### Assignment on Chapter 4a in “*Natural Language Processing with Python --- Analyzing Text with the Natural Language Toolkit” by* Steven Bird, Ewan Klein, and Edward Loper

You will find Chapter 4a posted at:

<http://www.nltk.org/book/ch04.html>

Section 4.11 lists exercises. You are asked to complete some of the exercises. On the Canvas, a file named

nltkChapter4a\_template.py

is posted. You are requested to use this template in formulating your code solutions.

Let *SimpleText* be the following text

*One day, his horse ran away quietly. The neighbors came to express their concern: "Oh, that's too bad. How are you going to work the fields now?" The farmer replied: "Good thing, Bad thing, Who knows?"*

*In a few days, his horse came back and brought another horse with her. Now, the neighbors were glad: "Oh, how lucky! Now you can do twice as much work as before!" The farmer replied: "Good thing, Bad thing, Who knows?"*

Let *TestText* be the text posted at URL[*https://www.cs.utexas.edu/~vl/notes/dijkstra.html*](https://www.cs.utexas.edu/~vl/notes/dijkstra.html)

To get text out of HTML, use a Python library called BeautifulSoup, available from <http://www.crummy.com/software/BeautifulSoup/> as prescribed in Dealing with HTML subsection in Chapter 3

Both of these variables are populated in the assignment template.

Complete the following of the exercises:

1. exercise 14. Use nltk.word\_tokenize() to tokenize the text. Assume each token returned by tokenizer to represent the word. Recall, that, for instance, “The” and “the” are the same words. In this question, use *TestText* for testing. Sort the resulting words alphabetically. Report first ten elements in your results.
2. exercise 17. Test your function on *TestText* with n=20, n=35, n=50, n=65.

a) How readable is the output for every n? (you don’t need to put entire outcome in your report. Instead, you can put few lines from each output.)

b) Use the function from part a on the following snippet text with n = 20. **Report the outcome** and describe your findings (how many words are taken off? What are those words?):

*to begin with I would like to thank the College of Natural Sciences for the most honouring Invitation to address its newest flock of Bachelors on this most festive day. I shall do my best.*

1. exercise 30. Print the output of your function for *SimpleText*. Compute compression using the following formula: length of the text resulting from the application of your function (considered as a string) divided by the length of the original text (considered as a string).

To clarify the notion of uniqueness point, consider a text of five words “all apes ate apples”. By [] we denote the uniqueness point of each word in this text: “ a[l]l ap[e]s a[t]e ap[p]les”. Thus, the output of your function for this text should be “al ape at app”.

A trie for this text follows:

a l l

p e s

p l e s

t e

Also, https://en.wikipedia.org/wiki/Cohort\_model

has a paragraph, which provides more verbal description of where a "uniqueness point" stems from. There a term "recognition point" is used for this concept at first. Below is the extract from that paragraph:

The cohort model consists of three stages: access, selection, and integration.[[7]](https://en.wikipedia.org/wiki/Cohort_model#cite_note-Gaskell_1997_613.E2.80.93656-7) Under this model, auditory lexical retrieval begins with the first one or two speech segments, or [phonemes](https://en.wikipedia.org/wiki/Phoneme), reach the hearer's ear, at which time the mental lexicon activates every possible word that begins with that speech segment.[[8]](https://en.wikipedia.org/wiki/Cohort_model#cite_note-packard288-8) This occurs during the "access stage" and all of the possible words are known as the cohort.[[9]](https://en.wikipedia.org/wiki/Cohort_model#cite_note-harley-9) The words that are activated by the speech signal but are not the intended word are often called "competitors."[[10]](https://en.wikipedia.org/wiki/Cohort_model#cite_note-ibrahim-10)  As more speech segments enter the ear and stimulate more neurons, causing the competitors that no longer match the input to be "kicked out" or to decrease in activation.[[8]](https://en.wikipedia.org/wiki/Cohort_model#cite_note-packard288-8)[[12]](https://en.wikipedia.org/wiki/Cohort_model#cite_note-12) The processes by which words are activated and competitors rejected in the cohort model are frequently called "activation and selection" or "recognition and competition." These processes continue until an instant, called the *recognition point*,[[8]](https://en.wikipedia.org/wiki/Cohort_model#cite_note-packard288-8) at which only one word remains activated and all competitors have been kicked out. This is also known as the *uniqueness point.*

Submit

* printed solutions to the assignment on a due date in the beginning of the class.
* file nltkChapter4a\_template.py (populated with your coding solutions) via Canvas 30 minutes before the class on a due date.