CS/INFO 4300 Language and Information, Spring 2016

Assignment 3 - "No one can be told what the Ranktrix is. You have to see it for yourself."

Due: Wednesday, March 9 11:59PM

This assignment must be completed individually.

In this assignment, we will explore an information retrieval concept called "relevance feedback." In this assignment, we will be looking at a dataset of movie transcripts and seeing if we can use user feedback to improve search queries. In particular, we will implement a system where a user can query for movies similar to a given one.

This assignment delves into pseudo-relevance feedback, ranking evaluation metrics, and the Rocchio algorithm for relevance feedback.

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 movie scripts in them.

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

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

- numpy
- matplotlib
- sklearn

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.