Machine Learning For NLP

PREPARED BY: AHMAD ALAA ALDINE

PRESENTED BY: AHMAD ALAA ALDINE

Overview

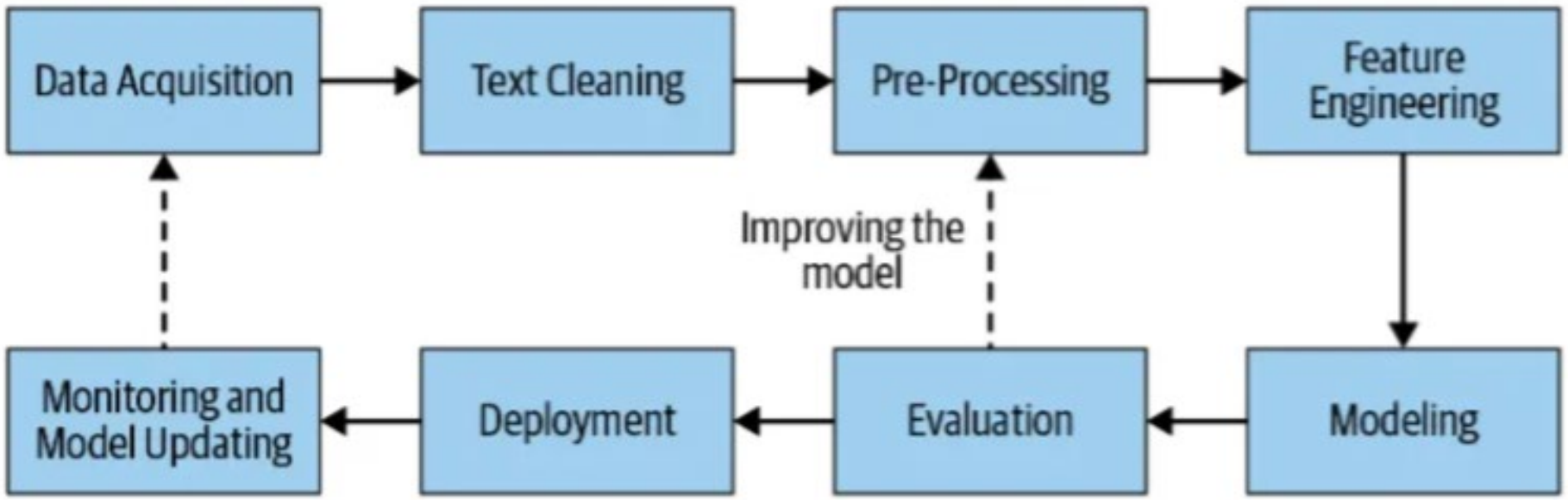
Rule-based NLP

- Time consuming and not scalable
- High precision but low recall

Machine learning brought automation and adaptability

- ML learns from data instead of relying on handcrafted rules
- Can adapt to different domains (news, health, tweets)
- Easier to update models as new data becomes available
- Formed the foundation for modern NLP pipelines

NLP Pipeline



Data Acquisition

Data Acquisition Techniques

1. Define Your Data Requirements
2. Utilize Existing Datasets
3. Utilize Databases
4. Domain-Specific Data Collection
5. Web Scraping
6. Utilize Existing APIs

Define Your Data Requirements

Before diving into data acquisition process, it is essential to clearly define your data requirements based on the specific NLP task you want to tackle.

Defining your data requirements will help you focus on collecting relevant and useful data.

Utilize Existing Datasets

One of the most efficient and fastest ways to acquire data for NLP is to utilize existing datasets that are publicly available.

Examples:

- Stanford Sentiment Treebank: labeled dataset for sentiment analysis
- Link: https://nlp.stanford.edu/sentiment/index.html
- Twenty Newsgroups Dataset: labeled dataset for document topic classification
- Link: http://qwone.com/~jason/20Newsgroups/

Utilize Databases

Example: Build a model to extract negative customer reviews that should be handled by the support team

1. Extract old reviews from databases
2. Label them (1: to be handled or 0: not to be handled)
3. Build a model
4. Use the model to extract reviews to be handled from new reviews



Limitation

Domain-Specific Data Collection

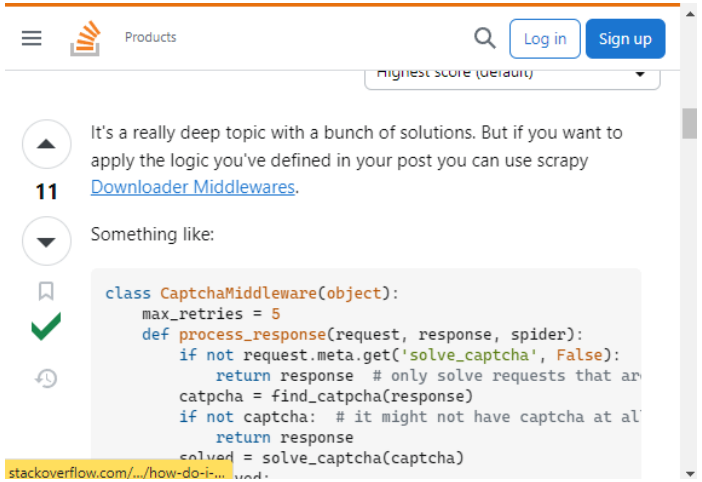
If your NLP task focuses on a specific domain or industry, it may be necessary to collect domain-specific data.

Example: developing a medical question-answering system

- Collect a dataset of medical documents or clinical notes
- Domain experts to guide and validate data

Web Scraping

Web scraping is the process of collecting information from the Internet (Webpages).





Challenges

- Variety: Every website is different
- Durability: Websites constantly change

Utilize Existing APIs

Application Programming Interfaces (APIs) – a set of online services that allow two applications to communicate

The number of APIs available for free has grown exponentially in the last decade.

In NLP, utilizing existing APIs is one of the best choices for collecting data

- Easy to use
- Well documented
- Data Quality and Consistency
- Legal and Ethical Compliance

APIs example:

- Twitter API
- News API
- Wikipedia API

Web Scraping Using Python

Web Scraping Python Libraries









BeautifulSoup is a popular library for parsing HTML and XML documents

- It is great for small to medium-scale web scraping tasks
- Easy to use for beginners
- Great documentation

Not suitable for scraping tasks that require advanced features such as:

- Handling JavaScript
- Navigating through the website



The BOSS of Python web scraping libraries

Scrapy is suitable for more complex and large-scale web scraping projects that require crawling multiple pages and handling complex website structures.

- Great documentation
- Asynchronous
- Low CPU and memory usage

Overkill for easy jobs:

- Not beginner-friendly



It was originally made for automated testing of web applications.

Selenium is useful when you need to interact with web elements (user authentication)

- Can do anything on a web page similar to a person
- Automate web browsers
- Easy to use for beginners

Not ideal for large projects:

- High CPU and memory usage
- Difficult to setup

Web Scraping Using BeautifulSoup

Loading Web Pages with 'request'

```
import requests

res = requests.get('https://en.wikipedia.org/wiki/Elon_Musk')

print(res.status_code)
print(res.text)
```

```
200
<!DOCTYPE html>
<html class="client-nojs vector-feature-language-in-header-enabled vector-feature-language-in-main-page-header-disabled vecto
r-feature-sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clientpref-1 vector-feat
ure-main-menu-pinned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-content-enabled vector-f
eature-zebra-design-disabled vector-feature-custom-font-size-clientpref-0 vector-feature-client-preferences-disabled vector-f
eature-typography-survey-disabled vector-toc-available" lang="en" dir="ltr">
<head>
<meta charset="UTF-8">
<title>Elon Musk - Wikipedia</title>
<script>(function(){var className="client-js vector-feature-language-in-header-enabled vector-feature-language-in-main-page-h
eader-disabled vector-feature-sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clie
ntpref-1 vector-feature-main-menu-pinned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-cont
ent-enabled vector-feature-zebra-design-disabled vector-feature-custom-font-size-clientpref-0 vector-feature-client-preferenc
es-disabled vector-feature-typography-survey-disabled vector-toc-available";var cookie=document.cookie.match(/(?:^|; )enwikim
wclientpreferences=([^;]+)/);if(cookie){cookie[1].split('%2C').forEach(function(pref){className=className.replace(new RegExp
('(^| )'+pref.replace(/-clientpref-\w+$|[^\w-]+/g,'')+'-clientpref-\\w+( |$)'),'$1'+pref+'$2');});}document.documentElement.c
lassName=className;}());RLCONF={"wgBreakFrames":false,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],
"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","February","March","April","May","June","July","August","Septembe
```

Extracting title with BeautifulSoup

```
from bs4 import BeautifulSoup

page = requests.get('https://en.wikipedia.org/wiki/Elon_Musk')
soup = BeautifulSoup(page.content, 'html.parser')
title = soup.title.text
print(title)
```

```
Elon Musk - Wikipedia
```

Extract body of page

```
page_body = soup.body
print(page_body)
```

```
<body class="skin-vector skin-vector-search-vue mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-Elon_Musk ro
otpage-Elon_Musk skin-vector-2022 action-view"><a class="mw-jump-link" href="#bodyContent">Jump to content</a>
<div class="vector-header-container">
<header class="vector-header mw-header">
<div class="vector-header-start">
<nav aria-label="Site" class="vector-main-menu-landmark" role="navigation">
<div class="vector-dropdown vector-main-menu-dropdown vector-button-flush-left vector-button-flush-right" id="vector-main-men
u-dropdown">
<input aria-haspopup="true" aria-label="Main menu" class="vector-dropdown-checkbox " data-event-name="ui.dropdown-vector-main
-menu-dropdown" id="vector-main-menu-dropdown-checkbox" role="button" type="checkbox"/>
```

Extract head of page

```
page_head = soup.head
print(page_head)
```

```
<head>
<meta charset="utf-8"/>
<title>Elon Musk - Wikipedia</title>
<script>(function(){var className="client-js vector-feature-language-in-header-enabled vector-feature-language-in-main-page-h
eader-disabled vector-feature-sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clie
ntpref-1 vector-feature-main-menu-pinned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-cont
ent-enabled vector-feature-zebra-design-disabled vector-feature-custom-font-size-clientpref-0 vector-feature-client-preferenc
es-disabled vector-feature-typography-survey-disabled vector-toc-available";var cookie=document.cookie.match(/(?:^|; )enwikim
wclientpreferences=([^;]+)/);if(cookie){cookie[1].split('%2C').forEach(function(pref){className=className.replace(new RegExp
('(^| )'+pref.replace(/-clientpref-\w+$|[^\w-]+/g,'')+'-clientpref-\\w+( |$)'),'$1'+pref+'$2');});}document.documentElement.c
lassName=className;}());RLCONF={"wgBreakFrames":false,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],
```

Select with BeautifulSoup

```
all_h1_tags = soup.select('h1')
first_h1 = all_h1_tags[0].text
print(first_h1)
```

```
Elon Musk
```

```
▼<tr>
  <th scope="row" class="infobox-label">Born</th>
 ▼<td class="infobox-data">
    <div style="display:inline" class="nickname">Elon Reeve Musk</div>
    <br>
   ▼<span style="display:none">
      " ("
      <span class="bday">1971-06-28</span>
      ") "
    </span>
    "June 28, 1971"
    <span class="noprint ForceAgeToShow"> (age 52)</span>
    <br>
   ▶<div style="display:inline" class="birthplace">…</div>
  </td>
</tr>
```

Extract text with specific tags and class attribute

```
all_birthday = soup.find_all("span", {"class": "bday"})
print(birthday[0].text)
```

```
1971-06-28
```

Extract all paragraph texts

```
all_p = soup.find_all('p')
print(len(all_p))
for i in range(10):
    print(all_p[i].text)
```

```
106


Elon Reeve Musk (/ˈiːlɒn/ EE-lon; born June 28, 1971) is a businessman and investor. Musk is the founder, chairman, CEO and chi
ef technology officer of SpaceX; angel investor, CEO, product architect and former chairman of Tesla, Inc.; owner, chairman and
CTO of X Corp.; founder of the Boring Company and xAI; co-founder of Neuralink and OpenAI; and president of the Musk Foundatio
n. He is the wealthiest person in the world, with an estimated net worth of US$219 billion as of November 2023[update], accordi
ng to the Bloomberg Billionaires Index, and $241 billion according to Forbes, primarily from his ownership stakes in Tesla and
SpaceX.[5][6]

Musk was born in Pretoria, South Africa, and briefly attended the University of Pretoria before immigrating to Canada at age 1
8, acquiring citizenship through his Canadian-born mother. Two years later, he matriculated at Queen's University at Kingston i
n Canada. Musk later transferred to the University of Pennsylvania, and received bachelor's degrees in economics and physics th
ere. He moved to California in 1995 to attend Stanford University. However, Musk dropped out after two days and, with his broth
er Kimbal, co-founded online city guide software company Zip2. The startup was acquired by Compaq for $307 million in 1999, and
with $12 million of the money he made, that same year Musk co-founded X.com, a direct bank. X.com merged with Confinity in 2000
to form PayPal.

In October 2002, eBay acquired PayPal for $1.5 billion, and that same year, with $100 million of the money he made, Musk founde
d SpaceX, a spaceflight services company. In 2004, he became an early investor in electric vehicle manufacturer Tesla Motors, I
nc. (now Tesla, Inc.). He became its chairman and product architect, assuming the position of CEO in 2008. In 2006, Musk helped
create SolarCity, a solar-energy company that was acquired by Tesla in 2016 and became Tesla Energy. In 2013, he proposed a hyp
erloop high-speed vactrain transportation system. In 2015, he co-founded OpenAI, a nonprofit artificial intelligence research c
ompany. The following year, Musk co-founded Neuralink—a neurotechnology company developing brain–computer interfaces—and the Bo
ring Company, a tunnel construction company. In 2022, he acquired Twitter for $44 billion. He subsequently merged the company i
nto newly created X Corp. and rebranded the service as X the following year. In March 2023, he founded xAI, an artificial intel
```

News API

Before using, create an account and get an api_key

```
from newsapi import NewsApiClient

# Init
newsapi = NewsApiClient(api_key='bea24a9017d145fc8656e2a83e30ad65')

# /v2/everything
all_articles = newsapi.get_everything(q='Artificial Intelligence',
                                      from_param='2023-11-25',
                                      to='2023-12-12',
                                      language='en',
                                      sort_by='relevancy')

print(len(all_articles['articles']))
```

```
100
```

```
print(all_articles['articles'][0]['title'])
```

```
The Future of Artificial Intelligence (AI): Opportunities, Challenges, and Impact on Society
```

```
print(all_articles['articles'][0]['content'])
```

```
The 2023 word of the year: artificial intelligence. And it’s pushing the boundaries of technology capabilities and human intell
igence.
As we venture further into the growth phase of artificial intel… [+13684 chars]
```