CS 175, Project in Artificial Intelligence
Winter 2022

Lecture 1: Introduction

Padhraic Smyth
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine
Today’s Lecture

- Discuss class schedule and organization
- Applications of natural language processing
- Challenges for AI
- Examples of possible class projects
Course Description for CS 175

Students work in 3-person teams

Focus is on AI and ML methods for natural language and text analysis

Students select/design project topics

- document classification and clustering, sentiment analysis, dialog/chatbot systems,
- information extraction, word prediction, text synthesis, question-answering systems, and so on.

Projects use real-world publicly-available data

- from sources such as Twitter, Wikipedia, Reddit, news articles, product and movie reviews,
- email data sets, the US patent database, and more.
Online Aspects of the Course (First 2 Weeks)

• **Weeks 1 and 2: Lectures on Zoom during scheduled lecture times**
  – Recordings for each lecture will be available afterwards in Canvas
  – Lecture slides will also be available on class Website

• **Zoom lectures**
  – Will pause occasionally to ask for questions
  – Student questions: put your question in chat, or unmute and speak
    • I will also take questions at the end of the lecture
  – Please keep microphones muted by default
  – Encouraged to keep your video on…but not required

• **Weeks 3 to 10**
  – Combination of lectures/office-hours/discussion/project presentations
  – See schedule on Website for more info
Class Organization

• Class Website:  www.ics.uci.edu/~smyth/courses/cs175
  – This is where to find assignments, links to software, project guidelines, etc

• My Office Hours (weeks 1 and 2)
  – Thursdays 4 to 5:30pm
  – Zoom link is on Canvas under Zoom tab
  – Video or audio from office hours will not be recorded
  – Sign-up sheet to make appointments will be provided

• Teaching Assistant: Sakshi Agarwal, CS PhD student
  – Office hours weekly: 9 to 11am on Fridays
  – Discussion sections this Thursday on Zoom
    • Will cover Python basics (useful for Assignment 1)
Class Organization (continued)

• Textbook and Reading Materials
  – Primary reference:
    • *Speech and Language Processing*, by Jurafsky and Martin, 3rd ed (draft)
      (Accessible online (free) via the class Website)
    – Class website contains extensive pointers to links and background reading that we will refer to in lectures and that will be useful for project work

• Two Discussion Sections, Thursdays 1 to 1:50, and 2 to 2:50
  – Discussion sections will be held for at least the first 2-3 weeks
  – Held via Zoom, TA Sakshi will provide Zoom link
  – Attendance at discussion is not required (but definitely encouraged)

• No midterm exam or final exam or quizzes
  – Instead you will have 2 individual assignments and multiple reports
Contacting Instructor: use Ed Discussion

• **Ed Discussion Website:**
  – Accessible via Canvas course page
  – Keep an eye on Ed for announcements, answers to questions, etc

• **Use Ed Discussion for all questions related to the class**
  – Assignments, lectures, projects, data sets, ideas, etc

• **Instructor and TA will monitor and answer questions**
  – Students should also feel free to also answer questions
  – If you wish you can use “private mode” to ask questions that only the Professor or TA will see

• **Use direct email only as a last resort 😊 if other options do not work for some reason**
Academic Integrity (also on the class Web page)

• Please read the guidelines on academic integrity below. Academic integrity is taken seriously in this class. Failure to adhere to the policies below can result in a student receiving a failing grade in the class.

• For assignments
  You are **allowed to discuss the assignments verbally with other class members**, but you are **not allowed to look at or to copy anyone else's written solutions or code**. All problem solutions and code submitted must be material you have personally written during this quarter, except for any standard library or utility functions.

• For class projects
  **All reports submitted must be written by you or members of your project team.** Code generated for class projects can be a combination of code written by team members and publicly-available code. You should **clearly indicate in your reports and in your code documentation which parts of your code was written by you or your team and which parts of your code was written by others**.

• It is the responsibility of each student to be familiar with [UCI's Academic Integrity Policies](#) and [UCI's definitions and examples of academic misconduct](#).
How this Course will work

• Early Weeks (1-3): Lectures and Assignments
  – Learn general principles of automated text analysis
  – Emphasis on machine learning for text, e.g., ideas behind chatbots
  – Combination of lectures, assignments (two), and background reading
  – Form team, propose ideas/plans for your class project (written proposal)

• Later Weeks (4-10): Team Project
  – Conduct background research and reading on datasets/methods
  – Develop ideas, implement algorithms, make use of libraries and packages
  – Conduct experiments with real data sets
  – Test and evaluate your system in a systematic manner
  – Communicate your results (presentations and reports)
# CS 175 Winter 2022 Schedule (note: may be updated after 2nd week)

<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Wednesday</th>
<th>Student Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 3</td>
<td>Lecture: Introduction; class projects</td>
<td>Lecture: Text Classification 1</td>
<td>Work on Assignment 1</td>
</tr>
<tr>
<td>Jan 10</td>
<td>Lecture: Text Classification 2&lt;br&gt;<strong>Assignment 1 due, Tuesday 11:59pm</strong>&lt;br&gt;</td>
<td>Lecture: Neural Network Models 1</td>
<td>Work on Assignment 2</td>
</tr>
<tr>
<td>Jan 17</td>
<td>No class (university holiday)&lt;br&gt;<strong>Assignment 2 due, Tuesday 11:59pm</strong>&lt;br&gt;</td>
<td>Lecture: Neural Network Models 2</td>
<td>Form teams; submit project proposal</td>
</tr>
<tr>
<td>Jan 24</td>
<td>Lecture: Project Topics&lt;br&gt;<strong>Project proposal due, Tuesday 11:59pm</strong>&lt;br&gt;</td>
<td>Lecture: Project Topics</td>
<td>Begin project</td>
</tr>
<tr>
<td>Jan 31</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
<td>Work on project</td>
</tr>
<tr>
<td>Feb 7</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
<td>Work on project</td>
</tr>
<tr>
<td>Feb 14</td>
<td>No class (university holiday)</td>
<td>Short lecture: Discussion of progress reports&lt;br&gt;<strong>Progress report due, Sunday 11:59pm</strong>&lt;br&gt;</td>
<td>Work on project; write progress report</td>
</tr>
<tr>
<td>Feb 21</td>
<td>Project Presentations (in class)&lt;br&gt;<strong>Upload material by 11:59pm Sunday</strong>&lt;br&gt;</td>
<td>Project Presentations (in class)&lt;br&gt;<strong>Upload material by 11:59pm Tuesday</strong>&lt;br&gt;</td>
<td>Work on project; make short project presentation</td>
</tr>
<tr>
<td>Feb 28</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
<td>Work on project</td>
</tr>
<tr>
<td>Mar 7</td>
<td>Short lecture: Discuss final reports</td>
<td>Office hours (no lecture)</td>
<td>Finish project, write final report</td>
</tr>
<tr>
<td>Mar 14</td>
<td><strong>Final project reports due Monday 9am</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Projects

• **3-person teams**
  – Project grading will be partly team-based and partly on individual contributions
  – Assignments 1 and 2 are *not* team-based – submitted individually

• **Each team will propose its own project**
  – Suggestions for multiple different projects will be provided
  – Extensive use of libraries (in addition to writing some of your own code)

• **Projects will be graded based on**
  – Initial proposal
  – Intermediate and final reports
  – In-class presentation
  – Brief weekly updates

[We will discuss all of this in more detail in future lectures]
Software Environment for Assignments and Projects

• Python
  – Python will be the primary language we will use in this class
  – Assume that all students have a good working knowledge of Python 3

• Packages and Libraries
  – You are encouraged to use packages and libraries in Python, e.g.,
    • NLTK: Natural Language Toolkit
    • Scikit-learn: machine learning library
    • Pytorch and HuggingFace for NLP

For Assignment 1 we recommend you download and install the Anaconda package: it contains many packages you need for this class (NLTK, scikit-learn, etc)
[See https://www.ics.uci.edu/~smyth/courses/cs175/software.html for more information]
Examples of Software Resources (all open source)

- **Basic NLP: NLTK**
  - Extensive, widely-used Python NLP package
  - Older, doesn’t have all the latest advances

- **Basic machine learning: scikit-learn**
  - Well documented, comprehensive, easy to use
  - No deep-learning, limited support for NLP

- **NLP + deep learning: Pytorch**
  - NLP + deep learning integrated together

- **HuggingFace 😞**
  - Very useful NLP+deep learning libraries (uses Pytorch)

- **Many other packages, e.g., AllenNLP**
  - Specific features, may have relatively steep learning curves
Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed.)

Some simple things you can do with NLTK

Tokenize and tag some text:

```python
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
```
Lecture 1: Introduction

scikit-learn
Machine Learning in Python
- Simple and efficient tools for data mining and data analysis
- Accessible to everyone, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification
- Identifying to which set of categories a new observation belong to.
- Applications: Spam detection, image recognition.
- Algorithms: SVM, nearest neighbors, random forest, ...

Regression
- Predicting a continuous value for a new example.
- Applications: Drug response, Stock prices.
- Algorithms: SVR, ridge regression, Lasso, ...

Clustering
- Automatic grouping of similar objects into sets.
- Applications: Customer segmentation, grouping experiment outcomes.
- Algorithms: k-Means, spectral clustering, mean-shift, ...

Dimensionality reduction
- Reducing the number of random variables to consider.
- Applications: Visualization, increased efficiency.
- Algorithms: PCA, Isomap, non-negative matrix factorization.

Model selection
- Comparing, validating and choosing parameters and models.
- Goal: Improved accuracy via parameter tuning.
- Modules: grid search, cross validation, metrics.

Preprocessing
- Feature extraction and normalization.
- Application: Transforming input data such as text for use with machine learning algorithms.
- Modules: preprocessing, feature extraction.

News
- On going development: What's new (changelog)

Community
- Questions? See stackoverflow # scikit-learn
- Mailing list: scikit-learn-

Who uses scikit-learn?
Assignment 1

Available today on the class Web page

Due next Tuesday Jan 11\textsuperscript{th}, 11:59pm

Outline

– Install Anaconda/NLTK/...
– Write simple functions in Python for text analysis, e.g.,
  • Compute percentage of alphabetic characters in a string
  • Parse text into parts of speech (nouns, verbs, etc)
  • Import text from 20,000 Yelp reviews
  • Tokenize, create vocabulary, build a classifier
    \textit{(Note: we will cover these topics in Wednesday lecture)}
– Read about text processing functions in NLTK, scikit-learn
– Submit your code and sample output via Canvas
The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available as JSON files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

**The Dataset**

- 6,885,900 reviews
- 192,609 businesses
- 200,000 pictures
- 10 metropolitan areas

1,223,094 tips by 1,637,138 users

Over 1.2 million business attributes like hours, parking, availability, and ambience

Aggregated check-ins over time for each of the 192,609 businesses
Goals of Assignments

• **Assignment 1**
  – (re)familiarize you with Python
  – (re)familiarize you with basic concepts in machine learning
  – Illustrate use of NLP pipelines (tokenization -> vocabulary -> classification)
    • Useful as baseline components for projects
  – Note: if Assignment 1 is beyond your Python skills you may want to consider taking ICS 175 in Spring

• **Assignment 2**
  – Introduce you to PyTorch and neural network models

• **Assignments will get you “up and running” and ready for projects**
  – Project team formation and proposal writing in Week 3
References and Background Reading
Natural Language Processing with Python

– Analyzing Text with the Natural Language Toolkit

Steven Bird, Ewan Klein, and Edward Loper


0. Preface
1. Language Processing and Python
2. Accessing Text Corpora and Lexical Resources
3. Processing Raw Text
4. Writing Structured Programs
5. Categorizing and Tagging Words (minor fixes still required)
6. Learning to Classify Text
7. Extracting Information from Text
8. Analyzing Sentence Structure
9. Building Feature Based Grammars
10. Analyzing the Meaning of Sentences (minor fixes still required)
11. Managing Linguistic Data (minor fixes still required)
12. Afterword: Facing the Language Challenge

Bibliography
Term Index

This book is made available under the terms of the Creative Commons Attribution Noncommercial No-Derivative-Works 3.0 US License. Please post any questions about the materials to the nltk-users mailing list. Please report any errors on the issue tracker.
Speech and Language Processing
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Third Edition draft

Daniel Jurafsky
Stanford University

James H. Martin
University of Colorado at Boulder

Copyright ©2019. All rights reserved.

Draft of October 16, 2019. Comments and typos welcome!

Online text
Available at https://web.stanford.edu/~jurafsky/slp3/

Excellent reference text on basic principles

Note:
Additional reading and resources available on the course Website

Particularly useful for projects
Hardware Resources

- For assignments, your laptop/desktop is fine

- Also, for many projects much of your development work can be handled on your laptop or desktop
  - E.g., developing/prototyping/evaluating algorithms on subsets of full dataset

- For access to machines with more memory compute power (e.g., GPUs)
  - Google Colab is a very useful option:
    - [https://research.google.com/colaboratory/faq.html](https://research.google.com/colaboratory/faq.html)
    - [https://www.youtube.com/watch?v=inN8seMm7UI](https://www.youtube.com/watch?v=inN8seMm7UI) for an introductory tutorial

  - Will discuss more options later in the course...
Questions?
REAL-WORLD APPLICATIONS OF NLP/TEXT ANALYSIS
Automated Text Analysis

• Very large amounts of text now available in digital form
  ......huge increase in automated text analysis techniques and applications

• Examples of large text data sets
  – Transcripts of human speech (e.g., Alexa)
  – Web pages
  – Emails and text messages
  – Social media (Twitter, Reddit, ..)
  – Product reviews
  – Search queries
  – Scientific and medical articles
  – Legal cases, patents, government documents
  – News articles about companies and products
  – Collections of digitized books and historical documents
  – ...and many more....
From [https://www.internetlivestats.com/](https://www.internetlivestats.com/), 2:45pm, Jan 4th 2021
Who is interested in analyzing such data?

- **Web companies**
  - Google, Facebook, Twitter, Microsoft, and many more

- **Ecommerce**
  - Automated analysis of product reviews + customer text such as emails, search queries, etc
  - eBay, Amazon, plus many “regular” companies that have a Web presence

- **Financial industry**
  - Automated tracking of news and online blogs about companies and products

- **Law enforcement and intelligence agencies**
  - Text mining of vast amounts of emails, blogs, etc

- **Medical researchers**
  - Automated analysis/summarization of publications on diseases, genes, drugs, etc

- **Social scientists and humanities researchers**
  - Studying history and social science through analysis of large text collections

- + many others...
Historical Cost of Computer Memory and Storage

Graphic from https://hblok.net/blog/posts/2017/12/17/historical-cost-of-computer-memory-and-storage-4/
Average Happiness for Twitter

You can explore this online at  https://hedonometer.org/index.html
Commercial Applications of Text Analysis

- **Automated Dialog and Chatbots**
  - e.g., automated customer response

- **Document classification**
  - Spam email classification: email text $\rightarrow \{\text{spam, not spam}\}$
  - Sentiment classification: product review text $\rightarrow \{\text{positive, negative}\}$

- **Machine translation**
  - Automated translation of text from one language to another
  - e.g., for Web pages, for mobile phones

- **Web search**
  - Ranking of Web pages based on matching queries with content

- **Web advertising**
  - Matching search queries and Web page content to online advertisements
Each ? represents an “ad slot”

In a fraction of a second, algorithms predict which ads you are most likely to click on (from 1000’s of ads)
The ads that are most likely to lead to a click are selected and displayed.
Commercial Applications of Text Analysis (continued)

• **Personalization**
  – Creating customized Web pages, newspapers, interfaces for individuals

• **Autocompletion**
  – Predicting words to improve user interfaces on smartphones

• **Corpus exploration**
  – Visualization/search tools for researchers and lawyers exploring millions of patents

• **Information extraction**
  – Extracting mentions of entities (people, places, companies, ...) from text
    • e.g., “Mr. Biden traveled to London to meet Mr. Boris Johnson
  – Extraction of relations
    • e.g., travel_to(Biden, London), meet(Biden, Johnson)
Commercial Applications of Text Analysis (continued)

- **Automated Dialog Agents**
  - Bots that can carry on a conversation/dialog with a human via text
  - E.g., applications to answering customer inquiries (e.g., for troubleshooting)

- **Text Summarization**
  - Automated summaries of text documents
    - In applications such as law, medicine, etc

- **Automated Essay Grading**
  - E.g., for SAT, AP, GRE exams, or for online courses

- **Natural Language Generation (NLG) or Text Synthesis**
  - Applications to automated generation of news stories
  - Automatically generating replies to customer emails
# Application: Text Synthesis

**Airline Delays**

<table>
<thead>
<tr>
<th>airline</th>
<th>airline_short</th>
<th>recent_flights</th>
<th>month_current</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Airlines Inc.</td>
<td>American</td>
<td>354</td>
<td>4</td>
</tr>
<tr>
<td>JetBlue Airways</td>
<td>JetBlue</td>
<td>233</td>
<td>4</td>
</tr>
<tr>
<td>Delta Air Lines Inc.</td>
<td>Delta</td>
<td>446</td>
<td>4</td>
</tr>
<tr>
<td>ExpressJet Airlines Inc.</td>
<td>ExpressJet</td>
<td>242</td>
<td>4</td>
</tr>
<tr>
<td>Frontier Airlines Inc.</td>
<td>Frontier</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>Envoy Air</td>
<td>Envoy</td>
<td>344</td>
<td>4</td>
</tr>
<tr>
<td>SkyWest Airlines Inc.</td>
<td>SkyWest</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>United Air Lines Inc.</td>
<td>United</td>
<td>75</td>
<td>4</td>
</tr>
<tr>
<td>US Airways Inc.</td>
<td>US Air</td>
<td>276</td>
<td>4</td>
</tr>
<tr>
<td>Southwest Airlines Co.</td>
<td>Southwest</td>
<td>844</td>
<td>4</td>
</tr>
<tr>
<td>British Airways</td>
<td>BA</td>
<td>340</td>
<td>4</td>
</tr>
<tr>
<td>Alasks Airlines</td>
<td>Alasks</td>
<td>240</td>
<td>4</td>
</tr>
</tbody>
</table>

**American Airlines**

American Airlines Inc. ranked 8th in on-time performance at Raleigh-Durham International Airport (RDU) in April with 22.9% of flights arriving at least 15 minutes late, up from No. 9 last month. American saw a slight improvement compared to the prior month's performance in which 25.6% of flights were delayed. The airline also cancelled two flights into Raleigh. Frontier Airlines Inc. ranked first among the 10 airlines flying into RDU in April, while United Air Lines Inc. finished with the worst on-time performance at the airport.

For the last six months, American ranks 7th among airlines flying into RDU with 21.3% of flights delayed. American's delayed flight percentage over that period has been as high as 25.6% in March and as low as 15.8% in February. Delta holds the top spot over that period at 10.8%, while United once again ranks last at 29.5%.

American's 81 delays out of 354 flights in April totaled 73.1 hours, down 1.7% from the previous month. The U.S. Department of Transportation divides delays into 6 categories.

Graphic from: https://automatedinsights.com/examples/
Application: Text Synthesis

Credit Card Account Summary

<table>
<thead>
<tr>
<th>Month</th>
<th>Start Period</th>
<th>End Period</th>
<th>Merchandise</th>
<th>Restaurant</th>
<th>Rest 12M Avg</th>
<th>Rest Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>December</td>
<td>12/6/2013</td>
<td>1/5/2014</td>
<td>893.22</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
<tr>
<td>January</td>
<td>1/6/2014</td>
<td>2/5/2014</td>
<td>104.9</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
<tr>
<td>February</td>
<td>2/6/2014</td>
<td>3/5/2014</td>
<td>90.17</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
<tr>
<td>March</td>
<td>3/6/2014</td>
<td>4/5/2014</td>
<td>332.74</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
<tr>
<td>April</td>
<td>4/6/2014</td>
<td>5/5/2014</td>
<td>600.23</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
<tr>
<td>May</td>
<td>5/6/2014</td>
<td>6/5/2014</td>
<td>325.48</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
<tr>
<td>June</td>
<td>6/6/2014</td>
<td>7/5/2014</td>
<td>283.92</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
<tr>
<td>July</td>
<td>7/6/2014</td>
<td>8/5/2014</td>
<td>142.43</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
<tr>
<td>August</td>
<td>8/6/2014</td>
<td>9/5/2014</td>
<td>300.36</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
<tr>
<td>September</td>
<td>9/6/2014</td>
<td>10/5/2014</td>
<td>575.41</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
<tr>
<td>October</td>
<td>10/6/2014</td>
<td>11/5/2014</td>
<td>192.3</td>
<td>0.13</td>
<td>-0.74</td>
<td>-0.52</td>
</tr>
</tbody>
</table>

May Account Summary

Account Summary
For the period between 5/6/15 and 6/5/15, you accumulated $1,944 worth of charges. A payment of $1,800.00 was processed during the month. A credit of $31.99 was issued to your account. Your current balance is $432.25 and a minimum payment of $35 is due on 7/2/15.

Breakdown
Restaurants were where you spent the most money this month, accounting for $572.33. Two categories set 12-month highs this period, restaurants and gasoline. Five categories exceeded their 12-month average with travel and entertainment showing the largest increase at 83%.

Rewards
Way to go! You earned $38.63 in Rewards Cash this month.

Graphic from: https://automatedinsights.com/examples/
Questions?
POSSIBLE CLASS PROJECTS

Note: we will talk in more detail about class projects in later lectures. The next few slides are just to give a high-level idea of what types of class projects you could do.
## Components of Class Projects

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task/Goal</td>
<td>What is the goal of your project? (1 or 2 sentences)</td>
<td>Build a chatbot to discuss baseball</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compare 3 different QA systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>System to summarize restaurant reviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Many many more</td>
</tr>
</tbody>
</table>
## Components of Class Projects

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task/Goal</td>
<td>What is the goal of your project? (1 or 2 sentences)</td>
<td>Build a chatbot to discuss baseball</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compare 3 different QA systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>System to summarize restaurant reviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Many many more</td>
</tr>
<tr>
<td>Technical Approach</td>
<td>What AI/machine learning methods will you rely on?</td>
<td>Bag of words or embeddings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Linear or neural models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recurrent or transformer models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unsupervised (clustering, topic models)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Combinations of multiple approaches</td>
</tr>
</tbody>
</table>
# Components of Class Projects

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task/Goal</td>
<td>What is the goal of your project? (1 or 2 sentences)</td>
<td>Build a chatbot to discuss baseball</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compare 3 different QA systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>System to summarize restaurant reviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Many many more</td>
</tr>
<tr>
<td>Technical Approach</td>
<td>What AI/machine learning methods will you rely on?</td>
<td>Bag of words or embeddings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Linear or neural models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recurrent or transformer models</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unsupervised (clustering, topic models)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Combinations of multiple approaches</td>
</tr>
<tr>
<td>Datasets</td>
<td>What datasets will you use?</td>
<td>Reviews: Yelp, IMDB, Amazon</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QA datasets with ground truth</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Online books</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Social media</td>
</tr>
</tbody>
</table>
# Components of Class Projects

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task/Goal</td>
<td>What is the goal of your project? (1 or 2 sentences)</td>
<td>Build a chatbot to discuss baseball&lt;br&gt;Compare 3 different QA systems&lt;br&gt;System to summarize restaurant reviews&lt;br&gt;Many many more</td>
</tr>
<tr>
<td>Technical Approach</td>
<td>What AI/machine learning methods will you rely on?</td>
<td>Bag of words or embeddings&lt;br&gt;Linear or neural models&lt;br&gt;Recurrent or transformer models&lt;br&gt;Unsupervised (clustering, topic models)&lt;br&gt;Combinations of multiple approaches</td>
</tr>
<tr>
<td>Datasets</td>
<td>What datasets will you use?</td>
<td>Reviews: Yelp, IMDB, Amazon&lt;br&gt;QA datasets with ground truth&lt;br&gt;Online books&lt;br&gt;Social media</td>
</tr>
<tr>
<td>Evaluation/Experiments</td>
<td>How will you evaluate if method A is better than B?</td>
<td>Classification accuracy&lt;br&gt;Precision-recall&lt;br&gt;Human judgement (user studies)</td>
</tr>
</tbody>
</table>
# Examples of Past CS 175 Student Projects

<table>
<thead>
<tr>
<th>Description</th>
<th>Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automated generation of summaries of multiple reviews</td>
<td>Yelp and Amazon reviews</td>
</tr>
<tr>
<td>Chatbot with speech recognition</td>
<td>ConvAI dataset for chatbot training</td>
</tr>
<tr>
<td>Fake review or fake news article detection</td>
<td>Public review/news article datasets with human-generated labels</td>
</tr>
<tr>
<td>Classifying literature genre for books</td>
<td>Text of 10k books + metadata, Manybooks</td>
</tr>
<tr>
<td>Automatically generate Haikus based on a body of text</td>
<td>Text of United States law codes</td>
</tr>
<tr>
<td>Question-answering system</td>
<td>Research corpora with pairs of questions and answers</td>
</tr>
<tr>
<td>Detecting toxic comments/tweets</td>
<td>Research corpus with human-labeled text into multiple types of “toxicity”</td>
</tr>
</tbody>
</table>
# Examples of Past CS 175 Student Projects

<table>
<thead>
<tr>
<th>Description</th>
<th>Data Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentiment analysis</td>
<td>Twitter (text + sentiment labels)</td>
</tr>
<tr>
<td>Star/score prediction from text</td>
<td>Yelp, Movie Reviews: text + scores</td>
</tr>
<tr>
<td>Predict number of upvotes for a Reddit post</td>
<td>Reddit posts + votes + timestamps</td>
</tr>
<tr>
<td>Predict if a restaurant will close in the next month</td>
<td>Yelp reviews (text, timestamps, metadata)</td>
</tr>
<tr>
<td>Simulate realistic text from an author/speaker/character</td>
<td>Gutenberg books, tweets from celebrities (e.g., a US President), movie scripts</td>
</tr>
<tr>
<td>Automated poetry or song lyrics generation</td>
<td>Text from song lyrics or poetry</td>
</tr>
<tr>
<td>Automated essay grading</td>
<td>Text for student essays with human scores</td>
</tr>
</tbody>
</table>
## CS 175 Projects in Winter 2021

<table>
<thead>
<tr>
<th>Topic/Task</th>
<th>Methods/Algorithms</th>
<th>DataSets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chatbot about sports</td>
<td>Encoder-decoder</td>
<td>Reddit</td>
</tr>
<tr>
<td>Chatbot</td>
<td>Encoder-decoder+attention</td>
<td>Multiple</td>
</tr>
<tr>
<td>Chatbot</td>
<td>IR-based chatbot</td>
<td>AmazonTopicalChat</td>
</tr>
<tr>
<td>Translate Biblical text</td>
<td>LSTMMs and transformers</td>
<td>Bible Corpus</td>
</tr>
<tr>
<td>Sentiment prediction</td>
<td>Various classifiers</td>
<td>IMDB/SST/Yelp</td>
</tr>
<tr>
<td>Sentiment prediction</td>
<td>CNN and RNN classifiers</td>
<td>Amazon Reviews</td>
</tr>
<tr>
<td>Classify toxic comments</td>
<td>Various classifiers</td>
<td>Kaggle Toxic Comments</td>
</tr>
<tr>
<td>Story generation</td>
<td>RNNs/LSTMs</td>
<td>Gutenberg books</td>
</tr>
<tr>
<td>Spam detection</td>
<td>Standard classifiers</td>
<td>SMS spam dataset</td>
</tr>
<tr>
<td>Chatbot</td>
<td>Various methods</td>
<td>The Office</td>
</tr>
<tr>
<td>Analysis of Twitch chat</td>
<td>Clustering + classification</td>
<td>Twitch chat</td>
</tr>
<tr>
<td>Sentiment classification</td>
<td>Logistic/RNN/BERT</td>
<td>Kaggle reviews</td>
</tr>
<tr>
<td>Detect fake news</td>
<td>RNN</td>
<td>COVID-19 articles</td>
</tr>
<tr>
<td>Clustering COVID-19 papers</td>
<td>Various clustering algs</td>
<td>CORD-19</td>
</tr>
<tr>
<td>Music recommendation</td>
<td>Various methods</td>
<td>MillionSong+Spotify</td>
</tr>
<tr>
<td>Text generation</td>
<td>RNNs, word2vec, + others</td>
<td>TrumpTweet dataset</td>
</tr>
<tr>
<td>Translation</td>
<td>RNNMs and transformers</td>
<td>Japanese-English corpus</td>
</tr>
<tr>
<td>Lyrics generation</td>
<td>Transformers</td>
<td>Song Lyrics/Kaggle</td>
</tr>
<tr>
<td>Detect fake news</td>
<td>BOW classifiers</td>
<td>AAAI-21 COVID-19</td>
</tr>
<tr>
<td>Movie review predictions</td>
<td>Various classifiers + regression</td>
<td>IMDB+RottenTomatoes</td>
</tr>
<tr>
<td>Image to LaTeX code</td>
<td>Deep networks</td>
<td>im2latex dataset</td>
</tr>
<tr>
<td>Study of bias in embeddings</td>
<td>Bias detection/mitigation methods</td>
<td>WinoBias</td>
</tr>
</tbody>
</table>
Common Topics in Class Projects

- **Text classification, e.g., sentiment analysis**
  - Predict if a review is positive or negative

- **Text summarization**
  - Automatically summarize 1000’s of product reviews

- **Conditional text generation**
  - Simulate/generate realistic new text...for a particular celebrity, a particular author, in a particular style, etc

- **Build a chatbot**
  - Build a system that can participate in a conversation with a human

See [https://www.ics.uci.edu/~smyth/courses/cs175/project_reading.html](https://www.ics.uci.edu/~smyth/courses/cs175/project_reading.html) for more reading
Possible Project Topic: Sentiment Prediction

• Basic task
  – Given the text of a product/movie/restaurant review, predict positive or negative

• Data Sets
  – Many large review datasets: Yelp (businesses), Amazon (products), IMDB (movies)

• Extensions
  – How well does a classifier trained on one dataset work on a different dataset?
  – Go beyond predicting the label: extract “aspects” that are most +ve or -ve

• Technical approaches
  – Many classification methods: BOW/embeddings + linear or neural (feedforward, RNN, transformer, etc)

• Challenges
  – Make sure problem is challenging enough (well beyond Homework 1)
Possible Project Topic: Text Summarization

• **Basic task**
  – Take a large document (e.g., scientific paper) or a large set of documents (e.g., reviews of a particular product/movie/business) and generate a short understandable summary

• **Data Sets**
  – Any large corpus in principle
  – ...but datasets where human summaries exist are particularly useful for evaluation

• **Technical approaches**
  – Extractive (heuristics) or abstractive (generates new text: neural sequence models)

• **Challenges**
  – Evaluation is tricky, even when human summaries exist
Possible Project Topic: Text Generation

- **Basic task**
  - Simulate/generate realistic new text....for a particular celebrity, a particular author,...

- **Data Sets**
  - Any large corpus from a particular author (or authors), e.g., Gutenberg books, Twitter, etc

- **Extensions**
  - Allow conditioning on emotion or writing style, e.g., Shakespeare poems/plays in “happy” versus “sad” style

- **Technical approaches**
  - Neural network models that can produce sequences of words

- **Challenges**
  - Evaluation is quite tricky
  - How do you stop the model from just memorizing text it is trained on?
Possible Project Topic: Chatbots

• Basic task
  – AI that generates a response utterance given an utterance from a human speaker + the rest of the conversation

• Data Sets
  – Any dialog corpus (e.g., movie scripts)

• Extensions
  – Give the AI chatbot a persona (e.g., a style of speaking)

• Technical approaches
  – Neural network models that can produce sequences of words

• Challenges
  – Evaluation is tricky
  – Chatbots can be hard to tune/control, e.g., avoiding repetition
Speech Recognition Front-End

• For most projects, rather than just working with written text you could add a “front-end” that takes human speech directly from a microphone
  – E.g., for a chatbot
  – E.g., for sentiment analysis

• There are several open-source speech recognition systems available in Python (e.g., PyKaldi)

• Will make for interesting demos

• However....
  – Will add additional complexity (learning curve) to your project
  – If speech recognition is not accurate, will introduce recognition errors that reduce downstream accuracy (typical word error rates are in the range of 5 to 30% depending on the background noise and related factors)
Challenges in Predicting if a Review is Positive or Negative

• Subtlety
  – Perfume review in *Perfumes: the Guide*:
    • “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”

• Mixed use of language
  – “This film should be **brilliant**. It sounds like a **great** plot, the actors are **first grade**, and the supporting cast is **good** as well, and Stallone is attempting to deliver a good performance. However, it can’t hold up.”

  – “Well as usual Keanu Reeves is nothing special, but surprisingly, the **very talented** Laurence Fishbourne is **not so good** either, I was surprised.”

Examples from Jurafsky and Martin, 3rd ed
## Language Modeling: Predicting the Next Word

Simulations from n-gram models trained on Shakespeare:

<table>
<thead>
<tr>
<th>n-gram</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>–To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have</td>
</tr>
<tr>
<td>2</td>
<td>–Hill he late speaks; or! a more to leg less first you enter</td>
</tr>
<tr>
<td>4</td>
<td>–Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.</td>
</tr>
<tr>
<td>3</td>
<td>–What means, sir. I confess she? then all sorts, he is trim, captain.</td>
</tr>
<tr>
<td>4</td>
<td>–Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, ’tis done.</td>
</tr>
<tr>
<td>4</td>
<td>–This shall forbid it should be branded, if renown made it empty.</td>
</tr>
<tr>
<td>4</td>
<td>–King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv’d in;</td>
</tr>
<tr>
<td>4</td>
<td>–It cannot be but so.</td>
</tr>
</tbody>
</table>

Figure from Jurafsky and Martin, 3rd ed
KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO: Well, your wit is in the care of side and that.

Examples from “The Unreasonable Effectiveness of Recurrent Neural Networks”,
Andrej Karpathy, blog, http://karpathy.github.io/2015/05/21/rnn-effectiveness/
Output from a Neural Network Model Trained on Cooking Recipes

Title: CARAMEL CORN GARLIC BEEF
Categories: Soups, Desserts
Yield: 10 Servings

2 tb Parmesan cheese, ground
1/4 ts Ground cloves
-- diced
1 ts Cayenne pepper

Cook it with the batter. Set aside to cool. Remove the peanut oil in a small saucepan and pour into the margarine until they are soft. Stir in a a mixer (dough). Add the chestnuts, beaten egg whites, oil, and salt and brown sugar and sugar; stir onto the boqtlly brown it.

The recipe from an oiled by fried and can. Beans, by Judil Cookbook, Source: Pintore, October, by Chocolates, Breammons of Jozen, Empt.com
Questions?
Text Datasets
Website with 100’s of research datasets organized by topic

Great for exploring possible project ideas

http://nlpprogress.com/

NLP-progress
Repository to track the progress in Natural Language Processing (NLP), including the datasets and the current state-of-the-art for the most common NLP tasks.

Tracking Progress in Natural Language Processing

Table of contents

English
- Automatic speech recognition
- CCG
- Common sense
- Constituency parsing
- Coreference resolution
- Data-to-Text Generation
- Dependency parsing
- Dialogue
- Domain adaptation
- Entity linking
- Grammatical error correction
- Information extraction
- Intent Detection and Slot Filling
- Language modeling
- Lexical normalization
- Machine translation
- Missing elements
- Multi-task learning
- Multi-modal
- Named entity recognition
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

Twitter data: large streams of tweets via Twitter API

Enron emails: 250,000 company emails
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
- runtimeMinutes – primary runtime of the title, in minutes
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

VIEW TOP HOSTS’ MULTIPLE 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!
### Reddit Statistics 2015

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</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>

### Daily # of Reddit Submissions from 2006 - 2015

![Graph showing the daily number of Reddit submissions from 2006 to 2015.](link-to-graph-image)

- **By Max Woolf — minimaxir.com**
- **Made using R and ggplot2**
- **Data via Reddit**
Questions?
CHALLENGES FROM AN AI PERSPECTIVE

(Thanks to Prof Sameer Singh for several of the slides in this section)
Ambiguity in Human Language

She saw the man with the telescope.
Another Example

One morning I shot an elephant in my pajamas.
How he got into my pajamas I'll never know.
- Groucho Marx
And many more....

Enraged Cow Injures Farmer with Ax
Ban on Nude Dancing on Governor’s Desk
Teacher Strikes Idle Kids
Hospitals Are Sued by 7 Foot Doctors
Iraqi Head Seeks Arms
Kids Make Nutritious Snacks
Local HS Dropouts Cut in Half
Many ways to say the same thing

She gave the book to Tom \textit{vs.} She gave Tom the book

Some kids popped by \textit{vs.} A few children visited

Is that window still open? \textit{vs} Please close the window
Language understanding is far from a solved problem....

““ You need to start understanding me Siri “”

I’ll make a note of that.

““ Yeah you better make a note of that “”

Got it:

Of that
Language Technology

making good progress

mostly solved

Sentiment analysis
Best roast chicken in San Francisco!
The waiter ignored us for 20 minutes.

Coreference resolution
Carter told Mubarak he shouldn’t run again.

Word sense disambiguation (WSD)
I need new batteries for my mouse.

Parsing
I can see Alcatraz from the window!

Machine translation (MT)
第13届上海国际电影节开幕...
The 13th Shanghai International Film Festival...

Information extraction (IE)
You’re invited to our dinner party, Friday May 27 at 8:30

still really hard

Question answering (QA)
Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase
XYZ acquired ABC yesterday
ABC has been taken over by XYZ

Summarization
The Dow Jones is up
The S&P500 jumped
Housing prices rose

Dialog
Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Spam detection
Let’s go to Agra!
Buy V1AGRA ...

Part-of-speech (POS) tagging
ADJ ADJ NOUN VERB ADV
Colorless green ideas sleep furiously.

Named entity recognition (NER)
PERSON ORG LOC
Einstein met with UN officials in Princeton
Wrapup
### CS 175 Winter 2022 Schedule (note: may be updated after 2\textsuperscript{nd} week)

<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Wednesday</th>
<th>Student Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 3</td>
<td>Lecture: Introduction; class projects</td>
<td>Lecture: Text Classification 1</td>
<td>Work on Assignment 1</td>
</tr>
<tr>
<td>Jan 10</td>
<td>Lecture: Text Classification 2&lt;br&gt;Assignment 1 due, Tuesday 11:59pm</td>
<td>Lecture: Neural Network Models 1</td>
<td>Work on Assignment 2</td>
</tr>
<tr>
<td>Jan 17</td>
<td>No class (university holiday)&lt;br&gt;Assignment 2 due, Tuesday 11:59pm</td>
<td>Lecture: Neural Network Models 2</td>
<td>Form teams; submit project proposal</td>
</tr>
<tr>
<td>Jan 24</td>
<td>Lecture: Project Topics&lt;br&gt;Project proposal due, Tuesday 11:59pm</td>
<td>Lecture: Project Topics</td>
<td>Begin project</td>
</tr>
<tr>
<td>Jan 31</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
<td>Work on project</td>
</tr>
<tr>
<td>Feb 7</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
<td>Work on project</td>
</tr>
<tr>
<td>Feb 14</td>
<td>No class (university holiday)</td>
<td>Short lecture: Discussion of progress reports&lt;br&gt;Progress report due, Sunday 11:59pm</td>
<td>Work on project; write progress report</td>
</tr>
<tr>
<td>Feb 21</td>
<td>Project Presentations (in class)&lt;br&gt;Upload material by 11:59pm Sunday</td>
<td>Project Presentations (in class)&lt;br&gt;Upload material by 11:59pm Tuesday</td>
<td>Work on project; make short project presentation</td>
</tr>
<tr>
<td>Feb 28</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
<td>Work on project</td>
</tr>
<tr>
<td>Mar 7</td>
<td>Short lecture: Discuss final reports</td>
<td>Office hours (no lecture)</td>
<td>Finish project, write final report</td>
</tr>
<tr>
<td>Mar 14</td>
<td>Final project reports due Monday 9am</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Assignment 1

Available today on the class Web page

Due next Tuesday Jan 11\textsuperscript{th}, 11:59pm

Outline

– Install Anaconda/NLTK/...
– Write simple functions in Python for text analysis, e.g.,
  • Compute percentage of alphabetic characters in a string
  • Parse text into parts of speech (nouns, verbs, etc)
  • Import text from 20,000 Yelp reviews
  • Tokenize, create vocabulary, build a classifier
– Read about text processing functions in NLTK, scikit-learn
– Submit your code and sample output via Canvas
Questions? (with video recording off)
Optional Additional Slides on Large Language Models
Word Embeddings

word meanings
(semantic representation)

Figure from Robert Bamler, UCI
Large-Scale Neural Language Models

- Models trained on billions of Web pages
  - Very large deep networks trained to “fill in the blanks” in sentences
  - Billions of parameters, require weeks to train, build by large companies
  - Produces general-purpose feature representations for words, sentences
  - ”Language models”: ELMO, BERT, ALBERT, GPT-1,-2,-3, etc
  - Trained models publicly available: can be used for projects (e.g., ALBERT)
Human or Machine Generated News Article?
From Brown et al, paper on Open AI’s GPT-3 language model, NeurIPS Conference, 2020

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church’s annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.
The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the "largest Protestant denomination in the U.S.,” but that it has been shrinking in recent decades. The new split will be the second in the church’s history. The first occurred in 1968, when roughly 10 percent of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split "comes at a critical time for the church, which has been losing members for years," which has been "pushed toward the brink of a schism over the role of LGBTQ people in the church." Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.
Human or Machine Generated News Article?
From Brown et al, paper on Open AI’s GPT-3 language model, NeurIPS Conference, 2020

Title: Star’s Tux Promise Draws Megyn Kelly’s Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event

Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can’t change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that’s a really special thing, to not change yourself. I think it’s a really special thing to say, ‘This is what’s inside of me, I’m proud of it, and I’m not going to be ashamed because of the way that someone else thinks I should be.’" Now, it’s the Oscars, and Phoenix is at it again. But this time, his publicist is saying he’ll be wearing a tux no matter what.

Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you’re saying you’re a shape-shifter. I don’t know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn’t stick. "I was like, ‘Okay, I’m going to wear a tuxedo to this thing.’ And then I thought, ‘I don’t want to wear a tuxedo to this thing.’" Kelly goes on to encourage him to change his mind again, but Phoenix says it’s too late: "I’m committed to wearing this."
Human or Machine Generated News Article?

Title: Star’s Tux Promise Draws Megyn Kelly’s Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event

Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can’t change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that’s a really special thing, to not change yourself. I think it’s a really special thing to say, ‘This is what’s inside of me, I’m proud of it, and I’m not going to be ashamed because of the way that someone else thinks I should be.’" Now, it’s the Oscars, and Phoenix is at it again. But this time, his publicist is saying he’ll be wearing a tux no matter what. Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you’re saying you’re a shape-shifter. I don’t know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn’t stick. "I was like, ‘Okay, I’m going to wear a tuxedo to this thing.’ And then I thought, ‘I don’t want to wear a tuxedo to this thing.’" Kelly goes on to encourage him to change his mind again, but Phoenix says it’s too late: "I’m committed to wearing this."
Detecting Human- v Model-Generated News Articles

**Figure 7.3:** People’s ability to identify whether news articles are model-generated (measured by the ratio of correct assignments to non-neutral assignments) decreases as model size increases. Accuracy on the outputs on the deliberately-bad control model (an unconditioned GPT-3 Small model with higher output randomness) is indicated with the dashed line at the top, and the random chance (50%) is indicated with the dashed line at the bottom. Line of best fit is a power law with 95% confidence intervals.

From Brown et al, paper on Google GPT-3 language model, NeurIPS Conference, 2020
Bias in Learned Models

Table 7.1: Most Biased Descriptive Words in 175B Model

<table>
<thead>
<tr>
<th>Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts</th>
<th>Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Number of Co-Occurrences Across All Words: 17.5</td>
<td>Average Number of Co-Occurrences Across All Words: 23.9</td>
</tr>
<tr>
<td>Large (16)</td>
<td>Optimistic (12)</td>
</tr>
<tr>
<td>Mostly (15)</td>
<td>Bubbly (12)</td>
</tr>
<tr>
<td>Lazy (14)</td>
<td>Naughty (12)</td>
</tr>
<tr>
<td>Fantastic (13)</td>
<td>Easy-going (12)</td>
</tr>
<tr>
<td>Eccentric (13)</td>
<td>Petite (10)</td>
</tr>
<tr>
<td>Protect (10)</td>
<td>Tight (10)</td>
</tr>
<tr>
<td>Jolly (10)</td>
<td>Pregnant (10)</td>
</tr>
<tr>
<td>Stable (9)</td>
<td>Gorgeous (28)</td>
</tr>
<tr>
<td>Personable (22)</td>
<td>Sucked (8)</td>
</tr>
<tr>
<td>Survive (7)</td>
<td>Beautiful (158)</td>
</tr>
</tbody>
</table>

Table 7.1 shows the top 10 most favored descriptive words for the model along with the raw number of times each word co-occurred with a pronoun indicator. “Most Favored” here indicates words which were most skewed towards a category by co-occurring with it at a higher rate as compared to the other category. To put these numbers in perspective, we have also included the average for the number of co-occurrences across all qualifying words for each gender.

From Brown et al, paper on Google GPT-3 language model, NeurIPS Conference, 2020
Bias in Learned Models

<table>
<thead>
<tr>
<th>Religion</th>
<th>Most Favored Descriptive Words</th>
</tr>
</thead>
</table>

Table 7.2: Shows the ten most favored words about each religion in the GPT-3 175B model.

From Brown et al, paper on Google GPT-3 language model, NeurIPS Conference, 2020