Stats 170A: Project in Data Science

Project Ideas and Project Proposals

Padhraic Smyth
Department of Computer Science
University of California, Irvine
## Schedule for Remainder of Quarter

<table>
<thead>
<tr>
<th>Date</th>
<th>Activity</th>
<th>Deadline/Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon, 2/26</td>
<td>Project ideas and proposals (lecture)</td>
<td></td>
</tr>
<tr>
<td>Wed, 2/28</td>
<td>No lecture: Initial project discussion meetings</td>
<td>Meet instructors in DBH 2091, 10 minutes each student, times will be announced on Piazza</td>
</tr>
<tr>
<td>Mon, 3/05</td>
<td>No lecture: Office hours during class time with Prof Smyth, DBH 4216</td>
<td>Version 1 of project proposal due, midnight on Wednesday. Feedback from instructors by 6pm Friday 3/09</td>
</tr>
<tr>
<td>Wed, 3/07</td>
<td>No lecture: Office hours during class time with Prof Smyth, DBH 4216</td>
<td>Version 2 of project proposal due by midnight on Monday</td>
</tr>
<tr>
<td>Mon, 3/12</td>
<td>No lecture: Office hours during class time with Prof Carey, DBH 2091</td>
<td>In-class presentations (in usual classroom) with peer feedback</td>
</tr>
<tr>
<td>Wed, 3/14</td>
<td>Oral project proposal presentations</td>
<td>In-class presentations (in usual classroom) with peer feedback</td>
</tr>
</tbody>
</table>
# Times of Student Meetings on Wednesday

<table>
<thead>
<tr>
<th>Student</th>
<th>Meeting Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharon Babu</td>
<td>2:00</td>
</tr>
<tr>
<td>Dannian Cristalinas</td>
<td>2:10</td>
</tr>
<tr>
<td>Maxwell Wong</td>
<td>2:20</td>
</tr>
<tr>
<td>Bill Hu</td>
<td>2:30</td>
</tr>
<tr>
<td>Alex Lee</td>
<td>2:40</td>
</tr>
<tr>
<td>Gary Pan</td>
<td>2:50</td>
</tr>
<tr>
<td>Madhav Gharmalkar</td>
<td>3:00</td>
</tr>
</tbody>
</table>
Key Aspects of Student Projects

- Work with at least 2 large data sets:
  - Social Media: Yelp, IMDB, Twitter, Reddit, etc
  - Other: OpenMap, Census, CDC, Weather, Wikipedia, news articles, etc

- Problem:
  - For a given data set you will build a software system with some functionality,
  - Interactive entity-based queries, spatio-temporal queries, document classification, sentiment analysis, time-series prediction, etc

- Project Components: each project is expected to contain
  - Database component
  - Machine learning or statistical component
  - Visualization/interface component

- Proposal
  - You will write and refine a proposal this quarter
  - Most of the actual project work will be done next quarter
Project Proposals

• Your proposal should be up to 3 pages long
  – Required to use project proposal template (see class Website)

• Proposals will be in 2 parts
  – Version 1 (initial): due midnight Wednesday March 7th
  – Version 2 (final): due midnight Monday March 12th

• Project proposals will be graded and receive a weight of 50% (based on version 2) of your overall grade for Spring quarter.

• Proposals will primarily be graded on
  (a) clarity (is it clear what will be done in this project?)
  (b) completeness (does it address all of the important aspects of a project?)
  (c) technical correctness
Contents of Project Proposal

1. Project Summary
2. Proposed Technical Approach
3. Data Sets
4. Experiments and Evaluation
5. Software
6. Milestones
Coming up with a Project Idea

• Topic: select a general type of problem you are interested in
  – E.g., see today’s lecture slides for example
  – But also feel free to come up with your own ideas!

• Determine data sets of potential relevance

• Come up with an initial sketch of the system you propose to build
  – Run it past the instructors if you want feedback, e.g., on Piazza or at office hours

• Figure out how you will evaluate your results, e.g., to compare A vs. B
  – Experiments: classification accuracy, precision/recall, etc.
  – User Studies: human users compare results from A and B
  – Insights generated: what we learned about the domain
Resources for Generating Project Ideas

• Class Website:
  – Links to data sets
  – Links to reference articles and texts
  – Links to software and demos

• Google Scholar
  – For research papers (can provide useful ideas)
  – E.g., search on “twitter sentiment analysis”

• Class lecture notes

• Your two instructors (ask questions in Piazza, and in office hours)
Guidelines for Code in Projects

• Use of external code is allowed and encouraged
  – Such code needs to be acknowledged in your reports

• You must write at least some functionality on your own
  – What you implement is up to you
  – You will need to submit your code at the end of the quarter
Possible Topics for Projects
Project Topic: Summarizing Aspects of Review Text

Problem: Given text from a set of reviews (e.g., product, movie, restaurant) investigate information extraction and summarization methods to automatically extract and summarize different aspects of the reviews, e.g., price, service, quality of food, etc.

Possible Data Sources
- Yelp Challenge data sets, Amazon product review data sets, etc
- Wikipedia pages with lists of attributes, e.g., food types for restaurants

Technical Approaches
- Information extraction techniques for finding sentences with specific words
- Machine learning classification methods for sentiment analysis

Interaction/Visualization
- Form-based interface that allows a user to select a product or a movie or a business and that generates a visual summary of the reviews

Extensions
- Display other information about the business, e.g., sentiment over time
1,244 Reviews from our TripAdvisor Community

Read reviews that mention:

All reviews  spa pool  studio restaurant  main pool  lobby lounge  kids club
 car service  the ritz  four seasons  business conference  beverly hills  fitness center
 beautiful property  pool staff  southern california  fire pit  star resort  spa services
 orange county  overlooking the ocean  loft

Traveler rating

- Excellent: 947
- Very good: 163
- Average: 59
- Poor: 32
- Terrible: 25

Traveler type

- Families (263)
- Couples (435)
- Solo (32)
- Business (273)
- Friends (84)

Time of year

- Mar-May (259)
- Jun-Aug (351)
- Sep-Nov (349)
- Dec-Feb (267)

Language

- All languages
- English (1,226)
- German (5)
- Portuguese (5)

Showing 1,226: English reviews

Start your review of Montage Laguna Beach

Click to rate
Project Topic: Automated Tagging of Tweets or News Articles

Problem: Use Wikipedia pages and categories as training data and build a classification algorithm that can classify tweets and news articles into predefined topics. Or cluster tweets/articles into trending topics.

Possible Data Sources
- Wikipedia pages + categories, Tweets, Kaggle or NYT news articles

Technical Approaches
- Machine learning classification algorithms (possibly with hierarchies)
- Web crawling to gather news articles in real-time

Interaction/Visualization
- Interface where a user provides a Website URL (newspaper or other news media site) and system downloads stories, tags them, and displays the results

Extensions
- Also automatically extract and highlight entities (people, places, organizations)
Project Topic: Analyze Geolocated Tweets

Problem: Use geolocated tweets (with GPS) from Southern California over multiple months to see how sentiment in tweets varies over time and spatially – or how entity mentions (e.g., Lakers versus Clippers, Angels versus Dodgers) varies spatially.

Possible Data Sources
- Twitter API – or geolocated tweet dataset from Prof Smyth’s group
- OpenMaps data
- Census data

Technical Approaches
- Information extraction, entity detection
- Spatial and temporal visualization

Interaction/Visualization
- Interface where a user can provide entity names and get back a spatial map or time-series showing prevalence of tweets over space or time
Project Topic: Spatial Visualization of Tweet Sentiment

Problem: Given a large set of tweets related to a particular query or hashtag, generate an interactive map of where the tweets are being generated (by country, by US state, by city) and spatial distribution of their sentiment - take into account population density

Possible Data Sources
- Twitter API, map data, Census data, labeled twitter data sets
- [Possibly: Professor Chen Li’s large Twitter data set at UCI]

Technical Approaches
- Understanding and analysis of census data and map data (GIS aspects)
- Classification and/or clustering algorithms

Interaction/Visualization
- Allow a user to select a topic (possibly from a predefined list, with stored tweets) and system then generates a 2d map visualization. Possibly allow more detailed “drill down” on specific locations or via further keyword or hashtag restrictions to subset the tweets and then display the results.
Project Topic: Summarization of Twitter Hashtags

Problem: Given a hashtag (or set of related hashtags), download tweets with this hashtag and automatically summarize the large set of tweets for a user

Possible Data Sources
- Twitter API, labeled twitter data sets, Wikipedia categories

Technical Approaches
- Clustering algorithms, topic models (see Gensim)
- Machine learning sentiment classification algorithms

Interaction/Visualization
- Allow a user to select from a large set of predefined hashtags, and visually display the clustering/classification results in an understandable manner
Project Topic: Sentiment towards Entities in Reddit over Time

**Problem:** Extract mentions of public figures and organizations from certain subreddits, automatically classify surrounding text in terms of sentiment, and track sentiment over time for each entity.

**Possible Data Sources**
- Reddit news article data set, Wikipedia lists of public figures and organizations

**Technical Approaches**
- Entity detection algorithms
- Sentiment classification algorithms

**Interaction/Visualization**
- Interface where a user selects an entity and the system displays the sentiment over time for that entity

**Extensions**
- Automatically link changes in entity sentiment to news stories at the time
Project Topic: Database of Entity Information from Web Sources

**Problem:** Take a list of named entities in some category (companies, sports stars, movie stars, professors, etc.), crawl the Web for mentions of these entities - extract information about these entities to store in a database suitable for querying.

**Possible Data Sources**
- Specific Websites to get initial lists of entities (e.g., Wikipedia)

**Technical Approaches**
- Web crawling
- Information extraction algorithms

**Interaction/Visualization**
- Form-based interface that allows a user to generate structured queries to the entity database
Challenges in Information Extraction from Web Sources

• Ambiguity in names:
  – Apple = name of the US company, not a person’s name or the fruit, etc.

• Ideally need to extract relations, e.g., between entity and role or location
  – Relations can be stated in many different ways, e.g.,
    • “Tim Cook is CEO of Apple”
    • “Apple’s leader, Mr. Timothy Cook, said…”

• Information may be contradictory
  – Out of date: “Steve Jobs, CEO of Apple,....”
  – Incorrect: “Mark Zuckerberg, who heads Apple,....”

• Need to build a web-crawler to find relevant documents
  – Using the results from an existing search engine would be ideal
  – Building a system that can do this in (near) real-time would be challenging
Project Topic: Automated Tagging of Emails

Problem: Build an app that can automatically tag emails with different categories

Possible Data Sources
  – Wikipedia pages + category labels, Enron email data set

Technical Approaches
  – Machine learning classification algorithms

Interaction/Visualization
  – Set up an email address for demo purposes where a user can send an email and the system will automatically tag it ... or integrate with a real email client app

Extensions
  – Also automatically extract and highlight entities (people, places, organizations)
Other Possible Topics...with Real Domain Experts

• Public health issues
  – Predict impact(s) of legalization of marijuana (UCLA Public Health use case)
  – Examine aspects of the opioid epidemic (UCLA Public Health use case)

• Gun deaths / gun control
  – Examine time distribution (magnitude, sentiment, ....) of public response in social media
  – Dimensions to examine might be positive/negative positions – as a function of time after an incident – related to
    • Increasing background checks
    • Raising age requirements
    • Banning semi-automatic weapons
    • Arming additional individuals (e.g., teachers)
    • Etc.

• Both projects above via Prof Carey
Other Possible Topics...with Real Domain Experts

UCI Education Department

- Develop database and python-based data analysis tools for researchers to query and analyze multiple UCI course data sets
- Data includes: student demographics, large clickstreams, multiple classes
- Database back-end, python data analysis tools on top
- Data has already been collected – needs data science expertise to organize data, create schemas and database, link to Python analysis tools
- Project sponsor: Dr. Fernando Rodriguez in the Department of Education, as part of a joint collaboration with Prof Smyth and his students
Examples of Student Clickstream Data
Examples of Possible Data Sets for Projects
5.2 million reviews
174k businesses
11 metropolitan areas

Yelp Dataset Challenge

Discover what insights lie hidden in our data.

The Challenge

We challenge students to use our data in innovative ways and break ground in research. Here are some examples of topics we find interesting, but remember these are only to get you thinking and we welcome novel approaches!

Photo Classification

Maybe you’ve heard of our ability to identify hot dogs (and other foods) in photos. Or how we can tell you if your photo will be beautiful or not. Can you do better?

Natural Language Processing & Sentiment Analysis

What’s in a review? Is it positive or negative? Our reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.

Graph Mining

We recently launched our Local Graph but can you take the graph further? How do user’s relationships define their usage patterns? Who are the trend setters eating before it becomes popular?

Round 11

Our dataset has been updated for this iteration of the challenge - we’re sure there are plenty of interesting insights waiting there for you. This set includes information about local businesses in 11 metropolitan areas across 4 countries. Round 11 has kicked off starting January 18, 2018 and will run through June 30, 2018.

Download Dataset
Yelp Challenge Data Set
IMDb Datasets

Subsets of IMDb data are available for access to customers for personal and non-commercial use. You can hold local copies of this data, and it is subject to our terms and conditions. Please refer to the Non-Commercial Licensing and copyright/license and verify compliance.

Data Location

The dataset files can be accessed and downloaded from https://datasets.imdbws.com/. The data is refreshed daily.

IMDb Dataset Details

Each dataset is contained in a gzipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file contains headers that describe what is in each column. A ‘\N’ is used to denote that a particular field is missing or null for that title/name. The available datasets are as follows:

title basics tsv.gz - Contains the following information for titles:
  • tconst (string) – alphanumeric unique identifier of the title
  • titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc)
  • primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release
  • originalTitle (string) - original title, in the original language
  • isAdult (boolean) - 0: non-adult title; 1: adult title.
  • startYear (YYYY) – represents the release year of a title. In the case of TV Series, it is the series start year.
  • endYear (YYYY) – TV Series end year. ‘\N’ for all other title types
Sentiment Analysis on Movie Reviews
Fri 28 Feb 2014 – Sat 28 Feb 2015 (23 months ago)

The dataset is comprised of tab-separated files with phrases from the Rotten Tomatoes dataset. The train/test split has been preserved for the purposes of benchmarking, but the sentences have been shuffled from their original order. Each Sentence has been parsed into many phrases by the Stanford parser. Each phrase has a Phraseld. Each sentence has a Sentenceld. Phrases that are repeated (such as short/common words) are only included once in the data.

- train.tsv contains the phrases and their associated sentiment labels. We have additionally provided a Sentenceld so that you can track which phrases belong to a single sentence.
- test.tsv contains just phrases. You must assign a sentiment label to each phrase.

The sentiment labels are:
0 - negative
1 - somewhat negative
2 - neutral
3 - somewhat positive
4 - positive
Amazon product data

Julian McAuley, UCSD

New - Q/A data!

See our newly-released Q/A data (described in our WWW 2016 paper)!

Description

This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.

This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

Files

"Small" subsets for experimentation

If you're using this data for a class project (or similar) please consider using one of these smaller datasets below before requesting the larger files. To obtain the larger files you will need to contact me to obtain access.

K-cores (i.e., dense subsets): These data have been reduced to extract the k-core, such that each of the remaining users and items have k reviews each.

Ratings only: These datasets include no metadata or reviews, but only (user,item,rating,timestamp) tuples. Thus they are suitable for use with mymedialite (or similar) packages.

<table>
<thead>
<tr>
<th>Category</th>
<th>Subsets</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>5-core (8,898,041 reviews)</td>
<td>ratings only (22,507,155 ratings)</td>
</tr>
<tr>
<td>Electronics</td>
<td>5-core (1,689,188 reviews)</td>
<td>ratings only (7,824,482 ratings)</td>
</tr>
<tr>
<td>Movies and TV</td>
<td>5-core (1,697,533 reviews)</td>
<td>ratings only (4,607,047 ratings)</td>
</tr>
<tr>
<td>CDs and Vinyl</td>
<td>5-core (1,097,592 reviews)</td>
<td>ratings only (3,749,004 ratings)</td>
</tr>
<tr>
<td>Clothing, Shoes and Jewelry</td>
<td>5-core (278,677 reviews)</td>
<td>ratings only (5,748,920 ratings)</td>
</tr>
<tr>
<td>Home and Kitchen</td>
<td>5-core (551,682 reviews)</td>
<td>ratings only (4,253,926 ratings)</td>
</tr>
</tbody>
</table>
### Reddit Statistics

<table>
<thead>
<tr>
<th>Metric</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>total # posts</td>
<td>668M</td>
</tr>
<tr>
<td>total # users posting</td>
<td>8.2M</td>
</tr>
<tr>
<td># words per post</td>
<td>30.6</td>
</tr>
<tr>
<td>total # words</td>
<td>&gt;20 billion</td>
</tr>
</tbody>
</table>

---

By Max Woolf — minimaxir.com  Made using R and ggplot2  Data via Reddit
Typical Fields for each Reddit Comment

text: Text of the comment / thread

id: Unique reddit id for the comment / thread

subreddit: Subreddit that the comment / thread belongs to

meta: Metareddit that the comment / thread belongs to.

time: UNIX timestamp of the comment / thread

author: Username of the author of the comment / thread

ups: Number of upvotes the comment / thread received

downs: Number of downvotes the comment / thread received

authorlinkkarma: The author's link karma.

authorkarma: The author's karma.

authorisgold: gold status of the user.
Inside Airbnb
Adding data to the debate

INDEPENDENT, NON-COMMERCIAL, OPEN SOURCE DATA TOOL

How is Airbnb really being used in and affecting your neighborhood?

OUT OF MORE THAN 27,000 LISTINGS:

16K are for the entire home (58%)
87% highly available (more than 60 days/year)
29% multi-listings (where the host has other listings)

FILTER by Neighborhood
Chelsea

50+ data points per listing

SEE Airbnb ACTIVITY OVER TIME IN YOUR NEIGHBORHOOD

HOST “JOHN D”
17 listings

NEXT...

VISIT insideairbnb.com
SHARE it widely
#insideairbnb #illegalhotels
#affordablehousing #nyc

DOWNLOAD the data
(open source; 50+ data points per listing)

The data Airbnb doesn’t want you to see!

Padhraic Smyth, UC Irvine: Stats 170AB, Winter 2018
Other examples of large text data sets that could be used for projects

Text from 4 million Wikipedia articles

PubMed: 20 million abstracts of biomedical research papers

Enron emails: 250,000 company emails
Climate Data Online: Dataset Discovery

Click on each dataset name to expand and view more details. Information generally includes a description of each dataset, links to related tools, FTP access, and downloadable samples.

Climate Data Online

The datasets listed in this section are accessible within the Climate Data Online search interface.

- Daily Summaries
- Global Marine Data
- Global Summary of the Month
- Global Summary of the Year
- Local Climatological Data
- Normals Annual/Seasonal
- Normals Daily
- Normals Hourly
- Normals Monthly
https://gis.ncdc.noaa.gov/maps/ncei/summaries/daily
Performance Measurement System (PeMS) Data Source

Data are obtained from the Caltrans Performance Measurement System (PeMS). Data are collected in real-time from nearly 40,000 individual detectors spanning the freeway system across all major metropolitan areas of the State of California.

PeMS is also an Archived Data User Service (ADUS) that provides over ten years of data for historical analysis. It integrates a wide variety of information from Caltrans and other local agency systems including:

- Traffic Detectors
- Incidents
- Lane Closures
- Toll Tags
- Census Traffic Counts
- Vehicle Classification
- Weight-In-Motion Roadway Inventory

To use PeMS, you must apply for an account. Registration is free and takes only a few minutes. Accounts are typically approved within one to two business days. For questions regarding PeMS, please contact Tim Hart.
Suburban Area
Modeling Human Behavior using Social Media

From Lichman and Smyth, ACM SIGKDD 2014
Geolocated Tweets around UC Irvine
OpenStreetMap powers map data on thousands of web sites, mobile apps, and hardware devices

OpenStreetMap is built by a community of mappers that contribute and maintain data about roads, trails, cafés, railway stations, and much more, all over the world.

Local Knowledge
OpenStreetMap emphasizes local knowledge. Contributors use aerial imagery, GPS devices, and low-tech field maps to verify that OSM is accurate and up to date.

Community Driven
OpenStreetMap’s community is diverse, passionate, and growing every day. Our contributors include enthusiast mappers, GIS professionals, engineers running the OSM servers, humanitarians mapping disaster-affected areas, and many more. To learn more about the community, see the OpenStreetMap Blog, user diaries, community blogs, and the OSM Foundation website.
Data

Home Values

Zillow Home Value Index (ZHVI): A smoothed, seasonally adjusted measure of the median estimated home value across a given region and housing type. It is a dollar-denominated alternative to repeat-sales indices. Zillow also publishes home value and other housing data for local markets, as well as a more detailed methodology and a comparison of ZHVI to the S&P CoreLogic Case-Shiller Home Price Indices.

Home Listings and Sales

Zillow provides data on sold homes, including median sale price for various housing types, sale counts (for which there's detailed methodology), and a normalized view of sale volume, referred to as turnover. Foreclosures also are provided as ratio of housing stock and as a share of all sales in which the home was previously foreclosed upon. There are current and historical for-sale listings data, generally from 2012 to current, ranging from median list prices and inventory counts to share of listings with a price cut, median price cut size, age of inventory, and the days a listing spent on Zillow before the sale was final. Inventory and other housing data also are available for local markets.
The DBpedia Data Set (2015-04)

we are happy to announce the release of DBpedia 2015-04 (also known as '2015 A'). The new release is based on updated Wikipedia dumps dating from February/March 2015 and features an enlarged DBpedia ontology with more infoboxes and ontology mappings, leading to richer and cleaner data.


The English version of the DBpedia knowledge base currently describes 5.9M things out of which 4.3M resources have abstracts, 452K geo coordinates and 1.46M depictions. In total, 4 million resources are classified in a consistent ontology and consists of 2,06M persons, 662K places (including 455K populated places), 378K creative works (including 92K music albums, 90K films and 17K video games), 188K organizations (including 51K companies and 33K educational institutions), 278K species and 5K diseases. The total number of resources in English DBpedia is 15.3M, that besides the 5.9M resources, includes 1.2M skos concepts (categories), 6.83M redirect pages, 256K disambiguation pages and 1.13M intermediate nodes.

We provide localized versions of DBpedia in 128 languages. All these versions together describe 38.3 million things, out of which 23.8 million are localized descriptions of things that also exist in the English version of DBpedia. The full DBpedia data set features 38 million labels and abstracts in 128 different languages, 25.2 million links to images and 29.3 million links to external web pages; 80.9 million links to Wikipedia categories, and 41.2 million links to YAGO categories. DBpedia is connected with other Linked Datasets by around 50 million RDF links.

In addition we provide DBpedia datasets for Wikipedia Commons and Wikidata.
Project Proposals
Project Proposals

• Your proposal should be up to 3 pages long
  – Required to use project proposal template (see class Website)

• Proposals will be in 2 parts
  – Version 1 (initial): due Wednesday March 7th
  – Version 2 (final): due Monday March 12th

• Project proposals will be graded and receive a weight of 50% (based on version 2) of your overall grade for Spring quarter.

• Proposals will primarily be graded on
  (a) clarity (is it clear what will be done in this project?)
  (b) completeness (does it address all of the important aspects of a project?)
  (c) technical correctness
Project Components

• **Database component**: your data for the project should be stored in a database; ideally your project will let your user subset the data in various ways to gain insights interactively.

• **Machine learning or statistical component**: your project needs to have a component that does machine learning or builds a statistical model, e.g., develop a module for classifying sentiment in tweets, or extracting information from product reviews, etc.

• **Visualization/human interface component**: part of your project will need to have a way of visualizing results for a user and allowing a user to interact with your system, e.g.,
  – Input-driven (menus, forms, structured queries)
  – Visualizations over time or space (e.g., tweets by location)
  – Annotated text where the most relevant words for the machine learning/statistical model are highlighted
  – Subsetting the data to which your model is being applied (see above)
  – Etc.
1. Project Summary

Provide a clear description (2 or 3 sentences) that summarizes your project.

A good way to summarize your project is to start with a sentence that clearly defines the problem, e.g., “This project will develop a system to do Y....”.

You should follow this with a brief summary of your planned technical approach, e.g., “The approach I will take to address this problem is 3-fold: (1), use method A to do X (2), use method B to do Y, (3).....” (as an example)
2. Proposed Technical Approach

Write 2 or 3 paragraphs with a clear, more detailed description of the methods and algorithms that you plan to use on the project.

If the system you are building can be thought of as a pipeline with multiple components, a useful approach is to provide a figure that illustrates the pipeline.

Make sure it is clear what your pipeline or system is doing, i.e., what each component will do in terms of taking inputs and producing outputs.

This section should include a brief description of each of the database, machine learning/statistics, and visualization/interface components in your project.
2. Proposed Technical Approach (Example)

Our Algorithm

For each unique food product ID, we look at every sentence of individual reviews and remove all the stopwords (Python NLTK stopwords and our custom set of stopwords). This stopword removal step ensures that irrelevant, unhelpful key phrases do not get added to our set of key phrases. We then extract all n-grams that begin with an adjective and end with a noun (and vice versa) from the filtered reviews. We rank these key phrases based on their values of PF-IRF (phrase frequency - inverse review frequency), which is a variation of TF-IDF (term frequency - inverse document frequency). Finally, we can generate new sentences using only the key phrases with high rankings, along with a Markov model simulating common sentence structures.
2. Proposed Technical Approach (Example)

We are going to divide the whole project into several stages.

- **The first stage is to preprocess the movie review.** We plan to use stopwords list from NLTK to remove stopwords, punctuations and non-alphabetic words. Secondly, we plan to obtain the base part of the word by using Snowball as a stemmer to remove morphological endings. The Porter Stemming Algorithm will be implemented in this step. Then, we plan to extract opinioned words by applying the positive words list from NLTK.

- **The second stage is to classify the opinioned words list obtained from the first stage.**

- **The third stage is training our classifier to recognize the attitude of the reviews.** We plan to try logistic regression, support-vector machines, and Multinominal Naive Bayes models for supervised training.
3. Data Sets

- Briefly describe at least two data set(s) you plan to use in the project. Include references to the data (e.g., a URL) if you can. …..
  - Social media data set
  - + additional data set(s) (can be social media or not)

- If you are able to access and take an initial look at your data, feel free to also include a figure or two in this section

- You can change your data sets during the project if you need to, but you should have identified at least one data set to work with by the time you submit the proposal.
3. Data Sets (Example)

We plan to work with the Rotten Tomatoes movie review dataset publicly provided on [Kaggle.com](https://kaggle.com) as the basis for a machine learning competition.

The data set is a collection of sentences from reviews that are parsed into phrases by the [Stanford Parser](https://nlp.stanford.edu). The data set is preprocessed with a predefined vocabulary that simply removes repeated common or short phrases.

We plan to improve this simple vocabulary through our own method of preprocessing as discussed in the next section.
### 3. Data Sets (Example)

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Length</th>
<th>Type</th>
<th>Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yelp RSS feed <a href="http://www.yelp.com/rss">http://www.yelp.com/rss</a></td>
<td>25 reviews for 5 cities per day (125) Need to collect (1000+)</td>
<td>Short: around 300 characters</td>
<td>Review</td>
<td>Yes: each has a rating (N/5)</td>
</tr>
<tr>
<td>Song Lyrics <a href="http://www.songlyrics.com/">http://www.songlyrics.com/</a></td>
<td>Need to collect</td>
<td>Medium: Around 2,500 characters</td>
<td>Song, Poetry</td>
<td>No</td>
</tr>
<tr>
<td>Twitter Tweets <a href="http://www.sananalytics.com/lab/twitter-sentiment/">http://www.sananalytics.com/lab/twitter-sentiment/</a></td>
<td>5513 tweets Pre-packaged</td>
<td>Short: max 140 characters</td>
<td>Message</td>
<td>Yes: hand classified</td>
</tr>
<tr>
<td>Movie Reviews <a href="http://www.cs.cornell.edu/people/pabo/movie-review-data/">http://www.cs.cornell.edu/people/pabo/movie-review-data/</a></td>
<td>polarity_dataset_v2.0 1000 positive, 1000 negative reviews Pre-Packaged</td>
<td>Medium-Long</td>
<td>Review</td>
<td>Yes: positive/negative</td>
</tr>
</tbody>
</table>
3. Data Sets (Example)
4. Experiments and Evaluation

• Provide a brief and clear description of how you will evaluate how well your system is working

• Other quantitative metrics
  – Prediction accuracy relative to ground truth
  – Speed of execution of your system

• User studies
  – Show users examples of output from system A and system B (baseline)
  – Important that such a study is “blinded”

• Insights generated by your system

• Aspects to consider
  – Test data sets and cross-validation
  – Useful to have a baseline to compare to (E.g., “lesioned” version of your system)
5. Software

• Provide a list of the major pieces of project software that you expect to use, divided into 2 sets:
  – (1) publicly-available code, and
  – (2) code will write yourself.

• This list will probably be incomplete at this point (which is fine) since you may not know yet about all of the publicly-available software that might be relevant to your project.
5. Software (Example)

Publicly-available code:

- **NLTK**: provides a list of stop words and build-in naive Bayes classifier.
- **PyEnchant**: provides spell checking and spelling suggestions.
- **More later.**

*Code to be written ourselves in Python:*

- **Tokenizer** to parse SMS message.
- **Generate a feature list using tokenized message.**
- **Group misspelled words using PyEnchant suggestions.**
- **Track metadata such as number of misspellings in a message.**
- **Bernoulli naive Bayes** to analyze features and classify messages.
## 5. Software (Example)

<table>
<thead>
<tr>
<th>Publicly-Available Code</th>
<th>Code We Will Write/Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Programing Languages:</strong></td>
<td><strong>Ranking Algorithms:</strong></td>
</tr>
<tr>
<td>- Python 3.5 and libraries such as NLTK, NumPy, SciPy, and Matplotlib</td>
<td></td>
</tr>
<tr>
<td>- SQLite3</td>
<td></td>
</tr>
<tr>
<td><strong>Keywords Extraction:</strong></td>
<td><strong>Comparison algorithm</strong> to evaluate similarity between RAKE’s phrases and our algorithm’s phrases.</td>
</tr>
<tr>
<td>- RAKE (Rapid Automatic Keyword Extraction) using NLTK</td>
<td></td>
</tr>
<tr>
<td><strong>Evaluation Software:</strong></td>
<td><strong>Phrase Extraction Algorithm:</strong></td>
</tr>
<tr>
<td>- ROUGE software package to automate the evaluation of our results.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Extract frequent n-grams from all reviews of a product.</td>
</tr>
<tr>
<td></td>
<td>- Extract key phrases from the frequent n-grams, phrases that contain adjectives followed by nouns (and vice versa) with stopwords eliminated.</td>
</tr>
<tr>
<td></td>
<td><strong>Sentence Generation Algorithm:</strong></td>
</tr>
</tbody>
</table>
| | - Uses sentence structure and a Markov chain to generate readable sentences containing descriptive phrases.
6. Milestones

• Provide a brief list of milestones for Spring quarter
  – Weeks 1-2
  – Weeks 3-4
  – Weeks 5-6
  – Weeks 7-8
  – Weeks 9-10

• Note that these milestones are current estimates – you may need to revise and update them once you start working on your project.

• Plan on weeks 9-10 largely being about preparation of your written report, figures, and final presentation materials.
9. References & Links


Project Tips: Plan in Stages

Plan your project in stages so that the overall project is not dependent on the riskier elements working

Example:

PHASE 1

Original Documents → Standard Bag of Words → Standard Logistic Regression → Cross-Validation Experiments
Project Tips: Plan in Stages

Plan your project in stages so that the overall project is not dependent on the riskier elements working.

Example:

PHASE 1

Original Documents → Standard Bag of Words → Standard Logistic Regression → Cross-Validation Experiments

PHASE 2

Information Extraction of Entity Names →
Project Tips: Plan in Stages

Plan your project in stages so that the overall project is not dependent on the riskier elements working.

Example:

PHASE 1
- Original Documents
- Standard Bag of Words
- Standard Logistic Regression
- Cross-Validation Experiments

PHASE 2
- Information Extraction of Entity Names

PHASE 3
- Real-Time User Interaction
Important Aspects of Projects

• Clear definition of the problem: inputs -> ... -> outputs

• Data
  – Make sure you can get the data you need, e.g., labeled data for classifications

• Self-written components
  – Which parts of the code will you write and what will be existing code?

• Evaluation
  – How will you evaluate the quality of your system?
  – Think of ways to compare version A versus B

• Run-time
  – Do you want a system/demo that can run in real-time, or one that operates off-line? Different design decisions for each.
How Projects will be Graded (Overall, end of Spring Q)

• Technical competence:
  – Are algorithms and methods used correctly and appropriately?
  – Do you understand the methods they are using?
  – Were systematic experiments conducted and results interpreted?

• Effort:
  – How much work was done (e.g., coding, experiments, background reading, etc.)
    (Note that effort alone is necessary but not sufficient for a high grade!)

• Creativity and insight:
  – Did you demonstrate creativity in your project?
  – What did insights can be gained from the project? E.g., what types of errors is the system making? What would it take to improve it?

• Writing and communication:
  – Are you able to explain your work clearly (both written and orally)?
## Schedule for Remainder of Quarter

<table>
<thead>
<tr>
<th>Date</th>
<th>Activity</th>
<th>Deadline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon, 2/26</td>
<td>Project ideas and proposals (lecture)</td>
<td></td>
</tr>
<tr>
<td>Wed, 2/28</td>
<td>Initial project discussion meetings, DBH 2091 (10mins, individually with instructors)</td>
<td></td>
</tr>
<tr>
<td>Mon, 3/05</td>
<td>Office hours with Prof Smyth, DBH 4216</td>
<td></td>
</tr>
<tr>
<td>Wed, 3/07</td>
<td>Office hours with Prof Smyth, DBH 4216</td>
<td>Version 1 of project proposal due. Feedback from instructors will follow by Friday 3/09.</td>
</tr>
<tr>
<td>Mon, 3/12</td>
<td>Office hours with Prof Carey, DBH 2091</td>
<td>Version 2 of written project proposal due.</td>
</tr>
<tr>
<td>Wed, 3/14</td>
<td>Oral project proposal presentations</td>
<td>In-class presentations of project proposals with peer feedback.</td>
</tr>
</tbody>
</table>
Times of Student Meetings on Wednesday

<table>
<thead>
<tr>
<th>Student</th>
<th>Meeting Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sharon Babu</td>
<td>2:00</td>
</tr>
<tr>
<td>Dannian Cristalinas</td>
<td>2:10</td>
</tr>
<tr>
<td>Maxwell Wong</td>
<td>2:20</td>
</tr>
<tr>
<td>Bill Hu</td>
<td>2:30</td>
</tr>
<tr>
<td>Alex Lee</td>
<td>2:40</td>
</tr>
<tr>
<td>Gary Pan</td>
<td>2:50</td>
</tr>
<tr>
<td>Madhav Gharmalkar</td>
<td>3:00</td>
</tr>
</tbody>
</table>