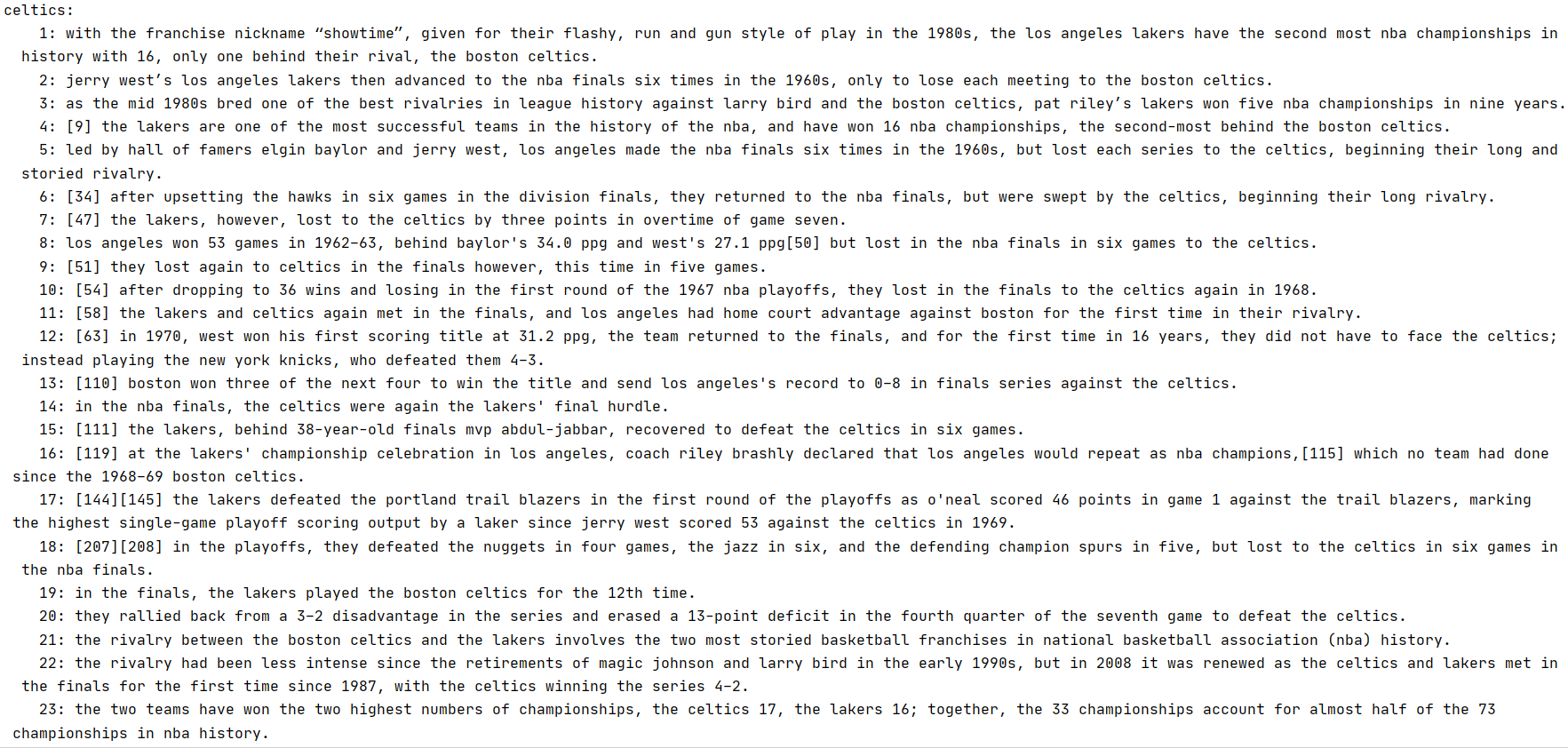
**Knowledge Base**

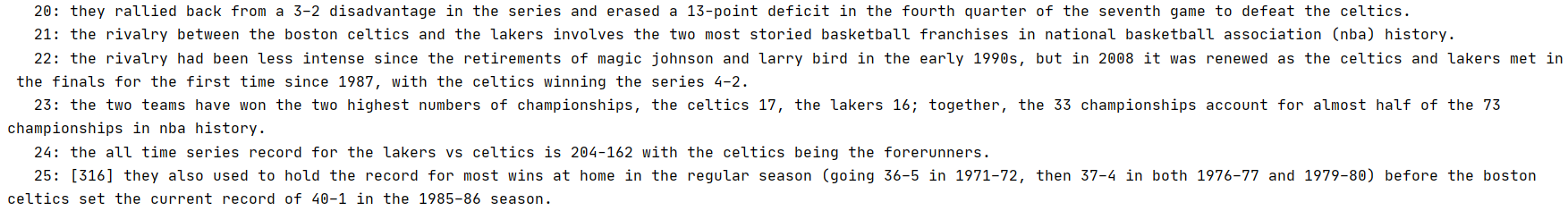
The way I created my knowledge base was to take all the important words from the collection of documents and create a dictionary. The key would be the word and the value will be a list. This list will have every sentence containing that specific word from all documents. For example, if we have 20 documents and each document has two sentences containing “nba” and if nba is an important word then it will have a list of 40 elements each element corresponding to a sentence that contains nba.

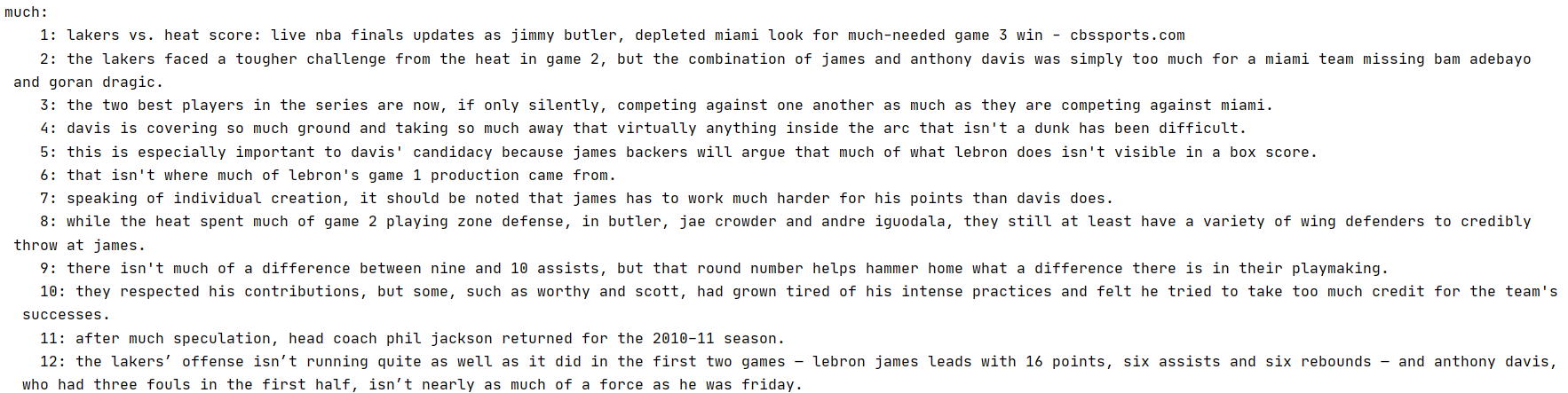
This knowledge base is then stored into a database file. Each important term will have its own table and the entries of the table will be the sentences containing the term from all documents. For example, if we have 41 important terms, then the database file will have 41 tables. So, if the word “nba” is an important term with 40 sentences, then there will be a table called “nba” with 40 entries each corresponding to one sentence.

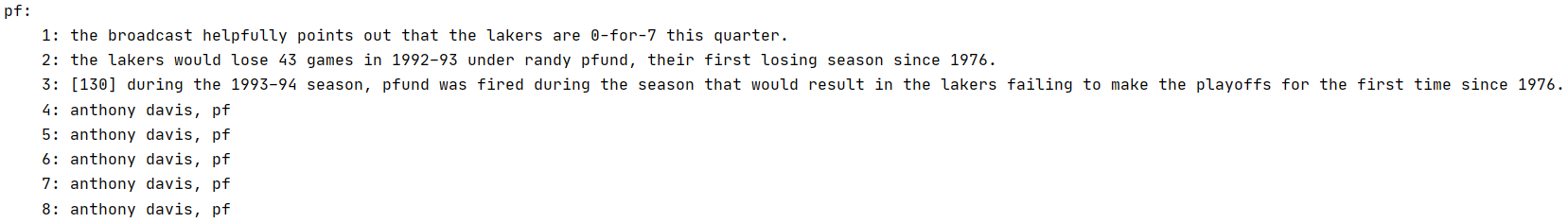
The following is the output of my knowledge base from the important words shown above.

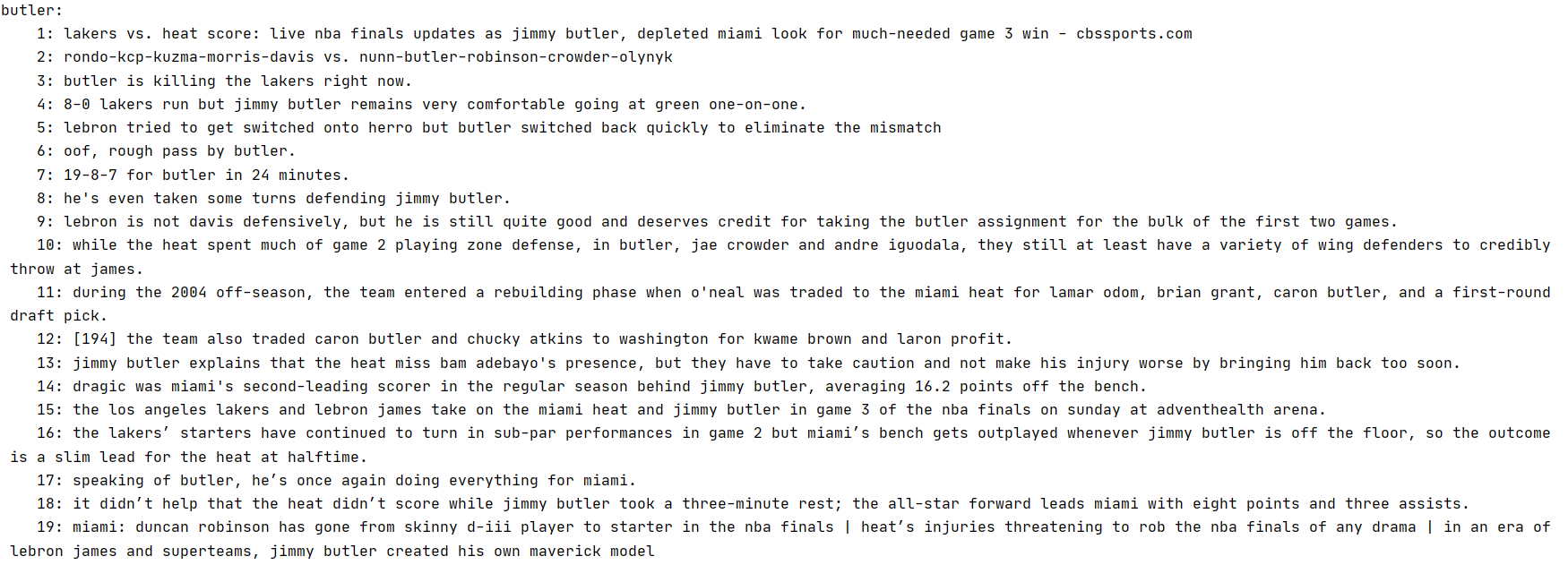


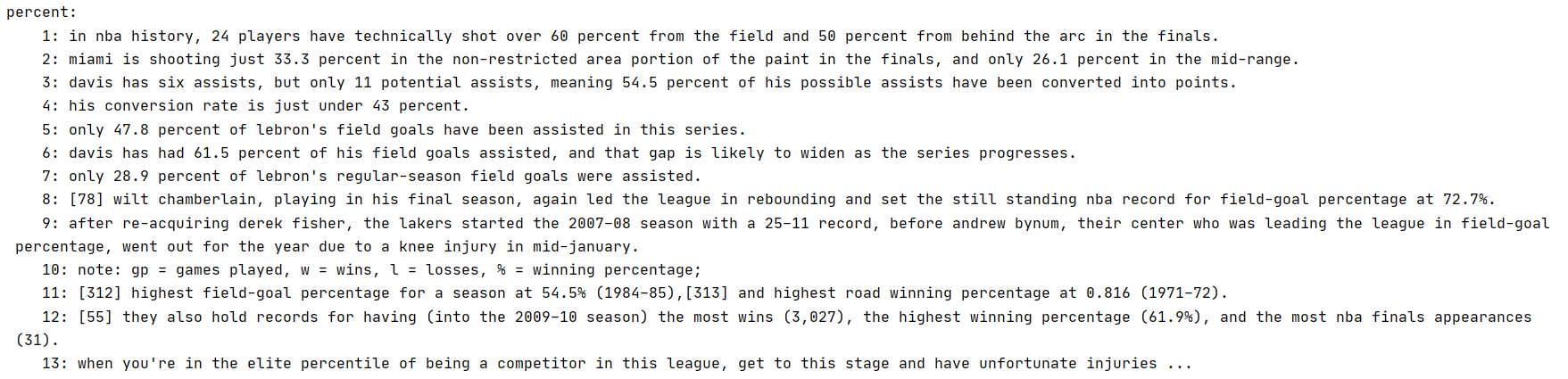












Note: This is not the complete knowledge base. The complete knowledge base with a limit of around 25 links is too big to fit all in this document. These are just some tables in the whole knowledge base database. If you would like to see the whole knowledge, go to the print\_results functions and uncomment the if statement in the last for loop. This will print the entire KB.

**Sample Dialog 1**

Hello, I am a chat-bot that loves to talk about the Los Angeles Lakers. Tell me what is your name?

My name is Alfredo.

Hello Alfredo. I do sure hope you are a fan of the Lakers. Are you a fan?

Yes, most definitely, born and raised in Los Angeles.

Awesome! That means we can talk about the Lakers all day long. Do you know the Celtics?

Of course, the rivals of the Lakers.

Did you know: los angeles won 53 games in 1962-63, behind baylor's 34.0 ppg and west's 27.1 ppg but lost in the nba finals in six games to the Celtics?

No, I did not know that.

I thought you said you were a fan of the Los Angeles Lakers. Are you a bandwagon?

No! I was born and raised in Los Angeles.

Okay, let me test you one more time. Did you know: with the franchise nickname "showtime", given for their flashy, run and gun style of play in the 1980s, the los angeles lakers have the second most nba championships in history with 16, only one behind their rival, the boston celtics?

Yes, I did know that. I hope they win another chip this year so it can be tied.

Ahh. So maybe you aren’t a bandwagon …

**Sample Dialog 2**

Hello, I am a chat-bot that loves to talk about the Los Angeles Lakers. Tell me what is your name?

My name is Alfredo.

Hello Alfredo. I do sure hope you are a fan of the Lakers. Are you a fan?

No, the Los Angeles Clippers convinced me to switch.

I respect that, but let me tell you something about the greatest basketball team, with the franchise nickname "showtime", given for their flashy, run and gun style of play in the 1980s, the los angeles lakers have the second most nba championships in history with 16, only one behind their rival, the boston celtics. Can you believe that?

No

Well believe it …

**Quick Note**

With the initial URL being a google search for the lakers, this means each run could potentially produce different results. Especially as the lakers are currently in the NBA Finals as of the making of this document. This means the google search is often being updated with new articles and information regarding the basketball team.

If the program is run multiple times and the database files created, are not deleted, then the URL database list will most likely increase with each run regardless on the limit. This because each run will have its set of URLs (abiding to the limit) and it will be added to the database file. However, the database file could potentially have the URLs from the previous run if the file was not deleted. It will get rid of any duplicates and add the new URLs to the database. This was done intentionally to have a large database of URLs from not just one run but multiple runs. If you would like to have URLs for that specific run, then after each run you will have to delete the database file containing the URLs.

The same thing applies to the knowledge base database file. Each run will produce different important terms and ultimately different sentences will be added to the current run knowledge base. This knowledge base will then be stored to the database file that could potentially have the previous knowledge base in it if it was not deleted. If there are repeated important terms then no table is created, duplicates sentences aren’t added to the table, and only new sentences are added to the table. This is, again, done intentionally to build a bigger knowledge base after each run. If you would like a knowledge base only for that run, then after each run delete the knowledge base database file.

However, the exception is the important terms database file. After each run, if there is an important terms database file it will delete it and create a new file for the current run. This is because the important terms are local to the current run containing a set of documents. The next run might contain different documents and thus potentially change the list of important terms. Thus, we do not want to combine them because they are a reflection of their own set of documents and each run could potentially have different documents. In this case, the important terms database file does not need to be deleted the program takes care of it.

If you have any questions, concerns, or comments, please feel free to contact me at my email: [axm180023@utdallas.edu](mailto:axm180023@utdallas.edu)

I would gladly explain any parts of my program. Thank you.