CS/INFO 4300 Language and Information, Spring 2016

Project milestone 1: Technical setup and proposal abstract

First deadline (piazza): Monday, March 21 at midnight
Second deadline (CMS): Thursday, March 24 at midnight

The purpose of the first milestone is to get you thinking about the final project, to form groups and to complete the technical setup.

1. Form groups of 3-4 students: In order to form groups, use piazza to pitch your ideas and/or make meaningful contributions to other people's ideas **by Monday March 21 midnight.** Even if you already have a group/idea post it on on CMS to avoid projects that are too similar. Form groups on CMS.

2. As a group, set up the project framework using the provided template and by following the documentation: https://github.com/CornellNLP/CS4300/blob/master/README.md

At this point you just need to change the project name and add the names and netids of the group members to personalize the template. Don't forget to add all your team-members to the github project, as well as cristiandnm (Prof. DNM's github account) as a user with admin rights.

At the end of this step your group should have a (public) github repository and a live heroku web-page with the running template.

3. Write up of a (roughly) one page project proposal abstract that refines the ideas discussed on piazza. Explain the goal of your proposed application, describe your application in terms of input, output, and discuss use cases. List possible data sources, and clearly discuss the role of the 3 components (information retrieval, machine learning, social).

We expect that the details of the project are likely to change as the semesters continues, but we want to see a general direction that would form the basis of our future discussions.

To be submitted on CMS on Thursday March 24 at midnight:

A PDF containing:

- a link to the Piazza discussion
- a link to your github repository and to your heroku app in PDF format
- The abstract from point 3.

Some possible datasets/APIs (many others out there):

Yelp reviews http://www.yelp.com/dataset_challenge

Twitter API https://dev.twitter.com/rest/public (Note very limited access)

Amazon Product Advertising API https://affiliate-program.amazon.com/gp/advertising/api/detail/main.html
(Rather limited access)

Movie scripts - matched to IMDB data
http://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

Wikipedia editors conversations
http://www.cs.cornell.edu/~cristian/Echoes_of_power_files/wikipedia.talkpages.README.v1.01.txt

StackExchange (and other Stack Overflow sites) http://blog.stackexchange.com/category/cc-wiki-dump/

A great resource of political data (including presidential/republican/democratic debates)
http://www.presidency.ucsb.edu/data.php

QUOTUS data (includes presidential addresses) - http://snap.stanford.edu/quotus/#data

Reddit RAOP data - http://cs.stanford.edu/~althoff/raop-dataset/

Reddit ChangeMyView data: https://chenhaot.com/pages/changemyview.html

Reddit dataset (large dataset of Reddit posts) https://chenhaot.com/data/multi-community/README.txt

Kaggle - https://www.kaggle.com/competitions

Stanford collection of datasets (only a few of them involve text): http://snap.stanford.edu/data/