CS 730: **Text Mining for Social Media** & Collaboratively Generated Content

Mondays 10am-12:30pm
Emerson E510 (Conference room)
Instructor: Eugene Agichtein ([http://www.mathcs.emory.edu/~eugene/](http://www.mathcs.emory.edu/~eugene/))
Lecture Plan

• What is Text Mining & Natural Language Processing?

• What is special about Social Media?

• Course topics & example applications

• Logistics & going forward
What is Text Mining?

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking ... of the extracted information together to form new facts or new hypotheses ...  

-- M. A. Hearst, 2003
Natural Language Processing

• Computers would be a lot more useful if they could handle our email, do our library research, talk to us ...

• Goals of NLP can be very ambitious:
  – True text understanding
  – Reasoning about text
  – Real-time participation in spoken dialogue

Dave Bowman: Open the pod bay doors, HAL.
HAL: I’m sorry Dave. I’m afraid I can’t do that.
Or goals can be very practical

• Computers understand natural language now!
  – Web search engines
  – Information extraction for automatic message processing
  – Machine translation services on the Web
  – Voice recognition on cell phones
  – Speech synthesis on telephones
  – Dialog systems on telephones
  – ....

• NL technologies aren’t yet perfect...
  – ...but they’re good enough to be useful to millions of people
  – ...every day
Areas of NLP

• Four main NL technologies
  – Information retrieval
  – Speech recognition and synthesis
  – Natural language processing
  – Machine translation

• Keep in mind:
  – They all look different at the surface level...
    ...but share techniques and tools
  – They’re different points along a spectrum of language understanding
“Traditional” Natural Language Processing

- **Phonetics/phonology/morphology**: what words (or subwords) are we dealing with?

- **Syntax**: What phrases are we dealing with? Which words modify one another?

- **Semantics**: What’s the literal meaning?

- **Pragmatics**: What should you conclude from the fact that I said something? How should you react?
The classic acid test for natural language processing.
Requires capabilities in both interpretation and generation.
About $10$ billion spent annually on human translation.
Why is NLP hard (continued)

- Natural language is:
  - highly ambiguous at all levels
  - complex and subtle
  - fuzzy, probabilistic
  - involves reasoning about the world
  - a key part of people interacting with other people (a social system):
    - persuading, insulting and amusing them

- But NLP can also be surprisingly easy sometimes:
  - rough text features can often do half the job
Ambiguity: Favorite Headlines

• Ban on Nude Dancing on Governor’s Desk
• Iraqi Head Seeks Arms
• Juvenile Court to Try Shooting Defendant
• Teacher Strikes Idle Kids
• Stolen Painting Found by Tree
• Kids Make Nutritious Snacks
• Local HS Dropouts Cut in Half
• Red Tape Holds Up New Bridges
• Man Struck by Lightning Faces Battery Charge
• Clinton Wins on Budget, but More Lies Ahead
• Hospitals Are Sued by 7 Foot Doctors
Reference Resolution

U: Where is *A Bug's Life* playing in *Mountain View*?
S: *A Bug's Life* is playing at the *Century 16 theater*.
U: When is it playing there?
S: It's playing at 2pm, 5pm, and 8pm.
U: I'd like 1 adult and 2 children for the first show.
   How much would that cost?

- **Knowledge sources:**
  - Domain knowledge
  - Discourse knowledge
  - World knowledge
Making Progress in NLP

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources

- The answer that’s been getting traction:
  - probabilistic models built from language data
    - \( P(\text{“maison”} \rightarrow \text{“house”}) \) high
    - \( P(\text{“L’avocat général”} \rightarrow \text{“the general avocado”}) \) low

- Some computer scientists think this is a new “A.I.” idea
  - But really it’s an old idea that was stolen from the electrical engineers....
## Estimates of daily content creation

[Ramakrishnan & Tomkins, IEEE Computer 2007]

<table>
<thead>
<tr>
<th>Content type</th>
<th>Amount produced / day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Published (books, magazines, newspapers)</td>
<td>3-4 Gb</td>
</tr>
<tr>
<td>Professional Web (paid creation, e.g., corporate Web site)</td>
<td>2 Gb</td>
</tr>
<tr>
<td><strong>User-generated (reviews, blogs, personal Web sites)</strong></td>
<td><strong>8-10 Gb</strong></td>
</tr>
<tr>
<td>Private text (instant messages, email)</td>
<td>3000 Gb (= 3Tb)</td>
</tr>
</tbody>
</table>
Goal of this Seminar

Using statistical/machine-learning approaches to NLP, how can we mine from the huge amounts of (social) content on the web?
Course Topics (Tentative)

Will study text mining to extract, filter, and organize knowledge from social media & collaboratively generated content.

- **Concept extraction and manipulation**
  - knowledge base construction and Explicit Semantic Analysis
- **Temporal analysis**
  - forecasting and topic evolution
- **Sentiment, objectivity analysis**
- **Information propagation**
- **Social network and community analysis**
- **Applications to search**
  - searching social media or using it to improve web search
Lots of content... Do we trust it?

On the Internet, nobody knows you are a dog.
The Role of Community in CGC
### CGC as Source of World Knowledge (AI Holy Grail)

**Before Social Web …**

- Encyclopædia Britannica
  - 65K articles (1768-2010)

- WordNet 3.0
  - 100K synsets (1985-2007)

- CYC
  - 4.6M assertions (1984-2007)

**Now**

- Wikipedia
  - 3M articles in English, 14M in 200+ languages (2001-2009)

- Wiktionary (English)
  - 1.4M entries (2002-2009)

- Yahoo! Answers
  - 1B Q&A pairs (2005-2009)

- Flickr
  - 4B images (2004-2009)

---

8/30/2010  
CS730: Text Mining for Social Media, F2010
Lecture Plan

✓ What is Text Mining & Natural Language Processing?

✓ What is special about Social Media?

✓ Course topics & example applications

➢ Logistics & going forward
Course Prerequisites & Logistics

• **Required:** Experience with > one of the following: NLP, IR, or Data Mining/Machine Learning, or **explicit permission** from the instructor.

• **Independence/self-motivation:** this is a seminar course, thus students are expected to lead the discussions/present the assigned readings.

• **Course grade:** based on in-class presentations (50%) and final project or survey paper (50%).
Course Logistics

- Course Webpage
  - [http://www.mathcs.emory.edu/~eugene/cs730/](http://www.mathcs.emory.edu/~eugene/cs730/)
  - Primary means of communication

- Assumed background:
  - Prior exposure to NLP/IR/Data Mining (at least one of)
  - Decent programming skills
  - Basic probability & stats, linear algebra
  - Ability to *independently* and *quickly* learn missing skills

- *Briefly* covers basic NLP methods (first 3-4 weeks).

- The rest of the course based on research/survey papers
Course Logistics (cont’d)

• Grading:
  – 50% class presentations
  – 50% final project/survey paper

• Expected workload:
  – Present/lead discussion of research paper every 2-3 weeks (depends on number of students)
Textbook (first 3-4 weeks)

- Manning & Schuetze, FSNLP
  

- Read online, or buy/borrow a copy
First 3-4 weeks of the course

• Language models (word, n-gram, ...)
• Statistical NLP:
  – Language models
  – Classification and sequence models
    • Part-of-speech tagging, entity tagging, information extraction
• Syntactic (probabilistic) parsing
• Semantic representation from text
• Part II: Applications (TBD)
Why statistical approach to NLP?

Are these sentences grammatical?

• John I believe Sally said Bill believed Sue saw.

• What did Sally whisper that she had secretly read?

• John wants very much for himself to win.

• That a serious discussion could arise here of this topic was quite unexpected.
Grammaticality as probabilistic phenomenon

- Consider (large) corpus of text (e.g., the WSJ)
- Build a model of observed grammatical constructions
- Can we predict whether a given sentence is grammatical?
- What is the (most likely) grammatical structure?
Language and Cognition as Probabilistic Phenomena

• World is filled with uncertainty

• Example: crossing a river (M&S 1.2.3)
  – Consider water flowing, weather, distance
  – Probably no piranhas or alligators in the area
  – Someone tells you “It’s only knee-deep if you walk towards that tree”
  – Integrate linguistic information into overall probabilities of whether it is safe to cross
Origins of LM: Speech

- Speech input is an acoustic wave form

“l” to “a” transition:
Speech (continued)

- Frequency gives pitch; amplitude gives volume
  - sampling at ~8 kHz phone, ~16 kHz mic (kHz = 1000 cycles/sec)

- Fourier transform of wave displayed as a spectrogram
  - darkness indicates energy at each frequency
Speech (continued)

- Time slices are translated into acoustic feature vectors (~15 real numbers per slice)

- Now we have to figure out a mapping from sequences of acoustic observations to words.
Noisy Channel Approach

- We want to predict a sentence given an acoustic sequence:
  \[ s^* = \arg \max_s P(s | A) \]

- The noisy channel approach:
  - Build a generative model of production (encoding)
    \[ P(A, s) = P(s) P(A | s) \]
  - To decode, we use Bayes' rule to
    \[ s^* = \arg \max_s P(s | A) \]
    \[ = \arg \max_s P(s) P(A | s) / P(A) \]
    \[ = \arg \max_s P(s) P(A | s) \]
  - Now, we have to find a sentence maximizing this product

- Why is this progress?

Accoustic seq. (given)

one of candidate sentences

Apply Bayes rule
LM Applications

- Handwriting recognition
  \[ P(\text{text} | \text{strokes}) \propto P(\text{text}) P(\text{strokes} | \text{text}) \]

- OCR
  \[ P(\text{text} | \text{pixels}) \propto P(\text{text}) P(\text{pixels} | \text{text}) \]

- Spelling Correction
  \[ P(\text{text} | \text{typos}) \propto P(\text{text}) P(\text{typos} | \text{text}) \]

- Translation?
  \[ P(\text{english} | \text{french}) \propto P(\text{english}) P(\text{french} | \text{english}) \]
“Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’”

- Warren Weaver (1955:18, quoting a letter he wrote in 1947)
Machine Translation Components

Language Model

source P(e) → e → channel P(f|e) → f

decoder

best e → argmax P(e|f) = argmax P(f|e)P(e)
e

Translation Model

observed f →
Levels of Transfer
Goal of Language Modeling

• Want to build models which assign scores to sentences.
  – $P(\text{I saw a van}) \gg P(\text{eyes awe of an})$
  – $P(\text{AT&T, how can I help you?}) \gg P(\text{80 and tea: ho May I hell Pooh?})$
  – Not really grammaticality: $P(\text{artichokes intimidate zippers}) \approx 0$

• One option: empirical distribution over sentences?
  – Problem: doesn’t generalize (at all)

• Two ways of generalizing
  – **Decomposition**: sentences generated in small steps which can be recombined in other ways
  – **Smoothing**: allow for the possibility of unseen events
Generalizability

• No loss of generality to break sentence probability down with the chain rule

\[ P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1}) \]

• Too many histories!
• N-gram solution: assume each word depends only on a short linear history

\[ P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_{i-k} \ldots w_{i-1}) \]
N-gram model

- Each word is predicted according to a conditional distribution based on a limited prior context

- Conditional Probability Table (CPT): $P(X|\text{both})$
  - $P(\text{of}|\text{both}) = 0.066$
  - $P(\text{to}|\text{both}) = 0.041$
  - $P(\text{in}|\text{both}) = 0.038$

- From 1940s onward (or even 1910s – Markov 1913)
- a.k.a. Markov (chain) models
**Unigram Models**

- Simplest case: unigrams
  \[ P(w_1 w_2 \ldots w_n) = \prod_i P(w_i) \]
- Generative process: pick a word, pick a word, …
- As a graphical model:

  ![Graphical representation of unigram models]

  - To make this a proper distribution over sentences, we have to generate a special STOP symbol last. (Why?)
- Examples:
  - [fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, dollars, quarter, in, is, mass.]
  - [thrift, did, eighty, said, hard, 'm, july, bullish]
  - [that, or, limited, the]
  - []
  - [after, any, on, consistently, hospital, lake, of, of, other, and, factors, raised, analyst, too, allowed, mexico, never, consider, fall, bungled, davison, that, obtain, price, lines, the, to, sass, the, the, further, board, a, details, machinists, the, companies, which, rivals, an, because, longer, oakes, percent, a, they, three, edward, it, currier, an, within, in, three, wrote, is, you, s., longer, institute, dentistry, pay, however, said, possible, to, rooms, hiding, eggs, approximate, financial, canada, the, so, workers, advancers, half, between, nasdaq]
Bigram Models

- Big problem with unigrams: $P(\text{the the the the}) \gg P(\text{I like ice cream})$

- Condition on last word:

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_{i-1})$$

- Any better?

  - [texaco, rose, one, in, this, issue, is, pursuing, growth, in, a, boiler, house, said, mr., gurria, mexico, 's, motion, control, proposal, without, permission, from, five, hundred, fifty, five, yen]
  - [outside, new, car, parking, lot, of, the, agreement, reached]
  - [although, common, shares, rose, forty, six, point, four, hundred, dollars, from, thirty, seconds, at, the, greatest, play, disingenuous, to, be, reset, annually, the, buy, out, of, american, brands, vying, for, mr., womack, currently, sharedata, incorporated, believe, chemical, prices, undoubtedly, will, be, as, much, is, scheduled, to, conscientious, teaching]
  - [this, would, be, a, record, november]
N-gram models = Markov models

- Deterministic FSMs with probabilities
  - broccoli: 0.002
  - fish: 0.1
  - chicken: 0.15
  - in: 0.01
  - for: 0.05
  - at: 0.03
  - for: 0.1

- No long distance dependencies
  - “The future is independent of the past given the present”
- No notion of structure or syntactic dependency
- But lexical
- (And: robust, have frequency information, ...
N-gram models (continued)

- Simplest linear graphical models
- Words are random variables, arrows are direct dependencies between them (CPTs)
N-order Markov Models

- First order Markov assumption = bigram
  \[ P(w_k|w_1 \ldots w_{k-1}) \approx P(w_k|w_{k-1}) = \frac{P(w_{k-1}w_k)}{P(w_{k-1})} \]
- Similarly, n-th order Markov assumption
- Most commonly, trigram (2nd order):
  \[ P(w_k|w_1 \ldots w_{k-1}) \approx P(w_k|w_{k-2}, w_{k-1}) = \frac{P(w_{k-2}w_{k-1}w_k)}{P(w_{k-2}, w_{k-1})} \]
Argument against n-gram models

- Relationships (say between subject and verb) can be arbitrarily distant and convoluted, as linguists love to point out:
  - The *man* that I was watching without pausing to look at what was happening down the street, and quite oblivious to the situation that was about to befall him confidently *strode* into the center of the road.
But n-gram models work in practice!

- That kind of thing doesn’t happen much
- Collins (1997):
  - 74% of dependencies (in the Penn Treebank – WSJ) are with an adjacent word (95% with one ≤ 5 words away), once one treats simple NPs as units:
  - Below, 4/6 = 66% based on words

The post office will hold out discounts
Sparsity

- Problems with n-gram models:
  - New words appear all the time:
    - Synaptitude
    - 132,701.03
    - fuzzificational
  - New bigrams: even more often
  - Trigrams or more – still worse!

- Zipf's Law
  - Types (words) vs. tokens (word occurrences)
  - Broadly: most word types are rare
  - Specifically:
    - Rank word types by token frequency
    - Frequency inversely proportional to rank
  - Not special to language: randomly generated character strings have this property
Elementary Information Theory
Measuring Model Quality

- **Word Error Rate (WER)**

  \[
  \text{Word Error Rate (WER)} = \frac{\text{insertions} + \text{deletions} + \text{substitutions}}{\text{true sentence size}}
  \]

  Correct answer: Andy saw a part of the movie
  Recognizer output: And he saw apart of the movie

  \[
  \text{WER: } \frac{4}{7} = 57\%
  \]

- **The “right” measure:**
  - Task error driven
  - For speech recognition
  - For a specific recognizer!

- **For general evaluation, we want a measure which references only good text, not mistake text**
Measuring Model Quality

- **The Shannon Game:**
  - How well can we predict the next word?
    - When I order pizza, I wipe off the ____
    - Many children are allergic to ____
    - I saw a ____
  - Unigrams are terrible at this game. (Why?)

- **The “Entropy” Measure**
  - Really: average cross-entropy of a text according to a model

  \[
  H(S | M) = \frac{\log_2 P_M(S)}{|S|} = \frac{\sum \log_2 P_M(s_i)}{\sum |s_i|}
  \]

  \[
  \sum \log_2 P_M(w_j | w_{j-1})
  \]
Measuring Model Quality

- Problem with entropy:
  - 0.1 bits of improvement doesn’t sound so good
  - Solution: perplexity

\[ P(S \mid M) = 2^{H(S \mid M)} = \sqrt[n]{\frac{1}{\prod_{i=1}^{n} P_M(w_i \mid h)} } \]

- Note that even though our models require a stop step, we typically don’t count it as a symbol when taking these averages.
Application: LM for Language Identification

• http://www.cs.jhu.edu/~jason/465/PDFSlides/lect01,3tr-ngram-gen.pdf
  – Pages 6-8
Application: Automatic CS Paper Generator

• Generates random Computer Science research papers, including graphs, figures, and citations.
  – Uses a hand-written context-free grammar to form all elements of the papers (word choice random)
  – Aim here is to maximize amusement, rather than coherence.
  – Auto-generate submissions to conferences that you suspect might have very low submission standards.

Readings for next meeting (9/13)

• FSNLP, Chapter 1, 2, & 3:
  http://cognet.mit.edu/library/books/view?isbn=0262133601

• Untangling Text Data Mining, M. Hearst, ACL 1999

• The Text Mining Handbook, Ch. 1: R. Feldman & J. Sanger
  http://books.google.com/books?id=3PcEoz48RBcC

Note: no meeting next Monday (9/6) -- due to Labor Day holiday