# Data Scraping for Natural Language Processing

Brought to you by:

smartcademy

# Introduction


# Singapore’s leading training provider for in-demand tech skills & career transformation

# Courses We Offer

1. Data Analytics
2. Digital Marketing
3. User Experience (UX) Design
4. Web/Mobile App Development

# Companies that **hire** our graduates











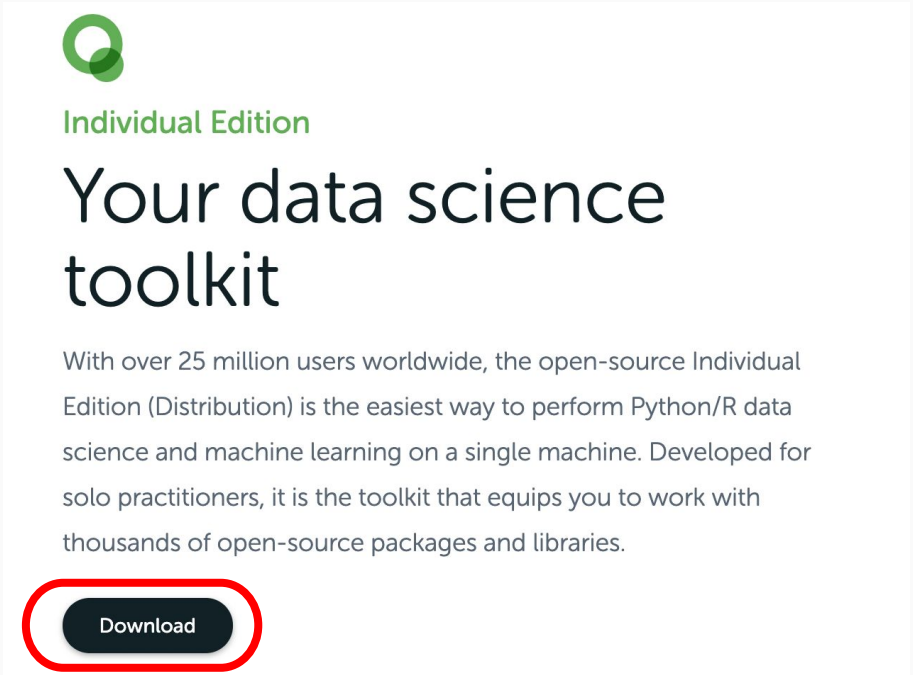

# Impacted

# > 6000+

# Agenda

1. Introduction to AI and NLP
2. Quick Recap on Key Data Scraping Concepts
3. Hands-on!

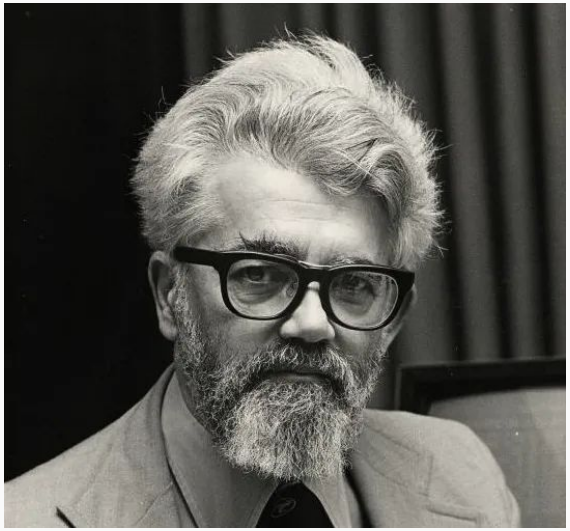
# Using Python




Visit https://www.anaconda.com/products/individual
to download a Python distribution / data science “platform”

# What is AI?

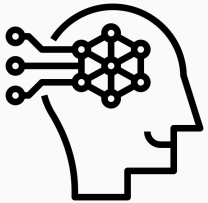
# Artificial Intelligence is the science and engineering of making intelligent machines



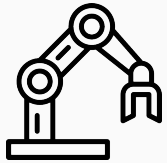
John Mccarthy

Natural Language Processing

Computer Vision

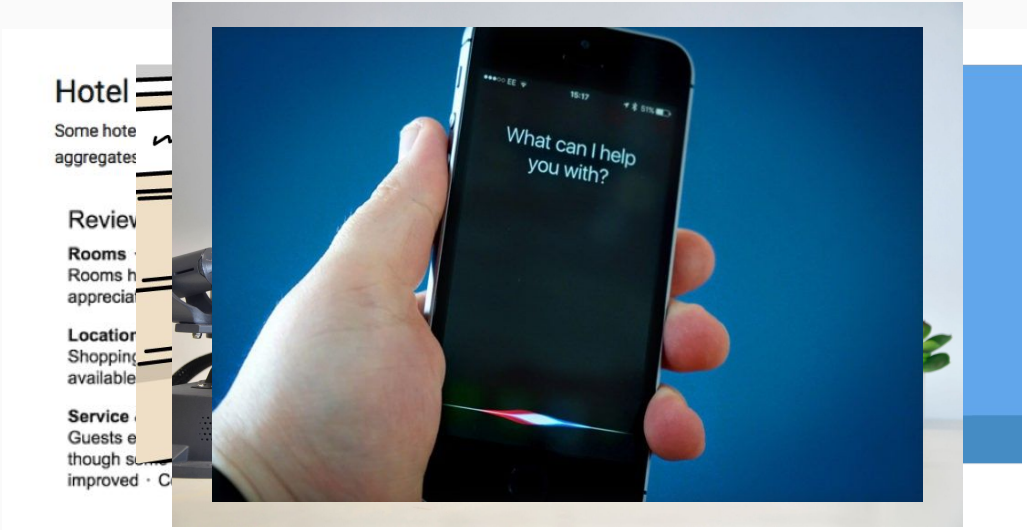


Machine Learning

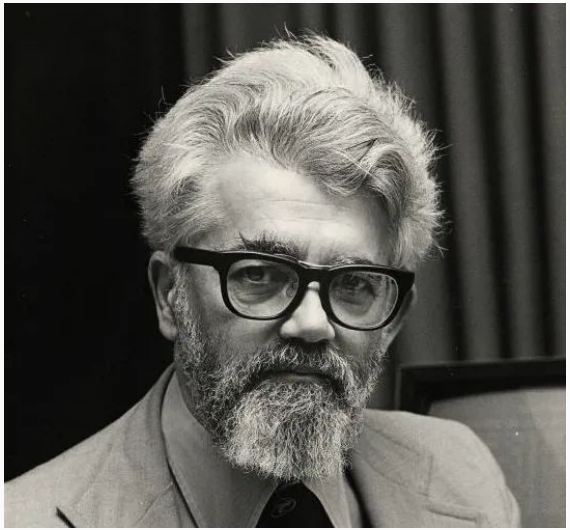


Automation & Robotics

Speech



# AI is everywhere



John Mccarthy

Natural Language Processing

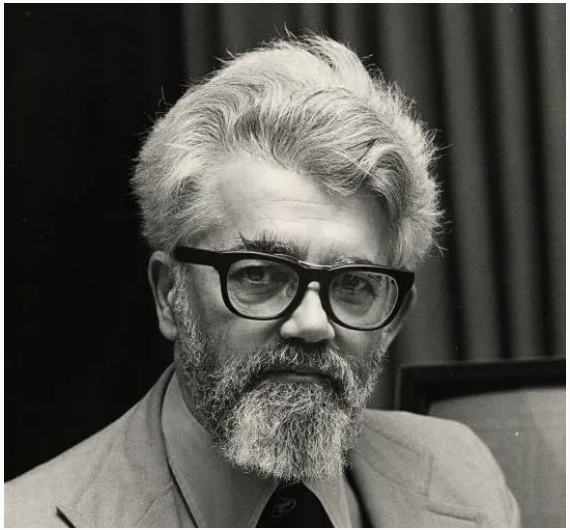
Computer Vision

Machine Learning

Automation & Robotics

Speech

# AI is everywhere



**John Mccarthy**

Aa

**Natural Language Processing**

ENGLISH

CHINESE

Natural Language Processing

PYTHON

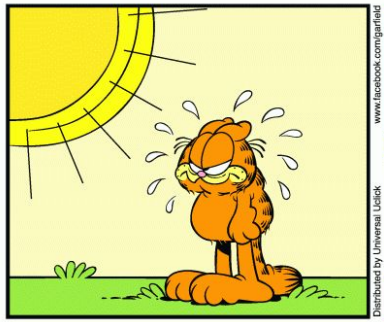
JAVA

# Natural Language Processing

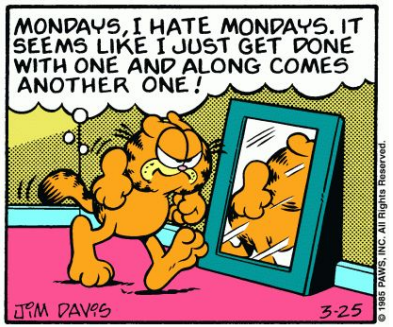
# Natural Language Processing

Garfield was trying to stay cool

GARFIELD WAS TRYING TO STAY COOL



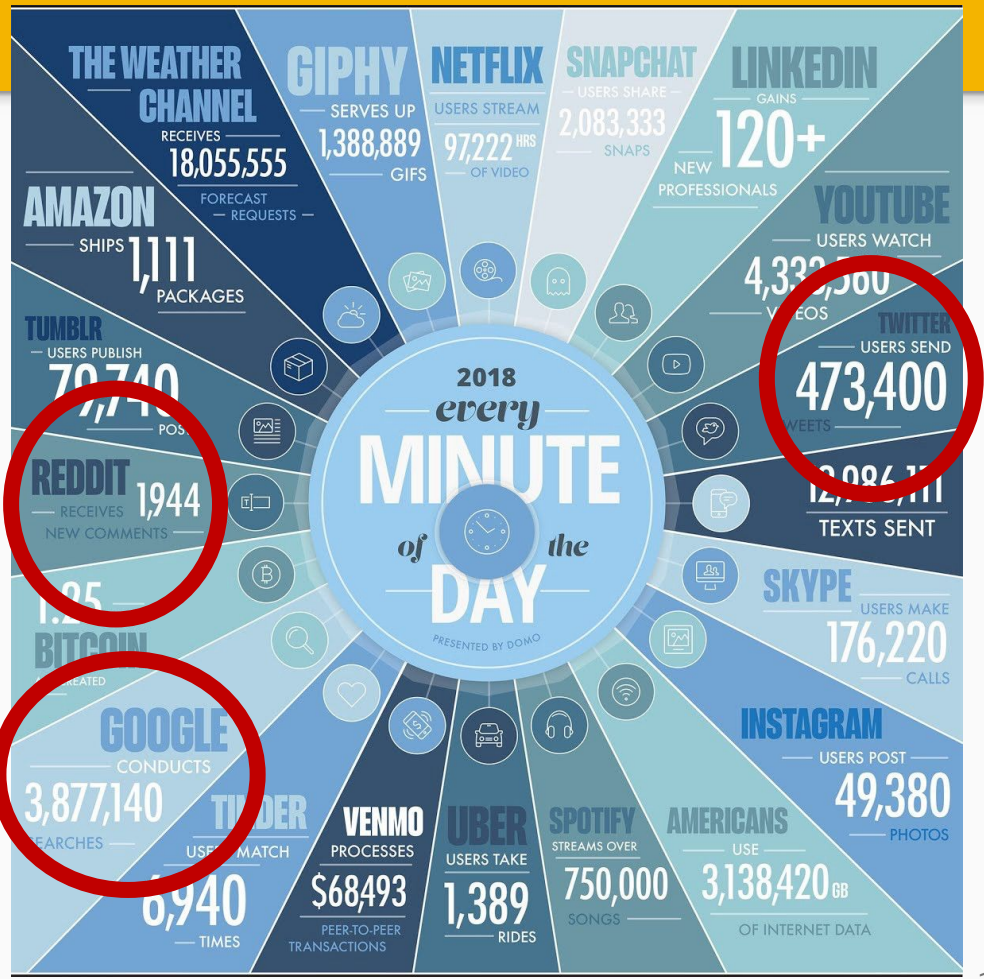

GARFIELD WAS TRYING TO STAY COOL




GARFIELD WAS TRYING TO STAY COOL

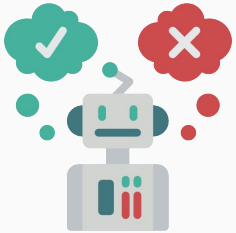
# WHY

- Natural Language
- Convey information between 2 people
- Structured Vs Unstructured Data
- NLP is the interdisciplinary field combining computer science and linguistics



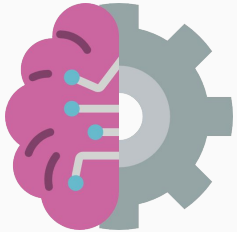

Source:

# Natural Language Processing - NLP



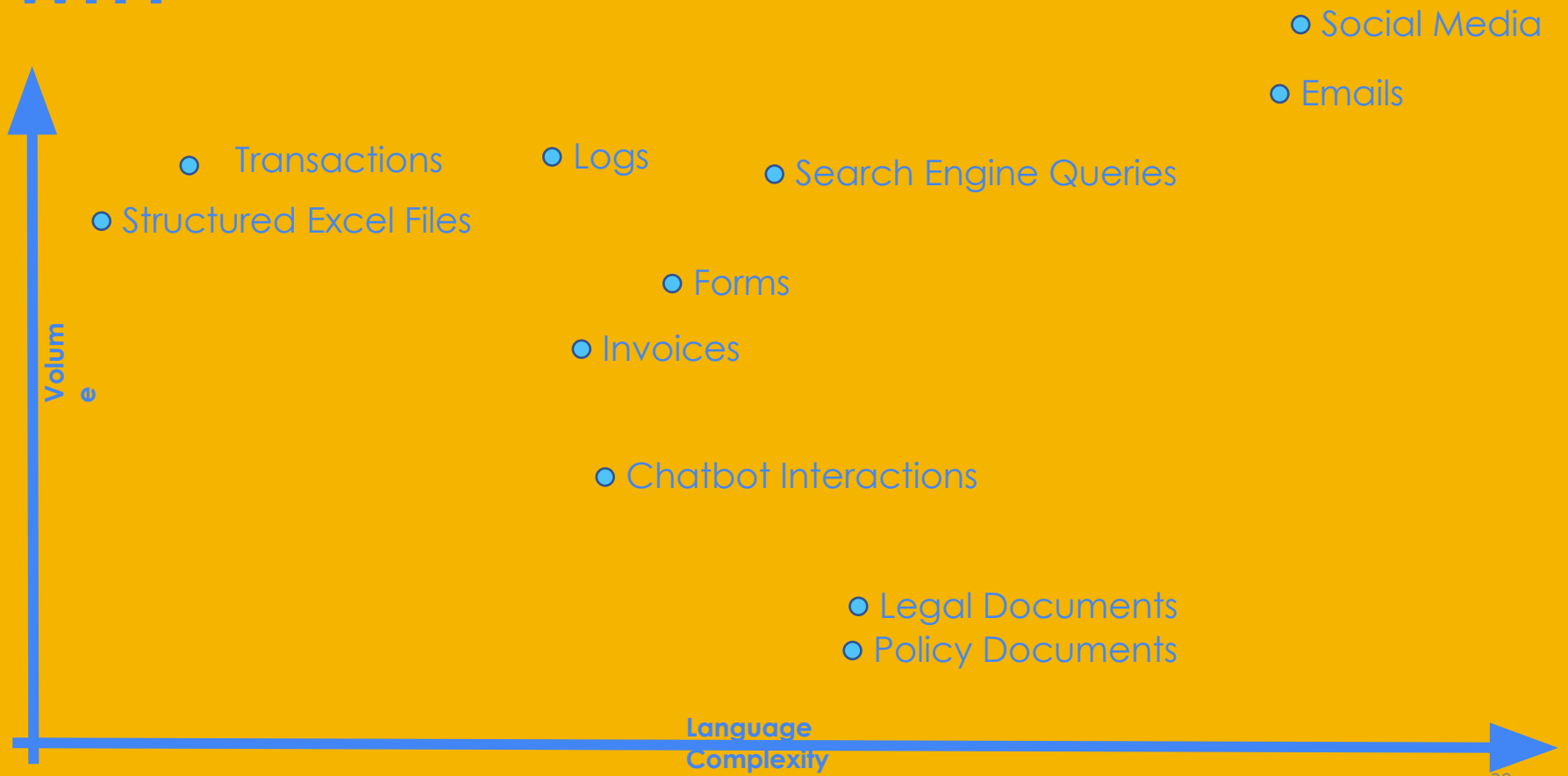
- MACHINE CAN INTERPRET HUMAN LANGUAGE
  - Facilitates the Human Machine Interaction
  - Enables the Machine to Machine Interaction

- DATA DRIVEN AND KNOWLEDGE DRIVEN
  - Machine Learning for data classification and generation
  - Semantic reasoning for data discovery and disambiguation



- SIMULATING HUMAN BRAIN
  - Current models performs well at individual task, still needs improvements for multiple tasks

# WHY


Social Media
Emails
Transactions
Logs
Search Engine Queries
Structured Excel Files
Forms
Invoices
Chatbot Interactions
Legal Documents
Policy Documents
Volume
Language Complexity

# Applications of NLP

1. Sentiment analysis
2. Chatbot
3. Speech recognition
4. Language Translation
5. Information retrieval/extraction
6. Advertisement matching

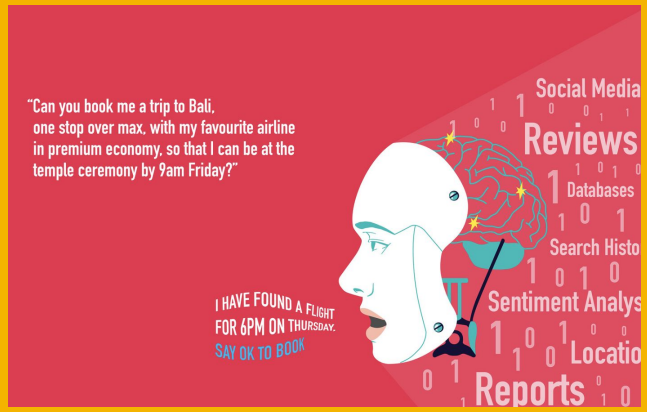
# Applications of NLP

1. Sentiment analysis
2. Chatbot
3. Speech recognition
4. Language Translation
5. Information retrieval/extraction
6. Advertisement matching

# Nordstrom digs into 5-star customer reviews and finds a shipping problem.

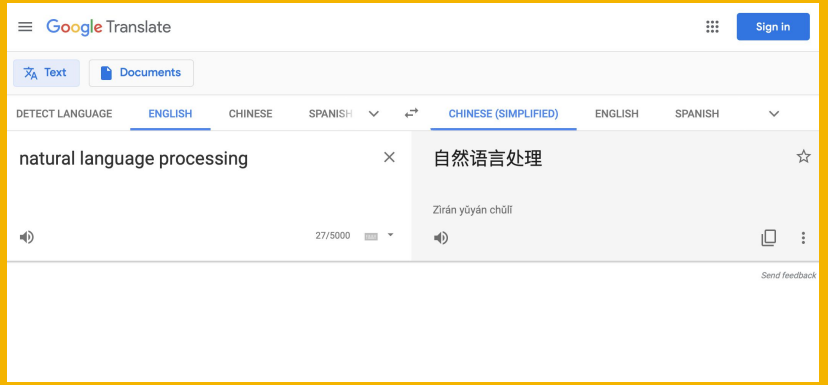
# Applications of NLP

1. Sentiment analysis
2. Chatbot
3. Speech recognition
4. Language Translation
5. Information retrieval/extraction
6. Advertisement matching



# Applications of NLP

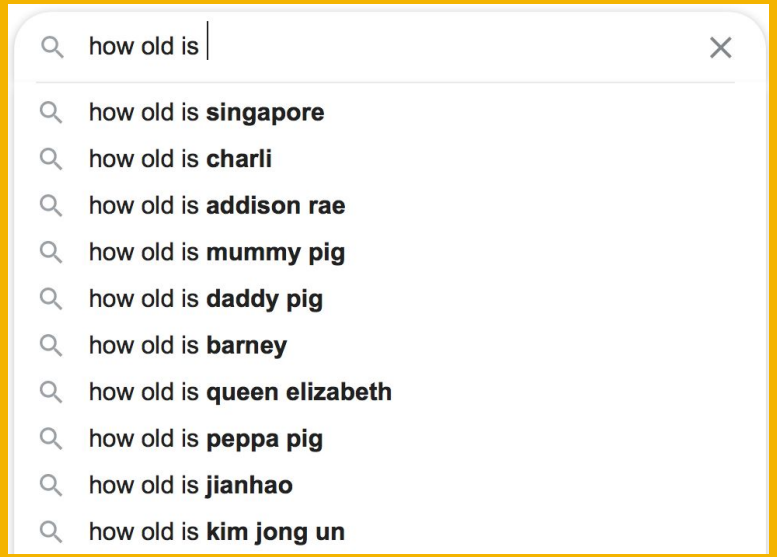
1. Sentiment analysis
2. Chatbot
3. Speech recognition
4. Language Translation
5. Information retrieval/extraction
6. Advertisement matching

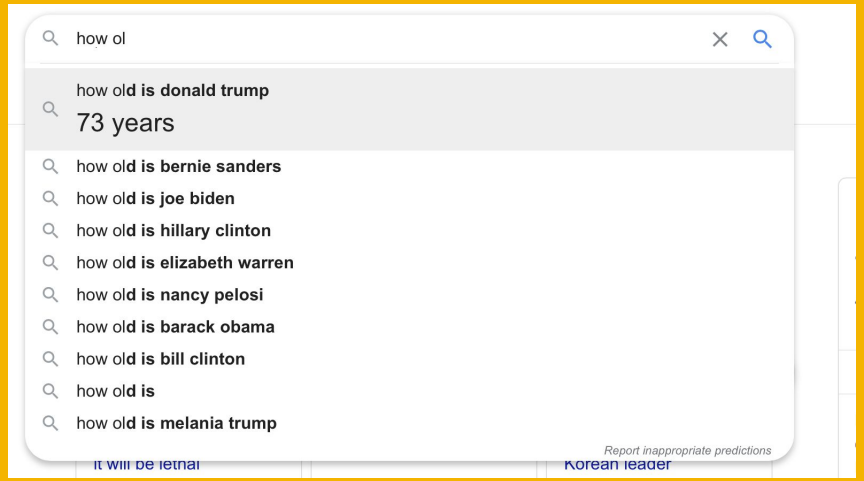


# Applications of NLP

1. Sentiment analysis
2. Chatbot
3. Speech recognition
4. Language Translation
5. Information retrieval/extraction
6. Advertisement matching

# E.g. Semantic search engine



# Walmart's semantic search engine increased conversion rates by 10-15%

# Applications of NLP

1. Sentiment analysis
2. Chatbot
3. Speech recognition
4. Language Translation
5. Information retrieval/extraction
6. Advertisement matching

# Natural Language Understanding

# TEXT NORMALISATION

**She bought 10 apples and 10 oranges from the nearby grocer .**

- CONVERTING ALL LETTERS TO LOWER OR UPPER CASE

**she bought 10 apples and 10 oranges from the nearby grocer .**

- CONVERTING NUMBERS INTO WORDS OR REMOVING NUMBERS

**she bought apples and oranges from the nearby grocer .**

- REMOVING PUNCTUATIONS, ACCENT MARKS AND OTHER DIACRITICS

**she bought apples and oranges from the nearby grocer**

- REMOVING WHITE SPACES

**she bought apples and oranges from the nearby grocer**

- REMOVING STOP WORDS, AND PARTICULAR WORDS

**bought apples oranges nearby grocer**

You can add your own Stop word. Go to your NLTK download **directory path** -> **corpora** -> **stopwords** -> update the stop word **file** depends on your language which one you are using. Here we are using english (stopwords.words('english')).

## PRE-PROCESSING

**She bought 10 red apples and 10 cans of coca cola from the nearby grocer.**

---

TOKENISATION

"bought" "red" "apples" "cans" "coca" "cola" "nearby" "grocer"

N-GRAMS

"red apples" "coca cola" "nearby grocer"

STEMMING

"bought" "appl" "can" "coca" "cola" "nearbi" "grocer"

PART OF SPEECH (POS) TAGGING

[('She', 'PRP'), ('bought', 'VBD'), ('10', 'CD'), ('apples', 'NNS'), ('and', 'CC'), ('10', 'CD'), ('cans', 'NNS'), ('of', 'IN'), ('coca', 'NN'), ('cola', 'NN'), ('from', 'IN'), ('the', 'DT'), ('nearby', 'JJ'), ('grocer', 'NN')]

NAMED ENTITY RECOGNITION

(S She/PRP bought/VBD 10/CD apples/NNS and/CC 10/CD cans/NNS of/IN (NP coca/NN) (NP cola/NN) from/IN (NP the/DT nearby/JJ grocer/NN))

# Tokenisation

Taking a text or set of text and breaking it up into its individual tokens (sentences, words, characters)

**She bought 10 red apples and 10 cans of coca cola from the nearby grocer.**

TOKENISATION **“bought” “red” “apples” “cans” “coca” “cola” “nearby” “grocer”**

- New York, Los Angeles, Singapore Management University

- **Language specific:**

Chinese: 地铁站
French: L'ensemble

- **Context is often missing: “can”**

# N-GRAMS

Sequence of N words, good for putting keywords into local context

**bought red apples cans coca cola nearby grocer**

**NGRAMS** "bought red" "red apples" "apples can" "coca cola" "nearby grocer"

**BIGRAMS** "Coca cola"

**TRIGRAMS** The Three Musketeers

**4-GRAMS** National University of Singapore

**5-GRAMS** etc

- Compression algorithms (the PPM variety especially) where the length of the grams depends on how much data is available for providing specific contexts.
- Approximate string matching (e.g. BLAST for genetic sequence matching)
- Predictive models (e.g. name generators)
- Speech recognition (phonemes grams are used to help evaluate the likelihood of possibilities for the current phoneme undergoing recognition)

# STEMMING & LEMMATISATION

Reduce inflectional forms and sometimes derivationally related forms of a word to a **common base form, to bring variant forms of a word together**

**She bought 10 red apples and 10 oranges from the nearby grocer.**

**STEM** “bought” “appl” “orang” “nearbi” “grocer”

**LEMMATIZE** “buy” “apple” “orange” “nearby” “grocer”

**SUFFIX**
**-ing**
**-ed**
**-es**
**-s**
**…**

```
application
 Stemming: applic Lemmatizing: application
applying
 Stemming: appli Lemmatizing: apply
applies
 Stemming: appli Lemmatizing: apply
applied
 Stemming: appli Lemmatizing: apply
apply
 Stemming: appli Lemmatizing: apply
apples
 Stemming: appl Lemmatizing: apples
apple
 Stemming: appl Lemmatizing: apple
```

**Porter**: Most commonly used stemmer, and provides Java support.

**Snowball**: Improvement over the Porter algorithm, even Porter admits it is better than his original algorithm. Slightly faster computation time than porter, with a fairly large community around it.
To view the entire algorithm: http://people.scs.carleton.ca/~armyunis/projects/KAPI/porter.pdf

# PART OF SPEECH TAGGING

Marking up a word in a corpus to a corresponding part of a speech tag, based on its context and definition

I **left** my keys in my **left** pocket.

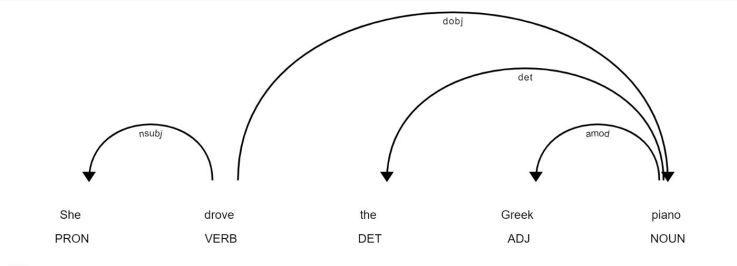
**PART OF SPEECH (POS) TAGGING**

[('I', 'PRP'), **('left', 'VBD'),** ('my', 'PRP$'), ('keys', 'NNS'), ('in', 'IN'), ('my', 'PRP$'), **('left', 'JJ'),** ('pocket', 'NN')]

Left - VBD verb, past tense
took

Left - JJ adjective

Building parse trees, which are used in building Named Entity Recognisers and extracting relations between words, helps in Syntactic and semantic analysis




Types:

1. Lexical Based Methods
2. Rule-Based Methods
3. Probabilistic Methods
4. Deep Learning Methods

# NAMED ENTITY Recognition

Identify all textual mentions of the named entities and classify them into pre-defined categories

**She bought 10 red apples and 10 cans of coca cola from the nearby grocer.**

**NAMED ENTITY RECOGNITION**

**(S She/PRP bought/VBD 10/CD apples/NNS and/CC 10/CD cans/NNS of/IN (NP coca/NN) (NP cola/NN) from/IN (NP the/DT nearby/JJ grocer/NN))**

**Stanford's Named Entity Recognizer** is based on an implementation of linear chain Conditional Random Field (CRF) sequence models. Model is only trained on instances of ***PERSON***, ***ORGANIZATION*** and ***LOCATION*** types.
Based on training data, the model will support different types of entities: https://spacy.io/api/annotation#section-named-entities

| Samples of Pre-defined categories | Examples |
|---|---|
| Names of people | Joan, Jeremy, Adam |
| Organisations | Accenture, Apple, GoJek |
| Locations | City Hall, Mount Fuji, |
| Expressions of times | June, 1980, 2008-03-10 |
| Percent | 100%, Twenty pct, |
| Monetary value | 18 Euros, $19, 600 Yen |

Each POS tag is attached to a single word, while NER tags can be attached to multiple words.

https://towardsdatascience.com/a-comparison-between-spacy-ner-stanford-ner-using-all-us-city-names-c4b6a54

# PRE-PROCESSING

**She bought 10 red apples and 10 cans of coca cola from the nearby grocer.**

| | |
|---|---|
| TOKENISATION | “bought” “red” “apples” “cans” “coca” “cola” “nearby” “grocer” |
| N-GRAMS | “red apples” “coca cola” “nearby grocer” |
| STEMMING | “bought” “appl” “can” “coca” “cola” “nearbi” “grocer” |
| PART OF SPEECH (POS) TAGGING | [('She', 'PRP'), ('bought', 'VBD'), ('10', 'CD'), ('apples', 'NNS'), ('and', 'CC'), ('10', 'CD'), ('cans', 'NNS'), ('of', 'IN'), ('coca', 'NN'), ('cola', 'NN'), ('from', 'IN'), ('the', 'DT'), ('nearby', 'JJ'), ('grocer', 'NN')] |
| NAMED ENTITY RECOGNITION | (S She/PRP bought/VBD 10/CD apples/NNS and/CC 10/CD cans/NNS of/IN (NP coca/NN) (NP cola/NN) from/IN (NP the/DT nearby/JJ grocer/NN)) |

# DOCUMENT TERM MATRIX

1

**ORIGINAL STATEMENT**

D1: Natural language processing is fun!
D2: Natural language processing is not fun!
D3: Drinking beer is fun!

2

**PROCESSED STATEMENT**

D1: natur languag process fun
D2: natur languag process fun
D3: drink beer fun

3

**VECTOR OUTPUT**

| | natur | languag | process | fun | drink | beer |
|---|---|---|---|---|---|---|
| D1 | 1 | 1 | 1 | 1 | | |
| D2 | 1 | 1 | 1 | 1 | | |
| D3 | | | | 1 | 1 | 1 |

Final vectors:
D1: (1,1,1,1,0,0)
D2: (1,1,1,1,0,0)
D3: (0,0,0,1,1,1)

# TERM FREQUENCY VS. TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY

- TERM FREQUENCY (TF)
  - Frequency of the term in the document
  - i.e. if the word appears twice, the frequency in the vector will be 2

- TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY (TF-IDF)
  - Words that appear across multiple documents are less important (less discriminative)
  - Give higher weightage to words that appear less
  - $IDF(W) = log\frac{N}{df(W)}$
  - N = Number of documents
  - df(W) = Number of documents the word appears in
  - $TF - IDF\ (W) = TF(W) \times IDF(W)$

$IDF(W) = log\frac{100}{20}$

$TF - IDF\ (W) = 25 \times log\frac{100}{20}$

100 movie reviews
20 on movie reviews
'Avengers' □ 25 times

# Getting Data

What are the possible sources of data?

# Possible Sources of Data

1. Primary Sources
   - Interviews
   - Surveys
   - Focus Groups
   - Sensors
     - Machines
     - Wearables
     - IoT
   - ...

2. Secondary Sources
   - Online Reviews
   - Online Comments
   - Social Media
   - Websites
   - ...

# Ctrl C Ctrl V

Small scale web scraping

WIKIPEDIA
The Free Encyclopedia

Not logged in Talk Contributions Create account Log in

Article Talk | Read View source View history | Search Wikipedia

# NASA

From Wikipedia, the free encyclopedia

Coordinates: 38°52′59″N 77°0′59″W

*For other uses, see NASA (disambiguation).*

The **National Aeronautics and Space Administration** (**NASA** /ˈnæsə/) is an independent agency of the U.S. federal government responsible for the civilian space program, as well as aeronautics and space research.[note 1]

NASA was established in 1958, succeeding the National Advisory Committee for Aeronautics (NACA). The new agency was to have a distinctly civilian orientation, encouraging peaceful applications in space science.[7][8][9] Since its establishment, most US space exploration efforts have been led by NASA, including the Apollo Moon landing missions, the Skylab space station, and later the Space Shuttle. NASA is supporting the International Space Station and is overseeing the development of the Orion spacecraft, the Space Launch System, Commercial Crew vehicles, and the planned Lunar Gateway space station. The agency is also responsible for the Launch Services Program, which provides oversight of launch operations and countdown management for uncrewed NASA launches.

NASA's science is focused on better understanding Earth through the Earth Observing System;[10] advancing heliophysics through the efforts of the Science Mission Directorate's Heliophysics Research Program;[11] exploring bodies throughout the Solar System with advanced robotic spacecraft such as *New Horizons*;[12] and researching astrophysics topics, such as the Big Bang, through the Great Observatories and associated programs.[13]

**Contents** [hide]

1 History

**National Aeronautics and Space Administration**


NASA seal

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

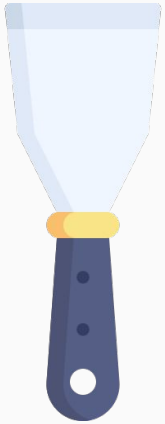
Contribute

Help
Learn to edit
Community portal
Recent changes
Upload file

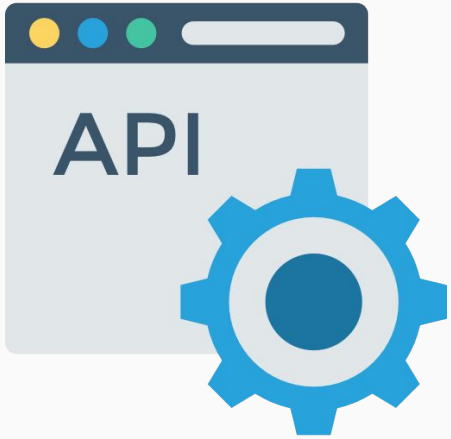
Tools

What links here
Related changes
Special pages
Permanent link

# Obtaining data from a webpage

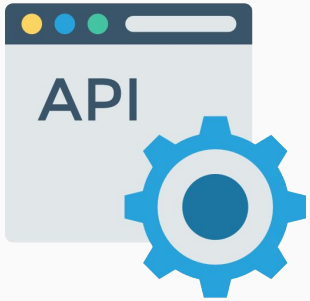


Web Scraper
Extract data from the website using a web scraping tool or software




Application Programming Interface (API)
Website gives you direct access to specific data

# Webscraper Vs API




| Webscraper | API |
| --- | --- |
| Done manually or by using software tools | Website needs to provide the API |
| May be illegal on some sites | Legal, but not all sites may provide |
| Can be programmed to collect the data you want | May not have all the data |
| F.O.C if done manually, but software tools may have a license | May require payment |

# Is Web Scraping Legal?

## Is Web Scraping and Crawling Legal in Singapore?

There is no specific law in Singapore that directly addresses whether web crawling or scraping is legal. However, crawling and scraping could possibly attract civil liability under existing contract law and copyright law, and even criminal liability under the Computer Misuse Act.

### Breach of website terms of use

When accessing a website, the user generally agrees to access it in accordance with the website's "terms of use" as stated on the site, and this forms a legally binding agreement.

If the terms of use prohibit crawling and scraping, doing so on the website would be a breach of this agreement. The website owner may then sue the bot operator for breach of contract, and claim monetary compensation for any loss suffered in relation to any downtime or slowdown the website may have incurred.

For example, when property listing platform 99.co first started operations, it used a web scraper to scrape rental listings from competing platform PropertyGuru for listing on its own platform. PropertyGuru regarded this as potential breaches of its website's Terms of Service and Acceptable Use Policy and infringement of its copyright in the listings, and requested 99.co to stop doing so.

Both platforms eventually settled the matter out of court in 2015, with 99.co signing an agreement not to substantially reproduce the content in PropertyGuru's website without its consent.

https://singaporelegaladvice.com/law-articles/legal-scrape-crawl-websites-data-singapore/

# Is Web Scraping Legal?


**The People's Court of Suiyang District in Central Henan Province imprisoned the Chinese software developer and his employer for three years in prison and a $70,260 fine (450,000 Yuan). the marketer used web scraping software to access data that was not publicly available**

# Is Web Scraping Legal?

POLICY \ US & WORLD \ TECH \

## Clearview AI hit with sweeping legal complaints over controversial face scraping in Europe

*The privacy watchdogs believe Clearview's image-scraping methods violate European laws*

By Ian Carlos Campbell | @soupsthename | May 27, 2021, 5:48am EDT

**The complaints filed in France, Austria, Greece, Italy, and the United Kingdom say that the company's method of documenting and collecting data — including images of faces it automatically extracts from public websites — violates European privacy laws. New York-based Clearview claims to have built "the largest known database of 3+ billion facial images."**

# How do we scrape data?

1. Extract the HTML
2. Parse the HTML
3. Extract the relevant the necessary data
4. Store/Transform the data

# Hands-on!