

Natural Language Processing

Artificial Intelligence @ Allegheny College

Janyl Jumadinova

October 20–22, 2021

Credit: NLP Stanford

Natural Language Processing

Understand, interpret and manipulate natural language

Question Answering: IBM's Watson

Won Jeopardy on February 16, 2011!

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



Bram Stoker

Information Extraction

Subject: **curriculum meeting**

Date: November 1, 2016

Event: Curriculum mtg

Date: Nov-1-2016

Start: 10:00am

End: 11:00am

Where: CC 103

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

It will be in CC 103 tomorrow from 10:00-11:00.

-Chris

Create new Calendar entry

Sentiment Extraction

2016 Election



Source: Washington Post

Machine Translation

The screenshot shows the Google Translate web interface. At the top is the Google logo. Below it is the 'Translate' heading. The interface is divided into two main sections: the source text area on the left and the target text area on the right. The source text area shows 'Happy Monday' in English, with a dropdown menu indicating the source language is 'English - detected'. The target text area shows the translation 'бактылуу Дүйшөмбү' in Kyrgyz, with a dropdown menu indicating the target language is 'Kyrgyz'. A blue 'Translate' button is visible between the two text areas. Below the target text, the Kyrgyz text 'бактылуу Дүйшөмбү' is repeated. At the bottom right of the target text area is a 'Suggest an edit' link. The interface also includes a 'Turn off instant translation' toggle and a user profile icon in the top right corner.

Google

Translate

Turn off instant translation

English Spanish French English - detected

English Spanish Kyrgyz Translate

Happy Monday

бактылуу Дүйшөмбү

бактылуу Дүйшөмбү

Suggest an edit

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy VIAGRA ... ✗

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?

Ambiguity makes NLP hard

Ambiguity makes NLP hard

- Teacher Strikes Idle Kids

Ambiguity makes NLP hard

- Teacher Strikes Idle Kids
- Red Tape Holds Up New Bridges

Ambiguity makes NLP hard

- Teacher Strikes Idle Kids
- Red Tape Holds Up New Bridges
- Juvenile Court to Try Shooting Defendant

Ambiguity makes NLP hard

- Teacher Strikes Idle Kids
- Red Tape Holds Up New Bridges
- Juvenile Court to Try Shooting Defendant
- Local High School Dropouts Cut in Half

Other NLP Difficulties

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 ...

Progress

- What tools do we need?
 - Knowledge about language
 - Knowledge about the world
 - A way to combine knowledge sources

- What tools do we need?
 - Knowledge about language
 - Knowledge about the world
 - A way to combine knowledge sources
- How we generally do this:
 - Probabilistic models built from language data
 - $P(\text{"maison"} \rightarrow \text{"house"}) \rightarrow \text{high}$
 - $P(\text{"L'avocat general"} \rightarrow \text{"the general avocado"}) \rightarrow \text{low}$

Basic Text Processing

Word tokenization

Every NLP task needs to do text normalization:

- ① **Segmenting/tokenizing words in running text**
- ② Normalizing word formats
- ③ Segmenting sentences in running text

How Many Words?

N = number of tokens

V = vocabulary = set of types

$|V|$ is the size of the vocabulary

Church and Gale (1990): $|V| > O(N^{1/2})$

	Tokens = N	Types = $ V $
Switchboard phone conversations	2.4 million	20 thousand
Shakespeare	884,000	31 thousand
Google N-grams	1 trillion	13 million

N - all words

V - distinct words

Basic Text Processing

Normalization

Every NLP task needs to do text normalization:

- ① Segmenting/tokenizing words in running text
- ② **Normalizing word formats**
- ③ Segmenting sentences in running text

Issues in Tokenization

- Finland's capital → Finland Finlands Finland's
- what're, I'm, isn't → What are, I am, is not
- Hewlett-Packard → Hewlett Packard
- state-of-the-art → state of the art
- Lowercase → lower-case lowercase lower case
- San Francisco → one token or two?

Issues in Tokenization

- Finland's capital → Finland Finlands Finland's
- what're, I'm, isn't → What are, I am, is not
- Hewlett-Packard → Hewlett Packard
- state-of-the-art → state of the art
- Lowercase → lower-case lowercase lower case
- San Francisco → one token or two?
- **Language Issues:** French, German, Japanese, Chinese,...

Issues in Tokenization

- Finland's capital → Finland Finlands Finland's
- what're, I'm, isn't → What are, I am, is not
- Hewlett-Packard → Hewlett Packard
- state-of-the-art → state of the art
- Lowercase → lower-case lowercase lower case
- San Francisco → one token or two?
- **Language Issues:** French, German, Japanese, Chinese,...

Normalization:

merging of different forms of a token into a canonical normalized form.

- ex.: "Mr.", "Mr", "mister", and "Mister" into a single form.

Basic Text Processing

Stemming

Every NLP task needs to do text normalization:

- ① Segmenting/tokenizing words in running text
- ② Normalizing word formats
- ③ **Segmenting sentences in running text**

Stemming

- Reduce terms to their stems in information retrieval
- **Stemming** is crude chopping of affixes language dependent
- Example: **automate(s)**, **automatic**, **automation** all reduced to **automat**.

*for example compressed
and compression are both
accepted as equivalent to
compress.*



for exampl compress and
compress ar both accept
as equal to compress

Porter's Algorithm

Most common English stemmer.

Step 1a

sses	→ ss	caresses	→ caress
ies	→ i	ponies	→ poni
ss	→ ss	caress	→ caress
s	→ ∅	cats	→ cat

Step 1b

(*v*)ing	→ ∅	walking	→ walk
		sing	→ sing
(*v*)ed	→ ∅	plastered	→ plaster
...			

Step 2 (for long stems)

ational	→ ate	relational	→ relate
izer	→ ize	digitizer	→ digitize
ator	→ ate	operator	→ operate
...			

Step 3 (for longer stems)

al	→ ∅	revival	→ reviv
able	→ ∅	adjustable	→ adjust
ate	→ ∅	activate	→ activ
...			

Sentence Segmentation

- !, ? are relatively unambiguous

Sentence Segmentation

- !, ? are relatively unambiguous
- Period "." is quite ambiguous
 - Sentence boundary
 - Abbreviations like Inc. or Dr.
 - Numbers like .02 or 4.3

Sentence Segmentation

- !, ? are relatively unambiguous
- Period "." is quite ambiguous
 - Sentence boundary
 - Abbreviations like Inc. or Dr.
 - Numbers like .02 or 4.3
- Build a binary classifier
 - Classifiers: hand--written rules, regular expressions, or machine--learning

Information Extraction (IE)

- Find and understand limited relevant parts of texts
- Gather information from many pieces of text
- Produce a structured representation of relevant information

Information Extraction

Goals:

- Organize information so that it is useful to people
- Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Information Extraction

Goals:

- Organize information so that it is useful to people
- Put information in a semantically precise form that allows further inferences to be made by computer algorithms

Roughly: **Who did what to whom when?**

Low-level information extraction

Google



google headquarters

🔍

All Maps Images Videos News More ▾ Search tools

About 731,000,000 results (0.53 seconds)

Google / Headquarters



Mountain View, CA

The image shows a Google search result for "google headquarters". It includes the Google logo, a search bar with the text "google headquarters", and a microphone icon. Below the search bar are tabs for "All", "Maps", "Images", "Videos", "News", "More", and "Search tools". The "All" tab is selected. The search results show "About 731,000,000 results (0.53 seconds)". The main result is titled "Google / Headquarters" and features two images: an aerial view of Mountain View, CA, and a map of the San Jose area with Mountain View highlighted. The map shows major highways like I-280, I-880, and SR-101, and labels for cities like Menlo Park, Palo Alto, Sunnyvale, Santa Clara, and San Jose. The text "Map data ©2016 Google" is visible at the bottom right of the map.

Named Entity Recognition (NER)

A very important sub-task: **find** and classify names in text

The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie, Rob Oakeshott, Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Named Entity Recognition (NER)

A very important sub-task: find and **classify** names in text

The decision by the independent MP **Andrew Wilkie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wilkie**, **Rob Oakeshott**, **Tony Windsor** and the **Greens** agreed to support **Labor**, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organization

Named Entity Recognition (NER)

The uses:

- Named entities can be indexed, linked, etc.
- Sentiment can be attributed to companies or products
- A lot of IE relations are associations between named entities
- For question answering, answers are often named entities

Named Entity Recognition (NER)

- Data $\{(c, d)\}$ of paired observations d and hidden classes c
- **Features** f are elementary pieces of evidence that link aspects of what we observe d with a category c that we want to predict

Named Entity Recognition (NER)

- Data $\{(c, d)\}$ of paired observations d and hidden classes c
- **Features** f are elementary pieces of evidence that link aspects of what we observe d with a category c that we want to predict

$$f_1(c, d) \equiv [c = \text{LOCATION} \wedge w_{-1} = \text{"in"} \wedge \text{isCapitalized}(w)]$$

$$f_2(c, d) \equiv [c = \text{LOCATION} \wedge \text{hasAccentedLatinChar}(w)]$$

$$f_3(c, d) \equiv [c = \text{DRUG} \wedge \text{ends}(w, \text{"c"})]$$

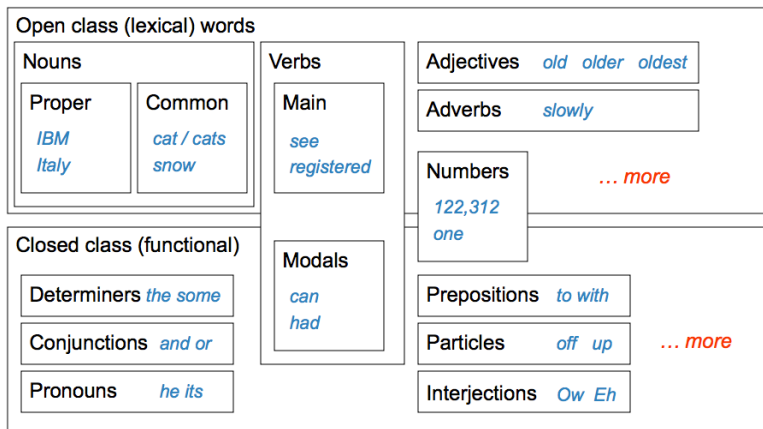
LOCATION
in Arcadia

LOCATION
in Québec

DRUG
*taking
Zantac*

PERSON
saw Sue

Parts of Speech (POS)



POS Tagging

Words often have more than one POS:

- The back door
- On my back
- Win the voters back
- Promised to back the bill

POS Tagging

Words often have more than one POS:

- The back door
- On my back
- Win the voters back
- Promised to back the bill



The **POS tagging problem** is to determine the POS tag for a particular instance of a word.

POS Tagging


- **Input:** Plays well with others
- **Ambiguity:** NNS/VBZ UH/JJ/NN/RB IN NNS
- **Output:** Plays/VBZ well/RB with/IN others/NNS


Penn Treebank Tag-set


Sentiment Analysis


**GoPro HERO5 Black**
\$399.00  In Stock. Ships from and sold by Amazon.com. Gift-wrap available.


Customer Reviews
★★★★☆ 95
3.9 out of 5 stars

5 star  59%

4 star  7%

3 star  10%

2 star  6%

1 star  18%

Share your thoughts with other customers

Write a customer review

See all verified purchase reviews

Top Customer Reviews

★★★★★ **Awesome Upgrade**
By Devin Stephens **WALL OF JAM** **TOP 50 REVIEWER** on October 4, 2016
Style Name: Basic

First Impressions: This review will be updated as I get more of a chance to play with this camera. I received it Oct 4th so still plenty of playing around to do! So far I have added 1 additional con, updated some of the pros, and added a few pros... I will continue to update!

Short Review: I have owned GoPro cameras since the Hero 2. I have recently been introduced to other brands that work just as well but I am loyal to GoPro and had to try out the new hero 5. I am glad I did because GoPro is now on top of their game again and this is now the best Action Camera I have ever owned... I used to hate all in one designs (Waterproofing being built into the camera) but I actually feel it works very well with this camera and provides much greater protection! No more "Naked" GoPros getting broken! (I have broke a few)

While the Hero 5 doesn't really have any major improvements in resolution or shooting specs, the improvements that have been made in design and software actually work well and really make the photo and videos appear much better! The most notable improvements include built in stabilization that works very well and much better audio! Audio is where action cameras have always been lacking, but GoPro has made a huge improvement here. GoPro has also improved their user interface which is a great improvement. Old models took a little while to learn, but not the Hero 5. I haven't seen an action camera this easy to use ever! So even though GoPro has carried over the same sensor from the Hero 4, this camera is very much worth the improvements

Pros:

- 1.) Video Quality - Is outstanding. The software changes they made, actually make the video quality look much better even though the same sensor was carried over

2. [Read more](#)

21 [Comments](#) | 177 people found this helpful. Was this review helpful to you? [Report abuse](#)

★★★★☆ **All the Pros, Cons and the Oks for this GoPro (H5)**
By **Honest Reviewer** on October 3, 2016
Style Name: Basic


Pros:

- Waterproof all the way. All ports are covered so there is no fear of them being submerged in the dirt.
- Design itself has smoother edges and overall has a professional feel.
- Voice command that beeps when you speak one of the 12 commands. It also takes into account your accent if you're in the US, UK, AUS, etc.

Get XFINITY® Internet
No term contract

XFINITY Internet
Not started at
\$19.99 a month / 12 months [Switch from X](#)

[xfinity](#)

At feedback 

Most Recent Customer Reviews

★★★★★ **Videos are great, photos are ok**
Videos are great, photos are ok. Took it to Zion National Park, was able to get some great videos of my hikes and took a ton of pictures. [Read more](#)
Published 12 hours ago by David Fattal


★★★★★ **Buy it.**
Awesome product, easy transaction, no worries. Funny part about it was the entire box was in Spanish but provided English directions inside. Not I even read the book. [Read more](#)
Published 14 hours ago by Jeremy T.

★★★★★ **so small yet so powerful**
so small yet so powerful. I love this little thing. I also got a GoPro Jawz Clamp Mount that works really well together. this is my first gopro and I am very impressed. [Read more](#)
Published 18 hours ago by Molloy

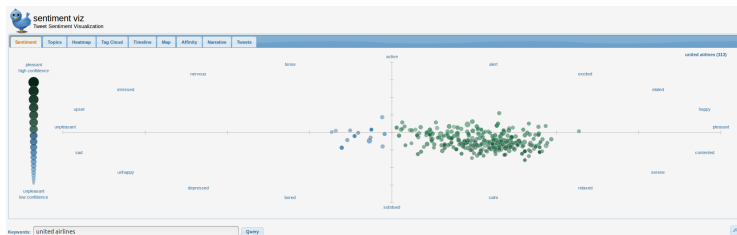
★★★★☆ **Single shot shutter too slow**
Single shot shutter speed too slow. 1/30 sec shutter speed is too slow for hand held portraits let alone sports and action shots. minimum is 1/60th and best is 1/125 or even 1/250. [Read more](#)
Published 19 hours ago by T. Rogers

★★★★☆ **Horrible connectivity, awful tech support**
Can't connect to one of the latest and most popular android phones, despite hours of different tries, getting a ridiculously horrible, clueless technical support from gopro. [Read more](#)
Published 1 day ago by Guy Z.

★★★★☆ **Total disappointment for a longtime Goopro Fanboy**
I have been hyped about the gopro5, yet I ended up in total disappointment. My gf bought me the camera as a gift. [Read more](#)
Published 1 day ago by Edward Sheng



Sentiment Analysis



- <https://www.nltk.org/howto/sentiment.html>
- <https://nlp.stanford.edu/sentiment/>
- <https://textblob.readthedocs.io/en/dev/>

Sentiment analysis has many other names

- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis

Sentiment Analysis

Sentiment analysis is the detection of **attitudes**

- “enduring, affectively colored beliefs, dispositions towards objects or persons”

Attitudes

- **Holder** (source) of attitude
- **Target** (aspect) of attitude

Attitudes

- **Holder** (source) of attitude
- **Target** (aspect) of attitude
- **Type** of attitude
 - From a set of types:
Like, love, hate, value, desire, etc.
 - Or (more commonly) simple weighted polarity:
positive, negative, neutral, together with strength

Attitudes

- **Holder** (source) of attitude
- **Target** (aspect) of attitude
- **Type** of attitude
 - From a set of types:
Like, love, hate, value, desire, etc.
 - Or (more commonly) simple weighted polarity:
positive, negative, neutral, together with strength
- **Text** containing the attitude
 - Sentence or entire document

Sentiment analysis

- **Simplest task:**

Is the attitude of this text positive or negative?

Sentiment analysis

- **Simplest task:**

Is the attitude of this text positive or negative?

- **More complex:**

Rank the attitude of this text from 1 to 5

Sentiment analysis

- **Simplest task:**

Is the attitude of this text positive or negative?

- **More complex:**

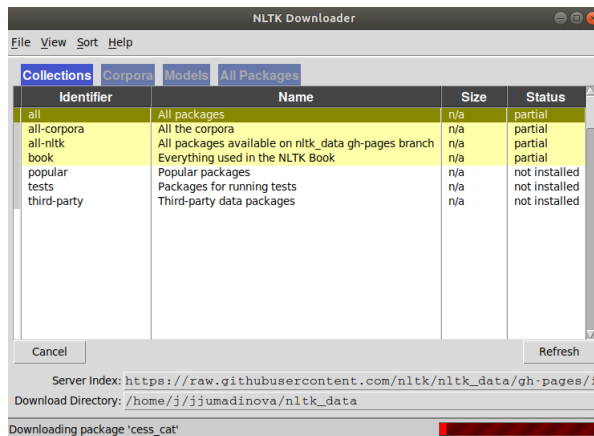
Rank the attitude of this text from 1 to 5

- **Advanced:**

Detect the target, source, or complex attitude types

NLTK

```
$ python3  
$ import nltk  
$ nltk.download()
```



NLTK Basic Pre-Processing

Tokenize using Python

- ① `urllib` module to crawl the webpage
- ② BeautifulSoup to clean the text with html tags
- ③ convert text into tokens using `split()` function

NLTK Basic Pre-Processing

Tokenize using Python

- ① `urllib` module to crawl the webpage
- ② BeautifulSoup to clean the text with html tags
- ③ convert text into tokens using `split()` function

Remove Stop Words

- ① get english stop words from `nlTK`
- ② remove stop words before plotting

NLTK Basic Pre-Processing

Tokenize using Python

- ① `urllib` module to crawl the webpage
- ② `BeautifulSoup` to clean the text with html tags
- ③ convert text into tokens using `split()` function

Remove Stop Words

- ① get english stop words from `nlk`
- ② remove stop words before plotting

Frequency Analysis

- ① `nlk's FreqDist` to calculate the frequency distribution
- ② `plot` function to produce a graph