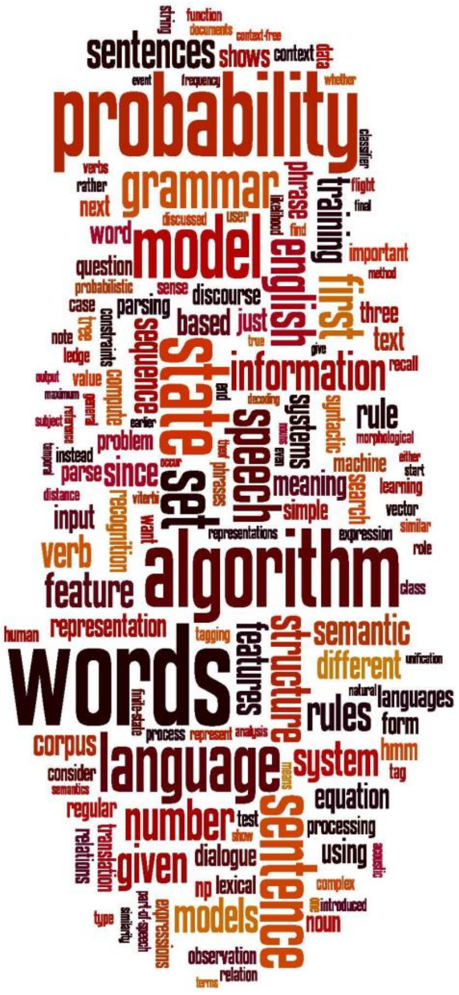


Introduction to NLP

Description

- Foundations of NLP
- Concepts and techniques for processing, understanding and characterizing textual content in natural languages
- Topics Include: Regular expressions and string matching, language modeling, text classification and sentiment analysis, vector representation and embeddings, sequence modeling, POS tagging, machine translation, and chatbots
- Programming in Python

Course Objectives

- To introduce the fundamental concepts and techniques in natural language processing
- To provide experience in the implementation and evaluation of algorithms
- To introduce NLP resources and application areas

Learning Outcomes

- Process and segment textual content for natural language processing
- Extract syntactic and semantic structure from text
- Understand and apply probabilistic and neural network sequence modeling techniques for text analytics
- Use language resources and corpora to implement NLP solutions
- Process multi-lingual and informal-language text in NLP solutions
- Implement and evaluate NLP solutions using libraries

Tentative Grading Criteria

• Assignments	15%
• Quizzes	10%
• Presentation	15%
• Mid-term	20%
• Final-Exam	40%

Google classroom: dxd4azj

Summarized Course Contents

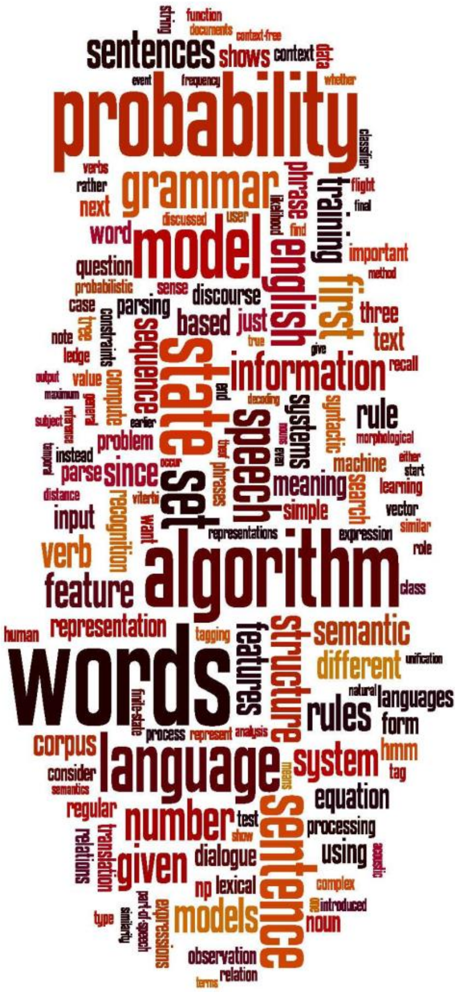
- Introduction and Motivation
- Text processing, regular expressions, edit distance
- Language Modeling with n-grams
- Text Classification and sentiment analysis
- Vector representations and embeddings
- HMMs, Sequential Neural networks
- Machine translation and language generation
- Self-Attention and Transformers, BERT
- LLMs, Prompt Engineering

Course Material

- Required Textbook
  - Speech and Language Processing, 3rd Edition, Jurafsky and Martin

- Recommended and Supplementary Text
  - Machine Learning for Text, C. Aggarwal, 2019
  - Natural Language Processing with Python, Bird and Klein, O’reilly Media, 2009

- Office Hours: Tue & Thur: 12PM to 1:30PM
- Email: hajra.waheed@nu.edu.pk



Introduction to NLP

What is Natural Language Processing?

What is NLP?

- Study of computational approaches for processing natural languages
  - Process → acquire, represent, store, understand, characterize
  - Natural languages→ human languages

- Other names
  - Computational Linguistics (CL)
  - Human language technologies (HLT)

Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF
WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

→ Bram Stoker

Watson gained widespread recognition in 2011 when it competed on and won the television game show Jeopardy! against human champions.

IBM Watson

https://www.youtube.com/watch?v=WFR3lOm_xhE

Information Extraction

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jurafsky

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

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

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

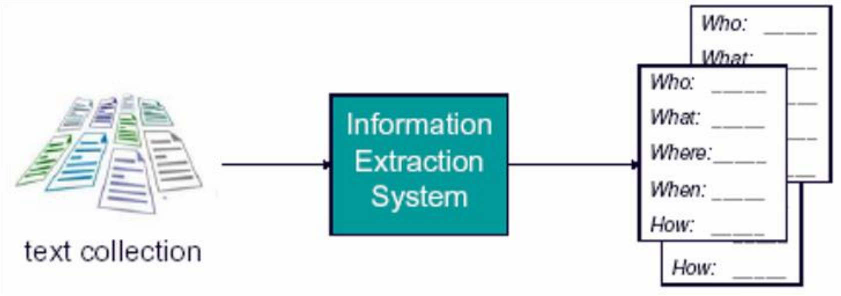
–Chris

▼

Create new Calendar entry

Information Extraction (IE)

- Identify specific pieces of information (data) in an unstructured or semi-structured text
- Transform unstructured information in a corpus of texts or web pages into a structured database (or templates)
- Applied to various types of text, e.g.
  - Newspaper articles
  - Scientific articles
  - Web pages
  - etc.



Source: J. Choi, CSE842, MSU

Information Extraction & Sentiment Analysis

AWributes:

- 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 plasIc and very light in weight you have to be very delicate in the handling of this camera

Machine Translation

- Fully automatic

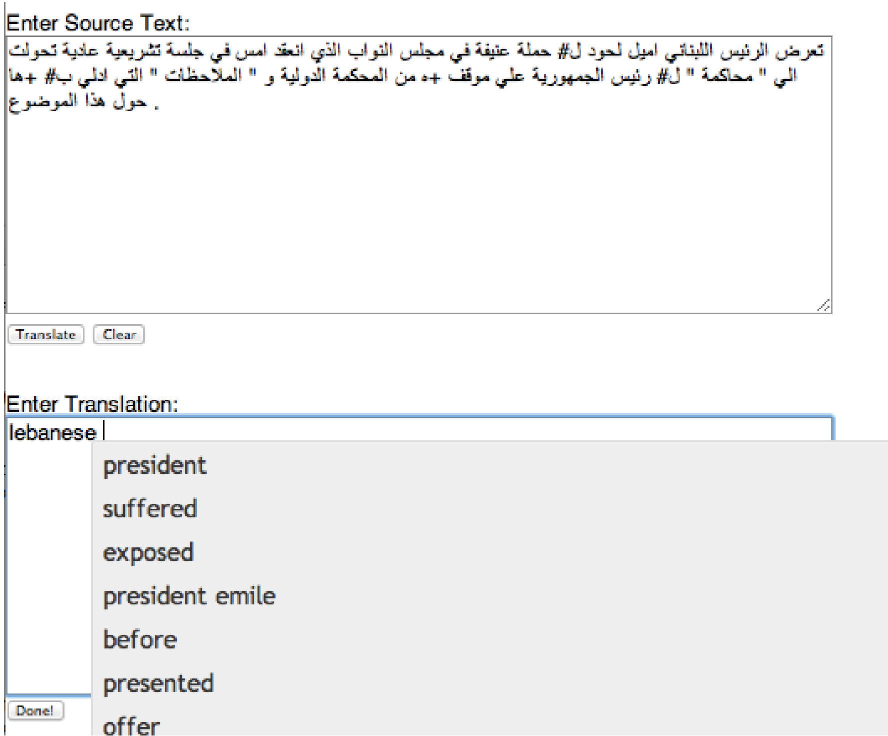
Enter Source Text:

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 .

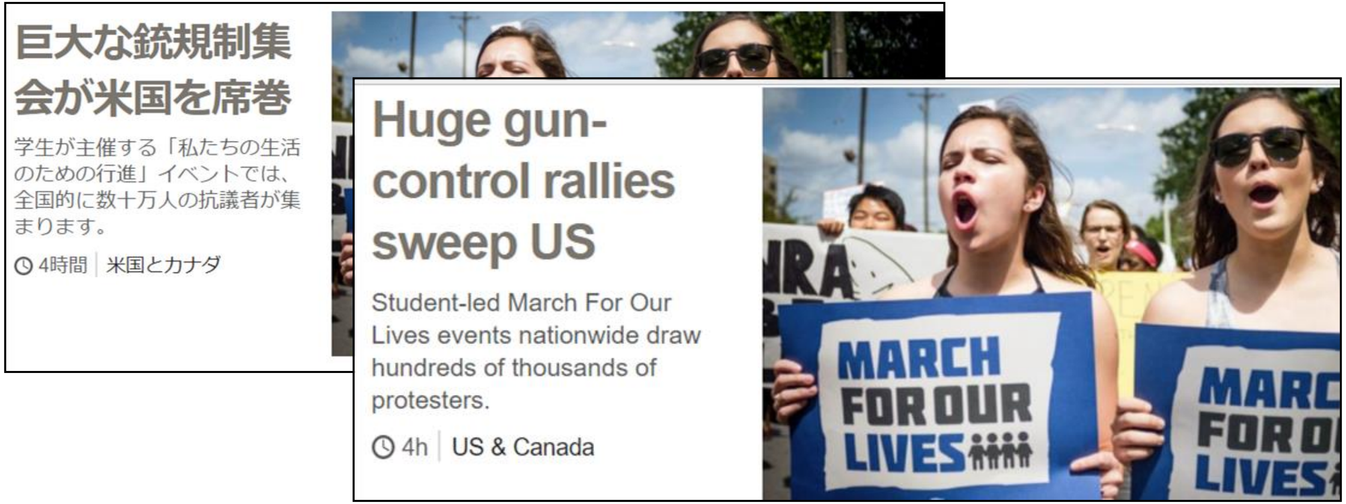
Enter Translation:

lebanese

president
suffered
exposed
president emile
before
presented
offer

Machine Translation

- The automatic translation of texts between languages is one of the oldest non-numerical applications in Computer Science.
- In the past 15 years or so, MT has gone from a niche academic curiosity to a robust commercial industry.



巨大な銃規制集会が米国を席巻

学生が主催する「私たちの生活のための行進」イベントでは、全国的に数十万人の抗議者が集まります。

4時間 | 米国とカナダ

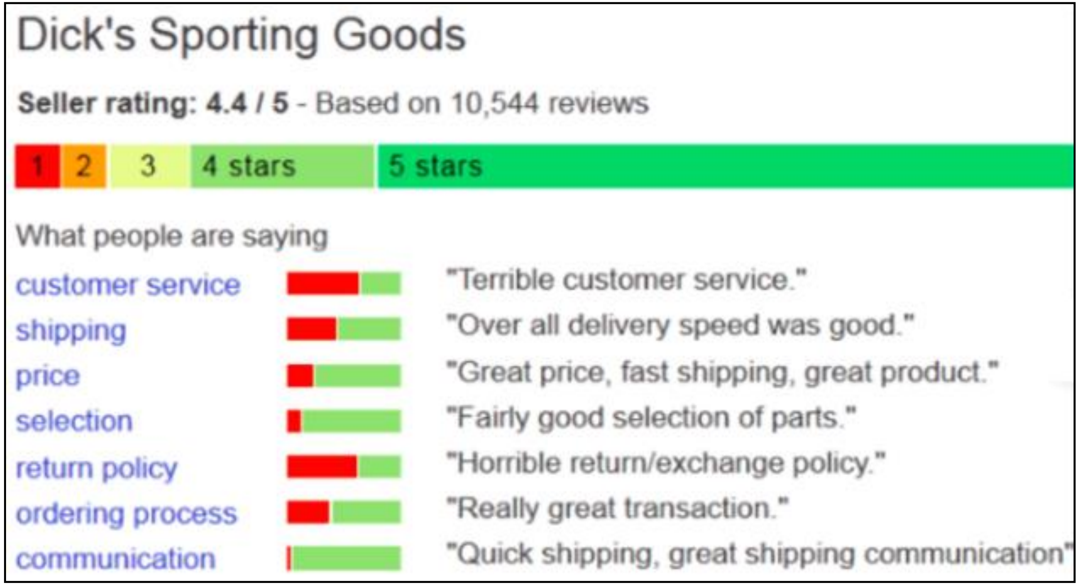
Huge gun-control rallies sweep US

Student-led March For Our Lives events nationwide draw hundreds of thousands of protesters.

4h | US & Canada

Text Analytics

- Data-mining of weblogs, microblogs, discussion forums, user reviews, and other forms of user-generated media.



Dick's Sporting Goods

Seller rating: 4.4 / 5 - Based on 10,544 reviews

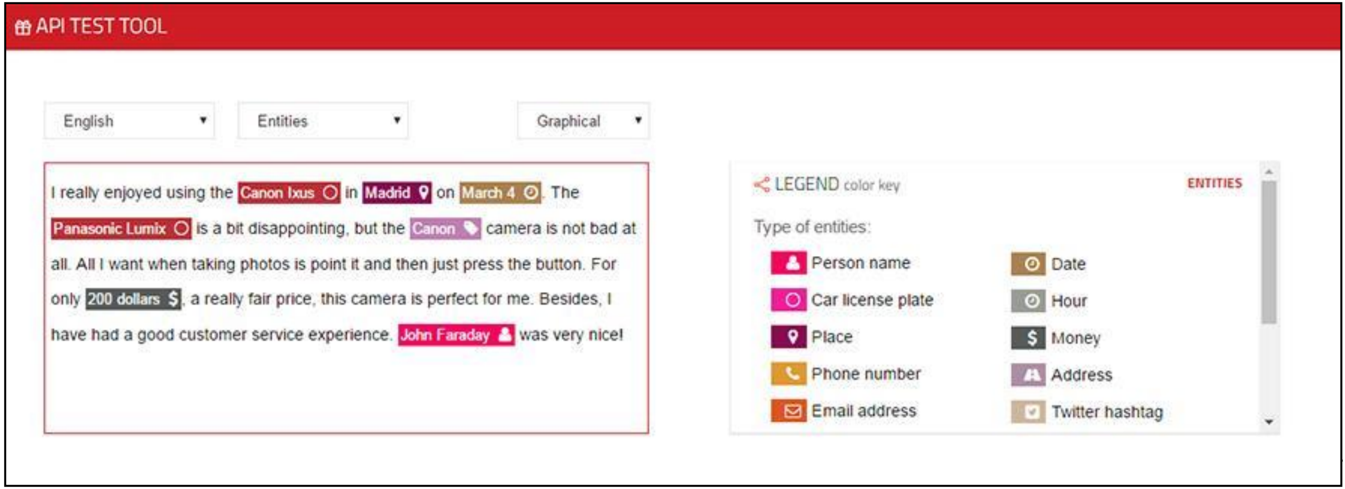
1 2 3 4 stars 5 stars

What people are saying

customer service	"Terrible customer service."
shipping	"Over all delivery speed was good."
price	"Great price, fast shipping, great product."
selection	"Fairly good selection of parts."
return policy	"Horrible return/exchange policy."
ordering process	"Really great transaction."
communication	"Quick shipping, great shipping communication"

Text Analytics (cont.)

- Typically this involves the extraction of **limited** kinds of semantic and pragmatic information from texts
  - Entity mentions
  - Concept identification
  - Sentiment



API TEST TOOL

English ▾ Entities ▾ Graphical ▾

I really enjoyed using the Canon Ixus in Madrid on March 4. The Panasonic Lumix is a bit disappointing, but the Canon camera is not bad at all. All I want when taking photos is point it and then just press the button. For only 200 dollars $, a really fair price, this camera is perfect for me. Besides, I have had a good customer service experience. John Faraday was very nice!

LEGEND color key — ENTITIES

Type of entities:

Person name	Date
Car license plate	Hour
Place	$ Money
☎ Phone number	Address
✉ Email address	Twitter hashtag

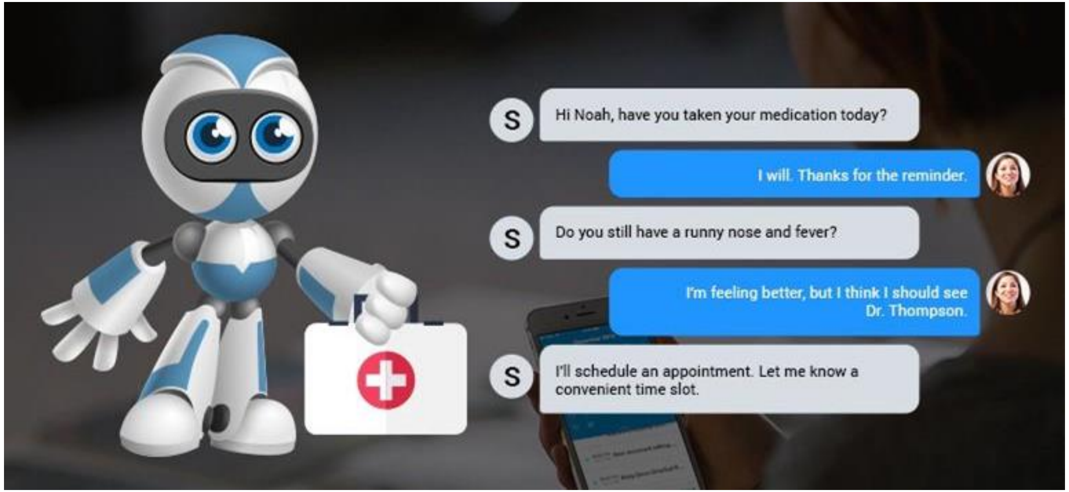
Demo

- Sentiment Analysis with Python NLTK Text Classification
  - http://text-processing.com/demo/sentiment/

- Tweet Sentiment Visualization Tool
  - https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

Conversational Agents

- Combine
  - Speech recognition/synthesis
  - Question answering
    - From the web and from structured information sources (freebase, dbpedia, yago, etc.)
  - Simple agent-like abilities
    - Create/edit calendar entries
    - Reminders
    - Directions
    - Invoking/interacting with other apps



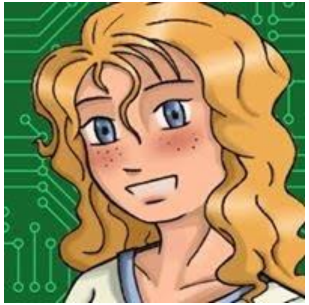
S: Hi Noah, have you taken your medication today?

I will. Thanks for the reminder.

S: Do you still have a runny nose and fever?

I'm feeling better, but I think I should see Dr. Thompson.

S: I'll schedule an appointment. Let me know a convenient time slot.



Mitsuku

English-language chatbot

Question Answering

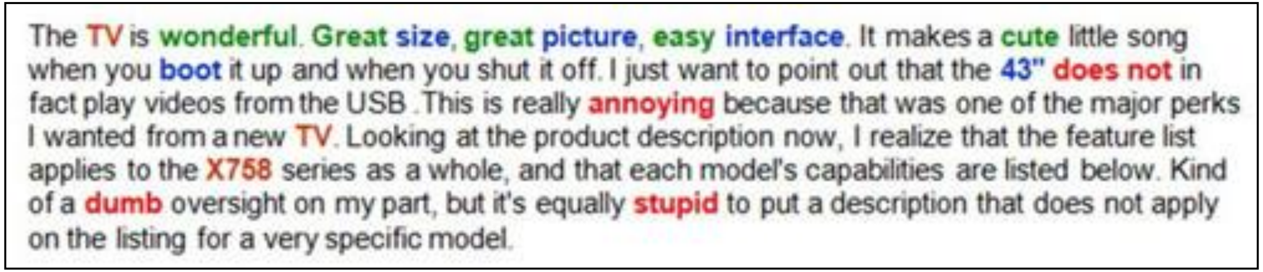
- Traditional *information retrieval* provides documents/resources that provide users with what they need to satisfy their information needs.
- *Question answering* on the other hand directly provides an answer to information needs posed as questions.

Text Mining Applications – Supervised

- Many typical **predictive modeling** or classification applications can be enhanced by incorporating textual data in addition to traditional input variables.
    - churning propensity models that include customer center notes, website forms, e-mails, and Twitter messages to predict customer attrition pattern
    - hospital admission prediction models incorporating medical records notes as a new source of information
    - insurance fraud modeling using adjustor notes
    - sentiment categorization (next page)
    - stylometry or forensic applications that identify the author of a particular writing sample

Sentiment Analysis

- The field of sentiment analysis deals with categorization (or classification) of opinions expressed in textual documents



The **TV** is **wonderful**. **Great size**, **great picture**, **easy interface**. It makes a **cute** little song when you **boot** it up and when you shut it off. I just want to point out that the **43"** **does not** in fact play videos from the USB .This is really **annoying** because that was one of the major perks I wanted from a new **TV**. Looking at the product description now, I realize that the feature list applies to the **X758** series as a whole, and that each model's capabilities are listed below. Kind of a **dumb** oversight on my part, but it's equally **stupid** to put a description that does not apply on the listing for a very specific model.

- Green color represents positive tone
- Red color represents negative tone
- Product features and model names are highlighted in blue and brown, respectively.

Language Technology

solved

Spam detection

Let's go to Agra! ✓

Buy AGRA … ✗

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! 👍

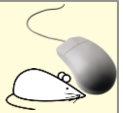
The waiter ignored us for 20 minutes. 👎

Coreference resolution

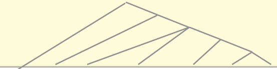
Carter told Mubarak he shouldn't run again.

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

making good progress?

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?

NLP Tasks

- NLP applications require several NLP analyses:
  - Word tokenization
  - Sentence boundary detection
  - Part-of-speech (POS) tagging
    - to identify the part-of-speech (e.g. noun, verb) of each word
  - Named Entity (NE) recognition
    - to identify proper nouns (e.g. names of person, location, organization; domain terminologies)
  - Parsing
    - to identify the syntactic structure of a sentence
  - Semantic analysis
    - to derive the meaning of a sentence

Part-Of-Speech (POS) Tagging

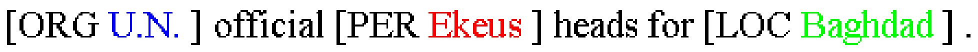
- POS tagging is a process of assigning a POS or lexical class marker to each word in a sentence (and all sentences in a corpus).

Input: `the lead paint is unsafe`

Output: `the/Det lead/N paint/N is/V unsafe/Adj`

Named Entity Recognition (NER)

- NER is to process a text and identify named entities in a sentence
  - e.g. “U.N. official Ekeus heads for Baghdad.”



[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad] .

Ambiguity makes NLP hard: “Crash blossoms”

100% REAL

Violinist Linked to JAL Crash Blossoms

Teacher Strikes Idle Kids

Red Tape Holds Up New Bridges

Hospitals Are Sued by 7 Foot Doctors

Juvenile Court to Try Shooting Defendant

Local High School Dropouts Cut in Half

Why else is natural language understanding difficult?

non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York--New Haven Railroad

the New-York New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene …

But that's what makes it fun!

Making progress on this problem…

- The task is difficult! What tools do we need?
  - Knowledge about language
  - Knowledge about the world
  - A way to combine knowledge sources
- How we generally do this:
  - probabilisIc models built from language data
    - P(“maison” → “house”)   high
    - P(“L’avocat général” → “the general avocado”)   low

Key Trends

- Learn a language from large corpora of text
- No labels required; try to predict words
- Language modeling is driving modern day NLP
- Traditional probabilistic language models (n-grams) to model learning based models
- Transformers race GPT, T5, BERT
- Feature representation and end-to-end learning
- Integrate corpus and knowledge-based information from raw input to final desired outcome
- Transfer Learning: train model on related task and apply to new task

Confluence of Fields

- Statistics and Probability
- Machine Learning/Artificial Intelligence
- Data Structure & Algorithms
- Linguistics
- Psychology

Basic Text Processing

Regular Expressions

Regular Expressions

Regular expressions, also known as regex, are a powerful tool for working with text data in NLP. They are used to search for patterns in text and can be used to perform a wide range of tasks, including:

R(text) outputs a set of strings

1.Tokenization: Regular expressions can be used to tokenize text data by splitting it into individual words or phrases. This is a common step in NLP preprocessing, as it allows us to analyze and work with text data at the level of individual words or tokens.

2.Data Cleaning: Regular expressions can be used to remove unwanted characters or formatting from text data. For example, regular expressions can be used to remove punctuation, special characters, or HTML tags from text data.

3.Information Extraction: Regular expressions can be used to extract specific information from text data. For example, regular expressions can be used to extract phone numbers, email addresses, or dates from text data.

Regular Expressions

4. **Text Classification**: Regular expressions can be used to extract features from text data that can be used for text classification. For example, regular expressions can be used to extract specific words or phrases from text data that are indicative of a particular class or category.
5. **Sentiment Analysis**: Regular expressions can be used to extract emoticons and emojis from text data. These emoticons and emojis can be used as features to predict the sentiment of the text data.
6. **Named Entity Recognition**: Regular expressions can be used to extract named entities from text.

Regular expressions

A formal language for specifying text strings

How can we search for any of these?

woodchuck
woodchucks
Woodchuck
Woodchucks

/Lahore/
/[Ll]ahore/

https://www.regexpal.com/99810

Regular Expressions: Disjunctions

Letters inside square brackets []

Pattern	Matches
[wW]oodchuck	Woodchuck, woodchuck
[1234567890]	Any digit

Ranges [A-Z]

Pattern	Matches	
[A-Z]	An upper case letter	Drenched Blossoms
[a-z]	A lower case letter	my
[0-9]	A single digit	Chapter 1: Down the Rabbit Hole

Regular Expressions: Disjunction

Woodchucks is another name for groundhog!

The pipe | for disjunction

<table>
<tr><th>Pattern</th><th>Matches</th></tr>
<tr><td>/groundhog|woodchuck/
/groundhog|woodchuck/i</td><td>Add i flag: case-insensitive</td></tr>
<tr><td>yours|mine</td><td>yours
mine</td></tr>
<tr><td>a|b|c</td><td>a, b, c</td></tr>
<tr><td>[gG]roundhog|[Ww]oodchuck</td><td></td></tr>
</table>

Regular Expressions: Disjunctions

Letters inside square brackets []

Pattern	Matches
[0-9]A[-/]	2A-, 3A/, 3A-,…
[a\-b]	a or b or -
Special Characters	
\.	matches "."
\+	Matches +
Specifying date Dd-mm-yy [0123][0-9][\-][0-9][\-][0-9][0-9]	

Regular Expressions: Negation in Disjunction

Negations [^Ss]

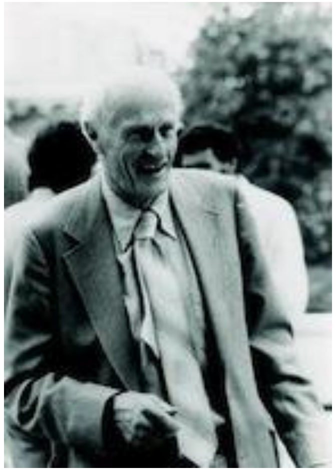
Carat means negation only when first in []

Pattern	Matches	
[^A-Z]	Not an upper case letter	Oyfn pripetchik
[^A-Za-z]	No caps no small letters	
[^Ss]	Neither 'S' nor 's'	reason
[^e^]	Neither e nor ^	Look here
a\^b	The pattern a carat b	Look up a^b now
a{3}	3 times a	aaa
a{3,6}	{min, max}	a can occur min 3 times or max times

The curly braces {} are called quantifiers

Wildcards: ?   *   +   .

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
[0-9]+		3,01, 000,011,0987654
[0-9]{2}	Two digits in a row	00,03,34…. **123abc456** Returns: 12, 45
beg.n	Exc. Special charc any can come	begin begun begun beg3n



Stephen C Kleene

Kleene *,   Kleene +

Regular Expressions: Anchors ^ $

Pattern	Matches
^[A-Z]	Find any charc A-Z that occurs in the start of line Palo Alto
^[^A-Za-z]	"1hello"
^[the] ^The	Finds t,h or e at the start of the string Top, eat, hat, The The occurs start of line
[\.$]	The end.
[\:$]	First occurrence of : is considered the end of line

Example

Find me all instances of the word “the” in a text.

`the` Misses capitalized examples

`[tT]he` `but might match incorrectly`
`other, theology`

`[^a-zA-Z][tT]he[^a-zA-Z] 1the2 etc..`
`Correct regex: \bthe`

\b denotes a word boundary, ensuring that "the" is matched as a whole word rather than as part of another word (e.g., "there")

\bthe\shistory: \s is used for space

the history..

Errors

The process we just went through was based on fixing two kinds of errors

- Matching strings that we should not have matched
- False positives (Type I)
  - faq (afaq is FP)
  - the (theory)
- Not matching things that we should have matched
  - False negatives (Type II)
  - Lahore (lahore is FN)
  - The (the is FN)

Errors cont.

In NLP we are always dealing with these kinds of errors.

Reducing the error rate for an application often involves two antagonistic efforts:

- Increasing accuracy or precision (minimizing false positives)
- Increasing coverage or recall (minimizing false negatives).

Summary

- Regular expressions play a surprisingly large role
  - Sophisticated sequences of regular expressions are often the first model for any text processing text
- For many hard tasks, we use machine learning classifiers
  - But regular expressions can be used as features in the classifiers
  - Can be very useful in capturing generalizations

Basic Text Processing

Regular Expressions