By popular request, we’re posting this as an “Official” forum to help find interesting projects and form groups.

***Using this is strictly optional - you’re welcome to form groups however you like!***

Post a project below, or review the listings to reach out to your fellow students! Please be respectful and don’t edit someone else’s project posting - use the comment feature instead.

You may also want to review the [Final Project Guidelines](https://docs.google.com/document/d/1IR-R8PsEGriB1sVNCwotSe52m5fiO3UYA6DmRtESmso/preview), or the [Course Roster](https://docs.google.com/document/d/1lF6BRxXzL85ZTENHg8OoJOvIjlH2BehH8yJVHpiOCYk/preview) for a listing of both sections.

# (optional) Project Ideas

**(sample) Build Strong AI**

Achieve human-level intelligence, ??, profit. Sure to win an A+, and maybe destroy the world.

References:

* [The Summer Vision Project](ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-100.pdf) (Papert, 1966)
* [Research Priorities for Robust and Beneficial Artificial Intelligence](http://futureoflife.org/data/documents/research_priorities.pdf) (Russell, et al. 2015)

**Automated Scoring of Written Content**

( Kyle Hamilton )

Challenge: to automate the assessments of online discussions in an educational setting

Goal: formative assessment (gauging progress), and summative assessments (scoring)

*Some Resources:*

<http://www.irrodl.org/index.php/irrodl/article/view/1857/3067>

<https://pdfs.semanticscholar.org/b89f/3d846b4f2d8ac91096efd93762d3cd773a0c.pdf>

<http://files.eric.ed.gov/fulltext/EJ768879.pdf>

<http://www.sciencedirect.com/science/article/pii/S0360131504000788>

<https://www.ets.org/research/topics/as_nlp/>

<http://nlp.stanford.edu/courses/cs224n/2013/reports/song.pdf>

*Possible Data Sources:*

<https://www.kaggle.com/c/asap-aes> (from the automated essay scoring competition)

~2,000 untagged posts from several recent courses on the topic of Cyber Security. I have a commitment from work that faculty and mentors will label some portion of this data to use as a test set.

**Classify Fluency Level or Grammatical Errors or Pronunciation Errors of Learner Speech (or Written Text)** (Ross Katz)

I am interested in utilizing a language learner corpus to identify the fluency level, grammatical errors, or pronunciation errors of speech from people who are learning a language. If learner speech corpora are difficult to find, we could utilize written corpora instead. [There are several corpora that we could explore using](https://www.uclouvain.be/en-cecl-lcworld.html) across languages, or we [could focus on English](http://ice-corpora.net/ice/download.htm). The [Cambridge Learner Corpus](http://www.cambridge.org/us/cambridgeenglish/about-cambridge-english/cambridge-english-corpus) seems like the right dataset, but I have not been able to discover a way to download. Ideally, this project would serve as the base components of a language learning game.

Related Papers:

<https://drive.google.com/file/d/0B1z75ZhM8TdcUFptTEh1ZFdqNUk/view?usp=sharing>

<https://drive.google.com/file/d/0B1z75ZhM8TdcNk1leUYwcGJGM3M/view?usp=sharing>

<https://drive.google.com/file/d/0B1z75ZhM8TdcNXNVX05yNnhIYjQ/view?usp=sharing>

<https://drive.google.com/file/d/0B1z75ZhM8TdcWGdiVDlGRHp0cVE/view?usp=sharing>

<https://arxiv.org/pdf/1607.06153v1.pdf>

<https://arxiv.org/pdf/1605.04278v2.pdf>

<https://arxiv.org/pdf/1603.09727v1.pdf>

<http://bspace.buid.ac.ae/bitstream/1234/664/1/90087.pdf>

----------

**(No longer open) ~~A first step toward people’s underlying “character”?~~**

(Derek S. Chan, [derekschan@ischool.berkeley.edu](mailto:derekschan@ischool.berkeley.edu))

1. **Motivation**: A long-term goal is to try to extract anonymous people’s motive/thought patterns (via their written text) to predict their harmful vs. helpful actions (e.g., physical or emotional violence vs. empathy). However, relevant datasets don’t seem readily available. So, below is a partially-related goal with imaginary data.
2. **Prediction task**: From a film or theater script, predict whether a character behaved violent physically.
3. **Inputs**: A character’s speaking lines and metadata (e.g., gender, estimated age of actress/actor at time of portrayal) in the script.
4. **Labels**: Character is violent physically at any point [yes=1, no=0].
5. **Some potential datasets**:
   1. Entire theatrical scripts can be read programmatically from HTML via Python.
      1. <http://www.imsdb.com/scripts/Taxi-Driver.html>
      2. [http://www.imsdb.com/scripts/Schindler's-List.html](http://www.imsdb.com/scripts/Schindler%27s-List.html)
      3. <http://www.imsdb.com/scripts/Help,-The.html>
      4. <http://www.imsdb.com/scripts/August-Osage-County.html>
      5. <http://www.imsdb.com/scripts/Training-Day.html>
   2. [Cornell University](https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html): Characters’ written text and gender
   3. [Carnegie Mellon University (CMU)](http://www.cs.cmu.edu/~ark/personas/): Characters’ metadata
   4. [UC Santa Cruz](https://nlds.soe.ucsc.edu/fc2): Entire theatrical scripts and genre metadata
   5. [Links](http://dailyscript.com/movie.html) to screenplays and corresponding IMDB character lists
   6. [Project Gutenberg plays](http://www.gutenberg.org/ebooks/search/?query=%28+Drama+%7C+Tragedies+%7C+Comedies+%29+!+bsxPlays) (many here, or a few via NLTK); [Simply Scripts (plays)](http://www.simplyscripts.com/plays_a_h.html)
6. **Pre-processing**: Manually label at least some characters as violent or non-violent
7. **Some related literature:**
   1. Best Paper Award: [Perceived or Not Perceived: Film Character Models for Expressive NLG](https://users.soe.ucsc.edu/~maw/papers/icids-v12.pdf) (2011)
   2. [Movie summary corpus and learning character personas](https://brenocon.com/blog/2013/05/movie-summary-corpus-and-learning-character-personas/) (2013)
   3. [Learning Latent Personas of Film Characters](http://www.cs.cmu.edu/~dbamman/pubs/pdf/bamman+oconnor+smith.acl13.pdf) (2012)
      1. “This character-centric perspective leads to two natural questions. First, can we learn what those standard personas are by how individual characters (who instantiate those types) are portrayed? Second, can we learn the set of attributes and actions by which we recognize those common types? How do we, as viewers, recognize a VILLIAN?”
   4. [Classifying Movie Scripts by Genre. . . . Using NLP-Based Features](http://nlp.stanford.edu/courses/cs224n/2008/reports/06.pdf) (2008)
   5. [Towards automatic detection of antisocial behavior from texts](http://www.aclweb.org/anthology/W11-3704) (2011)
   6. [Automatic Crime Prediction Using Events Extracted From Twitter Posts](https://preventviolentextremism.info/automatic-crime-prediction-using-events-extracted-twitter-posts) (2012)

**Mining valuable financial information from earning report/press release**

(Max Shen)

Use public available information (unstructured text) to predict financial market performance. Features include quarterly earning release and financial reports. Using text from these documents to predict change in volatility for a given period post earnings. Essentially create a robotic Sell Side Analyst with NLP and machine learning techniques.

<http://nlp.stanford.edu/courses/cs224n/2010/reports/aramz-naveed.pdf>

<http://www.businessinsider.com/san-francisco-fed-on-prattle-2016-9>

http://www.forbes.com/forbes/welcome/?toURL=http://www.forbes.com/sites/katinastefanova/2015/12/21/50-words-in-earnings-reports-that-mean-a-plumeting-stock-price/&refURL=https://www.google.com/&referrer=https://www.google.com/

<http://nlp.stanford.edu/pubs/lrec2014-stock.pdf>

https://www.gsam.com/content/gsam/us/en/individual/market-insights/gsam-insights/gsam-perspectives/2016/big-data/natural-language-processing.html

**Online Review Resonance**

(Kuan Lin)

A well-written short piece of text (tweets, online reviews, comments, etc) can go viral and bring sudden spike of attention to the relevant topic or product. The goal of this project is to characterize what features tend to increase resonance among the readers.

We will use Amazon review data, where the proxy of resonance will be the ratio and number of people finding the review useful.

<http://jmcauley.ucsd.edu/data/amazon/>

The project will compose two parts:

* Explore feature importance using common stylometry features ([some example](http://www2.tcs.ifi.lmu.de/~ramyaa/publications/stylometry.pdf))
* Construct a machine learning model using the explored stylometry features.
* Compare the above model with a black-box CNN model (like [this one](https://arxiv.org/abs/1408.5882)) for text classification.

**Yelp Dataset Challenge**

(Yi Jin)

Yelp provides the challenge dataset of 2.7M reviews and 649K tips from 86K businesses. The goal of this project is to analyze the reviews and predict the review’s rating based on the text alone.

<https://www.yelp.com/dataset_challenge>

**Customer Service Chat Text Sentiment**

(Rudy Ferrara, rferrara8@gmail.com)

Measuring sentiment about customer service agents (advisors) in an organization based on chat text.

How did the customer feel about how the advisor did over several categories?

How did the advisor actually do (not based on customer feedback) over the same categories (can the two be measured independently based on the text)?

Ability to Resolve/Help

Attitude

Empathy

Knowledge

Downtime-System not available for advisor

Chat Connectivity

Appointment Availability etc.

**Analyze the Analysts** --

An NLP system that automatic rates the performance of the analysts based on the sentiments of the articles they write

(Alan Wang yjawang@gmail.com)

The idea is to build a summary/extract/sentiment analysis model which takes in general financial news feeds and automatically and synthetically turn each article into one or multiple "ops" (opinion tuples) of the format <author, date, target, sentiment score, target time scope>, where 'target' can be stock tickers, companies, sector/industry, indexes, or general economy etc, and 'target time scope' is the time window of the author's sentiment against the target, (current/days to come, weeks to come, months to come, quarters to come, years to come etc), and finally, a positive sentiment is toward 'buy' and a negative sentiment is toward 'sell'.

After these ops have been constantly and accurately accumulated AND cross-referenced with the individual target's actual market performance, TWO intesting applications follow:

a. an objective rating system for the analyst can be drawn from how well the individual's ops match the actual market trend.

b. when new ops come in the system in each new day, can the system generate composite crowd-based ops which potential would out-rate all individuals?

Of course, a., in particular, b. is a bit beyond NLP, so this project should focus on the summary/extract/sentiment analysis model itself.

Moreover, on top of the sentiment analysis, the challenges are identifying the target(s) and the target time scope, which might exceed the scope of this project.

Finally, it is also interesting on varying the boundaries of potential ops. (sentence level? paragraph level? article level?) too.