CS/INFO 4300 Language and Information, Spring 2016

Assignment 1 - "Keeping Up with Social Information"

This assignment has **two parts.** Both parts are **individual** assignments.

In this assignment we will analyze transcripts from the reality TV shows "Keeping Up With The Kardashians" and try to uncover basic social information that is exhibited through language.

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

- This description,
- The IPython Notebook the assignment (contains both parts),
- An HTML version of each IPython notebook, for reading on other platforms,
- A folder with the raw, crawled HTML transcripts to be processed.

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).

Part 1 Due: Thursday, February 4, 12:00pm (noon)

This part of the assignment will deal with *preprocessing* of data. When analyzing social data, it is often the case that raw data is poorly formatted. The goal of this part is to extract and format information from raw HTML files into a format more suitable for analysis. Part 1 of the assignment addresses the following two tasks:

1. Processing the transcripts
2. Removing duplicates

Part 2 Due: Thursday, February 11, 12:00pm (noon)

Now that you've formatted the data, it's time for the exciting part! This part of the assignment focuses on basic NLP and social analysis of the text. In particular, there are two tasks:

3. Language analysis
4. Character interaction analysis

Logistics

This assignment is structured as an IPython Notebook that you will have to complete and submit via CMS by the due date. There will be seperate places to turn in the completed notebooks for part 1 and part 2. Both assignments are in the same notebook, however. When you are turning in part 1, you can turn in a notebook that has no answers filled in for part 2. For part 2, you should turn in your updated notebook with the part 2 solutions filled in, keeping your answers for part 1 intact.

Check in advance that you have access to CMS; if not, please email Jack (`jhessel@cs.cornell.edu`) and request access.

Required Software

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

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

- numpy
- matplotlib
- BeautifulSoup

You will also need to be familiar with regular expressions.