CS/INFO 4300 Language and Information, Spring 2015

Assignment 2 - "Search your transcripts. You will know it to be true."

Due: Tuesday, February 24, 5:00pm

This assignment can be completed in groups of 1 to 2. One group of 3 will be allowed in case of un-even matching. Indicate your groups at least one week before the deadline using CMS. Every group member is expected to know and understand their group submission completely.

In this assignment we will explore the tradeoffs of information retrieval systems by finding newspaper quotes from "Keeping Up With The Kardashians".

The assignment focuses on cosine similarity, inverted indexes and Edit Distance.

The assignment is structured as an IPython Notebook that you will have to complete and submit via CMS by the due date.

Documentation and tutorials for working with IPython Notebooks are available on the IPython Notebook website.

The bundled ZIP file is available on the course website, and contains:

- This description,
- The IPython Notebook with the assignment,
- An HTML version of the IPython notebook, for reading on other platforms,
- A JSON file with the processed transcripts.

The zip file is password protected; the password will be made available in class (and it can be obtained by emailing Vlad using your Cornell email).

Required libraries. These will be useful throughout the course, so it's worth getting accustomed to them:

- nltk
- numpy
- matplotlib
- python-Levenshtein

You should not use any additional libraries.

It is OK to take inspiration from the in-class demos (available on the course website); if you adapt any code from there you need to acknowledge it in your comments. Make sure to understand the code: more often than not, the in-class demo code is not directly applicable in the assignment.