Applied Natural Language Processing

Info 256
Lecture 1: Introduction (August 26, 2021)

David Bamman, UC Berkeley
Predictive text messaging

Will you order me a glass of c

“c”

coffee
champagne

q w e r t y u i o p

a s d f g h j k l

z x c v b n m

123 🎁 🎧 space return
London. Michaelmas term lately over, and the Lord Chancellor sitting in Lincoln's Inn Hall. Implacable November weather. As much mud in the streets as if the waters had but newly retired from the face of the earth, and it would not be wonderful to meet a Megalosaurus, forty feet long or so, waddling like an elephantine lizard up Holborn Hill.
Machine translation

Lasciate ogni speranza, voi ch'entrate

Abandon all hope, ye who enter here
Speech Recognition

“Alexa, how many cups are in a quart?”
who is the speaker of the house?

About 712,000,000 results (1.32 seconds)

Speaker of the United States House of Representatives / Officeholder

Nancy Pelosi

The current House speaker is Democrat Nancy Pelosi of California. She was elected to a fourth (second consecutive) term as speaker on January 3, 2021, the first day of the 117th Congress. She has led the Democratic Party in the House since 2003, and is the first woman to serve as speaker.

https://en.wikipedia.org › wiki › Speaker_of_the_United_...
NLP

If you’re interested in the core methods and algorithms, take Info 159/259 (NLP) instead.

- language modeling
- sequence labeling
- phrase-structure parsing
- dependency parsing
- dynamic programming
- MT
Applied NLP

How do we use the methodologies in NLP toward some end?
Software/Libraries

- NumPy
- SciPy
- scikit-learn
- HUGGING FACE
- spaCy
- PyTorch
- Keras
- NLTK
- TensorFlow
NLP is interdisciplinary

- Artificial intelligence
- Machine learning (ca. 2000—today); statistical models, neural networks
- Linguistics (representation of language)
- Social sciences/humanities (models of language at use in culture/society)
NLP is interdisciplinary

Computational Social Science

Adamic and Glance 2005

Computational Journalism

Change in insured Americans under the ACA,
NY Times (Oct 29, 2014)

Computational Humanities

Underwood 2018
Movie revenues

Input: text of movie review
Output: box office revenue

Joshi et al. (2010), “Movie Reviews and Revenues: An Experiment in Text Regression” (NAACL)
Geographical location

Input: tweet
Output: latitude, longitude

Wing and Baldridge (2011), “Simple supervised document geolocation with geodesic grids” (ACL)
Consumer sentiment

Input: tweets
Output: Gallup economic confidence score

O’Connor et al. (2010), “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series” (ICWSM)
Hiring practices

Input: job ads
Output: gender ratio of applicants
Enculturation

Input: employee emails
Output: promotion to manager, time to separation

Srivastava et al. (2017), "Enculturation Trajectories: Language, Cultural Adaptation, and Individual Outcomes in Organizations" (Management Science)
Respect

Input: transcripts of 981 OPD traffic stops (everyday interactions)

Output: measure of “respect” directed from officer to driver

Voigt et al. 2017, "Language from police body camera footage shows racial disparities in officer respect"
The fraction of words about female characters written by women from 1820 to 2000 is shown in the graph. The data is from Ted Underwood, David Bamman, and Sabrina Lee (2018), “The Transformation of Gender in English-Language Fiction,” (Cultural Analytics).
Measurement

This is fundamentally a problem of measurement: how do we design an algorithmic instrument that can transform a text into a quantity?
“TOM!” No answer. “TOM!” No answer. “What's gone with that boy, I wonder? You TOM!” No answer. The old lady pulled her spectacles down and looked over them about the room; then she put them up and looked out under them. She seldom or never looked through them for so small a thing as a boy; they were her state pair, the pride of her heart, and were built for “style,” not service--she could have seen through a pair of stove-lids just as well. She looked perplexed for a moment, and then said, not fiercely, but still loud enough for the furniture to hear: “Well, I lay if I get hold of you I'll--” She did not finish for she was bending down and punching under the bed with the broom. She resurrected nothing but the cat. “I never did see the beat of that boy!” She went to the open door and stood in it and looked out among the tomato vines and “jimpson” weeds that constituted the garden. No Tom. So she lifted up her voice at an angle calculated for distance and shouted: “Y-o-u-u TOM!” There was a slight noise behind her and she turned just in time to seize a small boy by the slack of his roundabout and arrest his flight. “There! I might 'a' thought of that closet. What you been doing in there?” “Nothing.” “Nothing! Look at your hands. And look at your mouth. What is that truck?” “I don't know, aunt.”
"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.
tom no answer tom no answer what’s gone with that boy, I wonder? you tom! no answer the old lady pulled her spectacles down and looked over them about the room.
"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.
"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.
"TOM!"

No answer.

"TOM!"

No answer.

"What's gone with that boy, I wonder? You TOM!"

No answer.

The old lady pulled her spectacles down and looked over them about the room.
“TOM!”
No answer.
“TOM!”
No answer.
“What’s gone with that boy, I wonder? You TOM!”
No answer.

The old lady pulled her spectacles down and looked over them about the room.

The old lady pulled her spectacles
TOM!
No answer.
TOM!
No answer.
What's gone with that boy, I wonder? You TOM!
No answer.
The old lady pulled her spectacles down and looked over them.

Temporal sequence
"TOM!"
No answer.
"TOM!"
No answer.
"What's gone with that boy, I wonder? You TOM!"
No answer.
The old lady pulled her spectacles down and looked over them about the room.
"TOM!"
No answer.
"TOM!"
No answer.
"What's gone with that boy, I wonder? You TOM!"
No answer.
The old lady pulled her spectacles down and looked over them about the room.
What makes language hard?

• Language is a complex social process

• Tremendous ambiguity at every level of representation

• Modeling it is AI-complete (requires first solving general AI)
What makes language hard?

• Speech acts (“can you pass the salt?)
  [Austin 1962, Searle 1969]

• Conversational implicature (“The opera singer was amazing; she sang all of the notes”).
  [Grice 1975]

• Shared knowledge (“Warren is running for election”)

• Variation/Indexicality (“This homework is wicked hard”)
  [Labov 1966, Eckert 2008]
Ambiguity

“One morning I shot an elephant in my pajamas”
Ambiguity

“One morning I shot an elephant in my pajamas”
Ambiguity

“One morning I shot an elephant in my pajamas”
I made her duck

[SLP2 ch. 1]

• I cooked waterfowl for her
• I cooked waterfowl belonging to her
• I created the (plaster?) duck she owns
• I caused her to quickly lower her head or body
• …
Information theoretic view

"One morning I shot an elephant in my pajamas"

Shannon 1948
Decoding

"One morning I shot an elephant in my pajamas"

representation

words

syntax

semantics

discourse

decode(encode(X))
“Raw” data

• We often want to make claims about the world using textual data.

• Data is not self-evident, neutral or objective

• Data is collected, stored, processed, mined, interpreted; each stage requires our participation.

• What is the process by which the data you have got to you?

Gitelman and Jackson (2013); D'Ignazio and Klein (2020)
Administrivia

• David Bamman
dbamman@berkeley.edu

• Office hours:
  • Wednesdays 10am-11am, Zoom
  • Thursdays 10am-11am, 314 South Hall

• Shefali Bhatia, TA
shefalibhatia@berkeley.edu
Info 256

• Each class period will be divided between:
  • a short lecture; and
  • in-class lab work using Jupyter notebooks

• Students must prepare for each class and submit homeworks before class.
Grading

• Homeworks (40%)
• Participation (10%)
• Group project (50%)
Late submissions

• All homeworks are due on the date/time specified, before each class. We’ll go over the homework in class, so **no late homeworks**.

• You can drop 3 homeworks.
Homeworks

• Homeworks will be frequent; you are free to discuss them at a high level with your classmates, but all coding must be done individually.

• If you use or build on others' code (e.g., from StackOverflow), you must cite its source.

• UC Berkeley code of conduct: http://sa.berkeley.edu/code-of-conduct
Participation

• Participation includes:
  • Coming to class and working in groups.
  • Peer assessment of homework and project deliverables.
  • Answering Piazza questions from your classmates
Course project

- Semester-long project (involving 1-3 students), involving natural language processing in support of an empirical research question.
  - Project proposal/literature review
  - Midterm report
  - 6-page final report, workshop quality
  - Project presentation
ACL 2021 workshops

- *SEM 2021: The 10th Joint Conference on Lexical and Computational Semantics
- 2nd International Workshop on Computational Approaches to Historical Language Change (LChange’21)
- Workshop on Natural Language Processing for Programming
- Third Workshop on Gender Bias for Natural Language Processing
- Workshop on Online Abuse and Harms
- 17th Workshop on Multiword Expressions (MWE 2021)
- 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)
- Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)
<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
<th>Readings</th>
<th>Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8/26</td>
<td>Introduction</td>
<td>NLTK 1</td>
<td>Nguyen et al. 2020</td>
</tr>
<tr>
<td>2</td>
<td>8/31</td>
<td>Words</td>
<td>NLTK 3; Potts</td>
<td>Manshel 2020; Fischer-Baum et al. 2020</td>
</tr>
<tr>
<td>3</td>
<td>9/2</td>
<td>Finding distinctive terms</td>
<td>Kilgarriff 2001 (up to p. 248); Monroe et al. 2009 (up to 3.3)</td>
<td>Jurafsky et al. 2014; Mosteller and Wallace 1964</td>
</tr>
<tr>
<td>3</td>
<td>9/7</td>
<td>Dictionaries</td>
<td>Stewart and Grimmer (2011) (up to section 5.2)</td>
<td>Lucy et al. 2020; Mendelsohn et al. 2020; Zhou and Jurgens 2020</td>
</tr>
<tr>
<td>3</td>
<td>9/9</td>
<td>Lexical semantics/word embeddings</td>
<td>SLP3 ch. 6 Gensim word2vec tutorial</td>
<td>Shechtman 2021; Soni et al. 2021; Kozlowski et al. 2019</td>
</tr>
<tr>
<td>4</td>
<td>9/14</td>
<td>Contextual embeddings</td>
<td>Smith 2020; Devlin et al. 2018</td>
<td>Bamman and Burns 2020</td>
</tr>
<tr>
<td>4</td>
<td>9/16</td>
<td>EDA: Text clustering</td>
<td>Blog post; Scikit-learn clustering</td>
<td>Nelson 2020</td>
</tr>
<tr>
<td></td>
<td>9/23</td>
<td>Text classification 1: logistic regression</td>
<td>NLTK 6; Scikit-learn tutorial</td>
<td>Zhang et al. 2018;</td>
</tr>
</tbody>
</table>
• Course homework will be on Github: https://github.com/dbamman/anlp21

• Sign up for an account right now if you don’t have one!
In class

- [https://github.com/dbamman/anlp21](https://github.com/dbamman/anlp21)
- anlp21/0.setup
- Install anaconda environment + libraries we’ll use frequently.