HLT
Homework 5: Web Scraping
Worth 200 points

Objective: Create a knowledge base scraped from the web. This knowledge base will be used in a later homework to create a chatbot that can carry on a limited conversation in a particular domain using the knowledge base, as well as knowledge it learns from the user.

- You may work alone if you prefer, or you can partner with one other person.
- Upload your code and all output files, zipped together.

Instructions

1. Build a web crawler function that starts with a URL representing a topic (a sport, your favorite film, a celebrity, a political issue, etc.) and outputs a list of at least 15 *relevant* URLs. The URLs can be pages within the original domain but should have a few outside the original domain.
2. Write a function to loop through your URLs and scrape all text off each page. Store each page's text in its own file.
3. Write a function to clean up the text. You might need to delete newlines and tabs. Extract sentences with NLTK's sentence tokenizer. Write the sentences for each file to a new file. That is, if you have 15 files in, you have 15 files out.
4. Write a function to extract at least 25 important terms from the pages using an importance measure such as term frequency, or tf-idf. First, it's a good idea to lower-case everything, remove stopwords and punctuation. Print the top 25-40 terms.
5. Manually determine the top 10 terms from step 4, based on your domain knowledge.
6. Build a searchable knowledge base of facts that a chatbot (to be developed later) can share related to the 10 terms. The "knowledge base" can be as simple as a Python dict which you can pickle. More points for something more sophisticated like sql.
7. In a doc: (1) describe how you created your knowledge base, include screen shots of the knowledge base, and indicate your top 10 terms; (2) write up a sample dialog you would like to create with a chatbot based on your knowledge base

Grading Rubric:

Element	Points
Step 1	50
Step 2	20
Step 3	20
Step 4	25
Step 5	10
Step 6	50
Step 7	25
Total	200