CS372, Spring Semester, 2020

School of Computing, KAIST

Homework #2

In the previous homework, we were supposed to search for a triple, (A, B, C), where an expression A is an intensified version of B with respect to the intensifier C. The success of the search depended much on the presence of the intensifier C, such as *very* or *highly*, which occurs with B in a corpus. Note that such intensifiers are mostly general-purpose, in the sense that they are not specific to a particular lexical item to modify. That is, we can use the adverb *very* to modify nearly any adjectives or adverbs.

For this homework, the goal is to find lexical items D that increase the intensity of a very restricted group of lexical items E that they modify, often to the extreme, such as *pitch* black, *dead* center, *deathly* sick, *stark* contrast, and so on. In particular, your task is to list as many pairs of such D and E as possible in an automated manner, ranking the output in a descending order of how restrictive such pairing is. As before, you may choose your own text corpora, but use techniques that can be implemented in Python and NLTK. You should not include any specific clause for particular lexical items, such as *dead* or *center*, in your program.

A Write a Python code to access some text corpora and to output ranked expressions.
B Include the first 100 expressions in a descending order of uniqueness for such pairing.
C Discuss your results, to argue why they are reasonably ranked, and to suggest how you can improve the quality of the results further.

All the requirements as underlined above must be composed by yourself and without help from anyone else, including the related resources on the Internet, if any. 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: 30 April (11:59pm, STRICT)

Homework Submission Guidelines

1. Submission files
    - $\boxed{A}$ CS372_HW2_code_[your ID].py
    - $\boxed{B}$ CS372_HW2_output_[your ID].csv for the ranked and initial 100 expressions
    - $\boxed{C}$ CS372_HW2_report_[your ID].docx

2. Remarks
    - Use only 1 page for your discussion $\boxed{C}$.
    - The code should include comments about your implemented idea.
    - For the implementation, external models are not allowed.
    - Your code should be runnable in our environment.
    - For the output, use slicing [:100] to produce the ranked and initial 100 expressions.
    - 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.