https://paulweston80.wordpress.com/2016/04/29/heteronyms-heterophones-homographs-homophones-and-homonyms/

heteronym is a subset of homographs

homographs: same spelling, different meaning, can be same pronunciation (homonym) or different pronunciation (heteronym)

heteronym: same spelling different meaning and pronunciation

CS372, Spring Semester, **2020**

School of Computing, KAIST

Homework #3

In the previous homework assignments, we searched for relevant pairs of information, such as (extol, praise highly), or (dead, center). For this homework, we will look into heteronyms and their pronunciations. A heteronym is "one of two or more homographs (such as a *bass* voice and *bass*, a fish) that differ in pronunciation and meaning" (Merriam-Webster Dictionary). Heteronyms may have the same parts-of-speech, as in bass (N), bow (V), tear (N), tear (V), wind (N), and wind (V), but they may also have different parts-of-speech, as in address (N, V), bass (Adj, N), conduct (N, V), frequent (Adj, V), rebel (N, V), tear (N, V) and wind (N, V).

The goal in this homework is first to search the internet or available corpora for sentences in English in which two or more heteronyms appear, and second to give pronouncing annotations to such heteronyms. Your program is supposed to output cited (i.e., together with the source information, such as the Brown Corpus, Times Live) sentences with suitable annotations on heteronyms, together with a CSV file that contains a ranked list of 30 cited sentences with annotations, where the ranking is computed as follows (with precedence 1>2>3):

1. Higher ranking is given to those sentences that contain more occurrences of the homographs, such as wind + wind + tear + tear, than to those sentences that contain fewer, such as wind + wind + tear.

2. Higher ranking is given to those sentences that contain multiple occurrences of homographs, such as wind + wind, than those sentences that contain heteronyms but not homographs, such as wind + tear.

3. Higher ranking is given to those sentences that contain heteronyms with the same part-of-speech information, such as tear + tear, than those with different part-of-speech information, such as PROduce + proDUCE.

You can devise your own ranking for all the other cases, such as preferring shorter sentences and/or sentences from the internet.

As before, you should use techniques that can be implemented in Python and NLTK.

A Write a Python code to access the internet and to output cited sentences with annotations.
B Include the first 30 sentences in an order as described above.
C Discuss your results, to explain how you addressed the goal, and to suggest how you can improve the quality of the results further. This document must be in English.

All the requirements as underlined above must be composed by yourself and without help from anyone else. Any similarity of the results will be flagged for plagiarism and, if found sufficiently similar, penalized, up to, but not limited to, a failure to this homework.

Deadline for uploading your homework at KLMS: 25 May (11:59pm, STRICT)

Homework Submission Guidelines

1. Submission files
   - A CS372_HW3_code_[your ID].py
   - B CS372_HW3_output_[your ID].csv for the ranked and initial 30 sentences that are cited for their sources and have pronouncing annotations for their heteronyms.
   - C CS372_HW3_report_[your ID].docx

2. Remarks
   - Use a maximum of 2 pages for your discussion C.
   - The code should include comments about your implemented idea.
   - For the implementation, you may use external models.
   - Your code should be runnable in our environment.
   - For the output, use slicing [:30] to produce the ranked and initial 30 sentences.
   - You may use any text corpora for the input.
   - Use 11pt font size and default margin/line spacing for your report.
   - Do not use a cover page.