Portfolio: Building a Corpus

Objectives:

- Understand the importance of corpora in NLP tasks
- Understand basic HTML
- Understand how web sites work
- Be able to do web scraping with Beautiful Soup or other APIs
- Create a web crawler

Turn in:

- Upload your .py code and report to eLearning for grading
- Upload your .py code and report to your portfolio, and create a link to on your index page
- This program should be written with an IDE

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 from each file. You might need to delete newlines and tabs first. 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
8. Create a link to the report and code on your index page

Be prepared to present your results to class:

- what was your starter site
- what kind of data did you get
- how did you clean up the data
- what were your top terms
- show us your knowledge base
- how might you use this data for a chatbot

Grading Rubric:

Element	Points
Web crawler code works	100
Text processing to get key terms and build knowledge base	50
Report	50
Total	200