**Portfolio Web Crawler Report**

1. Describe how you created your knowledge base
   * **My starter site was** [**https://golf.com**](https://golf.com)**. After the web scraping, I got 114 URLs and I was able to extract the text from each of those URLs and store each text in a text file. This text was very messy with lot of newlines, whitespaces, and tabs, which I tried to clean up. After this clean up, I used NLTK library to convert text from the scraped text files into list of sentences. I then performed preprocessing on the list of sentences to get list of tokens. I measured the term frequency of each token and put it in a dictionary consisting of the term and its total number of occurrences in the list of sentences. Using this dictionary, I was able to display the top 40 terms.**
   * **From the top 40 terms, I determine 10 terms that I thought were most relevant to the golfing domain.**
   * **The top 10 terms I chose were: "golf", "gear", "instruction", "bags", "woods", "clubs", "swing", "game", "tour", "drivers".**
   * **From the top 10 terms, I built a knowledge base Python dictionary where the term is the key and the list of sentences containing the term is the value.**
   * **I was not able to get a screenshot of the knowledge base from the output console since the console did not let me scroll to capture all of the knowledge base. You can find the knowledge base of each of the top 10 terms in Knowledge Base folder under the project folder.**
2. Write up a sample dialog you would like to create with a chatbot based on your knowledge base

**Bot: How can I help you today?**

**User: Name sports other than golf?**

**Bot: Skiing and baseball**

**User: Name a golf course?**

**Bot: SentryWorld**

**User: What is one good advice to give to talented juniors?**

**Bot: With talented juniors, it’s important to keep them focused on playing golf, not the golf swing.**

**User: What are some popular news topics about Tiger Woods?**

**Bot: ‘Tiger Woods apologizes for tampon joke he made mid-round with Justin Thomas.’, ‘Will Tiger Woods play the weekend?’, ‘How does Tiger Woods \*still\* generate so much ball speed?’**