Building of NLP Datasets

Topic 3

Dwi Intan Af'idah, S.T., M.Kom

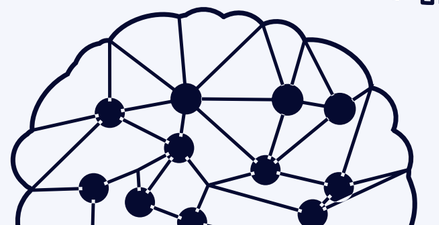
Agenda

01 Reminder about General Steps of NLP

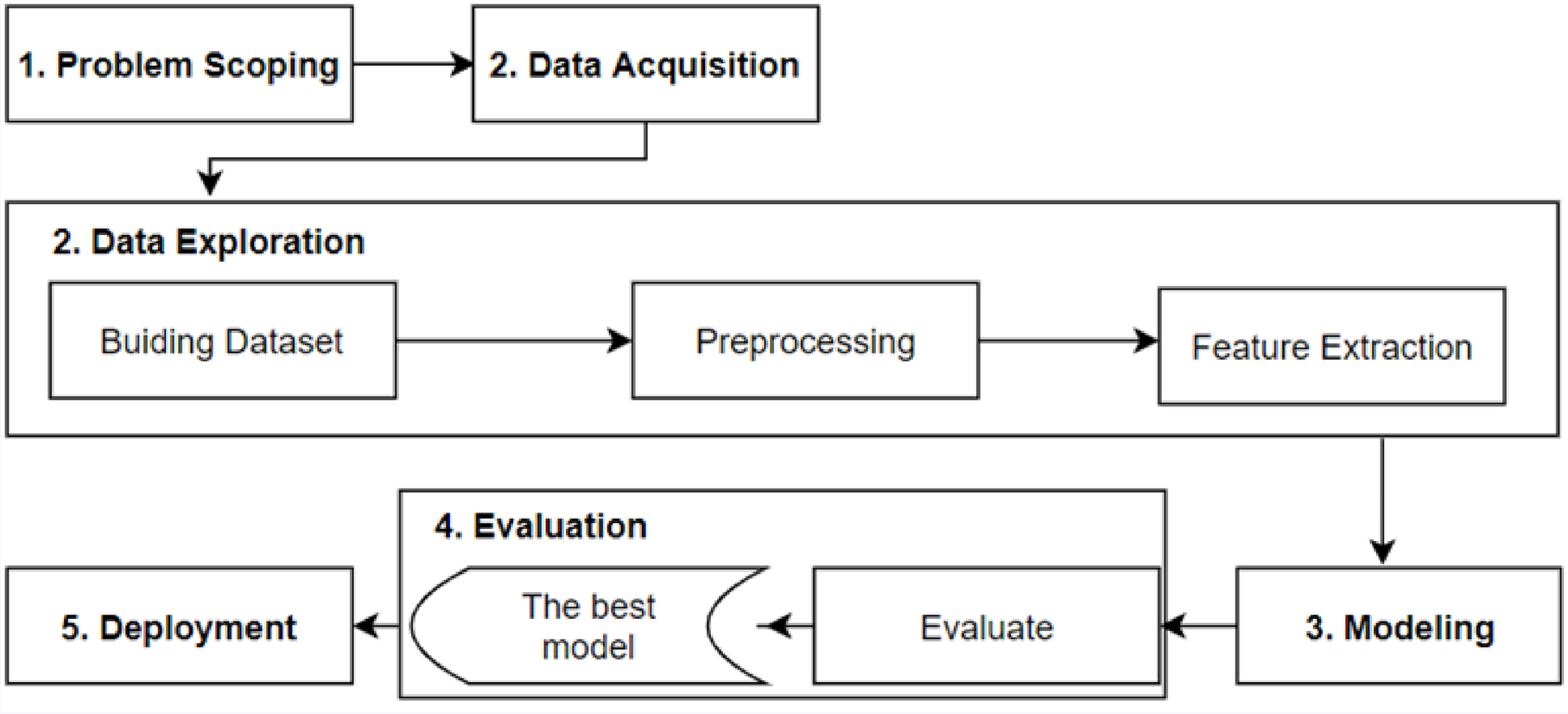
02 Crawling

03 Scraping

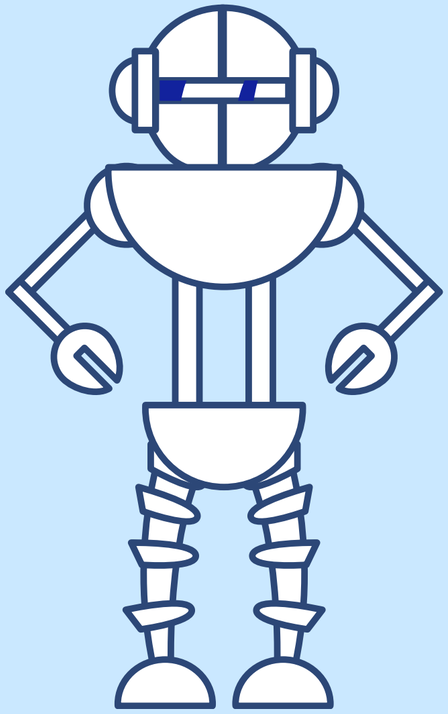
04 Basic Coding



General Steps of NLP



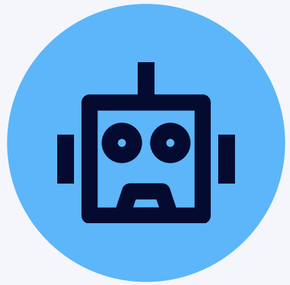
File for Exercise



https://bit.ly/bisaai-nlp

Understanding Datasets

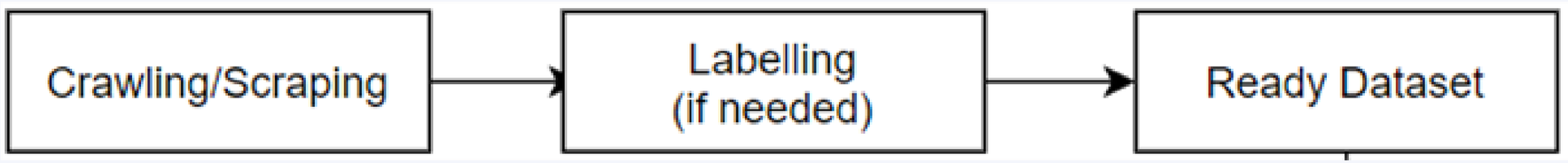
- Dataset adalah sekumpulan data yang disusun secara terstruktur.
- Biasanya, dataset dipresentasikan dalam bentuk tabel, alias baris dan kolom.
- Tiap baris dan kolom biasanya mewakili variable tertentu.
- Contohnya, kolom pertama merupakan data ulasan, sedangkan kolom kedua merupakan kolom label/kelas.



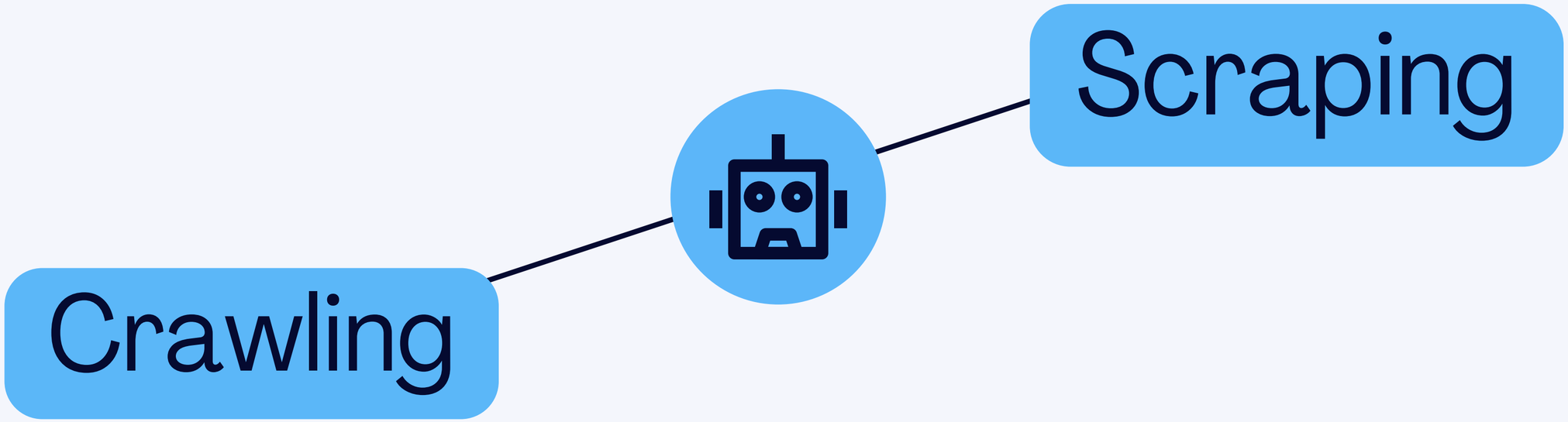
- Dataset pada penelitian NLP biasanya berupa data teks.
- Dataset pada riset NLP dapat menggunakan dataset dari:

1. Penelitian sebelumnya yang sejenis,
2. Membangun dataset sendiri.

Step of Building NLP Datasets



Collect Text Datasets

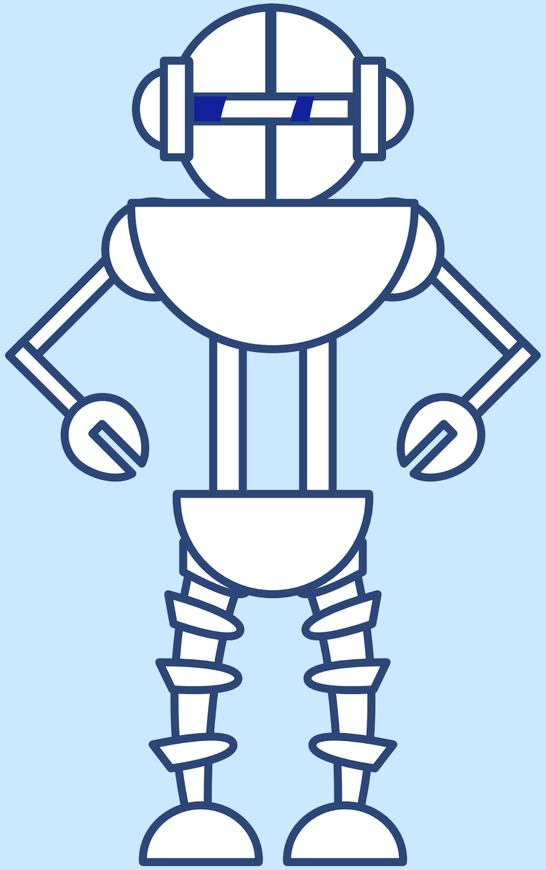


Crawling

- Penerapan crawling menggunakan automation program dan menggunakan Application Programming Interface (API) sebagai jalur komunikasi dalam mendapatkan data.
- Dengan API dapat dikumpulkan data lebih spesifik sesuai dengan link URL yang ada tanpa harus mengetahui element HTML pada sebuah website.

Crawling

Tweepy



Tweepy

Tweepy adalah *library* Python yang berguna untuk mengakses data Twitter melalui akses API (*Application Programming Interface*). Ini biasanya digunakan oleh para pengembang aplikasi untuk merancang atau menyempurnakan aplikasi yang membutuhkan data dari Twitter di dalamnya.

Library ini memberikan akses otomatis untuk menarik data dari Twitter melalui Python. Tidak hanya itu, kamu juga bisa mengatur akun Twitter mu menggunakan *library* ini seperti memberikan perintah untuk mengirimkan *tweet* baru, menghapus *tweet*, mengikuti ataupun berhenti mengikuti suatu akun.

Saat ini, Tweepy cenderung dimanfaatkan untuk membuat *bot* Twitter dengan tujuan tertentu ataupun melakukan analisis terhadap pengguna Twitter sendiri, utamanya sentimen yang terbentuk pada suatu kasus tertentu untuk mengetahui trennya.

Kampus Merdeka INDONESIA JAYA

Poltek Harber POLITEKNIK HARAPAN BERSAMA

Coding of Tweepy (1)

```
pip install tweepy
```

```
import tweepy
access_token = "1101429342041604096-brplhFporydhLgep9zE2JJD63Yehmc"
access_token_secret = "m0LewkNeSpBTH03m9jMFlxkEO3Vg8e36ReHnRN1Pj0MA0"
consumer_key = "N5qi9C8J4c1MSmL7mAeH2JW4v"
consumer_secret = "jjsz4HTeI1vUHtCxdp3PdnKDl7REBioqkhz9ossKiXj5AekuJJ"
```

```
tweets_for_csv = []

def get_tweets(username):
    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    api = tweepy.API(auth)

    limit = 100
    print('- Username : @'+username)

    for tweet in tweepy.Cursor(api.user_timeline,
    screen_name=username, include_rts=False,
    tweet_mode='extended').items(limit):
        actualTweet = re.sub(r'\s+', ' ', tweet.full_text)
        tweets_for_csv.append(
            [actualTweet])
```

Coding of Tweepy (2)

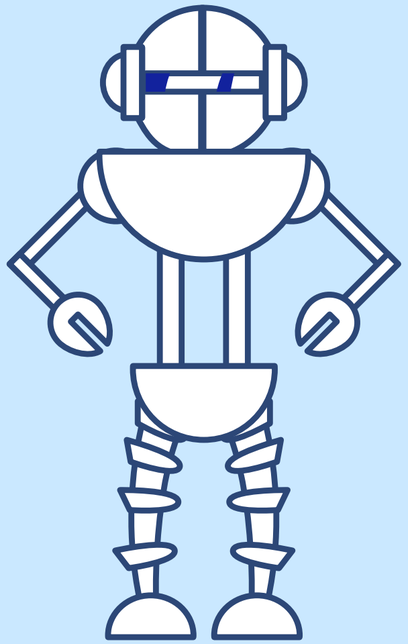
```python
import re
import csv

if __name__ == '__main__':
    users = ['CNNIndonesia']

    print("\nGet tweet from username ...")
    for user in users:
        get_tweets(user)

    outfile = "Topic 3-tweet-cnn-indonesia.csv"
    with open(outfile, 'w', newline='', encoding='utf-8') as csvfile:
        csvwriter = csv.writer(csvfile)
        csvwriter.writerow(["tweet"])
        csvwriter.writerows(tweets_for_csv)

    print("\nwriting to '" + outfile + "' complete.")
```

```python
import pandas as pd

dataset = pd.read_csv('Topic 3-tweet-cnn-indonesia.csv')
dataset[:5]
```

Scraping

- Scraping merupakan teknik mengumpulkan data pada sebuah website melalui proses ektraksi informasi menggunakan Hypertext Transfer Protocol (HTTP).
- Scraping dapat digunakan secara manual ataupun secara automation program.
- Namun untuk mendapatkan data kita perlu mengetahui element HTML ataupun XML pada sebuah website.
- Kemudian kita masukkan ke dalam program yang dibuat untuk mencari data sesuai nama id atau nama class dari element HTML tersebut.

Scraping



- Beautiful Soup
- Twint
- Selenium
- Googleplay Scrapper

Twint

- **Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter.**

- **The advantage of Twint is that you don’t need Twitter’s API to make TWINT work. Twint utilizes Twitter’s search operators to let you :**

1. scrape Tweets from specific users
2. scrape Tweets relating to certain topics
3. hashtags & trends
4. or sort out sensitive information from Tweets like e-mail and phone numbers.

```
pip install --upgrade git+https://github.com/twintproject/twint.git@origin/master#egg=twint
```

```
!pip install nest_asyncio
```

Coding of Twint (1)

```
import twint
import nest_asyncio

nest_asyncio.apply()
c = twint.Config()

c.Search = "Jokowi"
c.Since = "2022-1-1"
c.Until = "2022-1-31"
c.Limit = 100
c.Pandas = True

twint.run.Search(c)
```

Coding of Twint (2)

```
def column_names():
  return twint.output.panda.Tweets_df.columns

def twint_to_pd(columns):
  return twint.output.panda.Tweets_df[columns]

dataset = twint_to_pd(['tweet'])
print(dataset)

dataset.to_csv("Topic 3-tweet-jokowi.csv", index=False)
```

Beautiful soup (1)

Tentang BeautifulSoup

Kita sebenarnya bisa melakukan web scraping secara manual. Namun, dengan request url sederhana, komputer akan memberikan data HTML yang nampak membingungkan. Lihat contoh tampilannya di bawah ini.


```
'<!DOCTYPE html>\n<html class="client-nojs" lang="en" dir="ltr">\n<head>\n<meta charset="UTF-8"/>\n<title>Joko Widodo - Wiki
pedia</title>\n<script>document.documentElement.className="client-js";RLCONF={"wgBreakFrames":!1,"wgSeparatorTransformTabl
e":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","February","March","Apri
l","May","June","July","August","September","October","November","December"],"wgRequestId":"704ea0fd-86b8-4984-8de2-6376c01a
4a14","wgCSPNonce":!1,"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":!1,"wgNamespaceNumber":0,"wgPageName":"Joko_Wid
odo","wgTitle":"Joko Widodo","wgCurRevisionId":1030025186,"wgRevisionId":1030025186,"wgArticleId":29410367,"wgIsArticle":!
0,"wgIsRedirect":!1,"wgAction":"view","wgUserName":null,"wgUserGroups":["*"],"wgCategories":["CS1 Indonesian-language source
s (id)","CS1 maint: multiple names: authors list","Articles with short description","Short description is different from Wik
idata","Wikipedia pending changes protected pages","Use British English from August 2020","Use dmy dates from May 2018",\n"A
rticles containing Indonesian-language text","Commons category link from Wikidata","Wikipedia articles with GND identifier
s","Wikipedia articles with ISNI identifiers","Wikipedia articles with VIAF identifiers","Wikipedia articles with LCCN ident
ifiers","Wikipedia articles with PLWABN identifiers","Wikipedia articles with FAST identifiers","Wikipedia articles with SUD
OC identifiers","Wikipedia articles with WORLDCATID identifiers","Joko Widodo","1961 births","Gadjah Mada University alumn
i","Governors of Jakarta","Indonesian businesspeople","Indonesian Democratic Party of Struggle politicians","Indonesian engi
neers","Indonesian Muslims","Javanese people","Living people","Mayors of Surakarta","People from Surakarta","Presidents of I
ndonesia","Mayors of places in Indonesia"],"wgPageContentLanguage":"en","wgPageContentModel":"wikitext","wgRelevantPageNam
e":"Joko_Widodo","wgRelevantArticleId":29410367,"wgIsProbablyEditable":!0,"wgRelevantPageIsProbablyEditable":!0,"wgRestricti
onEdit":\n[],"wgRestrictionMove":[],"wgFlaggedRevsParams":{"tags":{"status":{"levels":-1}}},"wgStableRevisionId":103002518
6,"wgMediaViewerOnClick":!0,"wgMediaViewerEnabledByDefault":!0,"wgPopupsFlags":10,"wgVisualEditor":{"pageLanguageCode":"e
n","pageLanguageDir":"ltr","pageVariantFallbacks":"en"},"wgMFDisplayWikibaseDescriptions":{"search":!0,"nearby":!0,"watchlis
t":!0,"tagline":!1},"wgWMESchemaEditAttemptStepOversample":!1,"wgULSCurrentAutonym":"English","wgNoticeProject":"wikipedi
a","wgCentralAuthMobileDomain":!1,"wgEditSubmitButtonLabelPublish":!0,"wgULSPosition":"interlanguage","wgULSisCompactLinksEn
abled":!0,"wgGENewcomerTasksGuidanceEnabled":!0,"wgGEAskQuestionEnabled":!1,"wgGELinkRecommendationsFrontendEnabled":!1,"wgW
ikibaseItemId":"Q3318231"};RLSTATE={"ext.globalCssJs.user.styles":"ready","site.styles":"ready","noscript":"ready","user.sty
les":"ready","ext.globalCssJs.user":"ready","user":"ready","user.options":"loading","ext.flaggedRevs.icons":"ready",\n"oojs-
ui-core.styles":"ready","oojs-ui.styles.indicators":"ready","mediawiki.widgets.styles":"ready","oojs-ui-core.icons":"read
y","ext.cite.styles":"ready","skins.vector.styles.legacy":"ready","jquery.makeCollapsible.styles":"ready","ext.flaggedRevs.b
asic":"ready","ext.visualEditor.desktopArticleTarget.noscript":"ready","ext.uls.interlanguage":"ready","ext.wikimediaBadge
s":"ready","wikibase.client.init":"ready"};RLPAGEMODULES=["ext.cite.ux-enhancements","ext.scribunto.logs","site","mediawiki.
page.ready","jquery.makeCollapsible","mediawiki.toc","skins.vector.legacy.js","ext.flaggedRevs.advanced","ext.gadget.Referen
```

Beautiful soup (2)

Untuk memudahkan scraping, kita bisa menggunakan BeautifulSoup. BeautifulSoup adalah library Python yang digunakan untuk mengambil data HTML dan XML. BeautifulSoup berfungsi sebagai parser untuk memisahkan komponen-komponen HTML menjadi rangkain elemen yang mudah dibaca.

```
<!DOCTYPE html>

<html class="client-nojs" dir="ltr" lang="en">
<head>
<meta charset="utf-8"/>
<title>Joko Widodo - Wikipedia</title>
<script>document.documentElement.className="client-js";RLCONF={"wgBreakFrames":!1,"wgSeparatorTransformTable":["",""],"wgDig
itTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","February","March","April","May","Jun
e","July","August","September","October","November","December"],"wgRequestId":"704ea0fd-86b8-4984-8de2-6376c01a4a14","wgCSPN
once":!1,"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":!1,"wgNamespaceNumber":0,"wgPageName":"Joko_Widodo","wgTitl
e":"Joko Widodo","wgCurRevisionId":1030025186,"wgRevisionId":1030025186,"wgArticleId":29410367,"wgIsArticle":!0,"wgIsRedirec
t":!1,"wgAction":"view","wgUserName":null,"wgUserGroups":["*"],"wgCategories":["CS1 Indonesian-language sources (id)","CS1 m
aint: multiple names: authors list","Articles with short description","Short description is different from Wikidata","Wikipe
dia pending changes protected pages","Use British English from August 2020","Use dmy dates from May 2018",
"Articles containing Indonesian-language text","Commons category link from Wikidata","Wikipedia articles with GND identifier
s","Wikipedia articles with ISNI identifiers","Wikipedia articles with VIAF identifiers","Wikipedia articles with LCCN ident
ifiers","Wikipedia articles with PLWABN identifiers","Wikipedia articles with FAST identifiers","Wikipedia articles with SUD
OC identifiers","Wikipedia articles with WORLDCATID identifiers","Joko Widodo","1961 births","Gadjah Mada University alumn
i","Governors of Jakarta","Indonesian businesspeople","Indonesian Democratic Party of Struggle politicians","Indonesian engi
neers","Indonesian Muslims","Javanese people","Living people","Mayors of Surakarta","People from Surakarta","Presidents of I
ndonesia","Mayors of places in Indonesia"],"wgPageContentLanguage":"en","wgPageContentModel":"wikitext","wgRelevantPageNam
e":"Joko_Widodo","wgRelevantArticleId":29410367,"wgIsProbablyEditable":!0,"wgRelevantPageIsProbablyEditable":!0,"wgRestricti
onEdit":
```

Kampus Merdeka INDONESIA JAYA

Poltek Harber POLITEKNIK HARAPAN BERSAMA

Coding of BeautifulSoup 1

```python
# SCRAPING URL ADDRESS
import requests
import csv
from bs4 import BeautifulSoup

# Write ke file csv
csv_output = csv.writer(open('Topic 3-buku-url.csv', 'w', newline=''))

pages = []

# Collecting & parsing konten Web hlm 1-6
for i in range(1, 7):
    url = 'https://www.goodreads.com/list/show/2405.Buku_Non_Fiksi_Indonesia_Terbaik_Sepanjang_Masa' + str(i)
    pages.append(url)

for item in pages:
    page = requests.get(item)
    soup = BeautifulSoup(page.text, 'html.parser')

    # Find elemen class bookTitle di dalam class tableList
    novel_title_list = soup.find(class_='tableList js-dataTooltip')
    novel_title_list_items = novel_title_list.find_all(class_='bookTitle')

    # Get masing-masing title dan url dari class tableList
    for novel_title in novel_title_list_items:
        link = 'https://www.goodreads.com' + novel_title.get('href') + '?language_code=id'

        csv_output.writerow([link])
```

Coding of BeautifulSoup 2

```python
# SCRAPING REVIEW DATA FROM CSV FILE
import requests
from bs4 import BeautifulSoup
import csv

#Read input url dari file csv & write output review
with open('Topic 3-buku-url.csv', newline='') as f_urls, open('Topic 3-buku-ulasan.csv', 'w', newline='', encoding="utf-8") as f_output:
    csv_urls = csv.reader(f_urls)
    csv_output = csv.writer(f_output)
    csv_output.writerow(['Nama', 'Ulasan', 'Rating'])

    # Collecting & parsing konten web
    for line in csv_urls:
        r = requests.get(line[0]).text
        soup = BeautifulSoup(r, 'lxml')

        # Find elemen class review di dalam id bookReviews
        #novel_review_list = soup.find('div', {'id': 'bookReviews'})
        #novel_review_list_items = novel_review_list.find_all(class_='review')
        novel_review_list2 = soup.find_all('div', class_= "friendReviews elementListBrown")

        # Get masing-masing nama & review dari class bookReviews
        for novel_review in novel_review_list2:
            #name = novel_review.find(class_='user').get_text()
            review = novel_review.find(class_='readable').find('span', recursive=False).get_text()
            rating_element = novel_review.find('span', {'size': '15x15'})
            # Skip item review ketika elemen rating tidak ditemukan
            if rating_element == None:
                continue
            else :
                rating = rating_element.get_text()

            csv_output.writerow([review, rating])
```

Googleplay Scraper

Ada jutaan aplikasi, buku, dan film di Google Play Store, dan jumlahnya terus bertambah setiap hari. Menurut Penelitian AppBrainada sekitar 3 juta aplikasi pada kuartal pertama 2021. Miliaran komentar dimasukkan ke dalam jutaan aplikasi, buku, dan film ini setiap hari.

Peneliti, pengembang aplikasi, pakar pemasaran, dan pakar yang bekerja di berbagai bidang ingin mengorek dan memeriksa komentar yang dibuat di Google Play karena berbagai alasan.

Meskipun dimungkinkan untuk mengunduh komentar aplikasi yang Anda kembangkan dari Google Dashboard, Outscraper Pengikis Ulasan Google Play adalah alat yang tepat jika Anda mencari sistem di mana Anda dapat mengikis komentar untuk aplikasi lain tanpa batasan apa pun.

Outscraper memiliki Layanan Aplikasi Web dan API untuk Google Play Reviews scraping. Anda dapat langsung menggunakan Web App Scraper tanpa pengkodean apa pun, atau Anda dapat menggunakan API kami untuk menggunakannya di aplikasi / layanan Anda sendiri.

Coding of Googleplay Scraper 1

```python
from google_play_scraper import app

import pandas as pd

import numpy as np
```

```python
#Scrape desired number of reviews

from google_play_scraper import Sort, reviews

result, continuation_token = reviews(
    'com.telkom.mwallet',
    lang='id', # defaults to 'en'
    country='id', # defaults to 'us'
    sort=Sort.MOST_RELEVANT, # defaults to Sort.MOST_RELEVANT you can use Sort.NEWEST to get newst reviews
    count=5000, # defaults to 100
    filter_score_with= 1  # defaults to None(means all score) Use 1 or 2 or 3 or 4 or 5 to select certain score
)
```

Coding of Googleplay Scraper 2

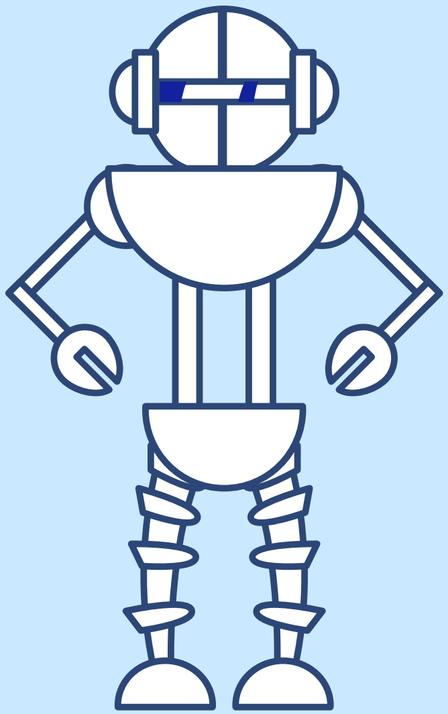
```
# Dataframe dengan nama
df_twt = pd.DataFrame(np.array(result),columns=['review'])

df_twt = df_twt.join(pd.DataFrame(df_twt.pop('review').tolist()))

df_twt.head()
```

```
len(df_twt.index) #count the number of data we got
```

```
my_df = df_twt[['userName', 'score','at', 'content']] #get userName, rating, date-time, and reviews only
```

```
my_df.to_csv("linkaja-rating1.csv", index = False)  #Save the file as CSV , to download: click the folder icon on the left.
#the csv file should be there.
```

Labelling of Dataset



Ulasan	Sentimen
Sistemnya harus lebih dibenahi,karna sering terjadi transaksi berhasil namun no tdk terdafta	0
Jaringan udah kenceng, udah hapus cache, hapus data, install ulang, ganti provider, ganti p	0
Saya mengalami masalah pada akun saya,tapi respon dana sangat2 lambat.. Akun saya di	0
Bagaimana sih dr smalam sampai skrng dana tidak bisa digunakan?? Saya sangat kecewa.	0
Saya jujur saja ...jgn pernah pakai aplikasi e-money ini... Ini aplikasi e-money terburuk mau	0
Buruk bener akun dana , tgl 10-12-2021 saya transfer dari akun dana ke mandiri belum mas	0
tambah ribet...padahal cuma mau upgread ke premium.. susah nya minta ampun.. tapi saya	0
Customer service dana care terburuk yg pernah ada!!!! Kirim uang engga sampai,dana ditah	0
Sangat memudahkan bnget	1
lumayan membantu..meskipun promo2 berkurang	1
Tingkat kn trus	1
Kamis, 23-09-2021, 16.00wib Saya transfer ke rek BCA dengan nominal ;Rp. 275.000 (TIDA	1
Mantap saya suka sekali aplikasiny	1
Gokiiilll buangeett nih apk OVO ayo download buruann gak bakal nyeseelll	1
Sangat membantu sekali	1
Tagihan begitu mudah	1
Sangat membantu sekali	1
Sangat membantu sekali	1
Bagus mudah dan aman	1
Tolong jaga keamanan	1

Refference

- Kedia, A., dan Rasu, M., 2020, Hands-On Python Natural Language Processing, Packt Publishing Ltd., Brimingham, UK, 35-41
- https://www.tweepy.org/
- https://github.com/twintproject/twint
- https://pypi.org/project/beautifulsoup4/
- https://pypi.org/project/google-play-scraper/

Thank You