Introduction to Natural Language Processing

Learning Objectives

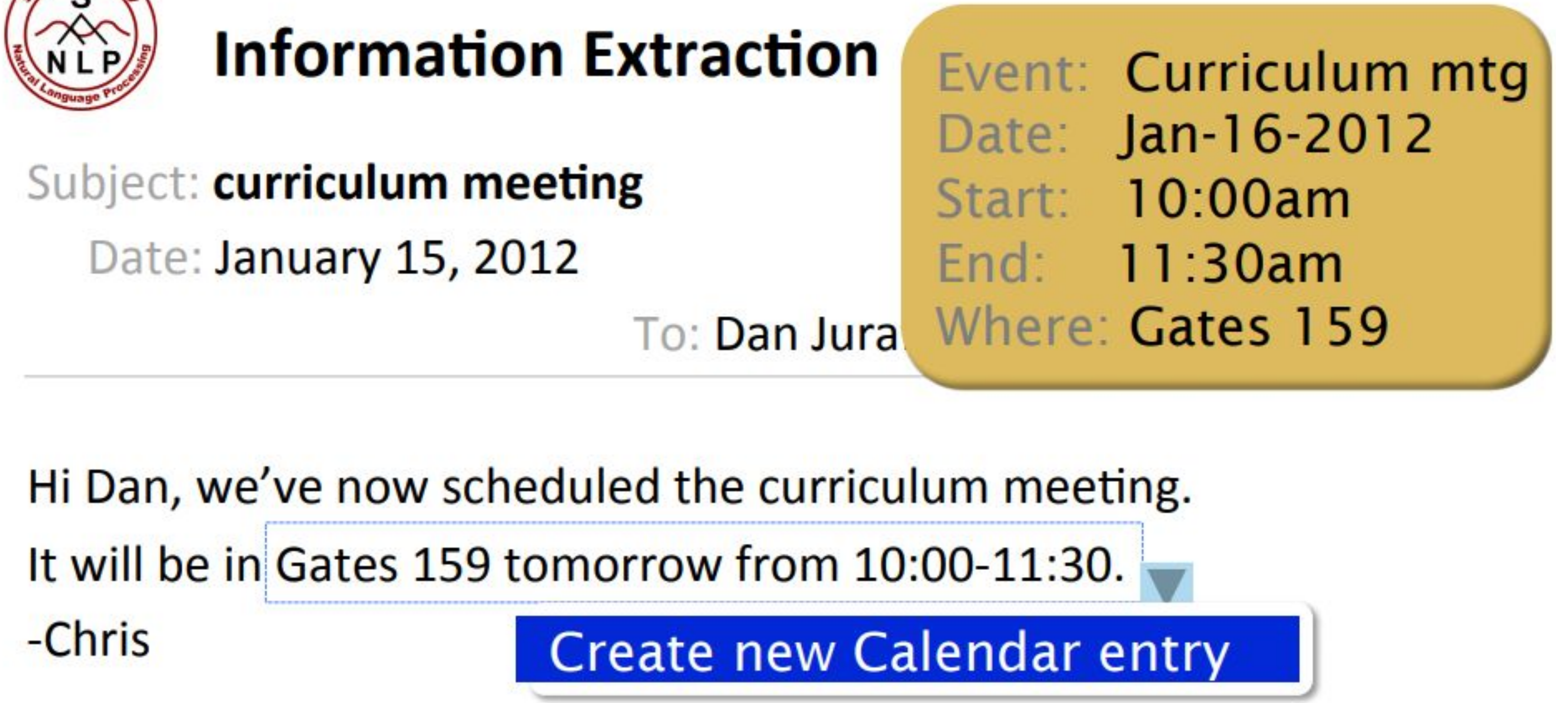
- Discuss the major tasks involved with natural language processing.
- Discuss, on a low level, the components of natural language processing.
- Identify why natural language processing is difficult.
- Demonstrate text classification.
- Demonstrate common text preprocessing techniques

What is Natural Language Processing?

- Using computers to process (analyze, understand, generate) natural human languages.
- Making sense of human knowledge stored as unstructured text.
- Building probabilistic models using data about a language.

What are some of the higher level task areas?

Information Extraction



Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Hi Dan, we've now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Event: Curriculum mtg

Date: Jan-16-2012

Start: 10:00am

End: 11:30am

Where: Gates 159

Information Extraction & Sentiment Analysis

Attributes:

- zoom
- affordability
- size and weight
- flash
- ease of use

Size and weight

- ✓ nice and compact to carry!
- ✓ since the camera is small and light, I
  around those heavy, bulky professio
- ✗ the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera

Machine Translation

- Fully automatic

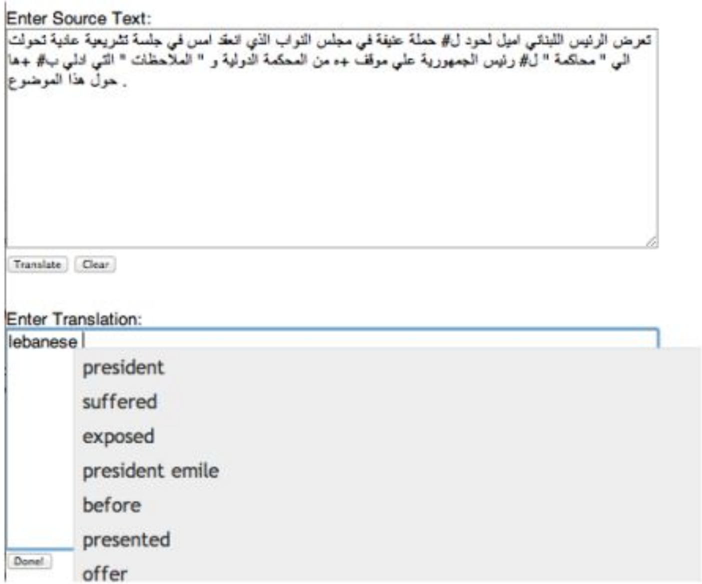
Enter Source Text:

这 不过 是 一 个 时间 的 问题 .

Translation from Stanford's *Phrasal*:

This is only a matter of time.

- Helping human translators



Enter Source Text:

تعرض الرئيس اللبناني اميل لحود ل# حملة عنيفة في مجلس النواب الذي انعقد امس في جلسة تشريعية عادية تحولت
الي " محاكمة " ل# رئيس الجمهورية علي موقف +ه من المحكمة الدولية و " الملاحظات " التي ادلي ب# +ها
حول هذا الموضوع .

Enter Translation:

lebanese

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

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

making good progress

Sentiment analysis

Best roast chicken in San Francisco! 👍

The waiter ignored us for 20 minutes. 👎

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

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

Party
May 27
add

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

Housing prices rose

⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

What are some of the Lower level Components?

What are some of the Lower level Components?

- **Tokenization:** Breaking text into tokens (words, sentences, n-grams)
- **Stop-word removal:** a/an/the
- **Stemming and lemmatization:** root word
- **TF-IDF:** word importance
- **Part-of-speech tagging:** noun/verb/adjective
- **Named entity recognition:** person/organization/location
- **Spelling correction:** "New Yrok City"
- **Word sense disambiguation:** "buy a mouse"
- **Segmentation:** "New York City subway"
- **Language detection:** "translate this page"
- **Machine learning:** specialized models that work well with text

Why is NLP hard?

- **Ambiguity**:
  - Hospitals Are Sued by 7 Foot Doctors
  - Juvenile Court to Try Shooting Defendant
  - Local High School Dropouts Cut in Half
- **Non-standard English:** text messages/ tweets
- **Idioms:** "throw in the towel"
- **Newly coined words:** "retweet"
- **Tricky entity names:** "Where is *A Bug's Life* playing?"
- **World knowledge:** "Mary and Sue are sisters", "Mary and Sue are mothers"

NLP Terms

Corpus: A collection of documents (or words)

Corpora: Plural of corpus

Bag-of-words: All possible words in the corpus

Text Vectorization: Converting all text in a corpus into numerical values

Countvectorizer: Converts each document into a set of words and their counts

Text Classification

- Predicting a category or topic from a text sample
  - Sentiment Analysis e.g. Positive or negative sentiment?
  - Category classification e.g. Sports or Business Story?
  - Rating
- Words are used as the features
- Numeric value is given to each word which could be the number of times they appear in a document
- Text is vectorized and referred to as bag-of-words

Dataset: Yelp Reviews

Countvectorizer

Doc 1: The quick brown fox jumped over the lazy dog
Doc 2: The lazy dog could not outrun the fox

INDEX

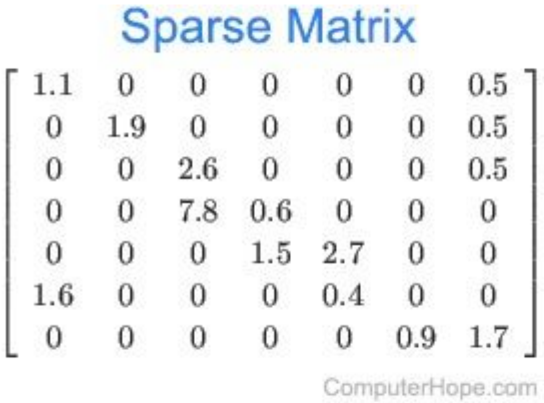
	0	1	2	3	4	5	6	7	8	9
	the	quick	brown	fox	jumped	over	lazy	dog	could	not
Doc 1	2	1	1	1	1	1	1	1	0	0
Doc 2	2	0	0	1	0	0	1	1	1	1

VOCABULARIES

jupyter

Sparse Matrix

A matrix which contains very few non-zero elements



Sparse Matrix

$$\begin{bmatrix} 1.1 & 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 1.9 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 2.6 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 7.8 & 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.5 & 2.7 & 0 & 0 \\ 1.6 & 0 & 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.9 & 1.7 \end{bmatrix}$$

ComputerHope.com

	the	quick	brown	fox	jumped	over	lazy	dog	could	not
Doc 1	0	0	0	1	1	0	1	0	0	0
Doc 2	1	0	0	0	0	1	0	0	1	0
Doc 3	0	1	0	0	1	0	0	0	0	0

- **Vectorizing text produces a sparse matrix**
- **A sparse matrix can be converted to the full form by calling .toarray() on the object**

N-Grams

Features which consist of *N* consecutive words

Text	My cat is awesome
1-gram	‘My’, ‘cat’, ‘is’, ‘awesome’
2-gram	‘My cat’, ‘cat is’, ‘is awesome’
3-gram	‘My cat is’, ‘cat is awesome’,

- **Ngram_range: the upper and lower boundary of ngrams**

EX: How many features do we get from the above examples with ngram_range=(1,3)

Stop Words

- Stop words are some of the most common words in a language
- They are used so that a sentence makes sense grammatically, such as prepositions and determiners, e.g., "to," "the," "and."
- they are so commonly used that they are generally worthless for predicting the class of a document
- They contribute noise to our model

Example:

Original Sentence: “The dog jumped over the fence”

After stop-word removal: “dog jumped over fence”

TextBlob

- provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Stemming and Lemmatization

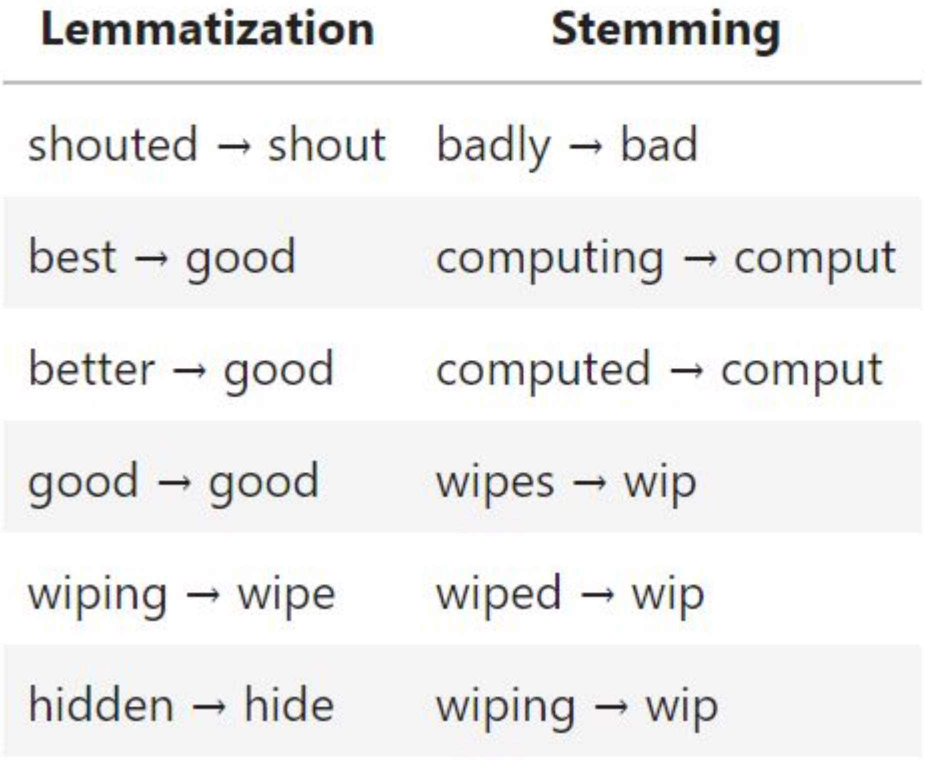
Stemming

- **Reducing a word to its base form.**
- **Removes common ending such as ‘ly’, ‘ing’, ‘s’, ‘es’, ‘ed’**
- **It helps in reducing the number of features**

Lemmatization

- **A more refined process that uses specific language and grammar rules to derive the root of a word**
- **It can be better than stemming e.g ‘best’ to ‘good’, ‘better’ to ‘good’**

Stemming and Lemmatization



Lemmatization	Stemming
shouted → shout	badly → bad
best → good	computing → comput
better → good	computed → comput
good → good	wipes → wip
wiping → wipe	wiped → wip
hidden → hide	wiping → wip

Term Frequency-Inverse Document Frequency (TF-IDF)

- TF-IDF computes the relative frequency with which a word appears in a document compared to frequency across all documents.
- It analyses the uniqueness of words between documents to find distinguishing characters.

Sentiment Analysis with TextBlob

- Understanding how positive or negative a review is.
- There are many ways in practice to compute a sentiment value. For example:
    - Have a list of "positive" words and a list of "negative" words and count how many occur in a document.
    - Train a classifier given many examples of "positive" documents and "negative" documents.
    - Use Generic models

Any Questions