Practical Methods for the Analysis of “Big Data”

Module 2: Natural Language Processing

Philip A. Schrodt

The Pennsylvania State University
schrodt@psu.edu

Workshop at the Odum Institute
University of North Carolina, Chapel Hill
20-21 May 2013

Topics: Module 2

Challenges of working with text

Text Sources

- New Social Media

Text Filtering

Complications in text analysis

NLP Suites

The NLP Processing Chain

Automated Coding

Named Entity Recognition/Resolution

Advantages of text as a data source

- Text is one of the primary methods of communicating political information
- The source material is intentional: it was created for some political purpose, either to *persuade, inform,* or *implement*
- Text is unaffected by the act of measurement
- Web-based text can be collected in near-real-time at very little cost
- Machine-assisted coding dramatically decreases any text analysis project, even when it is largely human coded
- A single individual can create an original, customized data set with little or no funding

Challenges of working with text

A bill to prohibit tobacco sales

A bill to prohibit gun sales

A bill to promote safe streets

A bill to promote smoking cessation

Use text similarities and differences to group these bills

Same but not relevant

A bill to prohibit tobacco sales

A bill to prohibit gun sales

A bill to promote safe streets

A bill to promote smoking cessation

Same and probably not relevant

A bill to prohibit tobacco sales

A bill to prohibit gun sales

A bill to promote safe streets

A bill to promote smoking cessation

Probably relevant but not the same

A bill to prohibit tobacco sales

A bill to prohibit gun sales

A bill to promote safe streets

A bill to promote smoking cessation

Resources for working with text

Library science	Automated indexing
Computational Linguistics	Automated translation and natural language processing generally
Psychology	Personality tests
Communications Studies	Content of popular culture —books, movie and television scripts
Education	Automated grading
Business	Automated evaluation of resumés, aptitude tests

Topics: Module 2

Challenges of working with text

Text Sources

- New Social Media

Text Filtering

Complications in text analysis

NLP Suites

The NLP Processing Chain

Automated Coding

Named Entity Recognition/Resolution

What about new social media?

Mainstream media

What about new social media?

Mainstream media

Internet and
new social media

New social media

The good

- Widely available to elites
- More or less uncensored in open many societies
- Should provide early information on changing sentiment prior to observing actual collective action

The bad

- No filters and mostly politically irrelevant:
- “Wanna getta pizza? ;)”
- Easily manipulated by anyone—business, government, NGOs—who wants to go to the trouble of doing so

The ugly

- No standardization of content

Utility in prediction

- Multiple studies show this seems to work in the 6 to 48 hour range

New Social Media

Police tell high school students to disguise their identity on FaceBook

+

Students choose the first country they recognize in an alphabetical list.... Afghanistan

---

OMG! Jihadis are checking out the junior prom!

Your turn: Just because NSM aren't useful for *me*....

Okay, let's come up with some better examples: this stuff is extremely interesting to a lot of people and is a quintessential example of "Big Data".

What can we do with it???

Topics: Module 2

Challenges of working with text

Text Sources

- New Social Media

Text Filtering

Complications in text analysis

NLP Suites

The NLP Processing Chain

Automated Coding

Named Entity Recognition/Resolution

Lexis-Nexis and Factiva

- No one has figured out how to automate the downloads of these
  - Instead, get your search string as precise as possible to reduce unnecessary downloads, then slog away...
- The LN search engine is very unpredictable: it is not designed for this sort of thing. Money (as in "DARPA") was insufficient to solve the problem
  - Note that this contradicts the advice immediately above about refining the search string
  - Supposedly LN is okay with bulk downloads so long as you do this sufficiently painfully
- Factiva is far more reliable but somewhat more awkward to use
  - Money could solve that problem, but more money than you probably have.
  - Factiva is apparently getting exceedingly upset with some institutions about bulk downloads
- All of this suggests a serious market failure vis a vis "data mining" that may be addressed at some point in the relatively near future/
  - ProQuest *may* be getting into this business

News aggregators and real-time news sources

- Google News
  - Watch this space...it would not be surprising to see them do something dramatic
- European Media Monitor
  - Extensive development until about five years ago but does not seem to have done anything recently
  - Very extensive NER systems

Your turn: Other Opportunities?

What else is out there on the web—remember, if you can read it, generally you can download it—that might be useful?

Any experience/stories where someone was able to make good use of web-based material? Where what seemed like a good idea didn't work out?

Are there opportunities to take advantage of **velocity** to do research that we could not do in the past?

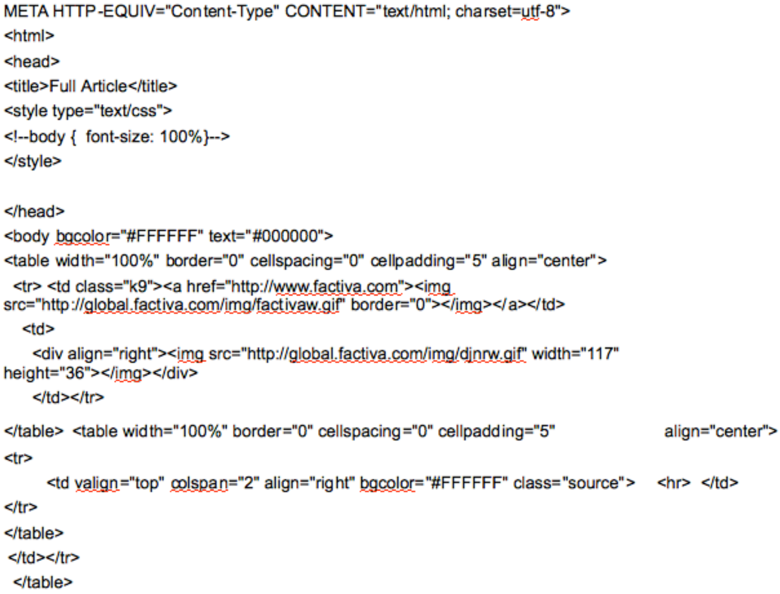
Can we identify can *trends* in the use of the conventional web that may open opportunities in the near future?

Formatting of source texts

How much work will be involved in getting the text into a form you can use?

- ASCII/UniCode text, for example news reports
- HTML, but web formats change frequently
  - Python has modules for removing HTML tags: don't try doing this just with regular expressions
- PDF
  - Try the tm module in R
- Scanned/OCR text
- Proprietary word processing formats (Word)
- New media sources such as blogs and tweets

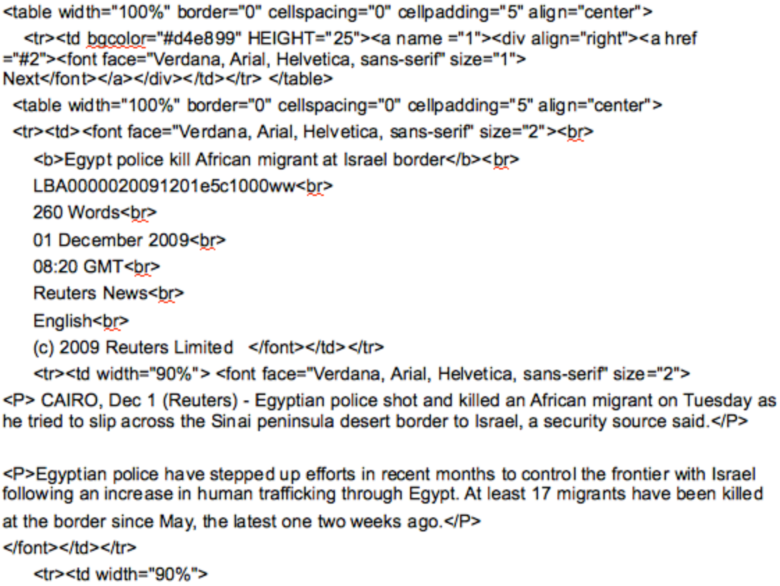
Text in HTML



```
META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=utf-8">
<html>
<head>
<title>Full Article</title>
<style type="text/css">
<!--body {  font-size: 100%}-->
</style>

</head>
<body bgcolor="#FFFFFF" text="#000000">
<table width="100%" border="0" cellspacing="0" cellpadding="5" align="center">
  <tr> <td class="k9"><a href="http://www.factiva.com"><img
src="http://global.factiva.com/img/factivaw.gif" border="0"></img></a></td>
    <td>
      <div align="right"><img src="http://global.factiva.com/img/djnrw.gif" width="117"
height="36"></img></div>
      </td></tr>
</table>  <table width="100%" border="0" cellspacing="0" cellpadding="5"                    align="center">
<tr>
         <td valign="top" colspan="2" align="right" bgcolor="#FFFFFF" class="source">    <hr>  </td>
</tr>
</table>
 </td></tr>
  </table>
```

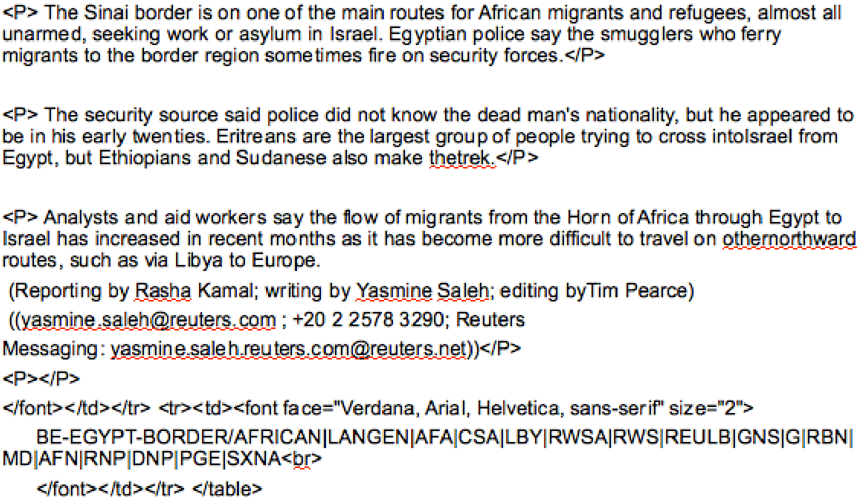
Text in HTML



```
<table width="100%" border="0" cellspacing="0" cellpadding="5" align="center">
   <tr><td bgcolor="#d4e899" HEIGHT="25"><a name ="1"><div align="right"><a href
="#2"><font face="Verdana, Arial, Helvetica, sans-serif" size="1">
Next</font></a></div></td></tr> </table>
 <table width="100%" border="0" cellspacing="0" cellpadding="5" align="center">
 <tr><td><font face="Verdana, Arial, Helvetica, sans-serif" size="2"><br>
    <b>Egypt police kill African migrant at Israel border</b><br>
    LBA0000020091201e5c1000ww<br>
    260 Words<br>
    01 December 2009<br>
    08:20 GMT<br>
    Reuters News<br>
    English<br>
    (c) 2009 Reuters Limited   </font></td></tr>
    <tr><td width="90%"> <font face="Verdana, Arial, Helvetica, sans-serif" size="2">
<P> CAIRO, Dec 1 (Reuters) - Egyptian police shot and killed an African migrant on Tuesday as
he tried to slip across the Sinai peninsula desert border to Israel, a security source said.</P>

<P>Egyptian police have stepped up efforts in recent months to control the frontier with Israel
following an increase in human trafficking through Egypt. At least 17 migrants have been killed
at the border since May, the latest one two weeks ago.</P>
</font></td></tr>
    <tr><td width="90%">
```

Text in HTML



```
<P> The Sinai border is on one of the main routes for African migrants and refugees, almost all
unarmed, seeking work or asylum in Israel. Egyptian police say the smugglers who ferry
migrants to the border region sometimes fire on security forces.</P>

<P> The security source said police did not know the dead man's nationality, but he appeared to
be in his early twenties. Eritreans are the largest group of people trying to cross intoIsrael from
Egypt, but Ethiopians and Sudanese also make thetrek.</P>

<P> Analysts and aid workers say the flow of migrants from the Horn of Africa through Egypt to
Israel has increased in recent months as it has become more difficult to travel on othernorthward
routes, such as via Libya to Europe.
 (Reporting by Rasha Kamal; writing by Yasmine Saleh; editing byTim Pearce)
 ((yasmine.saleh@reuters.com ; +20 2 2578 3290; Reuters
Messaging: yasmine.saleh.reuters.com@reuters.net))</P>
<P></P>
</font></td></tr> <tr><td><font face="Verdana, Arial, Helvetica, sans-serif" size="2">
    BE-EGYPT-BORDER/AFRICAN|LANGEN|AFA|CSA|LBY|RWSA|RWS|REULB|GNS|G|RBN|
MD|AFN|RNP|DNP|PGE|SXNA<br>
    </font></td></tr> </table>
```

Topics: Module 2

- Challenges of working with text
- Text Sources
  - New Social Media
- Text Filtering
- Complications in text analysis
- NLP Suites
- The NLP Processing Chain
- Automated Coding
- Named Entity Recognition/Resolution

Levels of analysis in text processing

Analytical Term	Linguistic Term	Methodology
Thematic	Lexical	Analysis of words and phrases. “bag of words”
Syntactic	Syntactic	Use grammatical rules to determine role of words
Network	Semantic	Use relationships between words to disambiguate meanings

Word Frequency in English

% of usage	# of words
40%	50
60%	2,300
85%	8,000
99%	16,000

Total words in American English: about 600,000
Total words in technical English (all fields): about 3-million

Functional Words

Very short words such as

- Articles: a an the
- Interogatives: who what when where why how
- Prepositions: to from at in above below
- Auxillary verbs: have has was were been
- Markers: by in at to (French de, German du, Arabic fi)
- Pronouns: I you he she him her his hers

In English, the specificity of a word is generally proportional to its length. These short will typically be in the stop word list, though a few longer words (e.g. "though" and "although") also will be stop words

Marker words have multiple uses: Random House College Dictionary lists 29 meanings for "by," 31 for "in," 25 for "to," and 15 for "for."

Disambiguation: “Bat”

Noun

- wooden (or aluminum) cylinder used in the game of baseball
- small flying mammal

Verb

- act of batting (“at bat”)
- blinking (“bat an eye”)

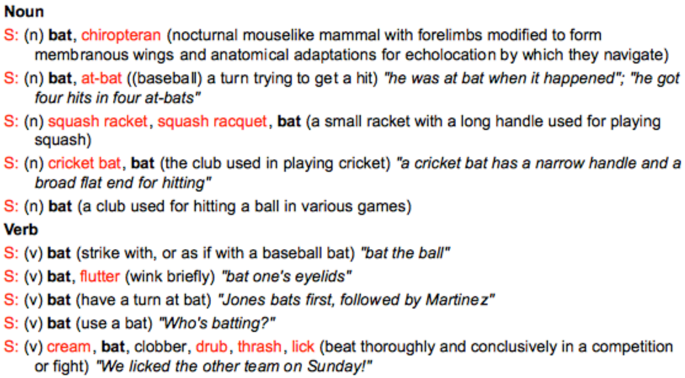
Idiomatic uses

- “go to bat for”: defending or interceding;
- “right off the bat”: immediately;
- “bats in the belfry”: commentary on an individual’s cognitive ability

Foreign phrases

- “bat mitzvah”: a girl’s coming-of-age ceremony (Hebrew).

WordNet word senses: “bat”



Noun

S: (n) **bat**, chiropteran (nocturnal mouselike mammal with forelimbs modified to form membranous wings and anatomical adaptations for echolocation by which they navigate)

S: (n) **bat**, at-bat ((baseball) a turn trying to get a hit) *"he was at bat when it happened"; "he got four hits in four at-bats"*

S: (n) squash racket, squash racquet, **bat** (a small racket with a long handle used for playing squash)

S: (n) cricket bat, **bat** (the club used in playing cricket) *"a cricket bat has a narrow handle and a broad flat end for hitting"*

S: (n) **bat** (a club used for hitting a ball in various games)

Verb

S: (v) **bat** (strike with, or as if with a baseball bat) *"bat the ball"*

S: (v) **bat**, flutter (wink briefly) *"bat one's eyelids"*

S: (v) **bat** (have a turn at bat) *"Jones bats first, followed by Martinez"*

S: (v) **bat** (use a bat) *"Who's batting?"*

S: (v) cream, **bat**, clobber, drub, thrash, lick (beat thoroughly and conclusively in a competition or fight) *"We licked the other team on Sunday!"*

Disambiguation

Any of these uses might be encountered in an English-language text. Multiple uses might be found in a single sentence:

> *"The umpire didn't bat an eye as Sarah lowered her bat to watch the bat flying around the pitcher."*

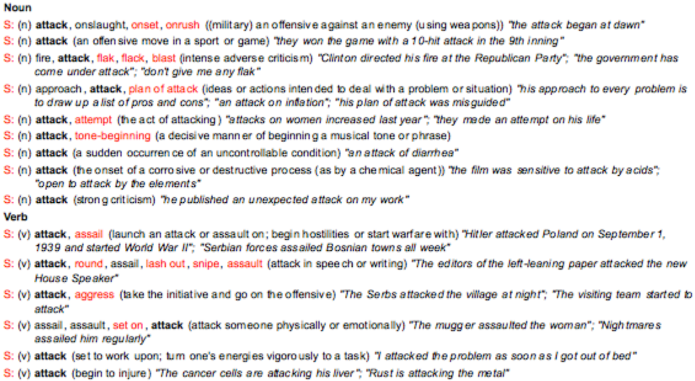
Words can also change from verbs to nouns without modification. Consider

- I plan to drive to the store, then wash the car.
- When John returned from the car wash, he parked his car in the drive.

In summary

> *"Verbing weirds language"*
> *Bill Watterson, Calvin and Hobbes*

WordNet word senses: “attack”



Noun

S: (n) **attack**, onslaught, onset, onrush ((military) an offensive against an enemy (using weapons)) *"the attack began at dawn"*

S: (n) **attack** (an offensive move in a sport or game) *"they won the game with a 10-hit attack in the 9th inning"*

S: (n) fire, **attack**, flak, flack, blast (intense adverse criticism) *"Clinton directed his fire at the Republican Party"; "the government has come under attack"; "don't give me any flak"*

S: (n) approach, **attack**, plan of attack (ideas or actions intended to deal with a problem or situation) *"his approach to every problem is to draw up a list of pros and cons"; "an attack on inflation"; "his plan of attack was misguided"*

S: (n) **attack**, attempt (the act of attacking) *"attacks on women increased last year"; "they made an attempt on his life"*

S: (n) **attack**, tone-beginning (a decisive manner of beginning a musical tone or phrase)

S: (n) **attack** (a sudden occurrence of an uncontrollable condition) *"an attack of diarrhea"*

S: (n) **attack** (the onset of a corrosive or destructive process (as by a chemical agent)) *"the film was sensitive to attack by acids"; "open to attack by the elements"*

S: (n) **attack** (strong criticism) *"he published an unexpected attack on my work"*

Verb

S: (v) **attack**, assail (launch an attack or assault on; begin hostilities or start warfare with) *"Hitler attacked Poland on September 1, 1939 and started World War II"; "Serbian forces assailed Bosnian towns all week"*

S: (v) **attack**, round, assail, lash out, snipe, assault (attack in speech or writing) *"The editors of the left-leaning paper attacked the new House Speaker"*

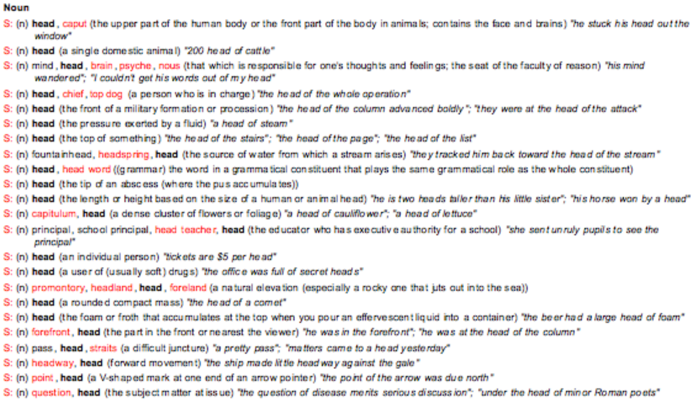
S: (v) **attack**, aggress (take the initiative and go on the offensive) *"The Serbs attacked the village at night"; "The visiting team started to attack"*

S: (v) assail, assault, set on, **attack** (attack someone physically or emotionally) *"The mugger assaulted the woman"; "Nightmares assailed him regularly"*

S: (v) **attack** (set to work upon; turn one's energies vigorously to a task) *"I attacked the problem as soon as I got out of bed"*

S: (v) **attack** (begin to injure) *"The cancer cells are attacking his liver"; "Rust is attacking the metal"*

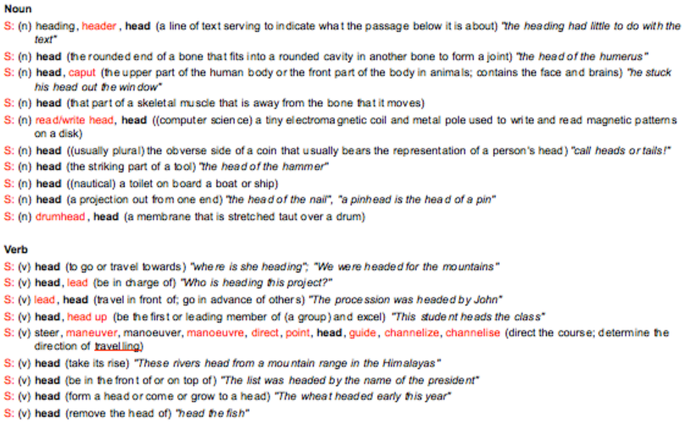
WordNet word senses: "head"



Noun

- S: (n) **head**, caput (the upper part of the human body or the front part of the body in animals; contains the face and brains) *"he stuck his head out the window"*
- S: (n) **head** (a single domestic animal) *"200 head of cattle"*
- S: (n) mind, **head**, brain, psyche, nous (that which is responsible for one's thoughts and feelings; the seat of the faculty of reason) *"his mind wandered"; "I couldn't get his words out of my head"*
- S: (n) **head**, chief, top dog (a person who is in charge) *"the head of the whole operation"*
- S: (n) **head** (the front of a military formation or procession) *"the head of the column advanced boldly"; "they were at the head of the attack"*
- S: (n) **head** (the pressure exerted by a fluid) *"a head of steam"*
- S: (n) **head** (the top of something) *"the head of the stairs"; "the head of the page"; "the head of the list"*
- S: (n) fountainhead, headspring, **head** (the source of water from which a stream arises) *"they tracked him back toward the head of the stream"*
- S: (n) **head**, head word ((grammar) the word in a grammatical constituent that plays the same grammatical role as the whole constituent)
- S: (n) **head** (the tip of an abscess (where the pus accumulates))
- S: (n) **head** (the length or height based on the size of a human or animal head) *"he is two heads taller than his little sister"; "his horse won by a head"*
- S: (n) capitulum, **head** (a dense cluster of flowers or foliage) *"a head of cauliflower"; "a head of lettuce"*
- S: (n) principal, school principal, head teacher, **head** (the educator who has executive authority for a school) *"she sent unruly pupils to see the principal"*
- S: (n) **head** (an individual person) *"tickets are $5 per head"*
- S: (n) **head** (a user of (usually soft) drugs) *"the office was full of secret heads"*
- S: (n) promontory, headland, **head**, foreland (a natural elevation (especially a rocky one that juts out into the sea))
- S: (n) **head** (a rounded compact mass) *"the head of a comet"*
- S: (n) **head** (the foam or froth that accumulates at the top when you pour an effervescent liquid into a container) *"the beer had a large head of foam"*
- S: (n) forefront, **head** (the part in the front or nearest the viewer) *"he was in the forefront"; "he was at the head of the column"*
- S: (n) pass, **head**, straits (a difficult juncture) *"a pretty pass"; "matters came to a head yesterday"*
- S: (n) headway, **head** (forward movement) *"the ship made little headway against the gale"*
- S: (n) point, **head** (a V-shaped mark at one end of an arrow pointer) *"the point of the arrow was due north"*
- S: (n) question, **head** (the subject matter at issue) *"the question of disease merits serious discussion"; "under the head of minor Roman poets"*

WordNet word senses: "head" continued



Noun

- S: (n) heading, header, **head** (a line of text serving to indicate what the passage below it is about) *"the heading had little to do with the text"*
- S: (n) **head** (the rounded end of a bone that fits into a rounded cavity in another bone to form a joint) *"the head of the humerus"*
- S: (n) **head**, caput (the upper part of the human body or the front part of the body in animals; contains the face and brains) *"he stuck his head out the window"*
- S: (n) **head** (that part of a skeletal muscle that is away from the bone that it moves)
- S: (n) read/write head, **head** ((computer science) a tiny electromagnetic coil and metal pole used to write and read magnetic patterns on a disk)
- S: (n) **head** ((usually plural) the obverse side of a coin that usually bears the representation of a person's head) *"call heads or tails!"*
- S: (n) **head** (the striking part of a tool) *"the head of the hammer"*
- S: (n) **head** ((nautical) a toilet on board a boat or ship)
- S: (n) **head** (a projection out from one end) *"the head of the nail", "a pinhead is the head of a pin"*
- S: (n) drumhead, **head** (a membrane that is stretched taut over a drum)

Verb

- S: (v) **head** (to go or travel towards) *"where is she heading"; "We were headed for the mountains"*
- S: (v) **head**, lead (be in charge of) *"Who is heading this project?"*
- S: (v) lead, **head** (travel in front of; go in advance of others) *"The procession was headed by John"*
- S: (v) **head**, head up (be the first or leading member of (a group) and excel) *"This student heads the class"*
- S: (v) steer, maneuver, manoeuver, manoeuvre, direct, point, **head**, guide, channelize, channelise (direct the course; determine the direction of travelling)
- S: (v) **head** (take its rise) *"These rivers head from a mountain range in the Himalayas"*
- S: (v) **head** (be in the front of or on top of) *"The list was headed by the name of the president"*
- S: (v) **head** (form a head or come or grow to a head) *"The wheat headed early this year"*
- S: (v) **head** (remove the head of) *"head the fish"*

Memes, idioms, metaphors and slang

Political text frequently uses distinct idiomatic phrases

“Right to life”, *“right to choice”*

Memes can have a high frequency for brief periods of time

“lipstick on a pig”
“top kill”, *“junk shot”*, *“Deep Horizon”*

Military metaphors are common in political (and sports) rhetoric

“Tea Party insurgency”, *“battleground state”*

OMG! WTF! Like IMHO slang expressions are common, and rapidly changing, in new media (lol…)

Style of language

- News reports and official documents are usually formal, syntactically-correct English
- Quotations and letters are a mix of formal and informal
- Open-ended responses range from formal to very fragmentary
- New media sources are often very informal and abbreviated
- Variants of English and changes in usage over time (e.g. slang, memes)
- Languages other than English: machine translation will work for some applications

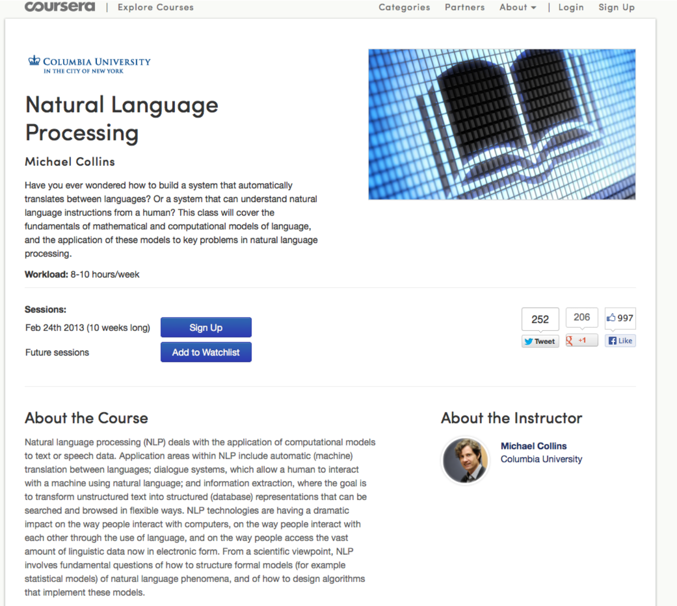
Topics: Module 2

- Challenges of working with text
- Text Sources
  - New Social Media
- Text Filtering
- Complications in text analysis
- NLP Suites
- The NLP Processing Chain
- Automated Coding
- Named Entity Recognition/Resolution

Major Open Source NLP Projects

- Open-NLP http://opennlp.apache.org/ [more of a list of other project now]
- GATE; http://gate.ac.uk/
- University of Illinois Cognitive Computation Group: http://cogcomp.cs.illinois.edu/page/software
- Stanford NLP Group: http://nlp.stanford.edu/software/index.shtml

LingPipe's "Competition" page (http://alias-i.com/lingpipe/web/competition.html) lists—as of March 2012—no fewer than 23 academic/open-source NLP projects, and 122 commercial projects.



Natural Language Processing

Michael Collins

Have you ever wondered how to build a system that automatically translates between languages? Or a system that can understand natural language instructions from a human? This class will cover the fundamentals of mathematical and computational models of language, and the application of these models to key problems in natural language processing.

Workload: 8-10 hours/week

Sessions:

Feb 24th 2013 (10 weeks long)

Future sessions

About the Course

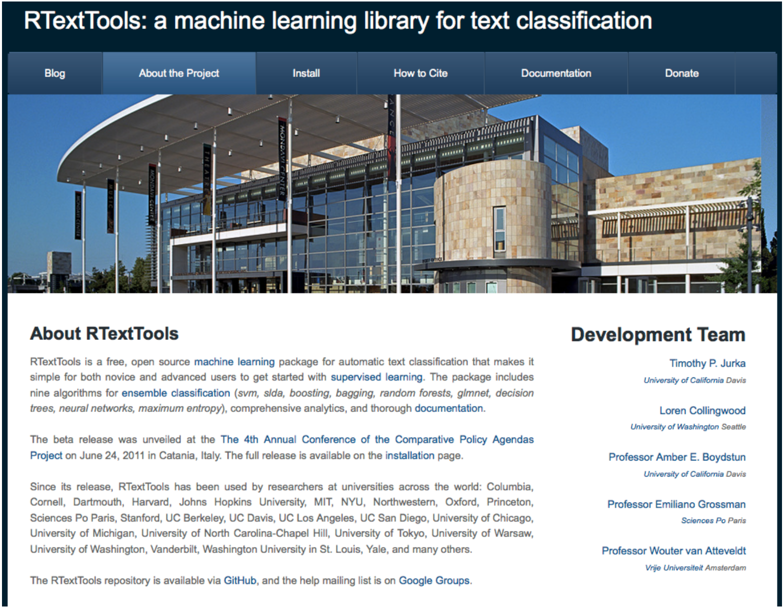
Natural language processing (NLP) deals with the application of computational models to text or speech data. Application areas within NLP include automatic (machine) translation between languages; dialogue systems, which allow a human to interact with a machine using natural language; and information extraction, where the goal is to transform unstructured text into structured (database) representations that can be searched and browsed in flexible ways. NLP technologies are having a dramatic impact on the way people interact with computers, on the way people interact with each other through the use of language, and on the way people access the vast amount of linguistic data now in electronic form. From a scientific viewpoint, NLP involves fundamental questions of how to structure formal models (for example statistical models) of natural language phenomena, and of how to design algorithms that implement these models.

About the Instructor

Michael Collins
Columbia University

tm: R text mining package

- Conversion to text from a number of formats, including XML, Word and pdf
- Filtering documents by query
- Stopword removal, stemming, tokenizing, whitespace removal
- Computation of term-document matrices, with optional dictionaries
- Assorted term frequency operations
- Dissimilarity metrics



RTextTools: a machine learning library for text classification

About RTextTools

RTextTools is a free, open source machine learning package for automatic text classification that makes it simple for both novice and advanced users to get started with supervised learning. The package includes nine algorithms for ensemble classification (*svm, slda, boosting, bagging, random forests, glmnet, decision trees, neural networks, maximum entropy*), comprehensive analytics, and thorough documentation.

The beta release was unveiled at the The 4th Annual Conference of the Comparative Policy Agendas Project on June 24, 2011 in Catania, Italy. The full release is available on the installation page.

Since its release, RTextTools has been used by researchers at universities across the world: Columbia, Cornell, Dartmouth, Harvard, Johns Hopkins University, MIT, NYU, Northwestern, Oxford, Princeton, Sciences Po Paris, Stanford, UC Berkeley, UC Davis, UC Los Angeles, UC San Diego, University of Chicago, University of Michigan, University of North Carolina-Chapel Hill, University of Tokyo, University of Warsaw, University of Washington, Vanderbilt, Washington University in St. Louis, Yale, and many others.

The RTextTools repository is available via GitHub, and the help mailing list is on Google Groups.

Development Team

Timothy P. Jurka
University of California Davis

Loren Collingwood
University of Washington Seattle

Professor Amber E. Boydstun
University of California Davis

Professor Emiliano Grossman
Sciences Po Paris

Professor Wouter van Atteveldt
Vrije Universiteit Amsterdam

Source: http://www.rtexttools.com/about-the-project.html

Duplicate detection: new stories

- Use of newswires
  - This is the reason for newswires
  - Detecting exact duplicates is easy
- Updating and corrections of previous stories
- News summaries
- Chronologies
- Multiple independent sources (Reuters, AFP, BBC)
- Except for exact duplicates, this is a difficult problem
  - "Near duplicate detection"
  - One-a-Day filter by event-day or location-day (GDELT)
- Duplication probably amplifies the signal you want: a story is more likely to be repeated if it is important

Topics: Module 2

Challenges of working with text

Text Sources

- New Social Media

Text Filtering

Complications in text analysis

NLP Suites

The NLP Processing Chain

Automated Coding

Named Entity Recognition/Resolution

Example

For purposes of illustration, consider the following initial sentences for a news story:

> *US Supreme Court Justice Stephen Breyer was robbed by a machete-wielding man at his Caribbean vacation home, a Supreme Court spokeswoman said.*
>
> *The robber broke into Judge Breyer's home on the island of Nevis around 21:00 EST (02:00 GMT) on Thursday.*
>
> *The Supreme Court justice was at home with his wife and guests, but no one was hurt, the spokeswoman said.*

Sentence delineation

Sentence delineation is a surprisingly difficult task due to

- abbreviations ending in periods
  - And note that the sentence “She lives on Main St.” ends in one period, not two.
- punctuation occurring inside sentences
- character strings that are not actually part of the sentence, particularly across multiple story formats.

In some languages, notably Chinese (or Latin before the Carolingian Renaissance), *word* boundaries can also be an issue

Stopwords

From tm.pdf
SMART stopwords from the SMART information retrieval system (obtained from http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop) (which coincides with the stopword list used by theMCtoolkit (http://www.cs.utexas.edu/users/ dml/software/mc/))

parts-of-speech markup

Parts-of-speech (POS) marking—or in the example below, a system that makes noun-verb distinctions and also classifies these into general categories.

```
US/noun.group Supreme/noun.group Court/noun.group Justice/noun.group
Stephen/noun.person Breyer/noun.person was robbed/verb.possession by
a machete-wielding man/noun.person at his/pronoun Caribbean
vacation/noun.artifact home/noun.artifact , a Supreme/noun.group
Court/noun.group spokeswoman/noun.person said/verb.communication .

The robber/noun.person broke/verb.communication into/verb.communicatio
Judge/noun.person Breyer/noun.person ’s home/noun.location on the
island/noun.object of Nevis/noun.location around 21:00 EST/noun.time o
Thursday/noun.time.
```

One of the major tasks of the TABARI dictionaries is noun-verb disambiguation: this issue accounts for much of their size.

Stemming

Many NLP systems use stemming—most frequently the Porter stemming algorithm for English (`http://tartarus.org/martin/PorterStemmer/`). This should both simplify and generalize the dictionaries.
An alternative approach, used in TABARI, is to automatically recognizing the regular forms of nouns and verbs.

Full parsing.

An assortment of full-parsers—as distinct from the shallow parser used in TABARI —are available, and the *TreeBank* parse format appears to be a fairly stable and standard output format. This allows a researcher could use the parser of his or her choice (notably some parser developed in the future) so long as these could produce *TreeBank*-formatted output.

```
(ROOT (S (S (NP (NNP US) (NNP Supreme) (NNP Court) (NNP Justice)
(NNP Stephen) (NNP Breyer)) (VP (VBD was) (VP (VBN robbed) (PP (IN by)
(NP (NP (DT a) (JJ machete-wielding) (NN man)) (PP (IN at) (NP (PRP$ h
(JJ Caribbean) (NN vacation) (NN home)))))))) (, ,) (NP (DT a)
(NNP Supreme) (NNP Court) (NN spokeswoman)) (VP (VBD said)) (. .)))
```

In event data coding most important contribution of the full parsing is insuring that the words identified as belonging to a verb phrase are in fact associated with that verb, and not with a subordinate clause or some other part of the sentence.

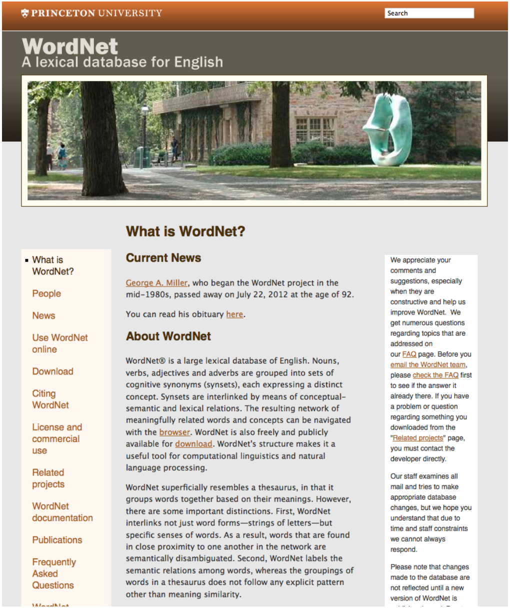
Pronoun and entity coreferencing

Some of the full-parsing systems provide pronoun and entity (“ coreferencing Alternatively, this can be provided in stand-alone coreferencing systems such as the ARK noun phrase coreferencer. (`http://www.ark.cs.cmu.edu/ARKref/`). Also Google “Hobbs anaphora resolution”: this is a rule-based algorithm which apparently works about 90% of the time for news texts.

```
<ref id="1" ent="1_4_8">US Supreme Court Justice Stephen Breyer</ref>
robbed by <ref id="2" ent="2">a machete-wielding man at
<ref id="3" ent="1_4_8">his</ref> <ref id="4" ent="3_7_46">Caribbean v
home</ref> , <ref id="5" ent="5_21">a Supreme Court spokeswoman</ref>

<ref id="6" ent="6_19">The robber</ref> broke into <ref id="8" ent="1_
Judge Breyer ’s</ref> <ref id="7" ent="3_7_46">home</ref> on
<ref id="9" ent="9">the island of Nevis</ref> around 21:00 EST on
<ref id="13" ent="13">Thursday</ref>.

<ref id="17" ent="1_4_8">The Supreme Court justice</ref> was at home w
<ref id="19" ent="1_4_8">his</ref> wife and guests, but <ref id="20" e
no one</ref> was hurt , <ref id="21" ent="5_21">the spokeswoman</ref>
```



WordNet
A lexical database for English

- What is WordNet?
- People
- News
- Use WordNet online
- Download
- Citing WordNet
- License and commercial use
- Related projects
- WordNet documentation
- Publications
- Frequently Asked Questions

What is WordNet?

Current News

George A. Miller, who began the WordNet project in the mid-1980s, passed away on July 22, 2012 at the age of 92.

You can read his obituary here.

About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

We appreciate your comments and suggestions, especially when they are constructive and help us improve WordNet. We get numerous questions regarding topics that are addressed on our FAQ page. Before you email the WordNet team, please check the FAQ first to see if the answer it already there. If you have a problem or question regarding something you downloaded from the "Related projects" page, you must contact the developer directly.

Our staff examines all mail and tries to make appropriate database changes, but we hope you understand that due to time and staff constraints we cannot always respond.

Please note that changes made to the database are not reflected until a new version of WordNet is

Source: http://wordnet.princeton.edu/

Topics: Module 2

Challenges of working with text

Text Sources

- New Social Media

Text Filtering

Complications in text analysis

NLP Suites

The NLP Processing Chain

Automated Coding

Named Entity Recognition/Resolution

Modes of reliability in text processing

- **Stability**—the ability of a coder to consistently assign the same code to a given text;
- **Reproducibility**—intercoder reliability;
- **Accuracy**—the ability of a group of coders to conform to a standard.

Source: Weber (1990:17)

In principle, it would be useful to know reproducibility

- Between coders at different phases of the project
- Between coders at multiple institutions if the project is decentralized

Textual Analysis By Augmented Replacement Instructions (TABARI)

- ANSI C++, approximately 14,000 lines of code
- Open-source (GPL)
- Unix, Linux and OS-X operating systems (gcc compiler)
- “Teletype” interface: text and keyboard
  - Easily deployed on a server
- Codes around 5,000 events per second on contemporary hardware
  - Speed is achieved through use of shallow parsing algorithms
  - Speed can be scaled indefinitely using parallel processing
- Standard dictionaries are open source, with around 15,000 verb phrases for events and 30,000+ noun phrases for actors
- Coded the 200-million event GDELT dataset without crashing

Advantages of automated coding

- Fast and inexpensive
- Transparent: coding rules are explicit in the dictionaries
- Reproducible: a coding system can be consistently maintained over a period of time without the "coding drift" caused by changing teams of coders.
- Coding dictionaries can be shared between institutions
- The coding of individual reports is not affected by the biases of individual coders. Dictionaries, however, can be so affected.
- It is possible to create rules for difficult technical and cultural vocabulary that is otherwise difficult to learn

Disadvantages of automated coding

- Automated thematic coding has problems with disambiguation
- Automated syntactic coding using shallow parsing makes errors on complex sentences by incorrectly identifying the object of the sentence.
- Requires a properly formatted, machine-readable source of text, therefore older paper and microfilm sources are difficult to code.
- Development of new coding dictionaries is time-consuming—KEDS/PANDA initial dictionary development required 2-labor-years. (Modification of existing dictionaries, however, requires far less effort)

Human and machine coding tradeoffs

- Machine coding uses only information that is explicit in the text; human coders are likely to use implicit knowledge of the situation.
- Machine coding is not affected by boredom and fatigue
- Human coders can more effectively interpret idiomatic and metaphorical text, provided they are familiar with the context
- Human coders can more effectively deal with complex subordinate phrases and other unexpected grammatical constructions

However...

- In very large textual databases that hve to be coded in near real time it is all irrelevant because automated coding is the only option

Human vs Machine Coding: Summary

Advantage to human coding	Advantage to machine coding
Small data sets	Large data sets
Data coded only one time at a single site	Data coded over a period of time or across projects
Existing dictionaries cannot be modified	Existing dictionaries can be modified
Complex sentence structure	Simple sentence structures
Metaphorical, idiomatic, or time- dependent text	Literal, present-tense text
Money available to fund coders and supervisors	Money is limited

Your turn: Opportunities—and not—for automated coding

Based on both the availability of source materials, and the requirements of the coding project, are there opportunities for automated coding

With a realistic assessment of the strengths and foibles of human coders, what are the sorts of coding (if any) where machines are likely to be better than humans?
Are there opportunities to take advantage of volume to compensate for lower accuracy?

What additional characteristics of automated coding would be needed to code things that currently cannot be coded by machine? Are there opportunities for machine assisted coding?
Anyone have stories/experience where machine-assisted coding has yielded significant increases in efficiency? Where it has ended up being more trouble than it is worth?

Insert “TABARI.coding.pdf” here

TABARI.coding.pdf

Improving JABARI Accuracy

- TABARI baseline: 56% precision, 54% recall
- Add Open-NLP Penn TreeBank parser: 68% precision, 35.4% recall
- Add GATE-Annie noun phrase synonyms, pronoun coreferencing, and default location agent resolution: 77% precision, 66.5% recall

Named Entity Recognition/Resolution

- Locating and classifying phrases into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.
- Examples: http://www.ldc.upenn.edu/Catalog/docs/LDC2005T33/BBN-Types-Subtypes.html.
- No general solution; approaches tend to be either
  - Rule and dictionary based, which requires manual development
  - Sequence-based machine-learning methods, specifically conditional random fields. These require an extensive set of marked-up examples
- *Name resolution* involves either
  - Differentiating two distinct entities which have the same name: "President Bush"
  - Combining multiple names of the same entity" "Obamacare" and "Affordable Care Act"
- Network models which associate a particular use of the name with other entities and/or time are frequently useful here.

Named Entity Recognition/Resolution

Most research on NER systems has been structured as taking an unannotated block of text, such as this one:

Jim bought 300 shares of Acme Corp. in 2006.

And producing an annotated block of text, such as this one:

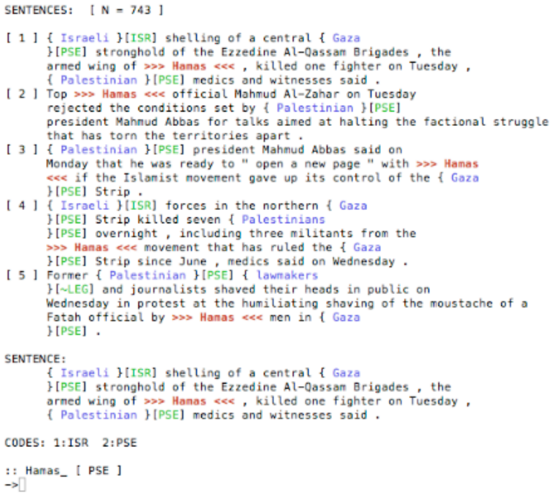
```
<ENAMEX TYPE="PERSON">Jim</ENAMEX>
  bought<NUMEX TYPE="QUANTITY">300</NUMEX>
   shares of
   <ENAMEX TYPE="ORGANIZATION">Acme Corp.</ENAMEX
    in <TIMEX TYPE="DATE">2006</TIMEX>.
```

In this example, the annotations have been done using so-called ENAMEX tags that were developed for the Message Understanding Conference in the 1990s.

State-of-the-art NER systems for English produce near-human performance. For example, the best system entering MUC-7 scored 93.39% of F-measure while human annotators scored 97.60% and 96.95%.

Source: http://en.wikipedia.org/wiki/Named_entity_recogniti

CodeCatcher



```
SENTENCES:  [ N = 743 ]

[ 1 ] { Israeli }[ISR] shelling of a central { Gaza
      }[PSE] stronghold of the Ezzedine Al-Qassam Brigades , the
      armed wing of >>> Hamas <<< , killed one fighter on Tuesday ,
      { Palestinian }[PSE] medics and witnesses said .
[ 2 ] Top >>> Hamas <<< official Mahmud Al-Zahar on Tuesday
      rejected the conditions set by { Palestinian }[PSE]
      president Mahmud Abbas for talks aimed at halting the factional struggle
      that has torn the territories apart .
[ 3 ] { Palestinian }[PSE] president Mahmud Abbas said on
      Monday that he was ready to " open a new page " with >>> Hamas
      <<< if the Islamist movement gave up its control of the { Gaza
      }[PSE] Strip .
[ 4 ] { Israeli }[ISR] forces in the northern { Gaza
      }[PSE] Strip killed seven { Palestinians
      }[PSE] overnight , including three militants from the
      >>> Hamas <<< movement that has ruled the { Gaza
      }[PSE] Strip since June , medics said on Wednesday .
[ 5 ] Former { Palestinian }[PSE] { lawmakers
      }[~LEG] and journalists shaved their heads in public on
      Wednesday in protest at the humiliating shaving of the moustache of a
      Fatah official by >>> Hamas <<< men in { Gaza
      }[PSE] .

SENTENCE:
      { Israeli }[ISR] shelling of a central { Gaza
      }[PSE] stronghold of the Ezzedine Al-Qassam Brigades , the
      armed wing of >>> Hamas <<< , killed one fighter on Tuesday ,
      { Palestinian }[PSE] medics and witnesses said .

CODES: 1:ISR  2:PSE

:: Hamas_ [ PSE ]
->
```