CS 175, Project in Artificial Intelligence
Winter 2020

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 in the real world

• Challenges for AI

• Ideas for possible class projects
Course Description for CS 175

Students in this project class will work in 3-person teams to develop artificial intelligence and machine learning algorithms and apply them to a range of different problems related to natural language and text analysis.

These problems can include, for example, document classification and clustering, sentiment analysis, dialog/chatbot systems, information extraction, word prediction, text synthesis, question-answering systems, and so on.

Projects can make use of 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.
The Turing Test for AI

Human or Computer?
Class Organization

- Class Website:  [www.ics.uci.edu/~smyth/courses/cs175](http://www.ics.uci.edu/~smyth/courses/cs175)
  - This is where to find assignments, links to software, project guidelines, etc

- My Office Hours (weeks 1 to 3)
  - Wednesdays, 4:30 to 6pm, DBH 4216

- Teaching Assistant: Daokun Jiang, CS PhD student
  - Office hours weekly: day/time TBD
  - Discussion section this Friday on Python basics (useful for Assignment 1)
Class Organization (continued)

• **Textbook and Reading Materials**
  – No official textbook
  – Several useful online texts: see class Webpage
  – Class Website will contain additional pointers to links and background reading that we will refer to in lectures and that will be useful for project work

• **Discussion Sections, Friday mornings, led by TA**
  – Discussion sections will be held for at least the first 2-3 weeks
  – Attendance at discussion is not required

• **No midterm or final exam**
  – but there is a progress report and final report
Contacting Instructor: use Piazza

- **Piazza Website:**
  - [https://piazza.com/uci/winter2020/compsci175/home](https://piazza.com/uci/winter2020/compsci175/home)
  - Keep an eye on Piazza for announcements, answers to questions, etc

- **Use Piazza for all offline 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: 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

• Later Weeks: Team Project
  – build a prototype software system for text analysis (weeks 4 to 10)
  – Propose an idea and plans for your class project (written proposal)
  – Do background research and reading
  – 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 2020 Schedule** (subject to minor updates during the quarter)

<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Wednesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 6</td>
<td>Lecture: Introduction; class projects</td>
<td>Lecture: Concepts in text analysis</td>
</tr>
<tr>
<td>Jan 13</td>
<td>Lecture: Text classification&lt;br&gt;Assignment 1 due, Monday 11:59pm</td>
<td>Lecture: Recurrent neural networks (RNNs)</td>
</tr>
<tr>
<td>Jan 20</td>
<td>No class (university holiday)&lt;br&gt;Assignment 2 due, Monday 11:59pm</td>
<td>Lecture: RNNS (continued); class projects</td>
</tr>
<tr>
<td>Jan 27</td>
<td>Lecture: Evaluation methods&lt;br&gt;Project proposal due, Monday 11:59pm</td>
<td>Office hours (no lecture)</td>
</tr>
<tr>
<td>Feb 3</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
</tr>
<tr>
<td>Feb 10</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
</tr>
<tr>
<td>Feb 17</td>
<td>No class (university holiday)</td>
<td>Short lecture: Discussion of progress reports&lt;br&gt;Progress report due, Sunday Feb 23 11:59pm</td>
</tr>
<tr>
<td>Feb 24</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
</tr>
<tr>
<td>Mar 2</td>
<td>Office hours (no lecture)</td>
<td>Short lecture: Discussion of final reports</td>
</tr>
<tr>
<td>Mar 9</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>
</tr>
<tr>
<td>Mar 16</td>
<td>Final project reports Sunday Mar 15, 11:59pm</td>
<td></td>
</tr>
</tbody>
</table>
Projects

- **3-person teams**
  - Project grading will be partly team-based and partly on individual contributions
  - Note that Assignments 1 and 2 are *not* team-based – these will be worked on and 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

[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 for NLP

For Assignment 1 you should download and install the Anaconda package: it contains many packages you need for this class (NLTK, scikit-learn, etc)
Software Resources

• 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

• Many other packages, e.g., AllenNLP
  – Specific features that may be useful for your projects
  – 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
```
Assignment 1

Available later today on the class Web page (will announce on Piazza)

Due next Monday Jan 13\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 as a single python file via EEE
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 and NLP
  – Illustrate use of 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 deep learning/recurrent networks

• Assignments will get you “up and running” and ready for projects
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/](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 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

- However, some projects may require access to machines with large main memory and/or large compute power (e.g., GPUs)

- We can work with you to find additional computing resources
  - E.g., Google Cloud credit for each student

- There are also resources you can find on your own, e.g.,
  - Google Colab
APPLICATIONS OF TEXT ANALYSIS

(Thanks to Prof Sameer Singh for several of the slides in the remainder of this presentation)
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
  – Web pages
  – Emails and text messages
  – Blogs and microblogs
  – 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....
1985: ~ $100k per gigabyte

2015: ~ $0.3 cents per gigabyte

source: mkomo.com
From [www.internetlivestats.com](http://www.internetlivestats.com), 11:50, Mar 9th 2017
Who is interested in analyzing such data?

- **Web companies**
  - Google, Facebook, Twitter, Microsoft, Yahoo!, 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
Google Search query = “beer”, over time
Google Search query = “beer”, over time
https://hedonometer.org/index.html
Tweets mentioning Coke (green) and Pepsi (red)

from chimpler.wordpress.com
The Google Books Project

- **Google has digitized over 8 million books**
  - Books from 40 university libraries around the world
  - 4.5 million in English, rest in other languages. 6% of all books ever published.
  - 500 billion words
  - Spans multiple centuries since 1500’s

- **Reading the books manually is impossible**
  - Reading only English-language entries since 2000, at the pace of 200 words/minute, with no sleep/food interruptions, would take 80 years!
Applications of Text Analysis

• Automated Dialog and Chatbots
  – e.g., automated customer response

• Document classification
  – Spam email classification: email text -> {spam, not spam}
  – Sentiment classification: product review 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.
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**
  - Developing visualization and search tools for researchers and lawyers exploring millions of patents

- **Information extraction**
  - Extracting mentions of entities (people, places, companies, ...) from text
    - e.g., “Mr. Trump traveled to London to meet Mr. Boris Johnson
  - Extraction of relations
    - e.g., travel_to(Trump, London), meet(Trump, Johnson)
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

<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>Alaska Airlines</td>
<td>Alaska</td>
<td>240</td>
<td>4</td>
</tr>
</tbody>
</table>

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:

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>
</tr>
</thead>
<tbody>
<tr>
<td>December</td>
<td>12/6/2013</td>
<td>1/5/2014</td>
<td>893.22</td>
</tr>
<tr>
<td>January</td>
<td>1/6/2014</td>
<td>2/5/2014</td>
<td>104.9</td>
</tr>
<tr>
<td>February</td>
<td>2/6/2014</td>
<td>3/5/2014</td>
<td>90.17</td>
</tr>
<tr>
<td>March</td>
<td>3/6/2014</td>
<td>4/5/2014</td>
<td>332.74</td>
</tr>
<tr>
<td>April</td>
<td>4/6/2014</td>
<td>5/5/2014</td>
<td>500.23</td>
</tr>
<tr>
<td>May</td>
<td>5/6/2014</td>
<td>6/5/2014</td>
<td>325.48</td>
</tr>
<tr>
<td>June</td>
<td>6/6/2014</td>
<td>7/5/2014</td>
<td>283.92</td>
</tr>
<tr>
<td>July</td>
<td>7/6/2014</td>
<td>8/5/2014</td>
<td>142.43</td>
</tr>
<tr>
<td>August</td>
<td>8/6/2014</td>
<td>9/5/2014</td>
<td>309.36</td>
</tr>
<tr>
<td>September</td>
<td>9/6/2014</td>
<td>10/5/2014</td>
<td>575.10</td>
</tr>
<tr>
<td>October</td>
<td>10/6/2014</td>
<td>11/5/2014</td>
<td>192.33</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/
CHALLENGES FROM AN AI PERSPECTIVE
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 vs. She gave Tom the book

Some kids popped by vs. A few children visited

Is that window still open? 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

Spam detection
Let's go to Agra!
Buy VIAGRA ...

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

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

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

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

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
Economy is good

Dialog
Where is Citizen Kane playing in SF?
Castro Theatre at 7:30. Do you want a ticket?
CLASS PROJECTS
Different Aspects of Projects

• **General Tasks/Areas**
  – Chatbots/dialog
  – Language modeling
  – Classification (e.g., sentiment analysis)
  – Summarization
  – Topic modeling
  – ...more

• **Techniques/Components**
  – Tokenization
  – Embeddings
  – RNNs/GRUs
  – Out-of-vocabulary methods
  – Speech recognition front-end
  – Transfer learning
  – ....more
Different Aspects of Projects

- **General Tasks/Areas**
  - Chatbots/dialog
  - Language modeling
  - Classification (e.g., sentiment analysis)
  - Summarization
  - Topic modeling
  - ...more

- **Techniques/Components**
  - Tokenization
  - Embeddings
  - RNNs/GRUs
  - Out-of-vocabulary methods
  - Speech recognition front-end
  - Transfer learning
  - ...more

- **Datasets**
  - Labeled or not?
  - Sequential or bag of words?
  - Preprocessing needed?
  - Already available or needs crawling?

- **Evaluation aspects**
  - Train/test/CV
  - Hyperparameter tuning
  - Metrics
    - Classification: accuracy, precision, F1,..
    - ROUGE, BLEU, etc
  - Error analysis
  - Visualization/explanation
  - User study evaluation

- **Style of Project**
  - Real-time system demo
  - Evaluation/comparison of different methods
  - Investigation of new approach (v baseline)

Notes:
1. These lists are suggestions: not exhaustive
2. By the time you write your proposal you should be able to describe each of these aspects of your project
## 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, 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>
Possible Projects using Document Classification

• Use Wikipedia pages and categories as training data and build a classification algorithm that can classify news articles

• Build a sentiment classification algorithm that can predict if a product or movie review is positive or negative

• Develop an algorithm that can automatically classify emails into an appropriate folder (e.g., for Gmail)

• Conduct a systematic study of how document length, sample size, or other factors affect the accuracy of document classifiers on standard data sets
Challenges in Predicting Review Sentiment

• 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:

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

Figure from Jurafsky and Martin, 3rd ed
Output from a Neural Network Model Trained on Shakespeare

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

MMMMMM---- Recipe via Meal-Master (tm) v8.05

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 boatly 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

MMMMMM

From https://gist.github.com/nylki/1efbaa36635956d35bcc
Generative Chatbot Architecture using Seq2Seq

Possible Projects using Document Clustering

• **Clustering of Documents:**
  - Takes a set of documents (each represented as a bag of words) and automatically clusters/groups the documents

• **Build an algorithm that can cluster news articles so that articles about the same news story end up in the same cluster**
  - Note that to do this well may require extraction of information about people and places and time from the articles

• **Develop a tool to download an individual’s email history (e.g., from Gmail) and to group emails into clusters on similar topics**
Other Ideas for Projects

• **Information Extraction:**
  - Extract names of products and companies from news articles
  - Extract names of food/attributes from restaurant reviews

• **Change in Language over Time:**
  - Develop an algorithm that can automatically identify key topics in US Patent data and track how these topics change over time

• The examples in these slides are just a partial list
  ...many other possibilities!
Word Embeddings

word meanings
(semantic representation)

Figure from Robert Bamler, UCI
Who Are Your Peers?

https://www.merriam-webster.com/dictionary/peer

Figure from Robert Bamler, UCI
Demo of Word Embeddings over Time
(Quicktime Video)
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)
Text Datasets
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

Twitter data: large streams of tweets via Twitter API
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?

Airbnb IN NYC

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

HOST “JOHN D”
17 listings

SEE Airbnb ACTIVITY OVER TIME IN YOUR NEIGHBORHOOD

2012
2013
2014

NEXT...

• VISIT insideairbnb.com
• SHARE it widely
  #insideairbnb #illegalthealts
  #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></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>

![Daily # of Reddit Submissions from 2006 - 2015](image)

By Max Woolf — minimaxir.com

Made using R and ggplot2

Data via Reddit
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 infobox to 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, 682K 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, 250K 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 29.8 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 [Wikimedia Commons and Wikidata](http://example.com).
Wrapup
### CS 175 Winter 2020 Schedule  
(subject to minor updates during the quarter)

<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Wednesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 6</td>
<td>Lecture: Introduction; class projects</td>
<td>Lecture: Concepts in text analysis</td>
</tr>
<tr>
<td>Jan 13</td>
<td>Lecture: Text classification</td>
<td>Lecture: Recurrent neural networks (RNNs)</td>
</tr>
<tr>
<td></td>
<td>Assignment 1 due, Monday 11:59pm</td>
<td></td>
</tr>
<tr>
<td>Jan 20</td>
<td>No class (university holiday)</td>
<td>Lecture: RNNS (continued); class projects</td>
</tr>
<tr>
<td></td>
<td>Assignment 2 due, Monday 11:59pm</td>
<td></td>
</tr>
<tr>
<td>Jan 27</td>
<td>Lecture: Evaluation methods</td>
<td>Office hours (no lecture)</td>
</tr>
<tr>
<td></td>
<td>Project proposal due, Sunday 11:59pm</td>
<td></td>
</tr>
<tr>
<td>Feb 3</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
</tr>
<tr>
<td>Feb 10</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
</tr>
<tr>
<td>Feb 17</td>
<td>No class (university holiday)</td>
<td>Lecture: Discussion of progress reports</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Progress report due, Sunday 11:59pm</td>
</tr>
<tr>
<td>Feb 24</td>
<td>Office hours (no lecture)</td>
<td>Office hours (no lecture)</td>
</tr>
<tr>
<td>Mar 2</td>
<td>Office hours (no lecture)</td>
<td>Lecture: Discussion of final reports</td>
</tr>
<tr>
<td>Mar 9</td>
<td>Project Presentations (in class)</td>
<td>Project Presentations (in class)</td>
</tr>
<tr>
<td></td>
<td>Upload material by 11:59pm Sunday</td>
<td>Upload material by 11:59pm Tuesday</td>
</tr>
<tr>
<td>Mar 16</td>
<td>Final project reports Sunday Mar 15, 11:59pm</td>
<td></td>
</tr>
</tbody>
</table>
Assignment 1

Available soon on the class Web page, will be announced on Piazza

Due by Monday Jan 13\textsuperscript{th}, 11:59pm

Outline

\begin{itemize}
  \item Install Anaconda/NLTK/...
  \item Write simple functions in Python for text analysis, e.g.,
    \begin{itemize}
      \item Compute percentage of alphabetic characters in a string
      \item Parse text into parts of speech (nouns, verbs, etc)
      \item Import text from 20,000 Yelp reviews
      \item Tokenize, create vocabulary, build a classifier
    \end{itemize}
  \item Read about text processing functions in NLTK, scikit-learn
  \item Submit your code as a single python file via EEE
\end{itemize}