### Assignment on Chapter 4b 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 4b posted at:

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

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

nltkChapter4b\_template.py

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

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

This variable is populated in the assignment template.

Complete the following of the exercises:

1. exercise 0 (0 is a dummy name in this case). Write a function to sort a list of WordNet synsets for proximity to a given synset in accordance with their shortest\_path\_distance(). Report what your function produces given the synsets minke\_whale.n.01, orca.n.01, novel.n.01, and tortoise.n.01, sorted with respect to right\_whale.n.01.
2. exercise 26. Report your findings for (c) for the 10th and 16th Catalan numbers. Template contains the parts utilizing timeit functionality required in (c).
3. exercise 32. Test your function on *TestText* with n=5, n=10, n=15, n=20, n=30. Does any of this number produce satisfactory/natural results? Include the output of your function for n=7.

Submit

* printed solutions to the assignment on a due date in the beginning of the class.
* file nltkChapter4b\_template.py (populated with your coding solutions) via Canvas 30 minutes before the class on a due date.