Text Mining & NLP

IST 407/707

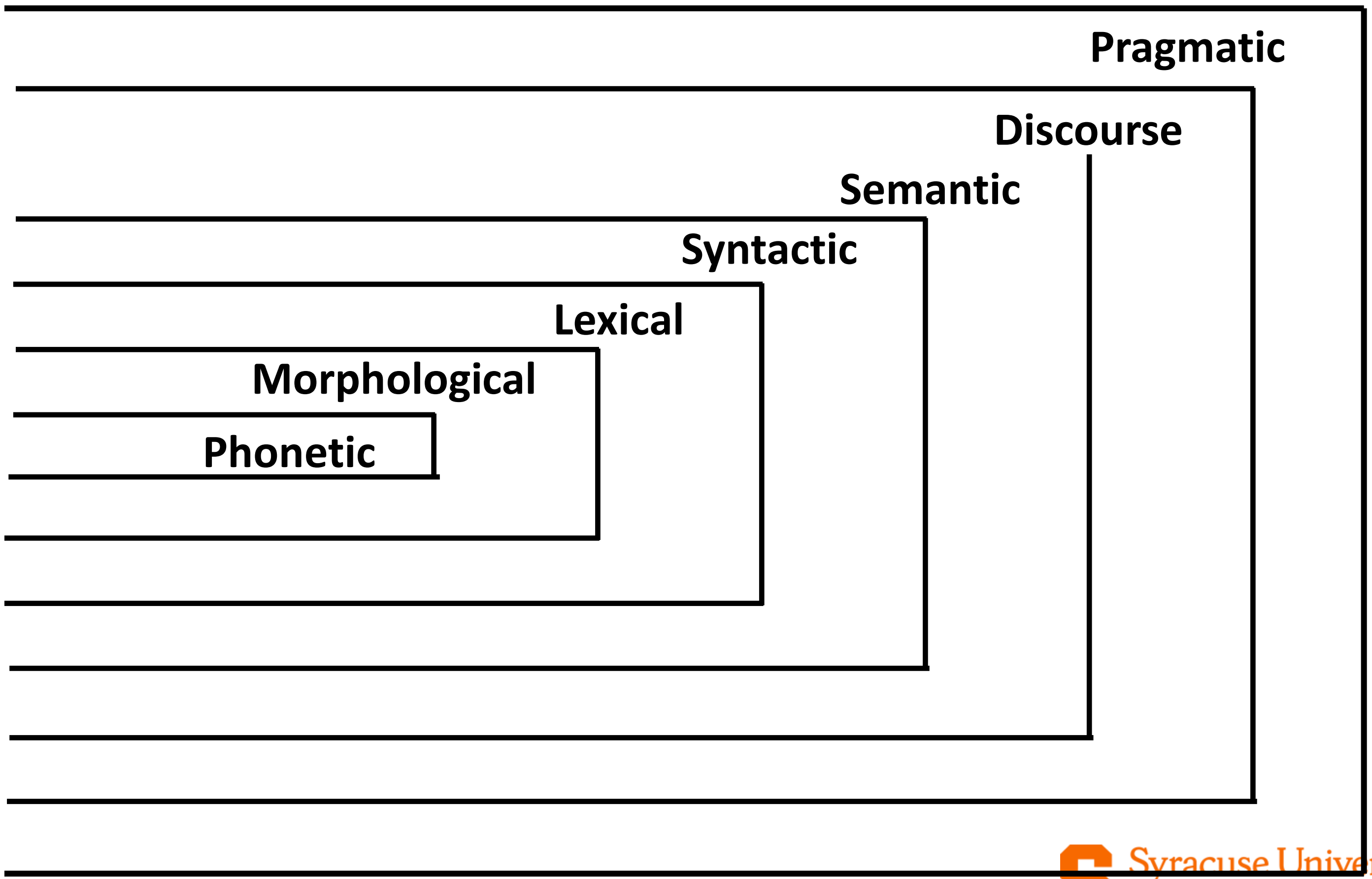
Syracuse University
School of Information Studies

Overview of Natural Language Processing (NLP) and Text Mining

A range of computational techniques for analyzing and representing naturally occurring texts for the purpose of achieving human-like language processing for a range of particular tasks or applications.

Levels of Language Analysis

Use the synchronic model to guide computational techniques to analyze text (as much as possible)



Natural Language as the User Interface

Goal is complete natural language understanding

- Enables computers to interact with humans with natural language

Most common current approach is to craft human/computer interfaces that are in terms that the computer can understand

- XML, drop down boxes, other forms of knowledge representation … cleverness is supplied by the human

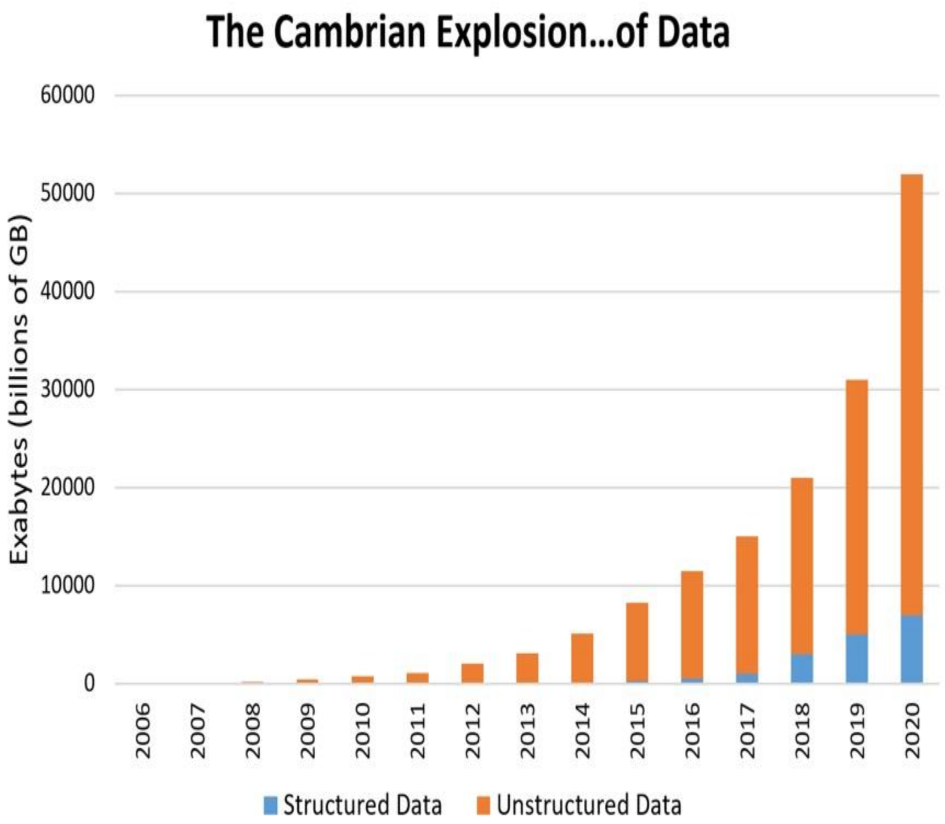
Nascent natural language interfaces are being deployed

- Apple's Siri, the Google Assistant, Amazon's Alexa

Need for Text Mining & NLP

Huge amounts of data

- Internet
- Intranet



Examples :

- Classify text into categories
- Index and search large texts
- Automatic translation of web documents in different languages
- Speech understanding
  - Understand phone conversations
- Information extraction
  - Extract useful information from resumes
- Automatic summarization
  - Condense 1 book into 1 page
  - Daily news summaries
- Question answering
- Knowledge acquisition
- Text generations / dialogues

Fields contributing to Text Mining & NLP

Linguistics
formal, structural models of language

Computer Science
internal representations of data and algorithms for efficient processing

Artificial Intelligence

computational theory of human language processing

Statistics

Frequencies and probabilities of linguistic patterns

Cognitive Psychology
human cognition in language

N.L.P

theoretical

applied

Two Sides of NLP: analysis and generation

1. paraphrase an input text
2. translate it to another language or representation
3. answer questions about it
4. draw inferences from it
5. phrase the results in natural language

Natural Language Processing

Language Analysis*

Language Generation

*Main emphasis in this lecture

Why is NLP so hard?

Seems simple for humans

- Usually quite unaware of the complexity of the language tasks they perform so effortlessly

Some reasons are

- Ambiguity
- Subtleties of meaning
  - Irony,
  - Sarcasm,
  - Humor,
  - Metaphor

Ambiguous Newspaper Headlines

"Stolen Painting Found by Tree"

"Local High School Dropouts Cut in Half"

"Red Tape Holds Up New Bridges"

"Hospitals are Sued by 7 Foot Doctors"

"Kids Make Nutritious Snacks"

- Examples collected by Chris Manning

How Does Text Mining Work ?

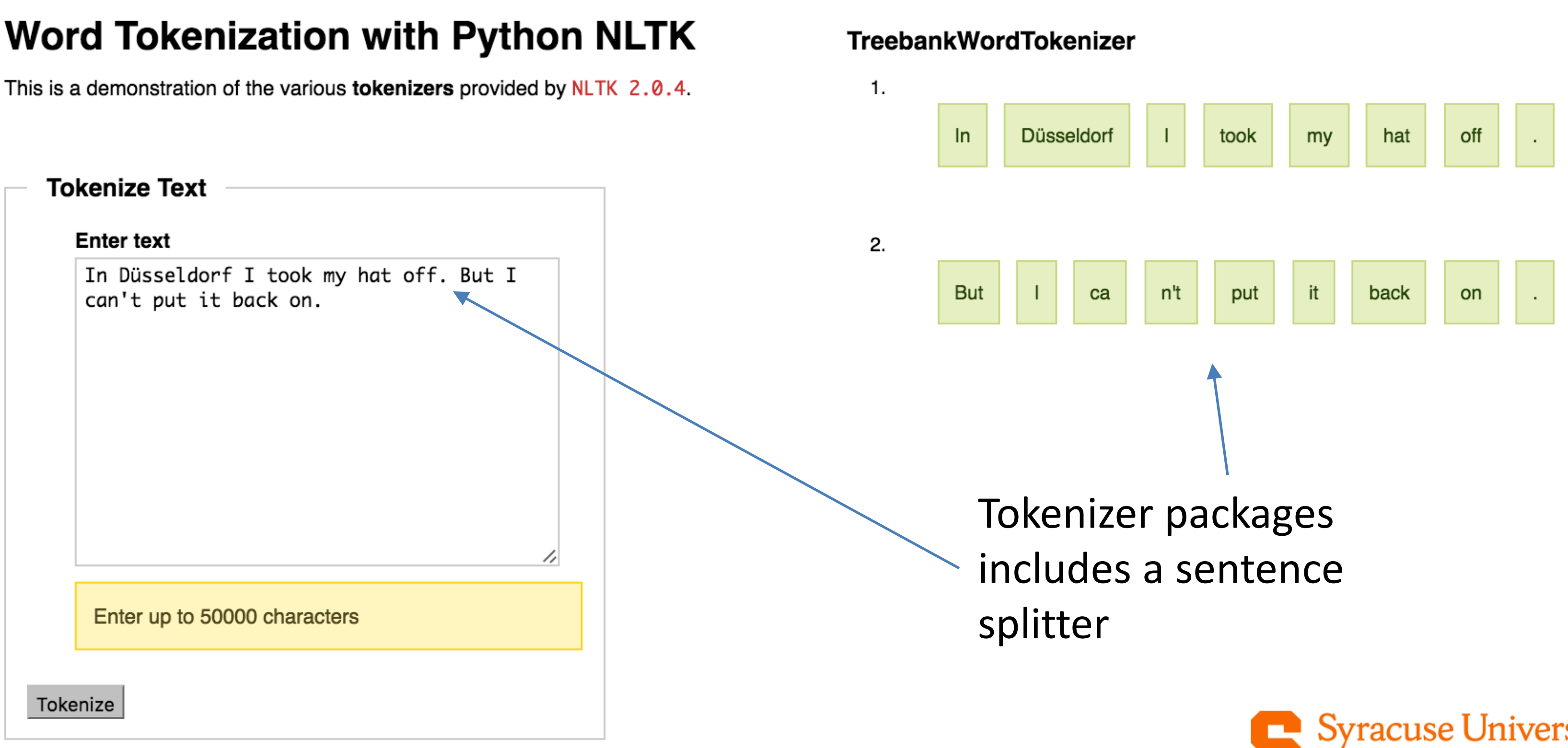
Text Representation/Vectorization

Computers can do only ONE thing, that is, COUNTING!

Convert text to numbers

Tokenization

A tokenizer has a set of rules about grouping characters into tokens



Tokenization rules

Compare how different tokenizers deal with the word "can't"

2.

But	I	ca	n't	put	it	back	on	.

2.

But	I	can	'	t	put	it	back	on	.

2.

But	I	can	't	put	it	back	on.

2.

But	I	can't	put	it	back	on.

Tokenization is not easy

Lowercase vs. uppercase

Words with inflected forms

- “dishwasher” vs. “dishwashers”

Words with multiple senses

- “There is a money **bank** near the river **bank.”**

Vectorization

How to Count Tokens

Many ways to convert documents into word vectors

Bag of Words (BOW)

- Boolean
- Term frequency
- Normalized term frequency
- Tf*idf

Problem is we lose order of words

Vectorization

Step 1: Use tokenizer to create a dictionary of unique words

1 “vector”
2 “number”
3 “text”
…

Step 2: represent how often each word occurs in each document: each word is an attribute/feature

	“vector”	“number”	“text”	…
Doc1	1	0	0	
Doc2	1	1	1	
doc3	1	0	1	

Values of word features

Boolean value: Only provide information on presence or absence

	"vector"	"number"	"text"	…
Doc1	1	0	0	
Doc2	1	1	1	
doc3	1	0	1	

Values of word features

Sometimes we need more information that simply presence vs absence, We want to know how often they occur

In these cases we use Word frequency: the number of word occurrences

	"vector"	"number"	"text"	…
Doc1	5	0	0	
Doc2	1	3	6	
doc3	2	0	8	

Values of word features

We have learned that some documents are longer than other. In order to adjust for different length documents we use Normalization

Normalized word frequency: word frequency normalized by the document length

	"vector"	"number"	"text"	…
Doc1	1	0	0	
Doc2	0.1	0.3	0.6	
doc3	0.2	0	0.8	

Values of word features

We can also use some weighting strategies. A popular weighting strategy is Tf-idf:

Tf*idf weighting

- Tf: term (word) frequency
- Df: document frequency, i.e, how many documents contain this term, e.g. 8 out of 100 documents -> 8/100
- Idf: inversed-document frequency, 100/8
- Tfidf=tf*log(idf)

We want to lower the weight for common words "Vector" and raise weight for rare words.

- Penalize Common words
- Emphases rare word

	"vector"	"number"	"text"
Doc1	1	0	0
Doc2	0.1	0.3	0.6
doc3	0.2	0	0.8

	"vector"	"number"	"text"
Doc1	0	0	0
Doc2	0	0.3*log3	0.6*log 1.5
doc3	0	0	0.8*log 1.5

Tf-idf

- Concept borrowed from information retrieval
- A “blind” weighting strategy for text classification

Reducing Vocabulary Size

Approaches to reduce the vocabulary size

- Stemming
- Case merging
- Removing stopwords
- word clustering

Stemming

Character of inflected language like English

Stemmer: remove postfixes to find the root(stem) form

- "Applied" and "application" -> "appli"
  - Stemmer tends to remove suffix and word not real

Lemmatizer: transform the root to a real word

- "Applied" and "application" -> apply
  - Transforms root to real word

NLTK Stemming Demo

Stem Text

Choose stemmer

Porter

Enter text

Stemming is funnier than a bummer says the sushi loving computer scientist

Enter up to 50000 characters

Stem

Stemmed Text

Stem is funnier than a bummer say the sushi love comput scientist

http://text-processing.com/demo/stem/

Stemming issues

How far should it go?

- "denormalization" -> denormalize -> denormal -> normal -> norm?

How accurate can it be?

- "bore"/ He wanted to bore a hole / He bore the students on his heart

How Useful is Stemming?

No consistent conclusion

- If nuances in different word forms matter, then don't use stemming.
- If you only care about the general stems, like in topic classification, then stemming is helpful.

Information retrieval (Nuance)

- Search "dishwasher" to know how it works
- Search "dishwashers" to shop around

Text categorization (Nuance)

- Future tense vs. past tense in company performance report
  - "Will do" vs. "have done"

Convert Uppercase to Lowercase?

Another example for reducing vocabulary size

Emily Dickinson’s poem

- “Joy” vs. “joy”
- “Love” vs. “love”

Uppercase

But pompous Joy Betrays us, as his first Betrothal Betrays a Boy.	The Treason of an Accent Might vilify the Joy - To breathe - corrode the rapture Of Sanctity to be	Boundlessness - Expanse cannot be lost - Not Joy, but a Decree Is Deity - His Scene, Infinity -

Capitalized Joy occurs in <u>abstract conversation</u>

Lowercase

Could she have guessed that it would be - Could but a Crier of the joy Have climbed the distant hill! -	I want to send you joy, I have half a mind to put up one of these dear little Robin's, and . . .	I cant believe you are coming - but when I think of it, and tell myself it's so, a wondrous joy comes over me, and my old fashioned life . . .

Lower case joy occur in <u>personalized conversation</u>

<u>In this situation, upper case and lower case do matter, but not always the case.</u>

Remove Stop Words

Observation: words occur in most documents are not useful of distinguishing documents

Stopwords are usually function words that bear no specific meaning, compared to content words

Search engines generally eliminate stop words like "and".

Example of the start of a stop word list

a
about
across
after
afterwards
again
against
all
alone
along
already
also
always

amongst
an
and
another
any
anyhow
anyone
anything
are
around
as
be
because

become
becoming
been
before
beforehand
behind
being
below
besides
between
beyond
but
can

Little Words Can Make a **Big** Difference

Little words can make big difference

Function words are useful for certain text mining tasks

- genre classification
- authorship attribution
- gender detection

Genre Classification

Goal is to Classify Document by Document Type (genere)

- Novel; Scientific Paper…

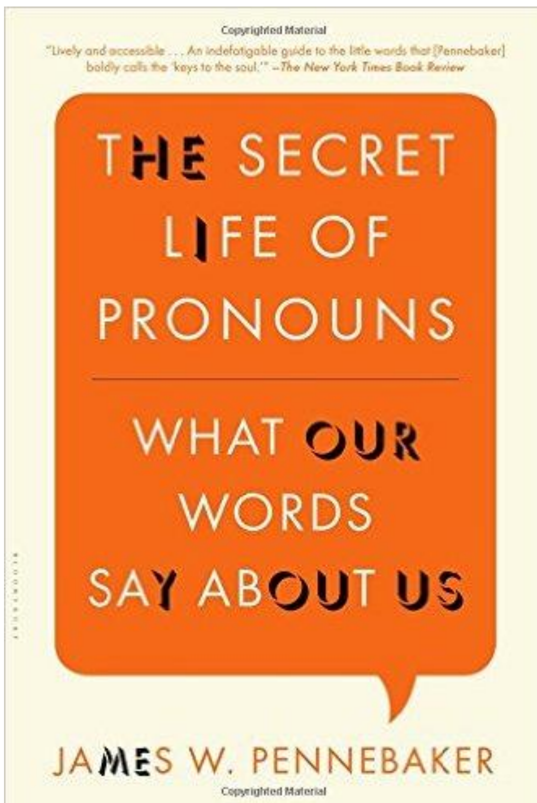
Example

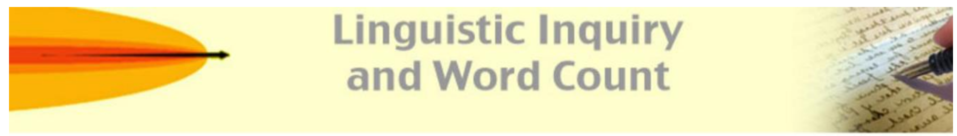
- Personal Homepage Identification (Riloff, 1995)
  - Top features “I” and “my”
  - <u>In these cases, do not remove stop words</u>

Riloff, E. (1995, July). Little words can make a big difference for text classification. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 130-136). ACM.

Personal Pronouns

How personal pronouns are used in a persons writing and speech can tell us a lot about a persons cognitive status





Testing LIWC Online

We understand completely. You are a student, a poor faculty member, or a researcher who wants to analyze a few cases without having to buy the LIWC program for almost $100. We've been there, and, because we know your plight, this page is for you. This is a no-frills page whereby you can enter text (by typing it directly or copying it from some other place and pasting it here) and get the basic LIWC output. All you have to do is enter the text file you want to analyze, press SUBMIT, and voila, we will give you feedback on some of the LIWC dimensions. That's the kind of people we are.

OK, we admit it. We aren't completely altruistic. We would like to keep a copy of your text files to add to our growing archive of 50,000+ files. To help us with our data, could you enter the age and gender of the author of the text (if you know it). If you don't know them or don't want to enter them, then choose 'No details' from the 'Gender of text author' selector.

Gender of author: ? *Age of author:* ?

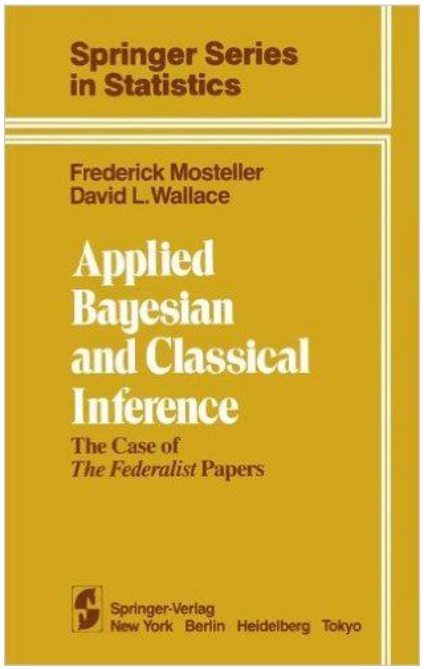
Type or paste the text you want analysed into the box below and then hit the submit button.

Submit text for online LIWC analysis

LIWC Tool – Widely used

Syracuse University
School of Information Studies

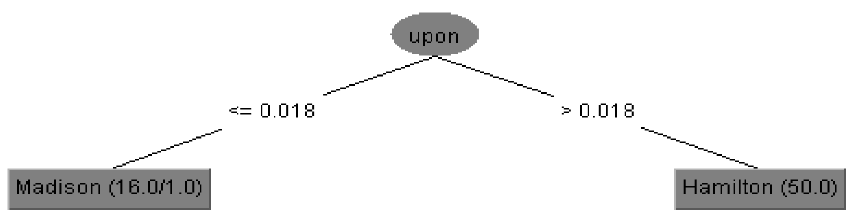
Function Words Used for Authorship Attribution



The Federalist Papers

This web-friendly presentation of the original text of the Federalist Papers (also known as The Federalist) was obtained from the e-text archives of Project Gutenberg. For more information, see About the Federalist Papers.

No.	Title	Author	Publication	Date
1	General Introduction	Hamilton	For the Independent Journal	- -
2	Concerning Dangers from Foreign Force and Influence	Jay	For the Independent Journal	- -
3	The Same Subject Continued: Concerning Dangers from Foreign Force and Influence	Jay	For the Independent Journal	- -
4	The Same Subject Continued: Concerning Dangers from Foreign Force and Influence	Jay	For the Independent Journal	- -
5	The Same Subject Continued: Concerning Dangers from Foreign Force and Influence	Jay	For the Independent Journal	- -
6	Concerning Dangers from Dissensions Between the States	Hamilton	For the Independent Journal	- -



https://www.congress.gov/resources/display/content/The+Federalist+Papers

Gender Classification in General Texts

TABLE 1 (*Continued*)

LIWC Dimension	*Examples*	*Female* *M*	*Female* *SD*	*Male* *M*	*Male* *SD*	*Effect Size (d)*
Pronouns		14.24	4.06	12.69	4.63	0.36
First-person singular	I, me, my	7.15	4.66	6.37	4.66	0.17
First-person plural	we, us, our	1.17	2.15	1.07	2.12	*ns*
Second person	you, you're	0.59	1.05	0.65	1.15	–0.06
Third person	she, their, them	3.41	3.45	2.74	3.01	0.20

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. Discourse Processes, 45(3), 211-236.

Gender Classification in Congress

Is the Text Written by Men or Women ?

Table 4 Gender differences in selected LIWC categories

LIWC dimension	Corpus	Female		Male		Effect size (*d*)	Result
		Mean	SD	Mean	SD		
Pronoun	NGHP	14.24	4.06	12.69	4.63	0.36	Disagree
	HS	7.55	0.01	7.69	0.01	−0.1	

Table 6 Gender differences in pronoun case use

Pronoun cases		Female		Male		Effect size (*d*)
		Mean	SD	Mean	SD	
Subjective	We	1.18	0.40	1.37	0.51	−0.39
	I	1.48	0.32	1.57	0.43	−0.21
Possessive	Our	0.76	0.30	0.58	0.28	0.64
	My	0.46	0.15	0.40	0.17	0.36
Objective	Us	0.22	0.10	0.22	0.10	0.00
	Me	0.15	0.07	0.15	0.08	−0.09

CongressWomen	Congressmen
“our community”	“Our enemy”
“our workforce”	“Our side”
“We honor”	“We ought”
“We share”	“We gave”

Yu, B. (2014). Language and gender in Congressional speech. Literary and Linguistic Computing, 29(1), 118-132.

NLP and text mining tasks

Topic modeling using **LDA**

Topic modeling using **LSA**

Sentiment analysis using **bag-of-words**

Overview of Natural Language Processing

NLP Application Areas

<u>Machine Translation</u> – conversion of text from one language to another

- Google, Yahoo and Bing all have language translators
- MT techniques use context , not just word for word substitution
- Often statistically based patterns of word usage and context

Google Translate

Syracuse University
School of Information Studies

NLP Application Areas

<u>Information Retrieval / Search Engines</u> – provision of documents containing requested information

- Google, many other search engines
- Use lowest levels of NLP to stem words, find phrases for indexing documents
- Users conform to keyword query restriction, but many search engines will now accept questions in natural language form

NLP Application Areas

<u>Information Extraction / Text-mining</u> – populating a structured database with specific bits of information found in text

- Competitive Intelligence analyzes news text and web blogs for
  - Names of people, companies and other entities
  - Relations between them, e.g. corporate roles, or events such as mergers

Weblog Analytics

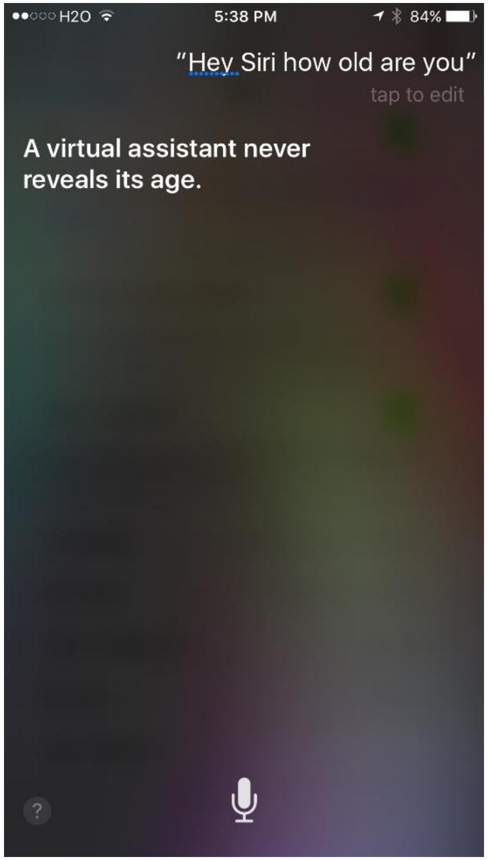
Data-mining of Weblogs, discussion forums, message boards, user groups, and other forms of user generated media

- Product marketing information
- Political opinion tracking
- Social network analysis
- Buzz analysis (what's hot, what topics are people talking about right now).

NLP Application Areas

Human-computer Interfaces –

information assistants,
chatbots,
automatic phone agents,
interactive querying of
databases



NLP Application Areas

Summarization – abstraction and condensation of text’s major points

- Current systems select a set of significant sentences from the document as a summary
- Example summarizer:
  - http://textsummarization.net/text-summarizer

NLP Application Areas

Question & Answering Systems – focused information provision

- Find answers to questions in documents or other resources
- Must be able to handle many different phrasings of desired answer and to provide justification

Watson

IBM's question answering system trained to play Jeopardy

Extensive development of NLP techniques

Syracuse University
School of Information Studies

Trends

- An enormous amount of knowledge is now available in machine readable form as natural language text
- Conversational agents are becoming an important form of human-computer communication
- Much of human-human communication is now mediated by computers

IBM Watson Sample Sites

Text Discovery

https://discovery-news-demo.ng.bluemix.net/

Natural Language Understanding

https://natural-language-understanding-demo.ng.bluemix.net/

Personality Insights

https://personality-insights-demo.ng.bluemix.net/

Project Debater
State of the Art NLP

- **Project Debater**
  - **Actual Debate**
    - https://www.youtube.com/watch?v=m3u-1yttrVw
      - **Opening Argument 10:50 – 15:30**
      - **Rebuttal 22:30 – 28:45**
      - **Summary 37:40 – 39:35**
    - **How does it work ?**
      - https://www.youtube.com/watch?time_continue=66&v=FmGNwMyFCqo
      - https://www.youtube.com/watch?time_continue=3&v=NSB06STBkdA
    - **How does this relate to what we will learn in this class?**

N. Slonim et al., "An autonomous debating system," Nature, vol. 591, no. 7850, pp. 379–384, Mar. 2021, doi: 10.1038/s41586-021-03215-w.