JAMES COOK UNIVERSITY
AUSTRALIA

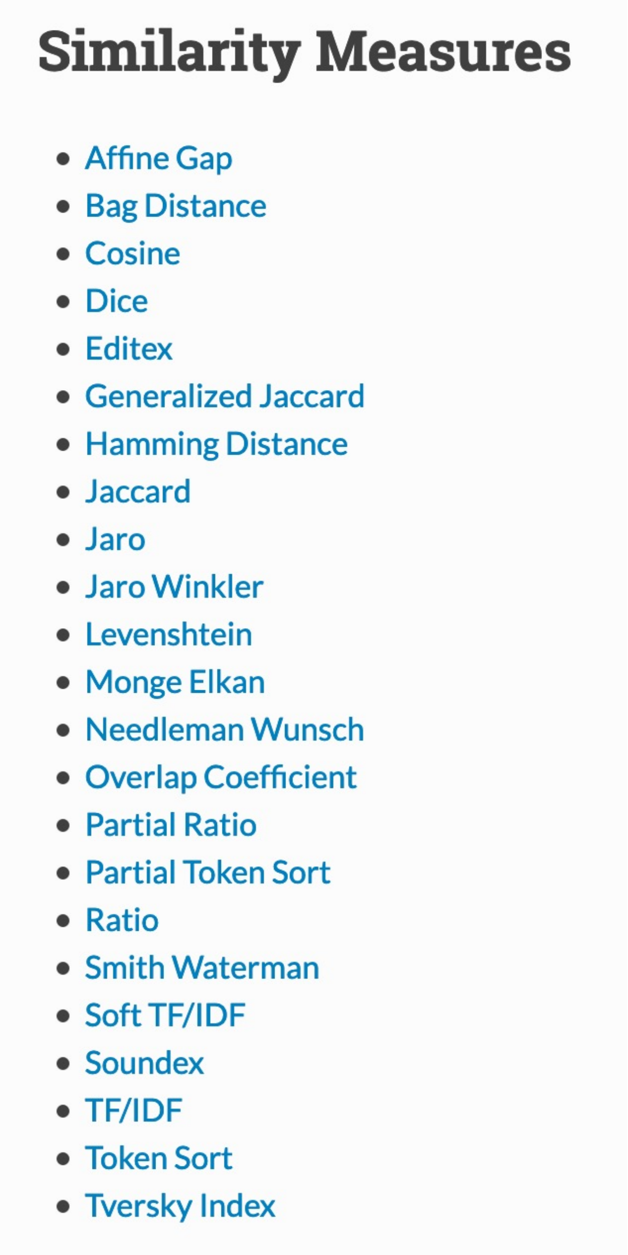
Week 4
MA5851 – Data Science Master Class 1
[Natural Language Processing]

Dr Mostafa Shaikh
mostafa.shaikh@jcu.edu.au

online.jcu.edu.au

Cairns
Singapore
Townsville

Text Similarity



Similarity Measures

- Affine Gap
- Bag Distance
- Cosine
- Dice
- Editex
- Generalized Jaccard
- Hamming Distance
- Jaccard
- Jaro
- Jaro Winkler
- Levenshtein
- Monge Elkan
- Needleman Wunsch
- Overlap Coefficient
- Partial Ratio
- Partial Token Sort
- Ratio
- Smith Waterman
- Soft TF/IDF
- Soundex
- TF/IDF
- Token Sort
- Tversky Index

https://anhaidgroup.github.io/py_stringmatching/v0.3.x/SimilarityMeasure.html

Topic

- ✓Information Extraction
  - ✓Chunking
  - ✓Chinking
- ✓Web Scraping – scapy and selenium web driver
- ✓BeautifulSoup
- ✓Demonstration – Similarity, Chunking, FOE data extraction
- ✓Demonstration – SLP 2
- ✓Demonstration – SLP 3

Information Extraction (IE)

- Identify specific pieces of information (data) in unstructured or semi-structured textual document.
- Transform unstructured information in a corpus of documents or web pages into a structured database.
- Applied to different types of text:
  - Newspaper articles
  - Web pages
  - Scientific articles
  - Newsgroup messages
  - Classified ads
  - Medical notes
  - Wikipedia (info boxes)..
- Focused on extracting information from news articles:
  - Terrorist events
  - Industrial joint ventures
  - Company management changes
- Information extraction of particular interest to the intelligence community (CIA, NSA).

Application Tasks of NLP

(1)Information Retrieval/Detection

To search and retrieve documents in response to queries for information

(2)Passage Retrieval

To search and retrieve part of documents in response to queries for information

(3)Information Extraction

To extract information that fits pre-defined database schemas or templates, specifying the output formats

(4) Question/Answering Tasks

To answer general questions by using texts as knowledgebase: Fact retrieval, combination of IR and IE

(5)Text Understanding

To understand texts as people do: Artificial Intelligence

Example of IE

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Acknowledgement: Prof. Tsuji; the University of Tokyo

Example of IE

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and “metal wood” clubs a month.

TIE-UP-1

Relationship: TIE-UP

Entities: “Bridgestone Sport Co.”

“a local concern”

“a Japanese trading house”

Joint Venture Company:

“Bridgestone Sports Taiwan Co.”

Activity: ACTIVITY-1

Amount: NT$200000000

ACTIVITY-1

Activity: PRODUCTION

Company:

“Bridgestone Sports Taiwan Co.”

Product:

“iron and ‘metal wood’ clubs”

Start Date:

DURING: January 1990

Example of IE

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Example of IE

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Example of IE

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

How it works!

Based on finite states automata (FSA)

set up
new Twaiwan dallors

1.Complex Words:
Recognition of multi-words and proper names

a Japanese trading house
had set up

2.Basic Phrases:
Simple noun groups, verb groups and particles

production of
20, 000 iron and
metal wood clubs

3.Complex phrases:
Complex noun groups and verb groups

[company]
[set up]
[Joint-Venture]
with
[company]

4.Domain Events:
Patterns for events of interest to the application
Basic templates are to be built.

5. Merging Structures:
Templates from different parts of the texts are merged if they provide information about the same entity or event.

Example of IE

Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and "metal wood" clubs a month.

TIE-UP-1
Relationship: TIE-UP
Entities: "Bridgestone Sport Co."
"a local concern"
"a Japanese trading house"
Joint Venture Company:
"Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
"Bridgestone Sports Taiwan Co."
Product:
"iron and 'metal wood' clubs"
Start Date:
DURING: January 1990

Example of IE

A German vehicle-firm executive was stabbed to death ….

……….

Jurgen Pfrang, 51, reportedly stumbled upon the robbers on the second floor of his Nanjing home early on Sunday.
The deputy general manager of Yaxing Benz, a Sino-German joint venture that makes buses and bus chassis in nearby Yangzhou, was hacked to death with 45 cm watermelon knives.

……….

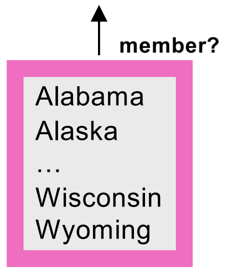
Crime-Type: Murder
Type: Stabbing
The killed: Name: Jurgen Pfrang
Age: 51
Profession: Deputy general manager
Location: Nanjing, China

Different template
for crimes

IE Techniques : Models

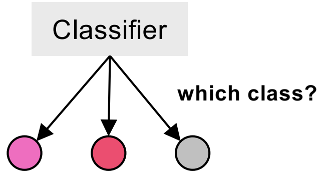
Lexicons

Abraham Lincoln was born in Kentucky.



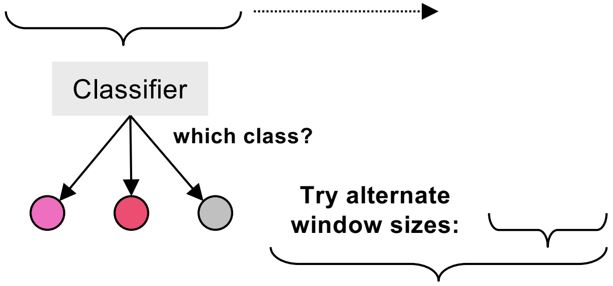
Classify Pre-segmented Candidates

Abraham Lincoln was born in Kentucky.

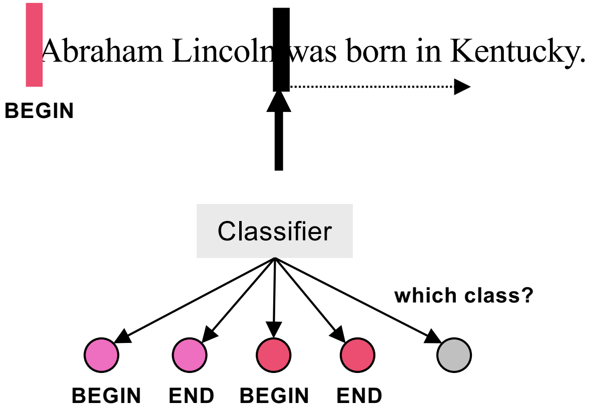


Sliding Window

Abraham Lincoln was born in Kentucky.

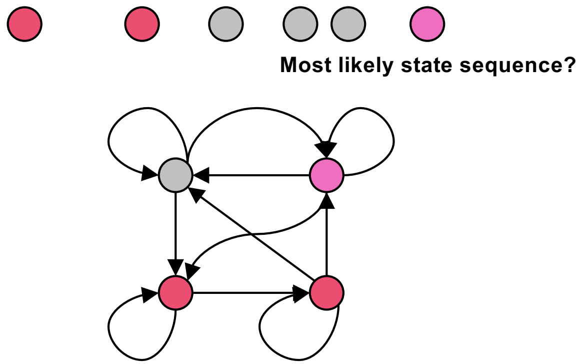


Boundary Models



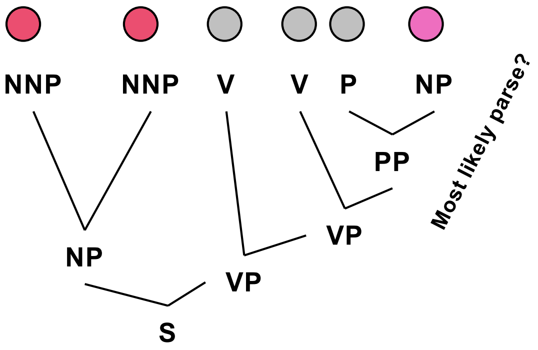
Finite State Machines

Abraham Lincoln was born in Kentucky.



Context Free Grammars

Abraham Lincoln was born in Kentucky.

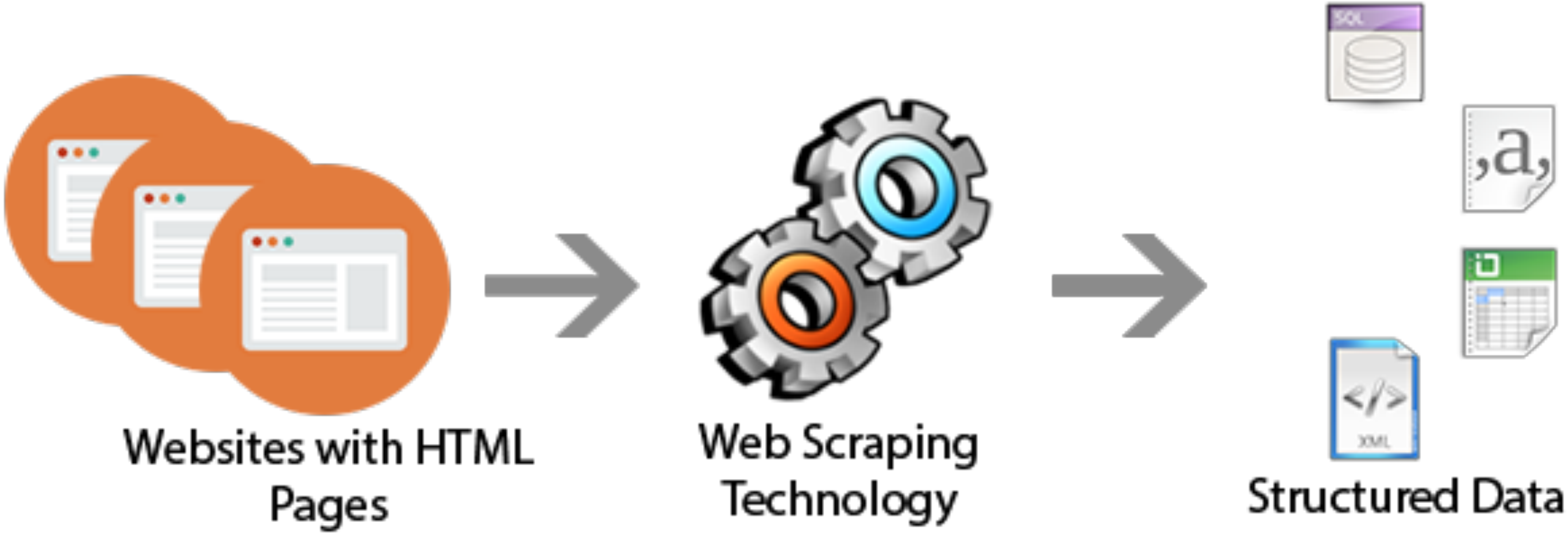


…and beyond

Any of these models can be used to capture words, formatting or both.

JAMES COOK
UNIVERSITY
AUSTRALIA

Web Scraping



Selenium

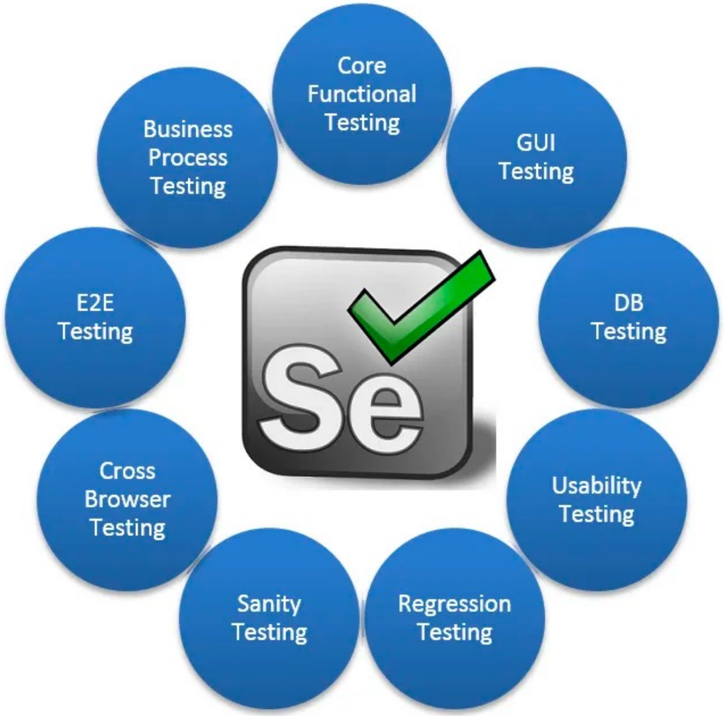
- ✓ Selenium IDE
- ✓ Selenium RC
- ✓ Selenium Grid
- ✓ Selenium Web Driver

<u>Individual Element</u>

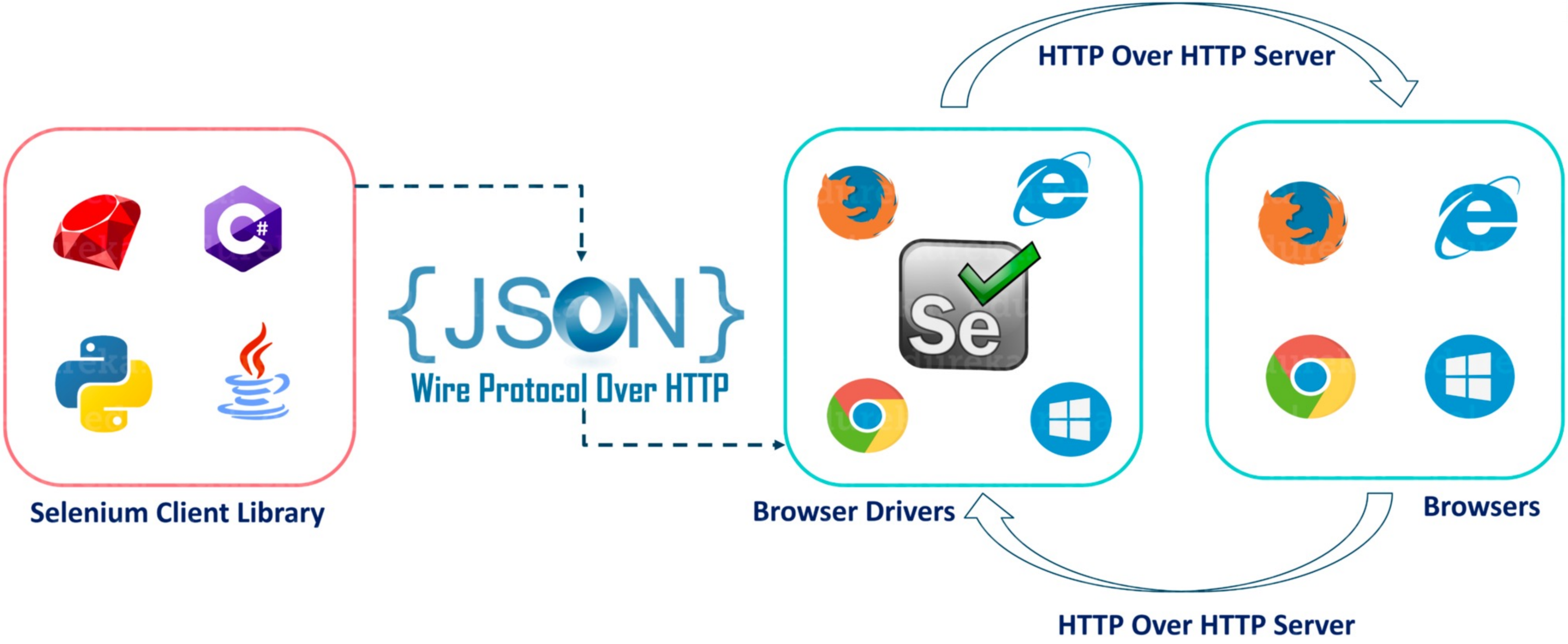
- *find_element_by_id*
- *find_element_by_name*
- *find_element_by_xpath*
- *find_element_by_link_text*
- *find_element_by_partial_link_text*
- *find_element_by_tag_name*
- *find_element_by_class_name*
- *find_element_by_css_selector.*

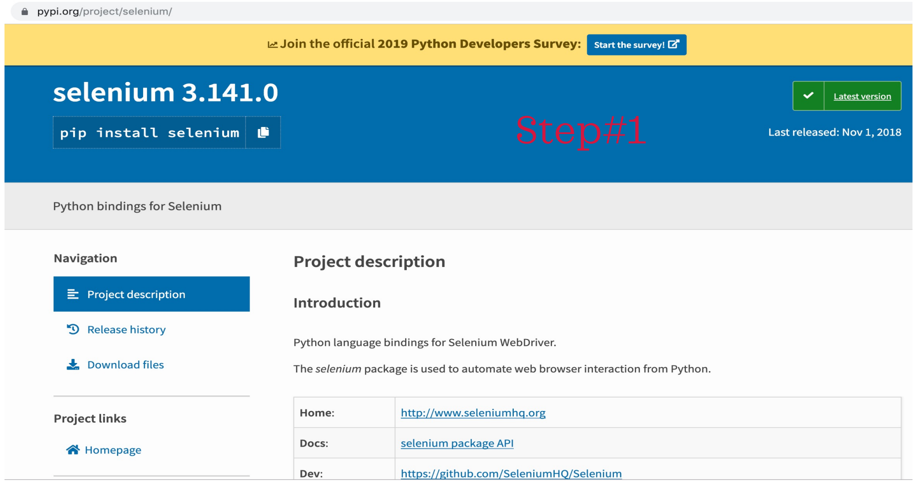
<u>Element List</u>

- *find_elements_by_name*
- *find_elements_by_xpath*
- *find_elements_by_link_text*
- *find_elements_by_partial_link_text*
- *find_elements_by_tag_name*
- *find_elements_by_class_name*
- *find_elements_by_css_selector.*



Selenium WebDriver Architecture





Step#1

pypi.org/project/selenium/

Join the official 2019 Python Developers Survey: Start the survey!

selenium 3.141.0

`pip install selenium`

Latest version

Last released: Nov 1, 2018

Python bindings for Selenium

Project description

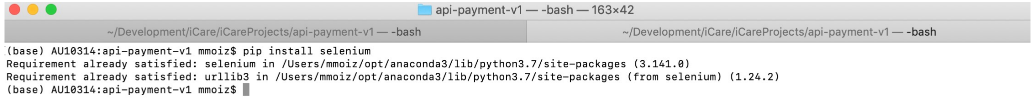
Introduction

Python language bindings for Selenium WebDriver.

The *selenium* package is used to automate web browser interaction from Python.

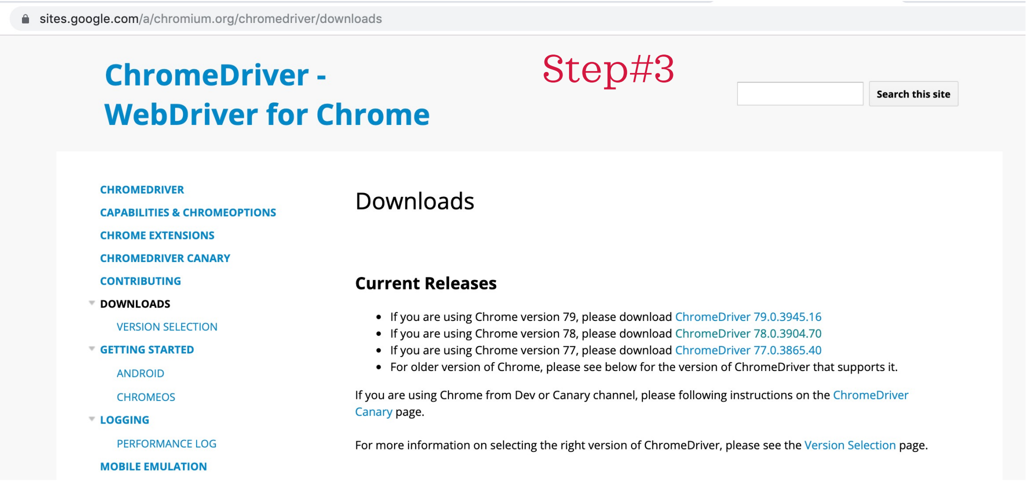
Home:	http://www.seleniumhq.org
Docs:	selenium package API
Dev:	https://github.com/SeleniumHQ/Selenium

Step#2



api-payment-v1 — -bash — 163×42
~/Development/iCare/iCareProjects/api-payment-v1 — -bash
~/Development/iCare/iCareProjects/api-payment-v1 — -bash

```
(base) AU10314:api-payment-v1 mmoiz$ pip install selenium
Requirement already satisfied: selenium in /Users/mmoiz/opt/anaconda3/lib/python3.7/site-packages (3.141.0)
Requirement already satisfied: urllib3 in /Users/mmoiz/opt/anaconda3/lib/python3.7/site-packages (from selenium) (1.24.2)
(base) AU10314:api-payment-v1 mmoiz$ 
```



Step#3

sites.google.com/a/chromium.org/chromedriver/downloads

ChromeDriver - WebDriver for Chrome

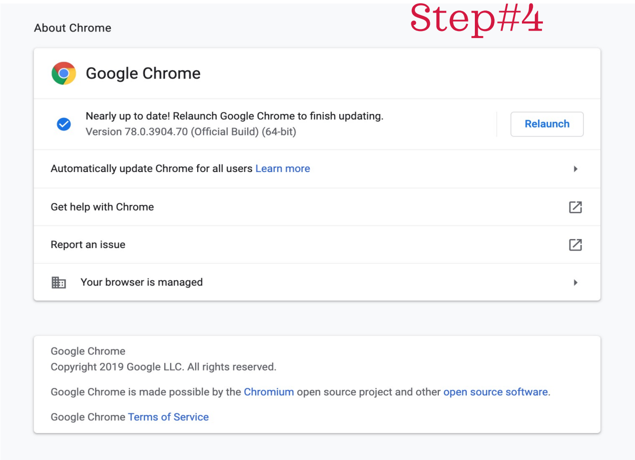
Downloads

Current Releases

- If you are using Chrome version 79, please download ChromeDriver 79.0.3945.16
- If you are using Chrome version 78, please download ChromeDriver 78.0.3904.70
- If you are using Chrome version 77, please download ChromeDriver 77.0.3865.40
- For older version of Chrome, please see below for the version of ChromeDriver that supports it.

If you are using Chrome from Dev or Canary channel, please following instructions on the ChromeDriver Canary page.

For more information on selecting the right version of ChromeDriver, please see the Version Selection page.



Step#4

About Chrome

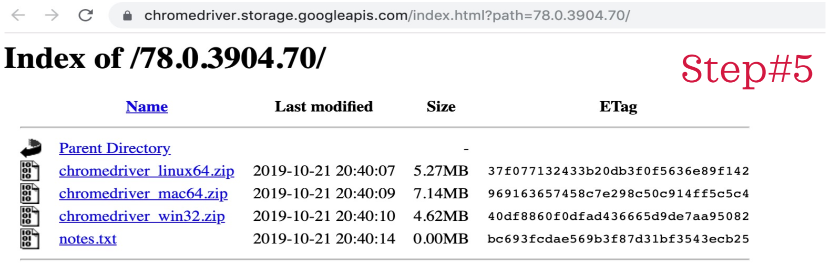
Google Chrome

Nearly up to date! Relaunch Google Chrome to finish updating.
Version 78.0.3904.70 (Official Build) (64-bit)

Google Chrome
Copyright 2019 Google LLC. All rights reserved.

Google Chrome is made possible by the Chromium open source project and other open source software.

Google Chrome Terms of Service



Step#5

chromedriver.storage.googleapis.com/index.html?path=78.0.3904.70/

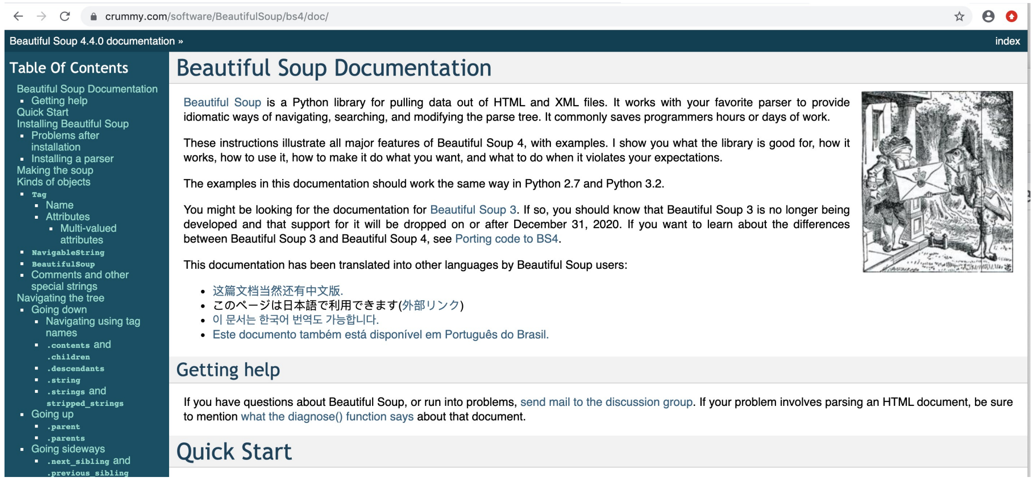
Index of /78.0.3904.70/

Name	Last modified	Size	ETag
Parent Directory		-	
chromedriver_linux64.zip	2019-10-21 20:40:07	5.27MB	37f077132433b20db3f0f5636e89f142
chromedriver_mac64.zip	2019-10-21 20:40:09	7.14MB	969163657458c7e298c50c914ff5c5c4
chromedriver_win32.zip	2019-10-21 20:40:10	4.62MB	40df8860f0dfad436665d9de7aa95082
notes.txt	2019-10-21 20:40:14	0.00MB	bc693fcdae569b3f87d31bf3543ecb25

BeautifulSoup

- BeautifulSoup is a popular Python library

- Transforms incoming texts to Unicode and outgoing versions to UTF-8.



crummy.com/software/BeautifulSoup/bs4/doc/
Beautiful Soup 4.4.0 documentation »
index

Table Of Contents

Beautiful Soup Documentation

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

These instructions illustrate all major features of Beautiful Soup 4, with examples. I show you what the library is good for, how it works, how to use it, how to make it do what you want, and what to do when it violates your expectations.

The examples in this documentation should work the same way in Python 2.7 and Python 3.2.

You might be looking for the documentation for Beautiful Soup 3. If so, you should know that Beautiful Soup 3 is no longer being developed and that support for it will be dropped on or after December 31, 2020. If you want to learn about the differences between Beautiful Soup 3 and Beautiful Soup 4, see Porting code to BS4.

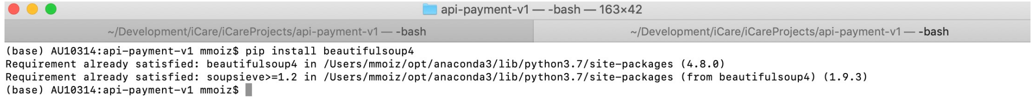
This documentation has been translated into other languages by Beautiful Soup users:

- 这篇文档当然还有中文版.
- このページは日本語で利用できます(外部リンク)
- 이 문서는 한국어 번역도 가능합니다.
- Este documento também está disponível em Português do Brasil.

Getting help

If you have questions about Beautiful Soup, or run into problems, send mail to the discussion group. If your problem involves parsing an HTML document, be sure to mention what the diagnose() function says about that document.

Quick Start

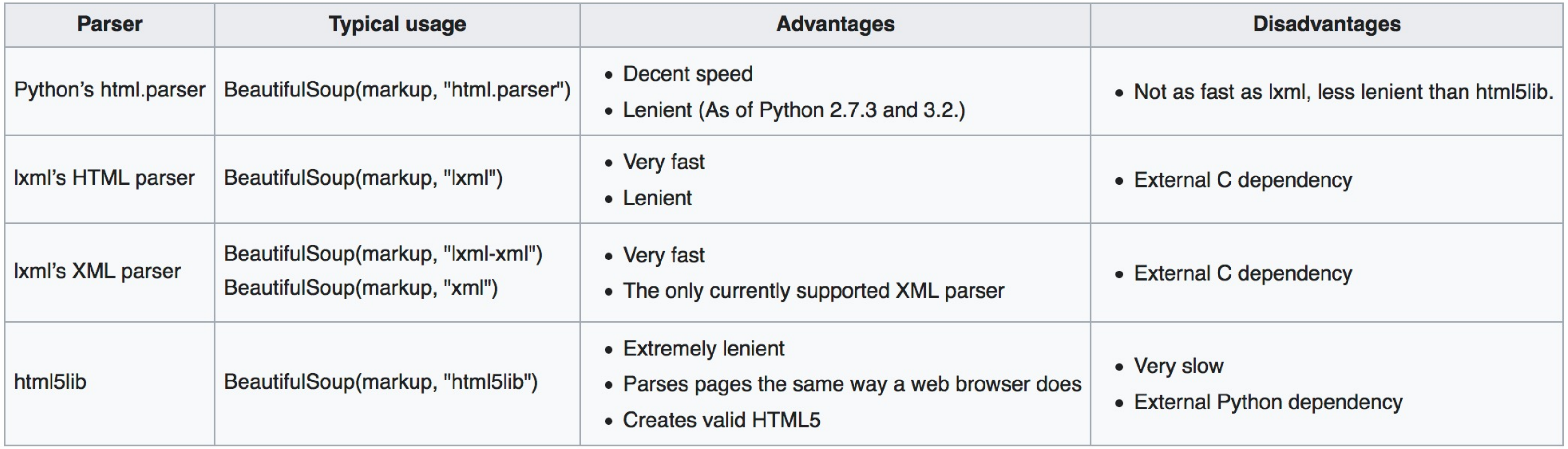


api-payment-v1 — -bash — 163×42
~/Development/iCare/iCareProjects/api-payment-v1 — -bash
~/Development/iCare/iCareProjects/api-payment-v1 — -bash

```
(base) AU10314:api-payment-v1 mmoiz$ pip install beautifulsoup4
Requirement already satisfied: beautifulsoup4 in /Users/mmoiz/opt/anaconda3/lib/python3.7/site-packages (4.8.0)
Requirement already satisfied: soupsieve>=1.2 in /Users/mmoiz/opt/anaconda3/lib/python3.7/site-packages (from beautifulsoup4) (1.9.3)
(base) AU10314:api-payment-v1 mmoiz$ 
```

Parsers usually used with Beautiful Soup.

1. **HTML parser** (Python's built in)
2. **LXML's HTML parser**
3. **HTML5 Lib**.



Parser	Typical usage	Advantages	Disadvantages
Python's html.parser	BeautifulSoup(markup, "html.parser")	• Decent speed • Lenient (As of Python 2.7.3 and 3.2.)	• Not as fast as lxml, less lenient than html5lib.
lxml's HTML parser	BeautifulSoup(markup, "lxml")	• Very fast • Lenient	• External C dependency

lxml's XML parser	BeautifulSoup(markup, "lxml-xml")
	BeautifulSoup(markup, "xml")
		• Very fast
		• The only currently supported XML parser
			• External C dependency

html5lib	BeautifulSoup(markup, "html5lib")	• Extremely lenient • Parses pages the same way a web browser does • Creates valid HTML5	• Very slow • External Python dependency

Use Cases

- ✓ Auto Job Finder and Resume Submitter

- ✓ Predicting Stock Prices on the basis of market sentiments and news reports

https://hotcopper.com.au/postview/
https://en.wikipedia.org/wiki/Nick_D%27Aloisio#Summly