Exercise Part 2

Q2.1 Write a program to do the following tasks:

1. Download the Web page of a given link and extract the text content of the page
2. Split the text into sentences and count the number of sentences
3. Split the text into tokens, and count the number tokens and number of unique tokens (i.e., token types)
4. Find lemmas (or stems) of the tokens and count lemma types
5. Do stemming on the tokens and count unique 'stemmed' tokens

You may use any tools, including nltk, LingPipe, and Stanford NLP software.

Q2.3 Find a document collection of reasonable size (e.g., 1000 documents), build an inverted index and using the inverted index to answer a few sample queries.

You may use any software packages for indexing and searching text.

Q2.2. Write a program (you may use third-party APIs), using Viterbi algorithm with the given HMM model, to compute the most likely weather sequences for each of the two following observation sequences. Sequence (A): 312312312 Sequence (B): 311233112