CS/INFO 4300 Language and Information, Spring 2015

Assignment 1 - "Keeping Up with Social Information"

Due: Tuesday, February 3, 5:00pm

This assignment is **individual**.

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 assignment has 4 major parts:

1. Processing the transcripts
2. Removing duplicates
3. Language analysis
4. Character interaction analysis

The assignment is structured as an IPython Notebook that you will have to complete and submit via CMS by the due date. Check in advance that you have access to CMS; if not, please email Vlad (`vlad@cs.cornell.edu`) and request access.

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 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 Vlad using your Cornell email).

Required 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 (covered in the prerequisite courses).