BMCS2013 Data Engineering

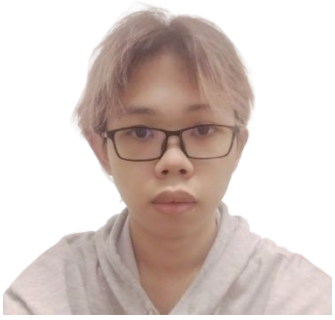
CLO2 ASSIGNMENT 202409

Assignment Title : **Data Engineering with Books Description in Google Books**

Team Reference : **Chiam Zi Wei**

Programme / Group : **RDS2S2 G2**

Student Name	**Student ID**	**Contribution**	**Signature**
Chiam Zi Wei	24PMR10315	33.34%	
Ang Yu Wen	24PMR10313	33.33%	
Lau Kit Lim	24PMR10318	33.33%	
TOTAL (must sum to 100%)		100%	

Chiam Zi Wei	Ang Yu Wen	Lau Kit Lim
		

Plagiarism Statement Form

I, Name	ANG YU WEN
Student ID	24PMR10313
Programme	Bachelor in Data Science (RDS)
Tutorial Group	2

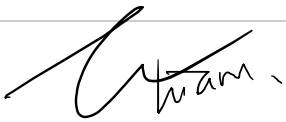
confirm that ALL submitted work for BMCS2013 Data Engineering are all my own work and is in my own words.

I	ANG YU WEN
acknowledge the use of AI generative technology.	
Signature :	
Date :	19/12/2024

Plagiarism Statement Form

I, Name	CHIAM ZI WEI
Student ID	24PMR10315
Programme	Bachelor in Data Science (RDS)
Tutorial Group	2

confirm that ALL submitted work for BMCS2013 Data Engineering are all my own work and is in my own words.

I	CHIAM ZI WEI
acknowledge the use of AI generative technology.	
Signature :	
Date :	19/12/2024

Plagiarism Statement Form

I, Name	LAU KIT LIM
Student ID	24PMR10318
Programme	Bachelor in Data Science (RDS)
Tutorial Group	2

confirm that ALL submitted work for BMCS2013 Data Engineering are all my own work and is in my own words.

I	LAU KIT LIM
acknowledge the use of AI generative technology.	
Signature :	
Date :	19/12/2024

Table of Contents

1. Data Collection and Preparation

1.1. Data Source(s)

1.1.1 Dewan Bahasa dan Pustaka (DBP)

The data source that we referenced for the initial scraping of this assignment is 'Dewan Bahasa dan Pustaka (DBP) (http://prpm.dbp.gov.my/)'.

DBP is a Malaysian government agency established in 1956 with the mission of advancing and promoting the Malay language and literature. It serves as Malaysia's official language authority for developing, standardizing, and preserving the Malay language while fostering literary and cultural heritage. DBP publishes dictionaries, academic materials, and literary works, provides language training, and conducts research to ensure the Malay language remains relevant and dynamic in various fields, including education, science, and technology. It also offers an extensive word database, thesaurus, and translation resources in Bahasa Malaysia, which are ideal for creating structured lexicons. The agency plays a pivotal role in upholding national identity and supporting the linguistic development of Malaysia.

DBP was established as Balai Pustaka in Johor Bahru on 22 June 1956 under the Malayan Ministry of Education. It was renamed Dewan Bahasa dan Pustaka during the Third Malay Literary and Language Congress in September 1956, with Royal Prof. Ungku Abdul Aziz playing a key role in its founding.

In 1957, DBP moved to Kuala Lumpur, and under the 1959 DBP Ordinance, it was granted a charter to oversee Malay language policies, promote the language, and engage in book publishing. On 31 January 1962, DBP relocated to its own building on Jalan Dewan Bahasa, designed by architect Lee Yoon Thim and featuring a mural by Ismail Mustam.

DBP later established regional offices in Bukit Mertajam and Kota Bharu (1999), and Johor Bahru (2003). It celebrated its 50th anniversary in 2006.

After World War II, Sarawak Crown officials established the Borneo Literature Bureau (BLB) in the 1940s-1960s to document oral literature, particularly in indigenous Bornean languages like Iban, fostering pride in native philosophies and knowledge systems.

In 1977, DBP opened offices in Kota Kinabalu and Kuching, taking over BLB's role. Allegations arose of DBP burying or burning BLB books, though some were rescued. DBP later explained that old books were disposed of when

expired, unneeded, or overstocked. Initially, DBP focused on publishing in the national language, Malay, citing policy and economic limitations, and declined to publish in regional languages, prioritizing Malay-language literary development.

DBP publishes the Kamus Dewan, a prestigious dictionary of the Malaysian national language. This dictionary is both descriptive and prescriptive, reflecting DBP's efforts to adapt Malay for technological and scientific advancements. DBP's role in developing and regulating the Malay language is comparable to similar language institutions worldwide, such as the Académie Française. (Wikipedia contributors, 2024)

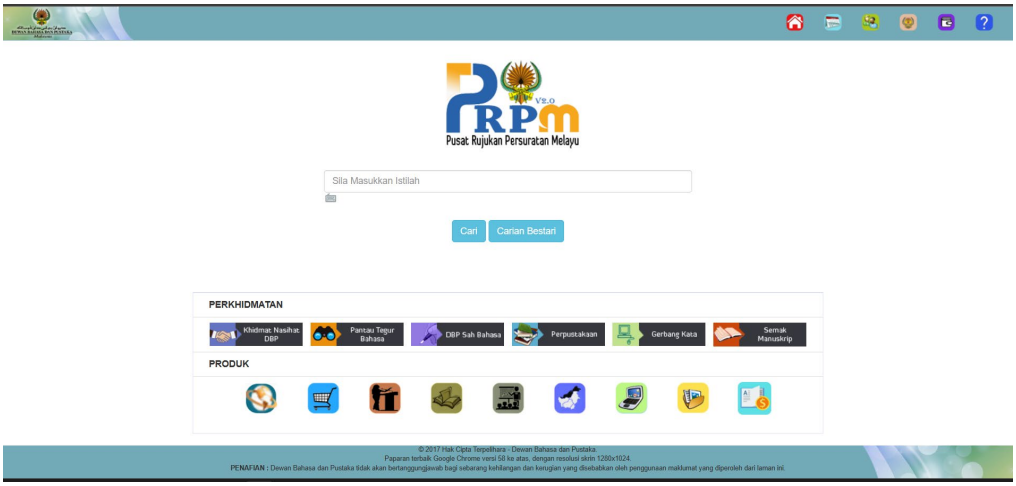


Fig. 1.1 *DBP Home Page*

1.1.2 Google Books

The data source that we referenced for the subsequent maintenance of this assignment is ‘Google Books (https://books.google.com/)’.

Google Books is a service by Google that allows users to search the full text of books and magazines. The content is scanned, converted to text using OCR, and stored in a digital database. Google Books come from two main sources:

1) The Partner Program, where publishers and authors provide books.

2) The Library Project, where books are provided by Google’s library partners.

Additionally, Google collaborates with magazine publishers to digitize magazine archives.

The Publisher Program, originally called Google Print, was launched in October 2004 at the Frankfurt Book Fair. The Google Books Library Project, which scans and digitizes library collections, was announced in December 2004.

The Google Books initiative has been praised for providing unprecedented access to a vast online collection of knowledge and promoting the democratization of knowledge. However, it has faced criticism for potential copyright violations and for errors in scanned texts caused by the OCR process.

As of October 2019, Google Books celebrated its 15th anniversary with over 40 million scanned titles. In 2010, Google estimated there were 130 million distinct titles globally and aimed to scan them all. However, scanning efforts in American academic libraries have slowed since the 2000s.

Google Books faced litigation, including the Authors Guild v. Google case, which was decided in Google's favor and nearly altered copyright practices for orphan works in the U.S. A 2023 study found that Google Books's digitization has led to increased sales of physical versions of the books. (Wikipedia contributors, 2024)

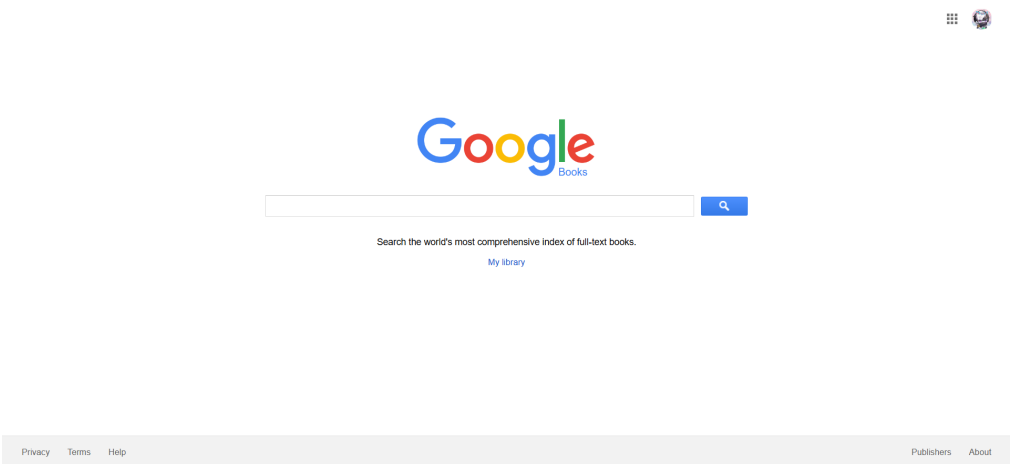


Fig. 1.2 *Google Books Home Page*

1.2. List of Python classes and Jupyter notebook(s) for data collection, cleaning and preprocessing:

Name of Python class(es) / notebooks	Author
web_scrapping.ipynb	Lau Kit Lim, Ang Yu Wen

web_scrapping.py data_preparation.ipynb	Chiam Zi Wei Lau Kit Lim, Ang Yu Wen

1.3. Code for Python Classes

1.3.1 web_scrapping.ipynb

Python Class for Web Scrapping

```
[1]: from python_script.web_scrapping import Word
```

Extract the Lexicon of a Word

```
w1 = Word('maksud')
w1.extractLexicon()
w1.getLexicon()
```

```
['maksud',
 6,
 'kata nama',
 ['1. tujuan, hajat, kehendak, niat',
  '2. erti yg tersimpul dlm sesuatu ayat (perbuatan dll)'],
 ['1. maksud aku ke sini ialah hendak berjumpa dgn engkau',
  '2. apakah maksud perkataanmu itu?'],
 ['tujuan',
  'hajat',
  'hasrat',
  'niat',
  'motif',
  'matlamat',
  'objektif',
  'keinginan',
  'kemahuan',
  'kehendak',
  'murad',
  'kasad',
  'anju',
  'hujung'],
 ['bermaksud', 'memaksudkan']]
```

Create a List to Store Lexicons

```
words = []
words.append(w1.getLexicon())
```

Expand Lexicons Size

```
import re
import string

prepared_keyword = w1.getDefinitionStr()
prepared_keyword = prepared_keyword.replace('\xad', ' ').replace('\xa0', ' ').replace('/', ' ')
prepared_keyword = prepared_keyword.translate(str.maketrans('', '', string.punctuation))
prepared_keyword = re.sub(r"\d+", "", prepared_keyword)

prepared_keyword = prepared_keyword.split(' ')
while '' in prepared_keyword:
    prepared_keyword.remove('')

print(prepared_keyword)
```

```
['ma', 'sud', 'maksud', 'Definisi', 'tujuan', 'hajat', 'kehendak', 'niat', 'aku', 'ke', 'sini', 'ialah', 'hendak', 'berjumpa', 'dgn', 'engkau', 'hatiku', 'pun', 'hendak', 'ke', 'sana', 'juga', 'erti', 'yg', 'tersimpul', 'dlm', 'sesuatu', 'ayat', 'per', 'buatan', 'dll', 'apakah', 'perkataanmu', 'itu', 'aku', 'masih', 'belum', 'tahu', 'lagi', 'aku', 'dipanggil', 'bermaksud', 'bertujuan', 'berhajat', 'berke', 'hendak', 'berniat', 'hendak', 'kami', 'hendak', 'menyerahkan', 'wang', 'itu', 'dgn', 'seberapa', 'segera', 'mengandungi', 'erti', 'apa', 'yg', 'diucapkan', 'sebentar', 'tadi', 'itu', 'tidak', 'demikian', 'memaksudkan', 'menghendaki', 'meminta', 'dialah', 'yg', 'dimaksudkan', 'hadir', 'dlm', 'seminar', 'itu', 'memberi', 'erti', 'yg', 'sebenar', 'hendak', 'menyampaikan', 'atau', 'menyatakan', 'sbg', 'mak', 'sud', 'Ali', 'masih', 'samarsamar', 'tentang', 'apa', 'yg', 'dimaksudkan', 'oleh', 'perempuan', 'tua', 'itu', 'apabila', 'ia', 'menerangkan', 'tentang', 'pendekatan', 'struktur', 'ia', 'pendekatan', 'formalistik', 'juga', 'termaksud', 'dikehendaki', 'dimaksudkan', 'Kamus', 'Dewan', 'Edisi', 'Keempat']
```

```
for word in prepared_keyword:
    w = Word(word)
    w.extractLexicon()

    words.append(w.getLexicon())

for word in ['tenaga', 'jelas', 'dunia', 'negara', 'tanah', 'asas']:
    w = Word(word)
    w.extractLexicon()
    prepared_keyword_temp = w.getDefinitionStr()
    prepared_keyword_temp = prepared_keyword_temp.replace('\xad', ' ').replace('\xa0', ' ').replace('/', ' ')
    prepared_keyword_temp = prepared_keyword_temp.translate(str.maketrans('', '', string.punctuation))
    prepared_keyword_temp = re.sub(r"\d+", "", prepared_keyword_temp)

    prepared_keyword_temp = prepared_keyword_temp.split(' ')
    while '' in prepared_keyword_temp:
        prepared_keyword_temp.remove('')

    for word_1 in prepared_keyword_temp:
        w1 = Word(word_1)
        w1.extractLexicon()

        words.append(w1.getLexicon())
```

```
print(len(words))
```

```
930
```

There are 930 non-cleansed records by scrapping information from 7 Malay vocabularies.

Data Processing by PySpark

Create Spark Session

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
```

```
24/12/14 12:11:57 WARN Utils: Your hostname, JORDAN. resolves to a loopback address: 127.0.1.1; using 10.255.255.254 instead (on interface lo)
24/12/14 12:11:57 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/12/14 12:11:57 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

Defines Columns' Heading and Create DataFrame

```
columns = ['vocab', 'word_length', 'part_of_speech', 'definitions', 'sentences_eg', 'synonyms', 'derived_words']
df = spark.createDataFrame(words, columns)
```

```
df.show()
```

```
+--------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|   vocab|word_length|part_of_speech|         definitions|         sentences_eg|            synonyms|       derived_words|
+--------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|  maksud|          6|     kata nama|[1. tujuan, hajat...|[1. maksud aku ke...|[tujuan, hajat, h...|[bermaksud, memak...|
|      ma|          2|              |                  []|                  []|                  []|                  []|
|     sud|          3|              |                  []|                  []|                  []|                  []|
|  maksud|          6|     kata nama|[1. tujuan, hajat...|[1. maksud aku ke...|[tujuan, hajat, h...|[bermaksud, memak...|
|Definisi|          8|              |                  []|                  []|                  []|                  []|
|  tujuan|          6|              |[1. arah, hala, h...|[1. tiap-tiap man...|                  []|                  []|
|   hajat|          5|     kata nama|[1. keinginan, ke...|[1. hampalah hati...|                  []|[berhajat, berhaj...|
|kehendak|          8|    kata tugas|                  []|                  []|[mahu, ingin, ber...|[menghendaki, keh...|
|    niat|          4|     kata nama|[1. maksud atau t...|[1. mematikan ora...|[maksud, tujuan, ...|[berniat, berniat...|
|     aku|          3|              |[1 membuat pernya...|[1. Dia tidak tah...|                  []|                  []|
|      ke|          2|              |                  []|                  []|                  []|                  []|
|    sini|          4|              |[1. = di sini di ...|[1. Mahkamah Ting...|                  []|                  []|
|   ialah|          5|     kata nama|[1. kata penghubu...|[1. punca kemalan...|                  []|[seia sekata, ber...|
|  hendak|          6|    kata tugas|[1 ingin mempunya...|[1. Budak itu hen...|[mahu, ingin, ber...|[menghendaki, keh...|
|berjumpa|          8|              |                  []|                  []|                  []|                  []|
|     dgn|          3|              |                  []|                  []|                  []|                  []|
|  engkau|          6|              |                  []|                  []|                  []|                  []|
|  hatiku|          6|              |                  []|                  []|                  []|                  []|
|     pun|          3|    kata tugas|[1 juga, 2 saja; ...|[1. Saya pun hend...|                  []|                  []|
|  hendak|          6|    kata tugas|[1 ingin mempunya...|[1. Budak itu hen...|[mahu, ingin, ber...|[menghendaki, keh...|
+--------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Transform All Invalid Datatype to Better Form

```
from pyspark.sql.functions import col, concat_ws

for cols in ['definitions', 'sentences_eg', 'synonyms', 'derived_words']:
    df = df.withColumn(cols, concat_ws(";", col(cols)))

df.show()
```

```
+--------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|   vocab|word_length|part_of_speech|         definitions|         sentences_eg|            synonyms|       derived_words|
+--------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|  maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|      ma|          2|              |                    |                    |                    |                    |
|     sud|          3|              |                    |                    |                    |                    |
|  maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|Definisi|          8|              |                    |                    |                    |                    |
|  tujuan|          6|              |1. arah, hala, ha...|1. tiap-tiap manu...|                    |                    |
|   hajat|          5|     kata nama|1. keinginan, keh...|1. hampalah hatin...|                    |berhajat;berhajat...|
|kehendak|          8|    kata tugas|                    |                    |mahu;ingin;berhas...|menghendaki;kehen...|
|    niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|maksud;tujuan;ing...|berniat;berniat-n...|
|     aku|          3|              |1 membuat pernyat...|1. Dia tidak tahu...|                    |                    |
|      ke|          2|              |                    |                    |                    |                    |
|    sini|          4|              |1. = di sini di t...|1. Mahkamah Tingg...|                    |                    |
|   ialah|          5|     kata nama|1. kata penghubun...|1. punca kemalang...|                    |seia sekata;beria...|
|  hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
|berjumpa|          8|              |                    |                    |                    |                    |
|     dgn|          3|              |                    |                    |                    |                    |
|  engkau|          6|              |                    |                    |                    |                    |
|  hatiku|          6|              |                    |                    |                    |                    |
|     pun|          3|    kata tugas|1 juga;2 saja; ju...|1. Saya pun henda...|                    |                    |
|  hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
+--------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Store the DataFrame in the form of CSV

```
df.coalesce(1).write.csv('lexicon_data', header=True, mode='overwrite')
```

```
from python_script.hdfs_commander import *

run_hdfs_command('hdfs dfs -mv lexicon_data/part-00000-*.csv lexicon_data/non_processed_data.csv')
run_hdfs_command('hdfs dfs -get lexicon_data/non_processed_data.csv lexicon_data/non_processed_data.csv')
```

Terminate Spark Session

```
spark.stop()
```

1.3.2 data_preparation.ipynb

Data Processing (using PySpark)

Create Spark Session

[20]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
```

Read the Non-Processed Data from CSV

[21]:
```
df = spark.read.csv('lexicon_data/non_processed_data.csv', header=True)
```

[22]:
```
df.show()
```

```
+--------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+
|   vocab|word_length|part_of_speech|         definitions|        sentences_eg|            synonyms|       derived_words|
+--------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+
|  maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|      ma|          2|         NULL|                NULL|                NULL|                NULL|                NULL|
|     sud|          3|         NULL|                NULL|                NULL|                NULL|                NULL|
|  maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|Definisi|          8|         NULL|                NULL|                NULL|                NULL|                NULL|
|  tujuan|          6|         NULL|1. arah, hala, ha...|1. tiap-tiap manu...|                NULL|                NULL|
|   hajat|          5|     kata nama|1. keinginan, keh...|1. hampalah hatin...|                NULL|berhajat;berhajat...|
|kehendak|          8|    kata tugas|                NULL|                NULL|mahu;ingin;berhas...|menghendaki;kehen...|
|    niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|maksud;tujuan;ing...|berniat;berniat-n...|
|     aku|          3|         NULL|1 membuat pernyat...|1. Dia tidak tahu...|                NULL|                NULL|
|      ke|          2|         NULL|                NULL|                NULL|                NULL|                NULL|
|    sini|          4|         NULL|1. = di sini di t...|1. Mahkamah Tingg...|                NULL|                NULL|
|   ialah|          5|     kata nama|1. kata penghubun...|1. punca kemalang...|                NULL|seia sekata;beria...|
|  hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
|berjumpa|          8|         NULL|                NULL|                NULL|                NULL|                NULL|
|     dgn|          3|         NULL|                NULL|                NULL|                NULL|                NULL|
|  engkau|          6|         NULL|                NULL|                NULL|                NULL|                NULL|
|  hatiku|          6|         NULL|                NULL|                NULL|                NULL|                NULL|
|     pun|          3|    kata tugas|1 juga;2 saja; ju...|1. Saya pun henda...|                NULL|                NULL|
|  hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
+--------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Convert all the Records in `vocab` Column to Lowercase

[23]:
```
from pyspark.sql.functions import *

# convert the data in vocab column to lowercase
df = df.withColumn(
    'vocab',
    lower(col('vocab'))
)

df.show()
```

```
+--------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+
|   vocab|word_length|part_of_speech|         definitions|        sentences_eg|            synonyms|       derived_words|
+--------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+
|  maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|      ma|          2|         NULL|                NULL|                NULL|                NULL|                NULL|
|     sud|          3|         NULL|                NULL|                NULL|                NULL|                NULL|
|  maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|definisi|          8|         NULL|                NULL|                NULL|                NULL|                NULL|
|  tujuan|          6|         NULL|1. arah, hala, ha...|1. tiap-tiap manu...|                NULL|                NULL|
|   hajat|          5|     kata nama|1. keinginan, keh...|1. hampalah hatin...|                NULL|berhajat;berhajat...|
|kehendak|          8|    kata tugas|                NULL|                NULL|mahu;ingin;berhas...|menghendaki;kehen...|
|    niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|maksud;tujuan;ing...|berniat;berniat-n...|
|     aku|          3|         NULL|1 membuat pernyat...|1. Dia tidak tahu...|                NULL|                NULL|
|      ke|          2|         NULL|                NULL|                NULL|                NULL|                NULL|
|    sini|          4|         NULL|1. = di sini di t...|1. Mahkamah Tingg...|                NULL|                NULL|
|   ialah|          5|     kata nama|1. kata penghubun...|1. punca kemalang...|                NULL|seia sekata;beria...|
|  hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
|berjumpa|          8|         NULL|                NULL|                NULL|                NULL|                NULL|
|     dgn|          3|         NULL|                NULL|                NULL|                NULL|                NULL|
|  engkau|          6|         NULL|                NULL|                NULL|                NULL|                NULL|
|  hatiku|          6|         NULL|                NULL|                NULL|                NULL|                NULL|
|     pun|          3|    kata tugas|1 juga;2 saja; ju...|1. Saya pun henda...|                NULL|                NULL|
|  hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
+--------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Study Anomaly in `part_of_speech` Column

[24]:
```
df.select('part_of_speech').distinct().show()
```

```
+--------------------+
|      part_of_speech|
+--------------------+
|sedap, lazat, ser...|
|empiris, saintifi...|
|segalanya, sekali...|
|bernafas, tumbuh,...|
|hal, perkara, per...|
|             adverba|
|           adjektif:|
|bersungguh-sunggu...|
|            adjektif|
|kemampuan, keupay...|
|       sinar, kilau,|
|           adjektif,|
|          kata nama,|
|had, takat, sempa...|
|sudah lazim, buka...|
|        tidak sampai|
|           kata nama|
|          kata tugas|
|                sisa|
|          kata kerja|
+--------------------+
only showing top 20 rows
```

In here, it is found that

- there exist records that are not part of speech
- some proper part of speech has punctuation joined behind

Hence, to standardize with the data, the following cleansing processes are done.

Remove Punctuation from the Columns

[25]:
```
df = df.withColumn(
    'part_of_speech',
    regexp_replace(col('part_of_speech'), "[^\w\s]", "")
)

df.select('part_of_speech').distinct().show()
```

```
+--------------------+
|      part_of_speech|
+--------------------+
|segalanya sekalia...|
|empiris saintifik...|
|kemampuan keupaya...|
|             adverba|
|            adjektif|
|hal perkara peris...|
|sudah lazim bukan...|
|  had takat sempadan|
|         sinar kilau|
|bernafas tumbuh m...|
|        tidak sampai|
|           kata nama|
|          kata tugas|
|sedap lazat seron...|
|bersungguhsungguh...|
|                sisa|
|          kata kerja|
|                NULL|
+--------------------+
```

Set the Values that are not Part of Speech to NULL

the allowed part of speech values include *kata nama* (noun), *kata tugas* or *kata kerja* (verb), *adjektif* (adjective), and *adverba* (adverb).

[26]:
```
allowed_pos = ['kata nama', 'kata tugas', 'kata kerja', 'adjektif', 'adverba']

# remove the data that is not considered part of speech
df = df.withColumn(
    'part_of_speech',
    when(col('part_of_speech').isin(allowed_pos), col('part_of_speech')).otherwise(None)
)

df.select('part_of_speech').distinct().show()
df.show()
```

```
+--------------+
|part_of_speech|
+--------------+
|       adverba|
|      adjektif|
|     kata nama|
|    kata tugas|
|    kata kerja|
|          NULL|
+--------------+
+--------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|   vocab|word_length|part_of_speech|         definitions|        sentences_eg|            synonyms|       derived_words|
+--------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|  maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|      ma|          2|          NULL|                NULL|                NULL|                NULL|                NULL|
|     sud|          3|          NULL|                NULL|                NULL|                NULL|                NULL|
|  maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|definisi|          8|          NULL|                NULL|                NULL|                NULL|                NULL|
|  tujuan|          6|          NULL|1. arah, hala, ha...|1. tiap-tiap manu...|                NULL|                NULL|
|   hajat|          5|     kata nama|1. keinginan, keh...|1. hampalah hatin...|                NULL|berhajat;berhajat...|
|kehendak|          8|    kata tugas|                NULL|                NULL|mahu;ingin;berhas...|menghendaki;kehen...|
|    niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|maksud;tujuan;ing...|berniat;berniat-n...|
|     aku|          3|          NULL|1 membuat pernyat...|1. Dia tidak tahu...|                NULL|                NULL|
|      ke|          2|          NULL|                NULL|                NULL|                NULL|                NULL|
|    sini|          4|          NULL|1. = di sini di t...|1. Mahkamah Tingg...|                NULL|                NULL|
|   ialah|          5|     kata nama|1. kata penghubun...|1. punca kemalang...|                NULL|seia sekata;beria...|
|  hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
|berjumpa|          8|          NULL|                NULL|                NULL|                NULL|                NULL|
|     dgn|          3|          NULL|                NULL|                NULL|                NULL|                NULL|
|  engkau|          6|          NULL|                NULL|                NULL|                NULL|                NULL|
|  hatiku|          6|          NULL|                NULL|                NULL|                NULL|                NULL|
|     pun|          3|    kata tugas|1 juga;2 saja; ju...|1. Saya pun henda...|                NULL|                NULL|
|  hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
+--------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Drop Duplicated Records

[27]:
```
# Drop duplicated records
df = df.dropDuplicates()
df.show()
print("Number of Unique Record:", df.count())
```

```
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|     vocab|word_length|part_of_speech|         definitions|        sentences_eg|            synonyms|       derived_words|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|     perlu|          5|    kata tugas|1. keharusan (yg ...|1. sembahyang dan...|mesti;wajib;harus...|memerlukan;keperluan|
|     waktu|          5|     kata nama|1. rangkaian atau...|1. secara rohania...|masa;jangka masa;...|sewaktu;sewaktu-w...|
|    mundur|          6|      adjektif|1. (berjalan, ber...|1. dia mundur set...|merosot;meleset;m...|memundurkan;kemun...|
|     sifat|          5|     kata nama|1. keadaan atau r...|1. adapun sifatny...|keadaan;rupa;tand...|bersifat;menyifatkan|
|      suci|          4|      adjektif|1. bersih (dr seg...|1. makanan yg suc...|bersih;murni;tida...|    bersuci;kesucian|
|     suara|          5|     kata nama|1. bunyi manusia ...|1. suara itu nyat...|     bunyi;nada;sora|bersuara;menyuarakan|
|   sanggup|          7|    kata kerja|1. bersedia (utk ...|1. tiada siapa yg...|bersedia;rela;sud...|menyanggupi;kesan...|
|    timbul|          6|    kata kerja|1. naik dr dlm ai...|1. ikan itu timbu...|naik;terapung;mel...|        menimbulkan|
|memutuskan|         10|    kata kerja|1. menjadikan put...|1. aneh sekali pe...|retas;rabit;relah...|memutuskan;terput...|
|      niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|maksud;tujuan;ing...|berniat;berniat-n...|
|     harga|          5|     kata nama|1. nilai sesuatu ...|1. kalau adapun b...|bayaran;belanja;p...|berharga;mengharg...|
| perempuan|          9|     kata nama|1. wanita, lwn la...|1. perempuan itu ...|wanita;kaum Hawa;...|       keperempuanan|
| bermaksud|          9|     kata nama|1. bertujuan, ber...|1. kami bermaksud...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|     kerja|          5|     kata nama|1. usaha (kegiata...|1. beberapa mingg...|khidmat;pekerjaan...|sekerja;bekerja;m...|
|    maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|menjelaskan|         11|      adjektif|1. menerangkan (l...|1. seorang pegawa...|terang;nyata;keta...|berjelas-jelas;me...|
|    hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
|penjelasan|         10|      adjektif|1. keterangan (yg...|1. Ali meminta pe...|terang;nyata;keta...|berjelas-jelas;me...|
|      tahu|          4|    kata kerja|1. maklum akan ke...|1. entah kapal pe...|maklum;faham;meng...|tahu-tahu;mengeta...|
|      jauh|          4|      adjektif|1. besar jaraknya...|1. rumahnya agak ...|tidak dekat;besar...|berjauh;menjauhi;...|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows

Number of Unique Record: 497
```

Check with the Vocabs that Definition is Undefined

[28]:
```
# check the records with no definition scrapped
null_def_df = df.filter(col('definitions').isNull())

null_def_df.show()
print("Number of Unique Record with No Definition:", null_def_df.count())
```

```
+-------------+-----------+--------------+-----------+------------+--------+--------------------+
|        vocab|word_length|part_of_speech|definitions|sentences_eg|synonyms|       derived_words|
+-------------+-----------+--------------+-----------+------------+--------+--------------------+
|         agar|          4|          NULL|       NULL|        NULL|    NULL|                NULL|
|        dewan|          5|          NULL|       NULL|        NULL|    NULL|                NULL|
|        harta|          5|     kata nama|       NULL|        NULL|    NULL|            berharta|
|      diatasi|          7|          NULL|       NULL|        NULL|    NULL|                NULL|
|    keupayaan|          9|     kata nama|       NULL|        NULL|    NULL|seupaya-upayanya;...|
|   penindasan|         10|          NULL|       NULL|        NULL|    NULL|                NULL|
|     rakannya|          8|          NULL|       NULL|        NULL|    NULL|                NULL|
|  mengajarkan|         11|          NULL|       NULL|        NULL|    NULL|                NULL|
|  berkumpulan|         11|          NULL|       NULL|        NULL|    NULL|                NULL|
|         drpd|          4|          NULL|       NULL|        NULL|    NULL|                NULL|
|          per|          3|          NULL|       NULL|        NULL|    NULL|                NULL|
|       apakah|          6|          NULL|       NULL|        NULL|    NULL|                NULL|
|   samarsamar|         10|          NULL|       NULL|        NULL|    NULL|                NULL|
| berkelayakan|         12|      adjektif|       NULL|        NULL|    NULL|selayaknya;melaya...|
|    menjelasi|          9|          NULL|       NULL|        NULL|    NULL|                NULL|
|        kimia|          5|          NULL|       NULL|        NULL|    NULL|                NULL|
|bergantiganti|         13|          NULL|       NULL|        NULL|    NULL|                NULL|
|   kakitangan|         10|     kata nama|       NULL|        NULL|    NULL|                NULL|
|    hutangnya|          9|          NULL|       NULL|        NULL|    NULL|                NULL|
|     wujudnya|          8|          NULL|       NULL|        NULL|    NULL|                NULL|
+-------------+-----------+--------------+-----------+------------+--------+--------------------+
only showing top 20 rows

Number of Unique Record with No Definition: 287
```

Redefine the Records to be Kept and Removed

The rules are described as follows:

Steps	Description
1	The words with and with no definition are separated
2	For all the word with no definition: • If the word is existed in the `derived_words` of the words with definition, then **keep** • If the word is not existed in the `derived_words` of the words with definition, then **delete**
3	Combine the two dataframes (with definition and with no definition)

The code segment of so processing is as follows.

[29]:
```
# Split the DataFrame into two parts: with definition and without definition
null_df = df.filter(col("definitions").isNull())
not_null_df = df.filter(col("definitions").isNotNull())

# If vocab is a derived_words of other vocab even if it does not have definition, keep.
# Else (no definition and is not derived_words of others), delete.
matched_df = null_df.alias("null_df").join(
    not_null_df.alias("not_null_df"),
    expr("instr(null_df.derived_words, not_null_df.vocab) > 0"),
    "inner"
).select("null_df.*")

df = not_null_df.union(matched_df)
df.show()
print("Number of Records after Definitions Cleansing:", df.count())
```

```
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|     vocab|word_length|part_of_speech|         definitions|        sentences_eg|            synonyms|       derived_words|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|      perlu|          5|    kata tugas|1. keharusan (yg ...|1. sembahyang dan...|mesti;wajib;harus...|memerlukan;keperluan|
|      waktu|          5|     kata nama|1. rangkaian atau...|1. secara rohania...|masa;jangka masa;...|sewaktu;sewaktu-w...|
|     mundur|          6|      adjektif|1. (berjalan, ber...|1. dia mundur set...|merosot;meleset;m...|memundurkan;kemun...|
|      sifat|          5|     kata nama|1. keadaan atau r...|1. adapun sifatny...|keadaan;rupa;tand...|bersifat;menyifatkan|
|       suci|          4|      adjektif|1. bersih (dr seg...|1. makanan yg suc...|bersih;murni;tida...|    bersuci;kesucian|
|      suara|          5|     kata nama|1. bunyi manusia ...|1. suara itu nyat...|     bunyi;nada;sora|bersuara;menyuarakan|
|    sanggup|          7|    kata kerja|1. bersedia (utk ...|1. tiada siapa yg...|bersedia;rela;sud...|menyanggupi;kesan...|
|     timbul|          6|    kata kerja|1. naik dr dlm ai...|1. ikan itu timbu...|naik;terapung;mel...|         menimbulkan|
| memutuskan|         10|    kata kerja|1. menjadikan put...|1. aneh sekali pe...|retas;rabit;relah...|memutuskan;terput...|
|       niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|maksud;tujuan;ing...|berniat;berniat-n...|
|      harga|          5|     kata nama|1. nilai sesuatu ...|1. kalau adapun b...|bayaran;belanja;p...|berharga;mengharg...|
|  perempuan|          9|     kata nama|1. wanita, lwn la...|1. perempuan itu ...|wanita;kaum Hawa;...|       keperempuanan|
|  bermaksud|          9|     kata nama|1. bertujuan, ber...|1. kami bermaksud...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|      kerja|          5|     kata nama|1. usaha (kegiata...|1. beberapa mingg...|khidmat;pekerjaan...|sekerja;bekerja;m...|
|     maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|menjelaskan|         11|      adjektif|1. menerangkan (l...|1. seorang pegawa...|terang;nyata;keta...|berjelas-jelas;me...|
|     hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
| penjelasan|         10|      adjektif|1. keterangan (yg...|1. Ali meminta pe...|terang;nyata;keta...|berjelas-jelas;me...|
|       tahu|          4|    kata kerja|1. maklum akan ke...|1. entah kapal pe...|maklum;faham;meng...|tahu-tahu;mengeta...|
|       jauh|          4|      adjektif|1. besar jaraknya...|1. rumahnya agak ...|tidak dekat;besar...|berjauh;menjauhi;...|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows

Number of Records after Definitions Cleansing: 227
```

After this steps, there remained only **227 records**.

Check again with the Duplicated Records

[30]:
```
# check if there is duplicated values
print('Distinct Record Count:', df.distinct().count())
```

```
Distinct Record Count: 221
```

We found that there are **221 distinct records**, which unmatched with the result of previous step.

Hence, we drop again duplicated records, and verify the data count.

[31]:
```
df = df.dropDuplicates()
df.show()
print('Number of Unique Records:', df.count())
```

```
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|     vocab|word_length|part_of_speech|         definitions|        sentences_eg|            synonyms|       derived_words|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|      perlu|          5|    kata tugas|1. keharusan (yg ...|1. sembahyang dan...|mesti;wajib;harus...|memerlukan;keperluan|
|      waktu|          5|     kata nama|1. rangkaian atau...|1. secara rohania...|masa;jangka masa;...|sewaktu;sewaktu-w...|
|     mundur|          6|      adjektif|1. (berjalan, ber...|1. dia mundur set...|merosot;meleset;m...|memundurkan;kemun...|
|      sifat|          5|     kata nama|1. keadaan atau r...|1. adapun sifatny...|keadaan;rupa;tand...|bersifat;menyifatkan|
|       suci|          4|      adjektif|1. bersih (dr seg...|1. makanan yg suc...|bersih;murni;tida...|    bersuci;kesucian|
|      suara|          5|     kata nama|1. bunyi manusia ...|1. suara itu nyat...|     bunyi;nada;sora|bersuara;menyuarakan|
|    sanggup|          7|    kata kerja|1. bersedia (utk ...|1. tiada siapa yg...|bersedia;rela;sud...|menyanggupi;kesan...|
|     timbul|          6|    kata kerja|1. naik dr dlm ai...|1. ikan itu timbu...|naik;terapung;mel...|         menimbulkan|
| memutuskan|         10|    kata kerja|1. menjadikan put...|1. aneh sekali pe...|retas;rabit;relah...|memutuskan;terput...|
|       niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|maksud;tujuan;ing...|berniat;berniat-n...|
|      harga|          5|     kata nama|1. nilai sesuatu ...|1. kalau adapun b...|bayaran;belanja;p...|berharga;mengharg...|
|  perempuan|          9|     kata nama|1. wanita, lwn la...|1. perempuan itu ...|wanita;kaum Hawa;...|       keperempuanan|
|  bermaksud|          9|     kata nama|1. bertujuan, ber...|1. kami bermaksud...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|      kerja|          5|     kata nama|1. usaha (kegiata...|1. beberapa mingg...|khidmat;pekerjaan...|sekerja;bekerja;m...|
|     maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|menjelaskan|         11|      adjektif|1. menerangkan (l...|1. seorang pegawa...|terang;nyata;keta...|berjelas-jelas;me...|
|     hendak|          6|    kata tugas|1 ingin mempunyai...|1. Budak itu hend...|mahu;ingin;berhas...|menghendaki;kehen...|
| penjelasan|         10|      adjektif|1. keterangan (yg...|1. Ali meminta pe...|terang;nyata;keta...|berjelas-jelas;me...|
|       tahu|          4|    kata kerja|1. maklum akan ke...|1. entah kapal pe...|maklum;faham;meng...|tahu-tahu;mengeta...|
|       jauh|          4|      adjektif|1. besar jaraknya...|1. rumahnya agak ...|tidak dekat;besar...|berjauh;menjauhi;...|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows

Number of Unique Records: 221
```

Remove Duplicated Substring in Definitions and Sentences Example

[32]:
```
import re

df_rows = df.collect()

vocabs = []
for row in df_rows:
    vocab = row[0]

    vocabs.append(vocab)

new_def = []
for row in df_rows:
    old_def = row[3]

    if old_def != None:
        while True:
            old_def = old_def.replace(';', '; ')
            out = re.sub(r'(?<!\S)(\S+(?:\s\S+)*)\s+\1(?!\S)', '\\1', old_def)
            if out == old_def:
                break
            old_def = out

    new_def.append(old_def)

new_sentences_eg = []
for row in df_rows:
    old_sentences = row[4]

    if old_sentences != None:
        while True:
            old_sentences = old_sentences.replace(';', '; ')
            out = re.sub(r'(?<!\S)(\S+(?:\s\S+)*)\s+\1(?!\S)', '\\1', old_sentences)
            if out == old_sentences:
                break
            old_sentences = out

    new_sentences_eg.append(out)

temp_df_data = []
for i in range(len(vocabs)):
    temp_df_data.append([vocabs[i], new_def[i], new_sentences_eg[i]])
```

[33]:
```
new_def_sent_df = spark.createDataFrame(temp_df_data, ["vocab", "new_def", "new_sentences"])

df = df.select('vocab', 'word_length', 'part_of_speech', 'synonyms', 'derived_words').join(new_def_sent_df, on='vocab')
df = df.withColumnRenamed('new_def', 'definitions').withColumnRenamed('new_sentences', 'sentences_eg')
df = df.select('vocab', 'word_length', 'part_of_speech', 'definitions', 'sentences_eg', 'synonyms', 'derived_words')
```

[34]:
```
df = df.dropDuplicates()
df.show()
```

```
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|     vocab|word_length|part_of_speech|         definitions|        sentences_eg|            synonyms|       derived_words|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|    sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|satu-satunya;cuma...|bersahaja;bersaha...|
|     kerja|          5|     kata nama|1. usaha (kegiata...|1. beberapa mingg...|khidmat;pekerjaan...|sekerja;bekerja;m...|
|      kaum|          4|     kata nama|1. puak, suku ban...|1. banyak kaum or...|suku;kabilah;taef...|   berkaum;perkauman|
|     jelas|          5|      adjektif|1. terang (perkat...|1. suara itu tida...|terang;nyata;keta...|berjelas-jelas;me...|
|     waktu|          5|     kata nama|1. rangkaian atau...|1. secara rohania...|masa;jangka masa;...|sewaktu;sewaktu-w...|
|      kuat|          4|      adjektif|1. banyak tenagan...|1. walaupun ia su...|gagah;berdaya;ber...|sekuat-kuatnya;be...|
|memaksudkan|         11|     kata nama|1. menghendaki, m...|1. dialah yg dima...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|      tahu|          4|    kata kerja|1. maklum akan ke...|1. entah kapal pe...|maklum;faham;meng...|tahu-tahu;mengeta...|
| merupakan|          9|     kata nama|1. membentuk (mem...|1. peti rahsia it...|paras;muka;wajah;...|rupanya;rupa-rupa...|
|     ramai|          5|      adjektif|1. riuh-rendah, h...|1. orang menyabun...|bising;riuh;hiruk...|seramai;beramai-r...|
|     penuh|          5|      adjektif|1. seluruhnya ber...|1. tong itu penuh...|berisi;tepu;padat...|sepenuhnya;memenu...|
|    rakyat|          6|     kata nama|1. seluruh pendud...|1. rakyat Malaysi...|warganegara;pendu...|          kerakyatan|
|      wang|          4|     kata nama|1. alat pertukara...|1. satuan wang ne...|duit;fulus;pitis;...|  berwang;mengewangi|
|    mundur|          6|      adjektif|1. (berjalan, ber...|1. dia mundur set...|merosot;meleset;m...|memundurkan;kemun...|
|penjelasan|         10|      adjektif|1. keterangan (yg...|1. Ali meminta pe...|terang;nyata;keta...|berjelas-jelas;me...|
|      niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|maksud;tujuan;ing...|berniat;berniat-n...|
| bermaksud|          9|     kata nama|1. bertujuan, ber...|1. kami bermaksud...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|   berniat|          7|     kata nama|1. bertujuan (aka...|1. aku pun tidak ...|maksud;tujuan;ing...|berniat;berniat-n...|
|      jiwa|          4|     kata nama|1. nyawa, roh (yg...|1. buku itu mence...|nyawa;roh;semanga...|    berjiwa;kejiwaan|
|menyatakan|         10|      adjektif|1. menjadikan nya...|1. di dlm buku in...|terang;jelas;keta...|menyatakan;ternya...|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Write Processed Data into a CSV File

[35]:
```
df.coalesce(1).write.csv('processed_data', header=True, mode='overwrite')
```

[36]:
```
from python_script.hdfs_commander import *

run_hdfs_command('hdfs dfs -mv processed_data/part-00000-*.csv lexicon_data/processed_data.csv')
run_hdfs_command('hdfs dfs -rm -R processed_data')
run_hdfs_command('hdfs dfs -get lexicon_data/processed_data.csv lexicon_data/processed_data.csv')
```

```
Output: Deleted processed_data
```

Terminate Spark Session

[37]:
```
spark.stop()
```

1.3.3 web_scrapping.py

```
import requests
from bs4 import BeautifulSoup as bs
import sys, os

class Word:

    def __init__(self, keyword):
        self.keyword = keyword
        self.lexicon = []
        self.definition_str = ''

    def getKeyword(self):
        return self.keyword

    def getLexicon(self):
        return self.lexicon

    def getDefinitionStr(self):
        return self.definition_str

    def extractLexicon(self):
        try:
            r =
       requests.get(f"https://prpm.dbp.gov.my/cari1?keyword={self.keyword}")
            soup = bs(r.content, 'html.parser')
            word_length = len(self.keyword)

             definition_soup = soup.find('div',
                                 class_='tab-pane fade in active')

            try:
                definition_soup.font.decompose()
            except:
                pass

            try:
                self.definition_str = definition_soup.get_text()
            except:
                pass

            def_list = []
            eg_list = []

            if self.definition_str.find('1') != -1:
                i = 1
                while True:
                    try:
                        def_start_index =
                            self.definition_str.index(f'{i:.0f}')
                        def_end_index =
                           len(self.definition_str[:def_start_index]) +
                           self.definition_str[def_start_index:].index(':')
```

```
                eg_start_index =
                   len(self.definition_str[:def_start_index]) +
                   self.definition_str[def_start_index:].index(':')
                eg_end_index =
                   len(self.definition_str[:eg_start_index]) +
                   self.definition_str[eg_start_index:].index(';')

                def_temp =
                   self.definition_str[def_start_index:def_end_index]
                   .replace(': ', '')
                eg_temp = f'{i}. ' +
                   self.definition_str[eg_start_index+2:eg_end_index]
                   .replace('~', self.keyword)

                while def_temp.find('\xad') != -1:
                    def_temp = def_temp.replace('\xad', '')
                while eg_temp.find('\xad') != -1:
                    eg_temp = eg_temp.replace('\xad', '')

                def_list.append(def_temp)
                eg_list.append(eg_temp)
                i += 1
            except Exception as e:
                break
    else:
        try:
            def_start_index = self.definition_str
                   .index(f'Definisi :')
            def_end_index =
                   len(self.definition_str[:def_start_index]) +
                   self.definition_str[def_start_index:].index(':')

            eg_start_index =
                   len(self.definition_str[:def_start_index]) +
                   self.definition_str[def_start_index:].index(':')
            eg_end_index = len(self.definition_str[:eg_start_index])
                   + self.definition_str[eg_start_index:].index('.')

            def_temp =
                   self.definition_str[def_start_index:def_end_index]
                   .replace(': ', '')
            eg_temp = f'{i}. ' +
                   self.definition_str[eg_start_index+2:eg_end_index]
                   .replace('~', self.keyword)

            while def_temp.find('\xad') != -1:
                def_temp = def_temp.replace('\xad', '')
            while eg_temp.find('\xad') != -1:
                eg_temp = eg_temp.replace('\xad', '')

            def_list.append(def_temp)
            eg_list.append(eg_temp)
        except:
            pass

    part_of_speech = ''
    synonyms = []
    derived_words = []
```

```
try:
    thesaurus_soup = soup.find('span',
        id='MainContent_SearchInfoTesaurus_lblTesaurus')
    thesaurus_str = thesaurus_soup.get_text()

    pos_start_index = thesaurus_str.index('(')
    pos_end_index = thesaurus_str.index(')')
    part_of_speech =
        thesaurus_str[pos_start_index+1:pos_end_index]

    thesaurus_str =
        thesaurus_str
        .replace(thesaurus_str[:pos_end_index+1], '')
    synonyms_str = thesaurus_str[:thesaurus_str
        .index('Kata Terbitan')]
    derived_words_str = thesaurus_str[thesaurus_str
        .index('Kata Terbitan : ') + 16:]

    i = 1
    while True:
        try:
            synonym_start_index =
               synonyms_str.index(f'{i}. Bersinonim dengan ')

            if synonyms_str.find(f'{i+1}.') != -1:
                synonym_end_index =
                     synonyms_str.index(f'{i+1}.')

                synonym_pure =
                     synonyms_str[synonym_start_index+21:synonym
                     _end_index].replace('\n', '')

                synonym_pure = synonym_pure.replace(':', ',')
                splitted_synonym = synonym_pure.split(', ')

                for i in range(len(splitted_synonym) - 1):
                    synonyms.append(splitted_synonym[i])

                synonyms_str =
                     synonyms_str.replace(synonyms_str[synonym_s
                     tart_index:synonym_end_index], '')
            else:
                synonym_pure =
                     synonyms_str[synonym_start_index+21:].repla
                     ce('\n','')

                synonym_pure = synonym_pure.replace(':', ',')
                splitted_synonym = synonym_pure.split(', ')

                for i in range(len(splitted_synonym) - 1):
                    synonyms.append(splitted_synonym[i])

            i += 1
        except:
            break

    while True:
        try:
            splitted_derived_words = derived_words_str
```

```
                                                .split(', ')

                    for i in range(len(splitted_derived_words) - 1):
                        derived_words.append(splitted_derived_words[i])
                    break
                except:
                    break
        except:
            pass

        try:
            self.lexicon.append(self.keyword)
            self.lexicon.append(word_length)
            self.lexicon.append(part_of_speech)
            self.lexicon.append(def_list)
            self.lexicon.append(eg_list)
            self.lexicon.append(synonyms)
            self.lexicon.append(derived_words)
        except:
            pass
    except Exception as e:
        exc_type, exc_obj, exc_tb = sys.exc_info()
        fname = os.path.split(exc_tb.tb_frame.f_code.co_filename)[1]
        print(exc_type, fname, exc_tb.tb_lineno)
```

1.4. Data Quantity

Before the scrapping of data from Google Books, we scrape data from DBP dictionary to act as a base lexicon for later lexicon analysis. We scrape the relevant texts from several words, namely ‘maksud’, ‘tenaga’, ‘jelas’, ‘negara’, ‘tanah’, and ‘asas’. With these base words, we successfully scraped an amount of **930** data. Note that these 930 data is not cleansed and processed.

Thus, followed by data cleansing and preparation, we perform processing such as convert records in columns to lowercase, remove anomalies in columns, remove punctuations, replace invalid records with NULL, drop duplicate records, etc. In detail, we check with the vocabularies in which the definitions are empty or null. This is usually caused by incomplete HTML structure or relevant pages not found (dictionary of a word undefined). Hence, the vocabularies are deleted from the lexicon dataframe.

Other than that, some of the vocabularies are derived from the other vocabularies. For example, the word ‘penjelasan’ is considered as one of the derived words of the word ‘jelas’. In this case, we do not remove the records as it may be useful in the future stages.

Thus, after proper data cleansing, we left about **219** data as base data.

2. Lexicon Creation

2.1. Data model design

In our project of this lexicon, we planned to build three types of data storage which respectively serves for different purposes. The three data storages are Hadoop Distributed File System (HDFS), MongoDB, and Neo4j Graph Database.

2.1.1 Hadoop Distributed File System (HDFS)

HDFS is the primary storage system used by a Hadoop application. It operates as a distributed file system and is designed to run on commodity hardware.It provides fault tolerance and high throughput data access to application data. Hence, HDFS is suitable for applications which have large data amounts.

HDFS has two main components, namely NameNodes and DataNodes. NameNode refers to the hardware with Linux Operating System and software. It provides the management of files, control of client's access to files, and files operations such as renaming, reading, writing, executing, etc. Meanwhile, DataNode is located at every node in a HDFS cluster. DataNodes help to control the data storage of the HDFS system by performing operations whenever client requests.

In our case, HDFS acts as a convenience in data retrieval due to its simplicity. For most of the dataframe, we store them in the form of comma separated values (csv) file. For example, the data we retrieved from google books are stored in the name of 'google_books_xxxx.csv' in HDFS.

2.1.2 MongoDB

MongoDB is a NoSQL database product developed by MongoDB Inc. It is a source-available, cross-platform, and document-oriented database application which utilizes JSON-like documents with optional schemas. (Wikipedia contributors, 2024b)

Fig. 2.1 *Logo of MongoDB*

```
_id: ObjectId('675e3e625084494eb96fd654')
seq_no : "01"
vocab : "asing"
word_length : 5
part_of_speech : null
definitions : "1. berlainan, terpisah, tersendiri; berasa diri ~ (di sesuatu
              tempat) …"
sentences_eg : "1. banyak bangsa asing di negeri ini;2. banyak bangsa asing di
               negeri …"
derived_words : null
synonyms : null
antonyms : null
freq_of_usage : 2
polarity : "negative"
subjectivity : "partially fact-driven"
subjectivity_score : 0.46875
```


Fig. 2.1 *Example of Data Entry in JSON-Like Structure in MongoDB*

In our case, we use MongoDB to serve as a primary storage of our data scraped from Google Books. Its simplicity allows us to retrieve and read data easily from the database. With the collaboration of Python codes, the process of data processing with MongoDB can be greatly simplified as the data obtained into Python is a list of dictionaries.

```
[{'_id': ObjectId('675a569b1ef52eda86b81bc1'),
  'antonyms': None,
  'definitions': '1. pergi bersama; 2. bersama-sama melakukan (ditimpa dll) '
                 'sesuatu, tidak terkecuali drpd (sesuatu kejadian, perbuatan, '
                 'dll), ikut serta; 3. patuh kpd (undang-undang, janji, dll), '
                 'taat kpd; 48;  berturut-turut, berturutan 1. tidak '
                 'ber-antara antara satu dgn lain, lepas satu-satu, '
                 'terus-menerus dgn teratur (tidak berselang), tidak '
                 'berhentinya, beruntun-runtun, bersam-bung-sambung; 5. '
                 'mengikut (berjalan dll), mengekori (dr belakang); 6. spt yg '
                 'dikhabarkan (diberitakan, dihebahkan, dll), sepanjang '
                 '(berita dll); 7. sesuai dgn, tidak bertentangan dgn, tidak '
                 'melanggar (syarat dll), selaras dgn; 8;  berturut-turut, '
                 'berturutan 1. tidak ber-antara antara satu dgn lain, lepas '
                 'satu-satu, terus-menerus dgn teratur (tidak berselang), '
                 'tidak berhentinya, beruntun-runtun, bersam-bung-sambung; 9. '
                 'ikut serta (melakukan sesuatu dll), turut serta; 10. '
                 'bergantung pd',
  'derived_words': None,
  'freq_of_usage': '6',
  'part_of_speech': None,
  'polarity': 'none',
  'sentences_eg': '1. saya turut ke mana dia pergi; 2. Ahmad yg turut terbunuh '
                  'oleh serangan hendap itu baru semalam ditemui mayatnya; 3. '
                  'sekiranya kamu turut akan nasihat ini berbahagialah kamu; '
                  '4. beliau telah kalah dlm tiga pilihan raya kecil kpd parti '
                  'lawannya secara turut; 5. dia berjalan menuju ruangan yg di '
                  'sebelah barat bangunan, aku turutnya dr belakang; 6. turut '
                  'yg diketahui sebanyak satu suku juta ringgit harga saham '
                  'telah terjual; 7. hari ini turut pendapat orang alim ialah '
                  'hari yg baik; 8. beliau telah kalah dlm tiga pilihan raya '
                  'kecil kpd parti lawannya secara turut; 9. Sultan Mansor '
                  'Shah pun turut menyebut nama Hang Tuah itu Laksamana; 10. '
                  'maju mundurnya dlm pelajarannya turut kerajinan-nya',
  'seq_no': '199',
  'subjectivity': 'fact-driven',
  'subjectivity_score': '0.18666666666666668',
  'synonyms': None,
  'vocab': 'turut',
  'word_length': '5'}]
```

Fig. 2.3 *Example of MongoDB Data Retrieved into Python Environment*

2.1.3 Neo4j Graph Database

Neo4j is a graph database management system (GDBMS) developed by Neo4j Inc. In Neo4j, there are two main components, namely **node** and **edge**. Each node represents a data point, storing all the information of the data. Edge connects node to node, indicating the relationship between two or more nodes, building a relationship network in the collection of nodes. (Wikipedia contributors, 2024c)

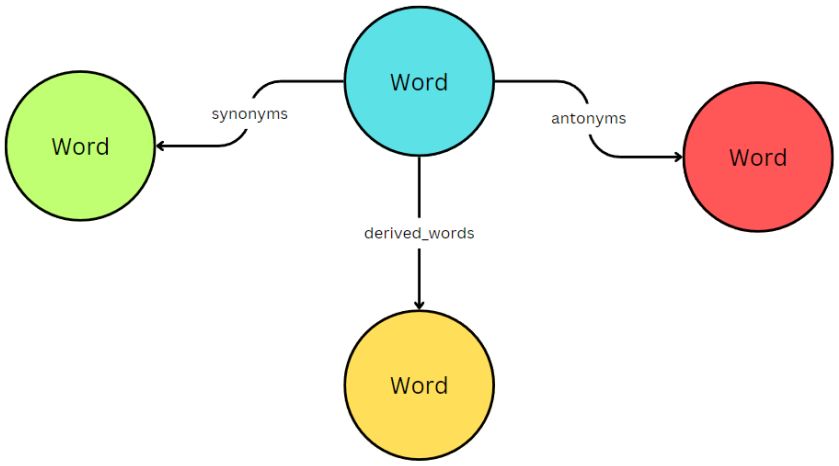


Fig. 2.4 *Data Model of the Lexicon in Neo4j Graph Database*

In our use case, each data is called a **Word**, presented in the form of a node. On this basis, nodes are related to each other with different associations such as **synonyms**, **antonyms**, and **derived words**. We use Neo4j to visualize the relationship between nodes clearer, allowing us to perform better descriptive analysis in the later stage.

2.2. Example of an entry in the data store

The table below shows an example of processed data entry in our data storage.

Column Heading	Datatype	Values
vocab	string	sahaja
word_length	integer	6
part_of_speech	string	kata tugas
definitions	string	1. yg satu-satunya, cuma, hanya, melainkan; 2. selalu (demikian), terus (demikian); 3. juga (demikian), juga (begitu); 4. barang ... pun, sebarang; 5. dgn mudah, dgn senang; 6. -lah (sbg anjuran), lebih baik; 7. sangat (utk mengeraskan kata sebelumnya), sekali; 8. sememangnya, memang; 9. sengaja, dgn niat sendiri; 10. spt biasa (tidak ditambah-tambah)
sentences_eg	string	1. tuan sahaja yg dapat menolong saya; 2. ia mengeluh sahaja; 3. muzik sama sahaja

		mustahaknya dgn bahasa atau kesusasteraan; 4. apa sahaja; 5. disembelihnya sahaja kijang yg tidak dapat berlari lagi itu; 6. makan sahaja apa yg ada; 7. mudah sahaja mengalahkan orang itu; 8. sahaja sebenarnyalah pekerjaan spt titah duli tuanku itu; 9. sahaja aku mengerjakan umrah dan ihram dengannya kerana Allah Taala; 10. dgn mata sahaja dapat dilihat bintang itu
synonyms	string	satu-satunya;cuma;hanya;melainkan;semata-mata;belaka
derived_words	string	bersahaja;bersahaja-sahaja;menyahajakan

Table 2.1 *Example of Processed Data Entry in Data Storage*

2.3. List of Python classes and Jupyter notebook(s) for lexicon:

Name of Python class(es) / notebooks	Author
database_interaction.ipynb	Chiam Zi Wei
lexicon_retrieval.ipynb	Lau Kit Lim

2.4. Code for Python Classes

2.4.1 database_interaction.ipynb

Database Interaction

Create a Spark Session and Read Content of File

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

df = spark.read.csv('lexicon_data/enriched_lexicon.csv', header=True)
df.show()
```

```
+---------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+-------------+--------+--------------------+--------------------+
|    vocab|word_length|part_of_speech|         definitions|         sentences_eg|        derived_words|            synonyms|     antonyms|freq_of_usage|polarity|        subjectivity| subjectivity_score|
+---------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+-------------+-------------+--------+--------------------+--------------------+
|  berniat|          7|     kata nama|1. bertujuan (aka...|1. aku pun tidak ...|berniat;berniat-n...|maksud;tujuan;ing...|         NULL|            0|    none|         fact-driven|                 0.0|
|   mundur|          6|      adjektif|1. (berjalan, ber...|1. dia mundur set...|memundurkan;kemun...|merosot;meleset;m...|  maju;maju|            0|    none|partially fact-dr...|0.30833333333333334|
|     kaum|          4|     kata nama|1. puak, suku ban...|1. banyak kaum or...|   berkaum;perkauman|suku;kabilah;taef...|         NULL|            5|    none|         fact-driven|                 0.1|
|     wang|          4|     kata nama|1. alat pertukara...|1. satuan wang ne...|  berwang;mengewangi|duit;fulus;pitis;...|         NULL|            6|    none|partially opinion...| 0.5777777777777777|
|    jelas|          5|      adjektif|1. terang (perkat...|1. suara itu tida...|berjelas-jelas;me...|terang;nyata;keta...|        kabur|            0|positive|         fact-driven|0.19166666666666665|
|    waktu|          5|     kata nama|1. rangkaian atau...|1. secara rohania...|sewaktu;sewaktu-w...|masa;jangka masa;...|         NULL|            0|    none|partially fact-dr...| 0.3712962962962963|
|    ramai|          5|      adjektif|1. riuh-rendah, h...|1. orang menyabun...|seramai;beramai-r...|bising;riuh;hiruk...|sunyi;sugul;lemba...|            1|positive|partially fact-dr...|0.49166666666666664|
|     kuat|          4|      adjektif|1. banyak tenagan...|1. walaupun ia su...|sekuat-kuatnya;be...|gagah;berdaya;ber...|lemah;rapuh;lemah...|            0|positive|partially fact-dr...|0.48975308641975307|
|    sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|bersahaja;bersaha...|satu-satunya;cuma...|        jarang|           13|    none|partially fact-dr...|0.45833333333333337|
|     kerja|          5|     kata nama|1. usaha (kegiata...|1. beberapa mingg...|sekerja;bekerja;m...|khidmat;pekerjaan...|         NULL|            0|    none|partially fact-dr...|0.27769360269360266|
|    rakyat|          6|     kata nama|1. seluruh pendud...|1. rakyat Malaysi...|          kerakyatan|warganegara;pendu...|pemerintah;ketua|           25|    none|partially fact-dr...|                0.26|
|     penuh|          5|      adjektif|1. seluruhnya ber...|1. tong itu penuh...|sepenuhnya;memenu...|berisi;tepu;padat...|kosong;sedikit;ku...|            4|positive|partially opinion...|                0.55|
| merupakan|          9|     kata nama|1. membentuk (mem...|1. peti rahsia it...|rupanya;rupa-rupa...|paras;muka;wajah;...|         NULL|            3|    none|      opinion-driven|                0.75|
|penjelasan|         10|      adjektif|1. keterangan (yg...|1. Ali meminta pe...|berjelas-jelas;me...|terang;nyata;keta...|        kabur|            0|    none|         fact-driven|                 0.0|
|     niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|berniat;berniat-n...|maksud;tujuan;ing...|         NULL|            0|    none|partially fact-dr...| 0.2857142857142857|
| bermaksud|          9|     kata nama|1. bertujuan, ber...|1. kami bermaksud...|bermaksud;memaksu...|tujuan;hajat;hasr...|         NULL|            1|    none|      opinion-driven|             0.84375|
|     tahu|          4|    kata kerja|1. maklum akan ke...|1. entah kapal pe...|tahu-tahu;mengeta...|maklum;faham;meng...|jahil;membiarkan|            0|    none|         fact-driven| 0.18571428571428572|
|memaksudkan|         11|     kata nama|1. menghendaki, m...|1. dialah yg dima...|bermaksud;memaksu...|tujuan;hajat;hasr...|         NULL|            0|negative|         fact-driven|               0.175|
|     masa|          4|     kata nama|1. waktu, ketika;...|1. masa dan tempa...|              semasa|waktu;ketika;deti...|         NULL|            3|    none|         fact-driven| 0.08333333333333333|
|     jauh|          4|      adjektif|1. besar jaraknya...|1. rumahnya agak ...|berjauh;menjauhi;...|tidak dekat;besar...|        dekat|            2|positive|partially opinion...| 0.6583333333333334|
+---------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+-------------+-------------+--------+--------------------+--------------------+
only showing top 20 rows
```

Create Connection to MongoDB

```
from pymongo import MongoClient
import pprint

uri = "mongodb+srv://jordan:jordan0551@cluster0.nzsyn.mongodb.net/?retryWrites=true&w=majority&appName=Cluster0"
client = MongoClient(uri)

db = client["lexicondb"]

print("Current database: " + db.name)
print(f"Collections in {db.name}: ")
pprint.pprint(db.list_collection_names())
```

```
Current database: lexicondb
Collections in lexicondb:
['inventory']
```

Store Data into MongoDB in Document

```
df_rows = df.collect()
inventory = db.inventory

data_list = []
for i in range(len(df_rows)):
    data_dict = {
        'seq_no'            : f'{i+1:02d}',
        'vocab'             : df_rows[i][0],
        'word_length'       : df_rows[i][1],
        'part_of_speech'    : df_rows[i][2],
        'definitions'       : df_rows[i][3],
        'sentences_eg'      : df_rows[i][4],
        'derived_words'     : df_rows[i][5],
        'synonyms'          : df_rows[i][6],
        'antonyms'          : df_rows[i][7],
        'freq_of_usage'     : df_rows[i][8],
        'polarity'          : df_rows[i][9],
        'subjectivity'      : df_rows[i][10],
        'subjectivity_score': df_rows[i][11]
    }

    if data_dict not in data_list:
        data_list.append(data_dict)

    if not inventory.find_one({'vocab': data_dict['vocab']}):
        inventory.insert_one(data_dict)
    else:
        print('Document with Vocab ' + data_dict['vocab'] + ' Existed!')
```

```
import random

pprint.pprint(list(inventory.find({'seq_no': f'{random.randint(1, 220)}'})))
```

```
[{'_id': ObjectId('675a569b1ef52eda86b81bc1'),
  'antonyms': None,
  'definitions': '1. pergi bersama; 2. bersama-sama melakukan (ditimpa dll) '
                 'sesuatu, tidak terkecuali drpd (sesuatu kejadian, perbuatan, '
                 'dll), ikut serta; 3. patuh kpd (undang-undang, janji, dll), '
                 'taat kpd; 48;  berturut-turut, berturutan 1. tidak '
                 'ber-antara antara satu dgn lain, lepas satu-satu, '
                 'terus-menerus dgn teratur (tidak berselang), tidak '
                 'berhentinya, beruntun-runtun, bersam-bung-sambung; 5. '
                 'mengikut (berjalan dll), mengekori (dr belakang); 6. spt yg '
                 'dikhabarkan (diberitakan, dihebahkan, dll), sepanjang '
                 '(berita dll); 7. sesuai dgn, tidak bertentangan dgn, tidak '
                 'melanggar (syarat dll), selaras dgn; 8;  berturut-turut, '
                 'berturutan 1. tidak ber-antara antara satu dgn lain, lepas '
                 'satu-satu, terus-menerus dgn teratur (tidak berselang), '
                 'tidak berhentinya, beruntun-runtun, bersam-bung-sambung; 9. '
                 'ikut serta (melakukan sesuatu dll), turut serta; 10. '
                 'bergantung pd',
  'derived_words': None,
  'freq_of_usage': '6',
  'part_of_speech': None,
  'polarity': 'none',
  'sentences_eg': '1. saya turut ke mana dia pergi; 2. Ahmad yg turut terbunuh '
                  'oleh serangan hendap itu baru semalam ditemui mayatnya; 3. '
                  'sekiranya kamu turut akan nasihat ini berbahagialah kamu; '
                  '4. beliau telah kalah dlm tiga pilihan raya kecil kpd parti '
                  'lawannya secara turut; 5. dia berjalan menuju ruangan yg di '
                  'sebelah barat bangunan, aku turutnya dr belakang; 6. turut '
                  'yg diketahui sebanyak satu suku juta ringgit harga saham '
                  'telah terjual; 7. hari ini turut pendapat orang alim ialah '
                  'hari yg baik; 8. beliau telah kalah dlm tiga pilihan raya '
                  'kecil kpd parti lawannya secara turut; 9. Sultan Mansor '
                  'Shah pun turut menyebut nama Hang Tuah itu Laksamana; 10. '
                  'maju mundurnya dlm pelajarannya turut kerajinan-nya',
  'seq_no': '199',
  'subjectivity': 'fact-driven',
  'subjectivity_score': '0.18666666666666668',
  'synonyms': None,
  'vocab': 'turut',
  'word_length': '5'}]
```


Create Connection with Neo4j Database

```
from neo4j import GraphDatabase

URI = "neo4j+s://14454b05.databases.neo4j.io"

neo4j_user = "neo4j"
neo4j_password = "Jag4fDqQ54xZFEEmXhHbZ7bSVbG-TBwCRutCQFsX9P4"

AUTH = (neo4j_user, neo4j_password)
driver = GraphDatabase.driver(URI, auth=AUTH)
```

```
driver.verify_connectivity()
```

Define Constraint for Data Insertion

the data inserted must be unique in the field of `vocab`.

```
driver.execute_query("CREATE CONSTRAINT FOR (w:Word) REQUIRE w.vocab IS UNIQUE")
```

```
EagerResult(records=[], summary=<neo4j._work.summary.ResultSummary object at 0x7fe2a6df6fe0>, keys=[])
```

Store Data into Neo4J Graph Database in Nodes

```
for d in data_list:
    try:
        summary = driver.execute_query(
            """
            CREATE (:Word { seq_no             : $seq_no,
                            vocab              : $vocab,
                            word_length        : $word_length,
                            part_of_speech     : $part_of_speech,
                            definitions        : $definitions,
                            sentences_eg       : $sentences_eg,
                            derived_words      : $derived_words,
                            synonyms           : $synonyms,
                            antonyms           : $antonyms,
                            freq_of_usage      : $freq_of_usage,
                            polarity           : $polarity,
                            subjectivity       : $subjectivity,
                            subjectivity_score : $subjectivity_score
            })
            """,
            seq_no=d['seq_no'],
            vocab=d['vocab'],
            word_length=d['word_length'],
            part_of_speech=d['part_of_speech'],
            definitions=d['definitions'],
            sentences_eg=d['sentences_eg'],
            derived_words=d['derived_words'],
            synonyms=d['synonyms'],
            antonyms=d['antonyms'],
            freq_of_usage=d['freq_of_usage'],
            polarity=d['polarity'],
            subjectivity=d['subjectivity'],
            subjectivity_score=d['subjectivity_score'],
            database_="neo4j"
        ).summary
    except:
        print('Record with Vocab ' + d['vocab'] + ' Existed!')
```

Define Relationship between Nodes

Synonym Relationship

```
df_rows = df.collect()

vocabs = []
for row in df_rows:
    vocabs.append(row[0])

for k in range(len(df_rows)):
    if df_rows[k][6] != None:
        synonyms = df_rows[k][6].split(';')

        for i in range(len(synonyms)):
            for j in range(len(vocabs)):
                if synonyms[i] == vocabs[j] and i != j:
                    summary = driver.execute_query(
                        """
                        MATCH (p:Word {vocab: $vocab}), (s:Word {vocab: $synonym})
                        CREATE (p) -[:SYNONYM]->(s)
                        """,
                        vocab = df_rows[k][0],
                        synonym = vocabs[j],
                        database_="neo4j"
                    ).summary
                    break
```

Antonym Relationship

```
for k in range(len(df_rows)):
    if df_rows[k][7] != None:
        antonyms = df_rows[k][7].split(';')

        for i in range(len(antonyms)):
            for j in range(len(vocabs)):
                if antonyms[i] == vocabs[j] and i != j:
                    summary = driver.execute_query(
                        """
                        MATCH (p:Word {vocab: $vocab}), (a:Word {vocab: $antonym})
                        CREATE (p) -[:ANTONYM]->(a)
                        """,
                        vocab = df_rows[k][0],
                        antonym = vocabs[j],
                        database_="neo4j"
                    ).summary
                    break
```

Derived Words Relationship

```
for k in range(len(df_rows)):
    if df_rows[k][5] != None:
        derived_words = df_rows[k][5].split(';')

        for i in range(len(derived_words)):
            for j in range(len(vocabs)):
                if derived_words[i] == vocabs[j] and i != j:
                    summary = driver.execute_query(
                        """
                        MATCH (p:Word {vocab: $vocab}), (dw:Word {vocab: $derived_word})
                        CREATE (p) -[:DERIVED_WORDS]->(dw)
                        """,
                        vocab = df_rows[k][0],
                        derived_word = vocabs[j],
                        database_="neo4j"
                    ).summary
                    break
```

2.4.2 lexicon_retrieval.ipynb

Lexicon Retrieval

[3]:
```
from pymongo import MongoClient
import pprint

uri = "mongodb+srv://jordan:jordan0551@cluster0.nzsyn.mongodb.net/?retryWrites=true&w=majority&appName=Cluster0"
client = MongoClient(uri)

db = client["lexicondb"]

collections = db.list_collection_names()
```

[47]:
```
search_keyword = input('Search a word: ').lower()
```

```
Search a word:  badan
```

[48]:
```
result_dict = []

for collection in collections:
    try:
        current_collection = db[collection]

        result_dict = list(current_collection.find({'vocab': f'{search_keyword}'}))

        if result_dict != []:
            break
    except:
        pass
```

[49]:
```
if result_dict != []:
    vocab = result_dict[0]['vocab']
    word_length = result_dict[0]['word_length']
    part_of_speech = result_dict[0]['part_of_speech'] if result_dict[0]['part_of_speech'] != None else '-'
    definitions = result_dict[0]['definitions'].split(';') if result_dict[0]['definitions'] != None else '-'
    sentences_eg = result_dict[0]['sentences_eg'].split(';') if result_dict[0]['sentences_eg'] != None else '-'
    derived_words = result_dict[0]['derived_words'] if result_dict[0]['derived_words'] != None else '-'
    synonyms = result_dict[0]['synonyms'] if result_dict[0]['synonyms'] != None else '-'
    antonyms = result_dict[0]['antonyms'] if result_dict[0]['antonyms'] != None else '-'
    freq_of_usage = result_dict[0]['freq_of_usage']
    polarity = result_dict[0]['polarity']
    subjectivity = result_dict[0]['subjectivity']
    subjectivity_score = result_dict[0]['subjectivity_score']

    print('=' * 120)
    print(f'Searching Result of Word : {search_keyword.upper()}')
    print('-' * 120)
    print(f'[ {"VOCAB": >20} ] {vocab}')
    print(f'[ {"LENGTH OF WORD": >20} ] {word_length}')
    print(f'[ {"PART OF SPEECH": >20} ] {part_of_speech}')
    print(f'[ {"DEFINITIONS": >20} ] ~\n')
    for i in definitions:
        print(f'{"":>10} {i.lstrip()}')

    print(f'\n[ {"EXAMPLES SENTENCES": >20} ] ~\n')
    for i in sentences_eg:
        print(f'{"":>10} {i.lstrip()}')
    print(f'\n[ {"DERIVED WORDS": >20} ] {derived_words}')
    print(f'[ {"SYNONYMS": >20} ] {synonyms}')
    print(f'[ {"ANTONYMS": >20} ] {antonyms}')
    print(f'[ {"USAGE FREQUENCY": >20} ] {freq_of_usage}')
    print(f'{"":>10} ** USAGE FREQUENCY is captured from 5 Malay essays\n')
    print(f'[ {"POLARITY": >20} ] {polarity}')
    print(f'[ {"SUBJECTIVITY": >20} ] {subjectivity} ({subjectivity_score:.4f})')
    print('=' * 120)
else:
    print('=' * 120)
    print(f'Searching Result of Word : {search_keyword.upper()}')
    print('-' * 120)
    print('Sorry (╥﹏╥)... The searched word is not found')
    print('=' * 120)
```

```
=================================================================================================================================
Searching Result of Word : BADAN
---------------------------------------------------------------------------------------------------------------------------------
[                VOCAB ] badan
[       LENGTH OF WORD ] 5
[       PART OF SPEECH ] -
[          DEFINITIONS ] ~

           1 semua bahagian tubuh manusia atau binatang yg dpt dilihat
           2 bahagian tubuh manusia atau binatang, tidak termasuk kepala, kaki ataupun tangan
           3 yg empunya tubuh
           diri sendiri. 4 bahagian utama drpd sesuatu
           4 bahagian utama drpd sesuatu

[   EXAMPLES SENTENCES ] ~

           1. badan orang itu besar serta gempal. 2 bahagian tubuh manusia atau binatang, tidak termasuk kepala, kaki ataupun tangan: badannya lebih pa
njang drpd kakinya. 3 yg empunya tubuh
           2. badannya lebih panjang drpd kakinya. 3 yg empunya tubuh
           3. badan kereta
           4. badan kereta

[        DERIVED WORDS ] -
[             SYNONYMS ] -
[             ANTONYMS ] -
[      USAGE FREQUENCY ] 0
           ** USAGE FREQUENCY is captured from 5 Malay essays

[             POLARITY ] none
[         SUBJECTIVITY ] fact-driven (0.0661)
=================================================================================================================================
```

3. Lexicon Enrichment (if any)

3.1. List of enrichment carried out

3.1.1 Antonyms

For each of the processed records, we try to identify whether there is any antonym(s) for the word. Similar to what we had done in finding synonyms, we scrape data from DBP and find the word 'berantonim dengan' in the thesaurus section to determine the existence of antonyms in a word. If the antonyms of the word exist, we add the value associated to the vocabulary in the new column named 'antonyms'.

Fig. 3.1 *Example of Thesaurus with Antonyms Existed.*

3.1.2 Frequency of Usage

From 5 essays written in Malay language available on the Internet, we scrap the whole contents of each essay. For each of the vocabulary, we study the total number of appearances in these 5 essays.

3.1.3 Sentiment

Afterwards, we study the sentiment expressed by each vocabulary by identifying the polarity and subjectivity of each word from the vocabulary itself and the provided example sentences.

We use the library TextBlob to identify the polarity and subjectivity of a vocabulary. In this library, polarity and subjectivity are both floating values, ranging from -1 to 1 for polarity, and 0 to 1 for subjectivity. For polarity, the sign indicates the tendency of the word in expressing a thing in positive or

negative emotion. For example, the word ‘atas’ has positive polarity, indicating that when people use ‘atas’ in speaking, it seems to present a positive expression. Instead of recording polarity as a floating value, we choose to save the polarity of a word in three different categories, namely *positive*, *none* (neutral), and *negative*.

Meanwhile for subjectivity, we record both in category and in numeric form. Subjectivity refers to the tendency of a word in providing a means of being someone’s opinion or a fact. In TextBlob, the nearer the subjectivity value to 1, the more the tendency of the word being driven by opinion. Oppositely, the nearer the value to 0, the more the tendency that the sentence formed by the word is expressing a fact. Here, we grade subjectivity by studying the average subjectivity in each vocabulary words, and categories the value into four classes, namely *fact-driven*, *partially fact-driven*, *partially opinion-driven*, and *opinion-driven* on ascending subjectivity values.

Example of Enriched Data

	vocab	word_length	part_of_speech	definitions	sentences_eg	derived_words	synonyms
1	berniat	7	kata nama	(akan), berkaul, bernazar	niat di lautan Tanjung Jati	niat-niat;meniatkan;terniat	angkaan;cadangan;nazar
2	mundur	6	adjektif	belakang, tertinggal, kolot	mundur, dan terkongkong	emundurkan;kemunduran	;susut;surut;berkurangan
3	kaum	4	kata nama	talian kekeluargaan (dgn)	; 5. dia itu kaum dgn saya	berkaum;perkauman	;bangsa orang;puak;suku
4	wang	4	kata nama	g bernilai satu pertiga tali	ayam seekor satu rupiah	berwang;mengewangi	duit;fulus;pitis;dukat;tikal
5	jelas	5	adjektif	as (pekerjaan, hutang, dll)	i boleh jelas pekerjaan ini	kan;kejelasan;penjelasan	;pasti;jali;curai;izhar;jaleh
6	waktu	5	kata nama	keadaan hari dll, suasana	berpeluh pd waktu malam	sewaktu;sewaktu-waktu	a;tempoh;zaman;tenggak
7	ramai	5	adjektif	ng), orang banyak, umum	6. ramai orang berkumpul	ai;meramaikan;keramaian	guh-lagah;becang-becok
8	kuat	4	adjektif	= kekuatan tenaga, daya	-kira; 9. kuat tarikan bumi	kuat;kekuatan;penguatan	sa;mampu;sanggup;boleh
9	sahaja	6	kata tugas	(tidak ditambah-tambah)	ja dapat dilihat bintang itu	aja-sahaja;menyahajakan	hkan;semata-mata;belaka
10	kerja	5	kata nama	pat dijalankan (dilakukan)	au tiada kerja oleh paman	gerjaan;pekerja;pengerja	ai;usaha;urusan;kegiatan
11	rakyat	6	kata nama	t tersebar luas kpd rakyat	ga istana sbg elit ilmuwan	kerakyatan	um;orang;warga;isi negeri
12	penuh	5	adjektif	, 4. banyak (ramai) sekali	pasar itu dgn perempuan	memenuhi;memenuhkan	etat;sendat;lejuh;merepuh
13	merupakan	9	kata nama	ukiskan, menggambarkan	i mempunyai daya berfikir	pa;rupawan;kerupawanan	t;tampang muka;air muka
14	penjelasan	10	adjektif	n (yg lebih jelas), huraian	sebab perbuatan adiknya	kan;kejelasan;penjelasan	;pasti;jali;curai;izhar;jaleh
15	niat	4	kata nama	(akan), bermaksud (akan)	ndak pulang ke rumahmu	niat-niat;meniatkan;terniat	angkaan;cadangan;nazar
16	bermaksud	9	kata nama	dak); 2. mengandungi erti	idak bermaksud demikian	ermaksud;memaksudkan	murad;kasad;anju;hujung
17	tahu	4	kata kerja	anggap sudah semestinya	la kerja yg dibekalkannya	etahuan;berpengetahuan	hui;sedar;mengakui;enga
18	memaksudkan	11	kata nama	nenyatakan (sbg maksud)	n oleh perempuan tua itu	ermaksud;memaksudkan	murad;kasad;anju;hujung
19	masa	4	kata nama	a dgn, pd ketika yg sama	dgn sejarah yg ditulisnya	semasa	;tenggak;sangkala;sekon
20	jauh	4	adjektif	sangat, terlampau sangat	Pulau Pinang drpd di sini	uh;menjauhi;menjauhkan	dekat;besar jaraknya;dura
21	selesai	7	adjektif	elesai dapat diselesaikan	si dasar kebangsaan kita	keselesaian;penyelesaian	edia;kamal;kafi;akmal;jap

Fig. 3.2 *Example of Enriched Data* (*row 1 - 20*, *column 1 - 7*)

antonyms	freq_of_usage	polarity	subjectivity	subjectivity_score
	0	none	fact-driven	0.0
maju;maju	0	none	partially fact-driven	0.30083333333333334
	5	none	fact-driven	0.1
	6	none	partially opinion-driven	0.5777777777777777
kabur	0	positive	fact-driven	0.19166666666666665
	0	none	partially fact-driven	0.3712962962962963
;unyi;sugul;lembap;sedikit	1	positive	partially fact-driven	0.49166666666666664
emah;rapuh;lemah;ringan	0	positive	partially fact-driven	0.48975308641975307
jarang	13	none	partially fact-driven	0.45833333333333337
	0	none	partially fact-driven	0.27769360269360266
pemerintah;ketua	25	none	partially fact-driven	0.26
ong;sedikit;kurang;sedikit	4	positive	partially opinion-driven	0.55
	3	none	opinion-driven	0.75
kabur	0	none	fact-driven	0.0
	0	none	partially fact-driven	0.2857142857142857
	1	none	opinion-driven	0.84375
jahil;membiarkan	0	none	fact-driven	0.18571428571428572
	0	negative	fact-driven	0.175
	3	none	fact-driven	0.08333333333333333
dekat	2	positive	partially opinion-driven	0.6583333333333334

Fig. 3.3 *Example of Enriched Data* (*row 1 - 20*, *column 8 - 12*)

3.2. List of Python classes and Jupyter notebook(s) for lexicon enrichment:

Name of Python class(es) / notebooks	Author
lexicon_enrichment.ipynb	Ang Yu Wen
antonym_finder.py	Lau Kit Lim, Ang Yu Wen
sentiment.py	Ang Yu Wen

3.3. Code for Python Classes

3.3.1 lexicon_enrichment.ipynb

Lexicon Enrichment

[1]:
```
from pyspark.sql.functions import *
from pyspark.sql.types import *
```

[2]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
```

```
24/12/14 12:14:22 WARN Utils: Your hostname, JORDAN. resolves to a loopback address: 127.0.1.1; using 10.255.255.254 instead (on interface lo)
24/12/14 12:14:22 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/12/14 12:14:23 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

User-Defined Function

`def findAntonym()` : to find the antonyms list of a word.

[3]:
```
from python_script.antonym_finder import findAntonym
```

Reorder the Columns

[4]:
```
df = spark.read.csv('lexicon_data/processed_data.csv', header=True)
df = df.select('vocab', 'word_length', 'part_of_speech', 'definitions', 'sentences_eg', 'derived_words', 'synonyms')
df.show()
```

```
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|      vocab|word_length|part_of_speech|         definitions|        sentences_eg|       derived_words|            synonyms|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|     sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|bersahaja;bersaha...|satu-satunya;cuma...|
|      kerja|          5|     kata nama|1. usaha (kegiata...|1. beberapa mingg...|sekerja;bekerja;m...|khidmat;pekerjaan...|
|       kaum|          4|     kata nama|1. puak, suku ban...|1. banyak kaum or...|   berkaum;perkauman|suku;kabilah;taef...|
|      jelas|          5|      adjektif|1. terang (perkat...|1. suara itu tida...|berjelas-jelas;me...|terang;nyata;keta...|
|      waktu|          5|     kata nama|1. rangkaian atau...|1. secara rohania...|sewaktu;sewaktu-w...|masa;jangka masa;...|
|       kuat|          4|      adjektif|1. banyak tenagan...|1. walaupun ia su...|sekuat-kuatnya;be...|gagah;berdaya;ber...|
|memaksudkan|         11|     kata nama|1. menghendaki, m...|1. dialah yg dima...|bermaksud;memaksu...|tujuan;hajat;hasr...|
|       tahu|          4|    kata kerja|1. maklum akan ke...|1. entah kapal pe...|tahu-tahu;mengeta...|maklum;faham;meng...|
|  merupakan|          9|     kata nama|1. membentuk (mem...|1. peti rahsia it...|rupanya;rupa-rupa...|paras;muka;wajah;...|
|      ramai|          5|      adjektif|1. riuh-rendah, h...|1. orang menyabun...|seramai;beramai-r...|bising;riuh;hiruk...|
|      penuh|          5|      adjektif|1. seluruhnya ber...|1. tong itu penuh...|sepenuhnya;memenu...|berisi;tepu;padat...|
|     rakyat|          6|     kata nama|1. seluruh pendud...|1. rakyat Malaysi...|          kerakyatan|warganegara;pendu...|
|       wang|          4|     kata nama|1. alat pertukara...|1. satuan wang ne...|  berwang;mengewangi|duit;fulus;pitis;...|
|     mundur|          6|      adjektif|1. (berjalan, ber...|1. dia mundur set...|memundurkan;kemun...|merosot;meleset;m...|
| penjelasan|         10|      adjektif|1. keterangan (yg...|1. Ali meminta pe...|berjelas-jelas;me...|terang;nyata;keta...|
|       niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|berniat;berniat-n...|maksud;tujuan;ing...|
|  bermaksud|          9|     kata nama|1. bertujuan, ber...|1. kami bermaksud...|bermaksud;memaksu...|tujuan;hajat;hasr...|
|    berniat|          7|     kata nama|1. bertujuan (aka...|1. aku pun tidak ...|berniat;berniat-n...|maksud;tujuan;ing...|
|       jiwa|          4|     kata nama|1. nyawa, roh (yg...|1. buku itu mence...|    berjiwa;kejiwaan|nyawa;roh;semanga...|
| menyatakan|         10|      adjektif|1. menjadikan nya...|1. di dlm buku in...|menyatakan;ternya...|terang;jelas;keta...|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Insert Antonyms for each Rows (if any)

[5]:
```
append_antonym_udf = udf(findAntonym, StringType())

df = df.withColumn(
    'antonyms', append_antonym_udf(col('vocab'))
)

df.show()
```

```
[Stage 2:>                                                          (0 + 1) / 1]
```

```
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+--------------------+
|      vocab|word_length|part_of_speech|         definitions|        sentences_eg|       derived_words|            synonyms|            antonyms|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+--------------------+
|     sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|bersahaja;bersaha...|satu-satunya;cuma...|              jarang|
|      kerja|          5|     kata nama|1. usaha (kegiata...|1. beberapa mingg...|sekerja;bekerja;m...|khidmat;pekerjaan...|                    |
|       kaum|          4|     kata nama|1. puak, suku ban...|1. banyak kaum or...|   berkaum;perkauman|suku;kabilah;taef...|                    |
|      jelas|          5|      adjektif|1. terang (perkat...|1. suara itu tida...|berjelas-jelas;me...|terang;nyata;keta...|               kabur|
|      waktu|          5|     kata nama|1. rangkaian atau...|1. secara rohania...|sewaktu;sewaktu-w...|masa;jangka masa;...|                    |
|       kuat|          4|      adjektif|1. banyak tenagan...|1. walaupun ia su...|sekuat-kuatnya;be...|gagah;berdaya;ber...|lemah;rapuh;lemah...|
|memaksudkan|         11|     kata nama|1. menghendaki, m...|1. dialah yg dima...|bermaksud;memaksu...|tujuan;hajat;hasr...|                    |
|       tahu|          4|    kata kerja|1. maklum akan ke...|1. entah kapal pe...|tahu-tahu;mengeta...|maklum;faham;meng...|    jahil;membiarkan|
|  merupakan|          9|     kata nama|1. membentuk (mem...|1. peti rahsia it...|rupanya;rupa-rupa...|paras;muka;wajah;...|                    |
|      ramai|          5|      adjektif|1. riuh-rendah, h...|1. orang menyabun...|seramai;beramai-r...|bising;riuh;hiruk...|sunyi;sugul;lemba...|
|      penuh|          5|      adjektif|1. seluruhnya ber...|1. tong itu penuh...|sepenuhnya;memenu...|berisi;tepu;padat...|kosong;sedikit;ku...|
|     rakyat|          6|     kata nama|1. seluruh pendud...|1. rakyat Malaysi...|          kerakyatan|warganegara;pendu...|    pemerintah;ketua|
|       wang|          4|     kata nama|1. alat pertukara...|1. satuan wang ne...|  berwang;mengewangi|duit;fulus;pitis;...|                    |
|     mundur|          6|      adjektif|1. (berjalan, ber...|1. dia mundur set...|memundurkan;kemun...|merosot;meleset;m...|          maju;maju|
| penjelasan|         10|      adjektif|1. keterangan (yg...|1. Ali meminta pe...|berjelas-jelas;me...|terang;nyata;keta...|               kabur|
|       niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|berniat;berniat-n...|maksud;tujuan;ing...|                    |
|  bermaksud|          9|     kata nama|1. bertujuan, ber...|1. kami bermaksud...|bermaksud;memaksu...|tujuan;hajat;hasr...|                    |
|    berniat|          7|     kata nama|1. bertujuan (aka...|1. aku pun tidak ...|berniat;berniat-n...|maksud;tujuan;ing...|                    |
|       jiwa|          4|     kata nama|1. nyawa, roh (yg...|1. buku itu mence...|    berjiwa;kejiwaan|nyawa;roh;semanga...|        jasad;zahir|
| menyatakan|         10|      adjektif|1. menjadikan nya...|1. di dlm buku in...|menyatakan;ternya...|terang;jelas;keta...|               kabur|
+----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Calculate the Frequency of Usage from FIVE (5) Malay Essay

[6]:
```
import requests
from bs4 import BeautifulSoup as bs
import string

urls = [
    "https://infopelajar.com.my/karangan-buli-siber-kesan-kepada-emosi/",
    "https://infopelajar.com.my/karangan-jenayah-siber/",
    "https://infopelajar.com.my/karangan-rumah-terbuka-di-malaysia/",
    "https://infopelajar.com.my/karangan-rasuah-ancaman-kepada-negara/",
    "https://infopelajar.com.my/contoh-karangan-pidato/"
]

url_contents = []
for url in urls:
    r = requests.get(url)
    soup = bs(r.content, 'html.parser')

    entry_content = soup.find('div', class_='entry-content')
    entry_content.find('div', class_='cta-box').decompose()
    essay_str = entry_content.get_text().replace('\n', ' ')
    essay_str = essay_str.translate(str.maketrans('', '', string.punctuation))

    url_contents.append(essay_str)

full_content_str = ' '.join(url_contents)
words_in_content = full_content_str.lower().split(' ')

while '' in words_in_content:
    words_in_content.remove('')
```

[7]:
```
df_rows = df.collect()

freq_of_usage = []

for row in df_rows:
    count = 0
    vocab = row[0]

    for item in words_in_content:
        if item == vocab:
            count = count + 1

    freq_of_usage.append([vocab, count])
```

[8]:
```
freq_df = spark.createDataFrame(freq_of_usage, ["vocab", "freq_of_usage"])

df = df.join(freq_df, on='vocab')
df.show()
```

```
+----------+----------+--------------+--------------------+--------------------+--------------------+--------------------+--------------------+------
-------+
|      vocab|word_length|part_of_speech|         definitions|        sentences_eg|       derived_words|            synonyms|            antonyms|freq_o
f_usage|
+----------+----------+--------------+--------------------+--------------------+--------------------+--------------------+--------------------+------
-------+
|     sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|bersahaja;bersaha...|satu-satunya;cuma...|              jarang|
13|
|      kerja|          5|     kata nama|1. usaha (kegiata...|1. beberapa mingg...|sekerja;bekerja;m...|khidmat;pekerjaan...|                    |
0|
|       kaum|          4|     kata nama|1. puak, suku ban...|1. banyak kaum or...|   berkaum;perkauman|suku;kabilah;taef...|                    |
5|
|      jelas|          5|      adjektif|1. terang (perkat...|1. suara itu tida...|berjelas-jelas;me...|terang;nyata;keta...|               kabur|
0|
|      waktu|          5|     kata nama|1. rangkaian atau...|1. secara rohania...|sewaktu;sewaktu-w...|masa;jangka masa;...|                    |
0|
|       kuat|          4|      adjektif|1. banyak tenagan...|1. walaupun ia su...|sekuat-kuatnya;be...|gagah;berdaya;ber...|lemah;rapuh;lemah...|
0|
|memaksudkan|         11|     kata nama|1. menghendaki, m...|1. dialah yg dima...|bermaksud;memaksu...|tujuan;hajat;hasr...|                    |
0|
|       tahu|          4|    kata kerja|1. maklum akan ke...|1. entah kapal pe...|tahu-tahu;mengeta...|maklum;faham;meng...|    jahil;membiarkan|
0|
|  merupakan|          9|     kata nama|1. membentuk (mem...|1. peti rahsia it...|rupanya;rupa-rupa...|paras;muka;wajah;...|                    |
3|
|      ramai|          5|      adjektif|1. riuh-rendah, h...|1. orang menyabun...|seramai;beramai-r...|bising;riuh;hiruk...|sunyi;sugul;lemba...|
1|
|      penuh|          5|      adjektif|1. seluruhnya ber...|1. tong itu penuh...|sepenuhnya;memenu...|berisi;tepu;padat...|kosong;sedikit;ku...|
```

```
4|
|     rakyat|          6|     kata nama|1. seluruh pendud...|1. rakyat Malaysi...|          kerakyatan|warganegara;pendu...|    pemerintah;ketua|
25|
|       wang|          4|     kata nama|1. alat pertukara...|1. satuan wang ne...|  berwang;mengewangi|duit;fulus;pitis;...|                    |
6|
|     mundur|          6|      adjektif|1. (berjalan, ber...|1. dia mundur set...|memundurkan;kemun...|merosot;meleset;m...|           maju;maju|
0|
| penjelasan|         10|      adjektif|1. keterangan (yg...|1. Ali meminta pe...|berjelas-jelas;me...|terang;nyata;keta...|               kabur|
0|
|       niat|          4|     kata nama|1. maksud atau tu...|1. mematikan oran...|berniat;berniat-n...|maksud;tujuan;ing...|                    |
0|
|  bermaksud|          9|     kata nama|1. bertujuan, ber...|1. kami bermaksud...|bermaksud;memaksu...|tujuan;hajat;hasr...|                    |
1|
|    berniat|          7|     kata nama|1. bertujuan (aka...|1. aku pun tidak ...|berniat;berniat-n...|maksud;tujuan;ing...|                    |
0|
|       jiwa|          4|     kata nama|1. nyawa, roh (yg...|1. buku itu mence...|    berjiwa;kejiwaan|nyawa;roh;semanga...|         jasad;zahir|
1|
| menyatakan|         10|      adjektif|1. menjadikan nya...|1. di dlm buku in...|menyatakan;ternya...|terang;jelas;keta...|               kabur|
0|
+-----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+--------------------+------
-------+
only showing top 20 rows
```

Assign Sentiment to each Word

[9]:
```
from python_script.sentiment import *
```

[10]:
```
import re

df_rows = df.collect()

sentiments = []

for rows in df_rows:
    vocab = rows[0]
    sentences = rows[4]

    trans_vocab = translateFromMyToEn(vocab)

    polarity = getPolarityOf(trans_vocab)
    if sentences != None:
        trans_sentences = []
        clean_sentences = re.sub('\d.', '', sentences)
        splitted_sentences = clean_sentences.split(';')

        for sente in splitted_sentences:
            trans_sentences.append(translateFromMyToEn(sente))

        total_subjectivity = 0
        sentences_count = 0
        for sente in trans_sentences:
            total_subjectivity = total_subjectivity + getSubjectivityValue(sente)
            sentences_count = sentences_count + 1

        subjectivity_score = total_subjectivity / sentences_count
        subjectivity = getSubjectivityFrom(subjectivity_score)
    else:
        subjectivity_score = getSubjectivityValue(trans_vocab)
        subjectivity = getSubjectivityOf(trans_vocab)

    sentiments.append([vocab, polarity, subjectivity, subjectivity_score])
```

[11]:
```
sentiment_df = spark.createDataFrame(sentiments, ["vocab", "polarity", "subjectivity", "subjectivity_score"])

df = df.join(sentiment_df, on='vocab')
df = df.dropDuplicates()
df.show()
```

```
+----------+----------+-------------+--------------------+--------------------+--------------------+--------------------+--------------------+------
-------+--------+--------------------+------------------+
|      vocab|word_length|part_of_speech|         definitions|        sentences_eg|       derived_words|            synonyms|            antonyms|freq_o
f_usage|polarity|        subjectivity| subjectivity_score|
+----------+----------+-------------+--------------------+--------------------+--------------------+--------------------+--------------------+------
-------+--------+--------------------+------------------+
|    berniat|         7|     kata nama|1. bertujuan (aka...|1. aku pun tidak ...|berniat;berniat-n...|maksud;tujuan;ing...|                    |
0|    none|         fact-driven|                 0.0|
|     mundur|         6|      adjektif|1. (berjalan, ber...|1. dia mundur set...|memundurkan;kemun...|merosot;meleset;m...|           maju;maju|
0|    none|partially fact-dr...|0.30083333333333334|
|       kaum|         4|     kata nama|1. puak, suku ban...|1. banyak kaum or...|   berkaum;perkauman|suku;kabilah;taef...|                    |
5|    none|         fact-driven|                 0.1|
|       wang|         4|     kata nama|1. alat pertukara...|1. satuan wang ne...|  berwang;mengewangi|duit;fulus;pitis;...|                    |
6|    none|partially opinion...| 0.5777777777777777|
|      jelas|         5|      adjektif|1. terang (perkat...|1. suara itu tida...|berjelas-jelas;me...|terang;nyata;keta...|              kabur|
0|positive|         fact-driven|0.19166666666666665|
|      waktu|         5|     kata nama|1. rangkaian atau...|1. secara rohania...|sewaktu;sewaktu-w...|masa;jangka masa;...|                    |
0|    none|partially fact-dr...| 0.3712962962962963|
|      ramai|         5|      adjektif|1. riuh-rendah, h...|1. orang menyabun...|seramai;beramai-r...|bising;riuh;hiruk...|sunyi;sugul;lemba...|
1|positive|partially fact-dr...|0.49166666666666664|
|       kuat|         4|      adjektif|1. banyak tenagan...|1. walaupun ia su...|sekuat-kuatnya;be...|gagah;berdaya;ber...|lemah;rapuh;lemah...|
0|positive|partially fact-dr...|0.48975308641975307|
|     sahaja|         6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|bersahaja;bersaha...|satu-satunya;cuma...|              jarang|
13|    none|partially fact-dr...|0.45833333333333337|
|      kerja|         5|     kata nama|1. usaha (kegiata...|1. beberapa mingg...|sekerja;bekerja;m...|khidmat;pekerjaan...|                    |
0|    none|partially fact-dr...|0.27769360269360266|
|     rakyat|         6|     kata nama|1. seluruh pendud...|1. rakyat Malaysi...|          kerakyatan|warganegara;pendu...|    pemerintah:ketua|
25|    none|partially fact-dr...|                0.26|
|      penuh|         5|      adjektif|1. seluruhnya ber...|1. tong itu penuh...|sepenuhnya;memenu...|berisi;tepu;padat...|kosong;sedikit;ku...|
4|positive|partially opinion...|                0.55|
|  merupakan|         9|     kata nama|1. membentuk (mem...|1. peti rahsia it...|rupanya;rupa-rupa...|paras;muka;wajah;...|                    |
3|    none|      opinion-driven|                0.75|
| penjelasan|        10|      adjektif|1. keterangan (yg...|1. Ali meminta pe...|berjelas-jelas;me...|terang;nyata;keta...|              kabur|
0|    none|         fact-driven|                 0.0|
|       niat|         4|     kata nama|1. maksud atau tu...|1. mematikan oran...|berniat;berniat-n...|maksud;tujuan;ing...|                    |
0|    none|partially fact-dr...| 0.2857142857142857|
|  bermaksud|         9|     kata nama|1. bertujuan, ber...|1. kami bermaksud...|bermaksud;memaksu...|tujuan;hajat;hasr...|                    |
1|    none|      opinion-driven|             0.84375|
|       tahu|         4|    kata kerja|1. maklum akan ke...|1. entah kapal pe...|tahu-tahu;mengeta...|maklum;faham;meng...|    jahil;membiarkan|
0|    none|         fact-driven|0.18571428571428572|
|memaksudkan|        11|     kata nama|1. menghendaki, m...|1. dialah yg dima...|bermaksud;memaksu...|tujuan;hajat;hasr...|                    |
0|negative|         fact-driven|               0.175|
|       masa|         4|     kata nama|1. waktu, ketika;...|1. masa dan tempa...|             semasa|waktu;ketika;deti...|                    |
3|    none|         fact-driven|0.08333333333333333|
|       jauh|         4|      adjektif|1. besar jaraknya...|1. rumahnya agak ...|berjauh;menjauhi;...|tidak dekat;besar...|              dekat|
2|positive|partially opinion...| 0.6583333333333334|
+----------+----------+-------------+--------------------+--------------------+--------------------+--------------------+--------------------+------
-------+--------+--------------------+------------------+
only showing top 20 rows
```

[12]:
```
df = df.filter(df['vocab'] != 'kepentingan')
df.count()
```

[12]:
```
219
```

Write into CSV File

[13]:
```
df.coalesce(1).write.csv('enriched_lexicon', header=True, mode='overwrite')
```

[14]:
```
from python_script.hdfs_commander import *

run_hdfs_command('hdfs dfs -mv enriched_lexicon/part-00000-*.csv lexicon_data/enriched_lexicon.csv')
run_hdfs_command('hdfs dfs -rm -R enriched_lexicon')
run_hdfs_command('hdfs dfs -get lexicon_data/enriched_lexicon.csv lexicon_data/enriched_lexicon.csv')
```

```
Output: Deleted enriched_lexicon
```

Terminate Spark Session

[15]:
```
spark.stop()
```

3.3.2 antonym_finder.py

```
import re
import requests
import string
from bs4 import BeautifulSoup as bs

def findAntonym(word: str):
    r = requests.get(f"https://prpm.dbp.gov.my/cari1?keyword={word}")
    soup = bs(r.content, 'html.parser')

    try:
        thesaurus_soup = soup.find('span',
              id='MainContent_SearchInfoTesaurus_lblTesaurus')

        if thesaurus_soup:
            thesaurus_str = thesaurus_soup.get_text()
            thesaurus_str = re.sub('\d', ' ', thesaurus_str)

            try:
                derived_word_index = thesaurus_str.index('Kata Terbitan')
                thesaurus_str = thesaurus_str[:derived_word_index]
            except:
                pass

            antonym_list = []
            while thesaurus_str.find('Berantonim dengan') != -1:
                match = re.search(r'Berantonim dengan (\w+)', thesaurus_str)
                if match:
                    antonym_index = thesaurus_str.index('Berantonim dengan')
                    antonym = match.group(1)

                    thesaurus_str = thesaurus_str[antonym_index+17:]
                    antonym_list.append(antonym)

        return ';'.join(antonym_list)
    except:
        return ''
```

3.3.3 sentiment.py

```
from textblob import TextBlob
from deep_translator import GoogleTranslator

def translateFromMyToEn(s):
    translated_s = GoogleTranslator(source='ms', target='en').translate(s)
    return translated_s

def getPolarityValue(s):
    return TextBlob(s).sentiment.polarity

def getPolarityOf(s):
    pol = getPolarityValue(s)

    if pol < 0:
        return 'negative'
    elif pol == 0:
        return 'none'
    else:
        return 'positive'

def getSubjectivityValue(s):
    return TextBlob(s).sentiment.subjectivity

def getSubjectivityFrom(val):
    if val >= 0 and val < 0.25:
        return 'fact-driven'
    elif val >= 0.25 and val < 0.5:
        return 'partially fact-driven'
    elif val >= 0.5 and val < 0.75:
        return 'partially opinion-driven'
    else:
        return 'opinion-driven'

def getSubjectivityOf(s):
    sub = getSubjectivityValue(s)

    return getSubjectivityFrom(sub)
```

4. Lexicon Analysis and Evaluation

4.1. Results of Analysis (if any)

All the results of descriptive analysis in graphical representation please look for content in section **4.5 Code for Python Classes**. The following is the summary of descriptive analysis that we had done for the lexicon.

4.1.1 Distribution of Usage Frequency

As a definition of usage frequency, the phrase ‘rarely used’ is defined as words that are used less than 15 times. The phrase 'regularly used’ is defined as words that are used between 16 to 30 times. The phrase ‘frequently used’ is defined as words that are used more than 30 times.

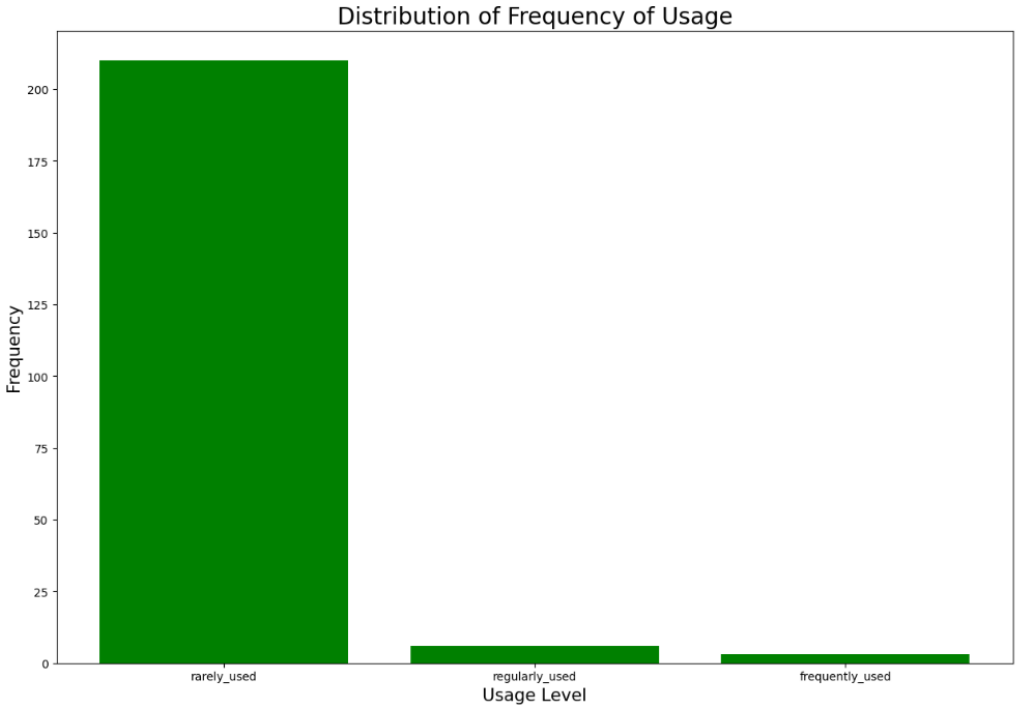


Fig. 4.1 *Distribution of Frequency of Usage in Bar Chart*

With the 219 words in the lexicon, with the aid of a bar chart, we can conclude that most of the words in the lexicon are rarely used with a word count of 210 words. We can also conclude that only a few words are regularly and frequently used. Both word counts sit at 6 regularly used words and 3 frequently used words.

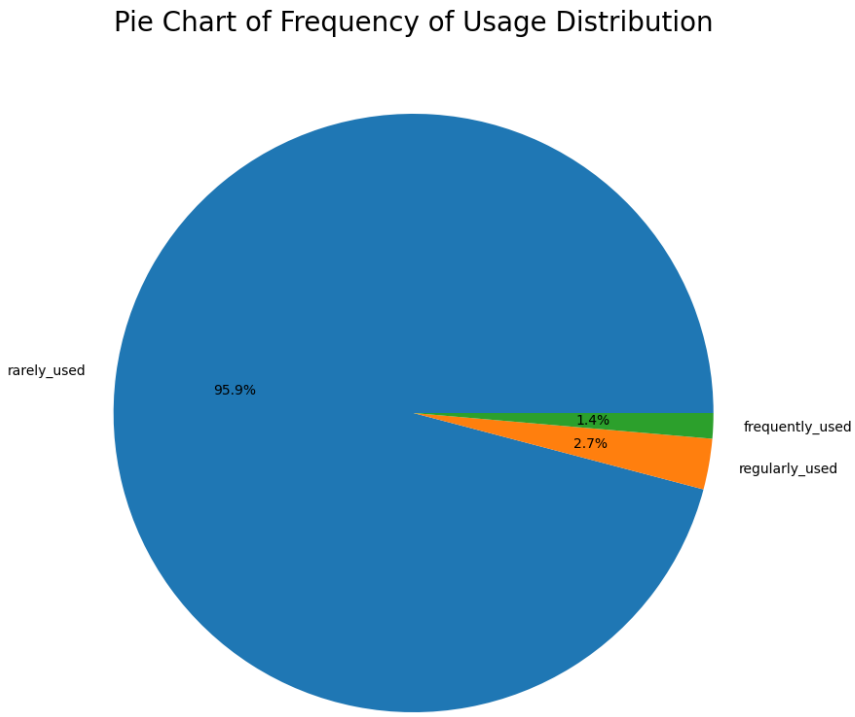


Fig. 4.2 *Pie Chart that Represents the Distribution of Frequency of Usage*

With the aid of a pie chart we can observe that 95.9% of words in the lexicon are rarely used. Followed up with 2.7% of words that are used regularly and 1.4% of words that are used frequently.

4.1.2 Distribution of Length of Words

Length of the words in the lexicon usually varies (from very short to very long). To study with the distribution of length of words in the words data captured, a histogram is plotted. The histogram is setted with minimum value of 3 and maximum value of 14, with every class (bin) of size 1 unit. The following is the result of the visualization.

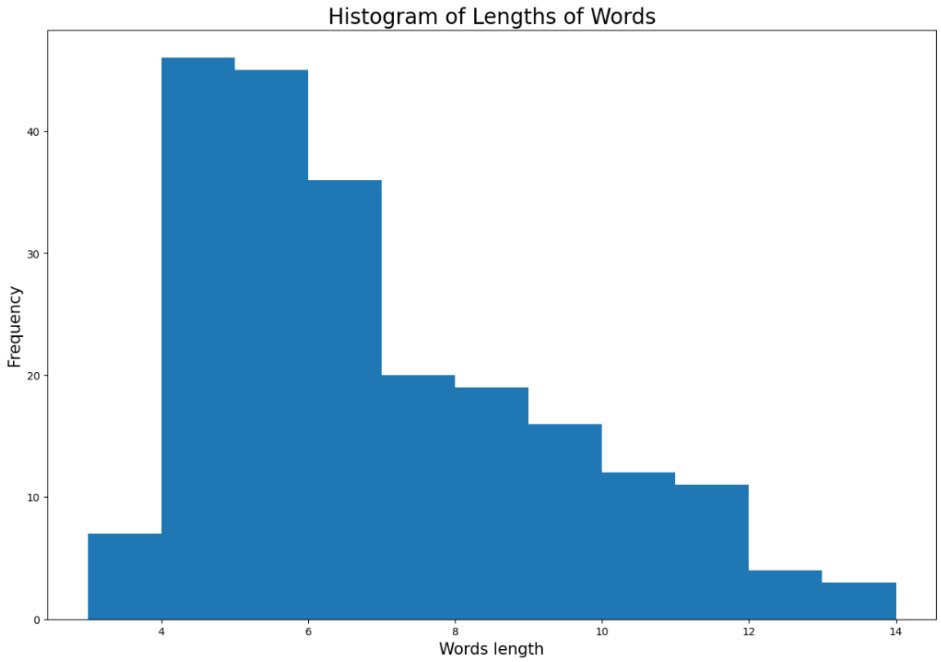


Fig. 4.3 *Histogram of Words Length Distribution*

From the graph plotted above, it is easily observed that the distribution of length of words is highly skewed to the right. To be more precise, most of the Malay vocabularies have their length concentrated around 4 to 6. This can be easily observed from the graph as the bar at the word lengths from 4 to 6 is the highest, and even producing a great difference with the neighbouring bars.

Moreover, the center of the distribution may have regression between the length of words and frequency of the appearance, as the decrement from one to the next is considered quite uniform. In other words, if given a set of data, we can have the opportunity to estimate how frequent the specific length of words exists in the data set, making the data predictable.

4.1.3 Prefixes and Suffixes Analysis

Prefixes are added to words to change the meaning of the root word. Some examples of prefixes in the Malay language are as follows, (‘me’, ‘meng’, ‘memper’, ‘be’, ‘ber’, ‘ke’, ‘pe’, ‘per’, ‘peng’, ‘pen’, ‘se’, ‘di’). Suffixes on the other hand are added to words to make the word more grammatically sensible in a sentence. Some examples of suffixes in the Malay language are as follows, (‘an’, ‘kan’, ‘i’, ‘nya’, ‘pun’).

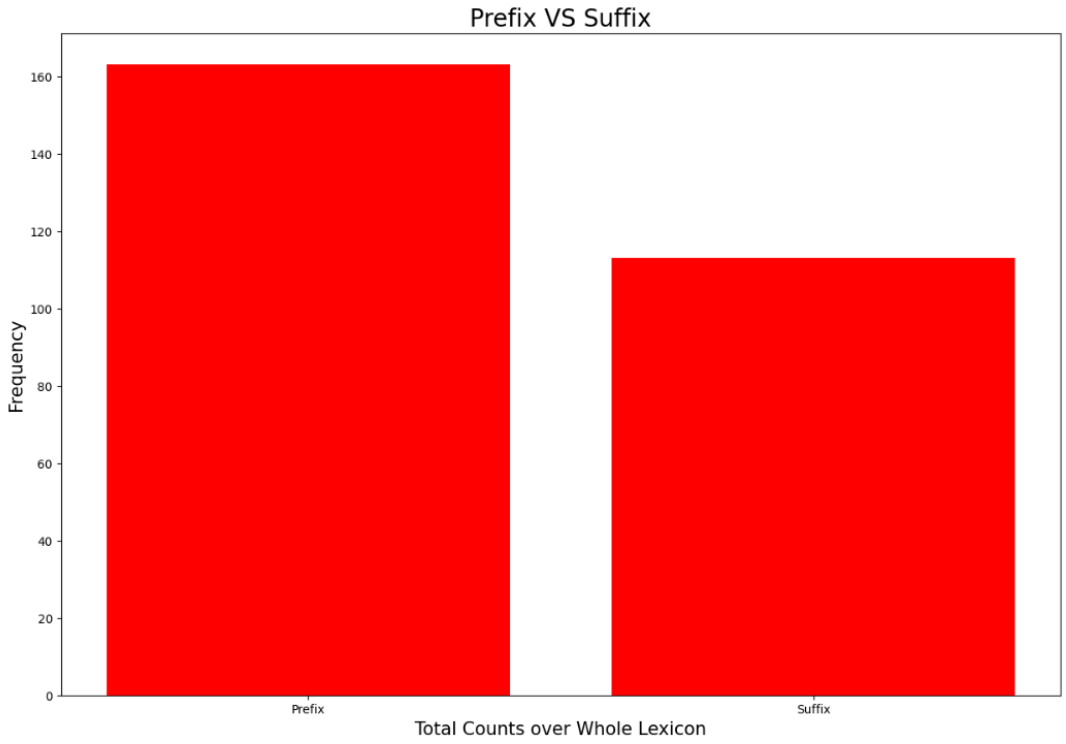


Fig. 4.4 *Distribution of Number of Prefix and Suffix Used in Bar Chart*

With the aid of a bar chart, we can observe that the number of prefixes is higher than the number of suffixes, and we can conclude that the usage of prefixes is more frequent than the usage of suffixes.

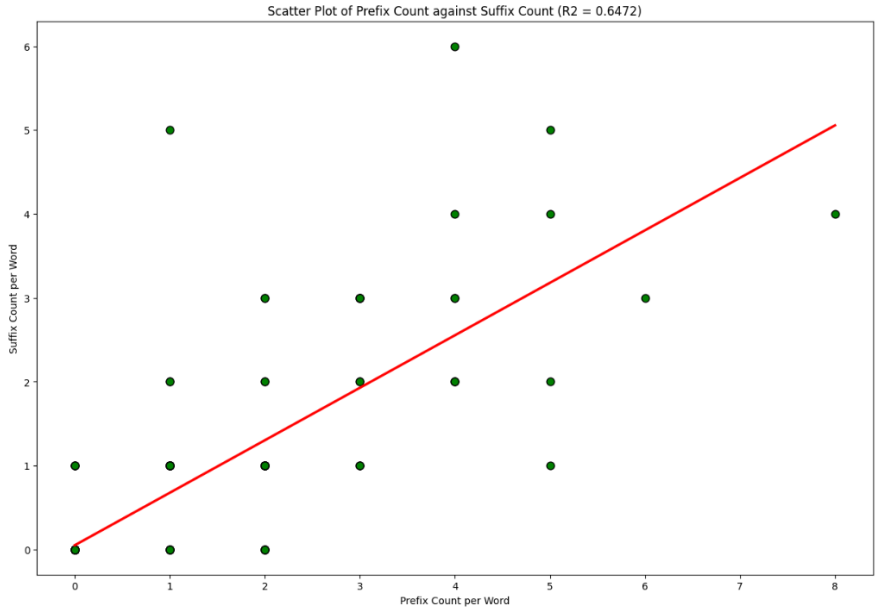


Fig. 4.5 *Scatter Plot of Prefix Count Against Suffix Count*

With the aid of a scatter plot and a linear regression, through the implemented formula we calculated that the coefficient of determination (R2) of the prefix count against suffix count is 0.6479, which implies that the association of the

appearance of a prefix and a suffix to both appear in the same word is relatively strong.

4.1.4 Synonyms Relationship Network

After that, we want to study the synonyms relationship within the words in one data set. To be more specific, we are interested in the existence of synonyms of a word as another row of record in the same data set. Thus, a network that uses vocabulary words as nodes is constructed as shown in the diagram.

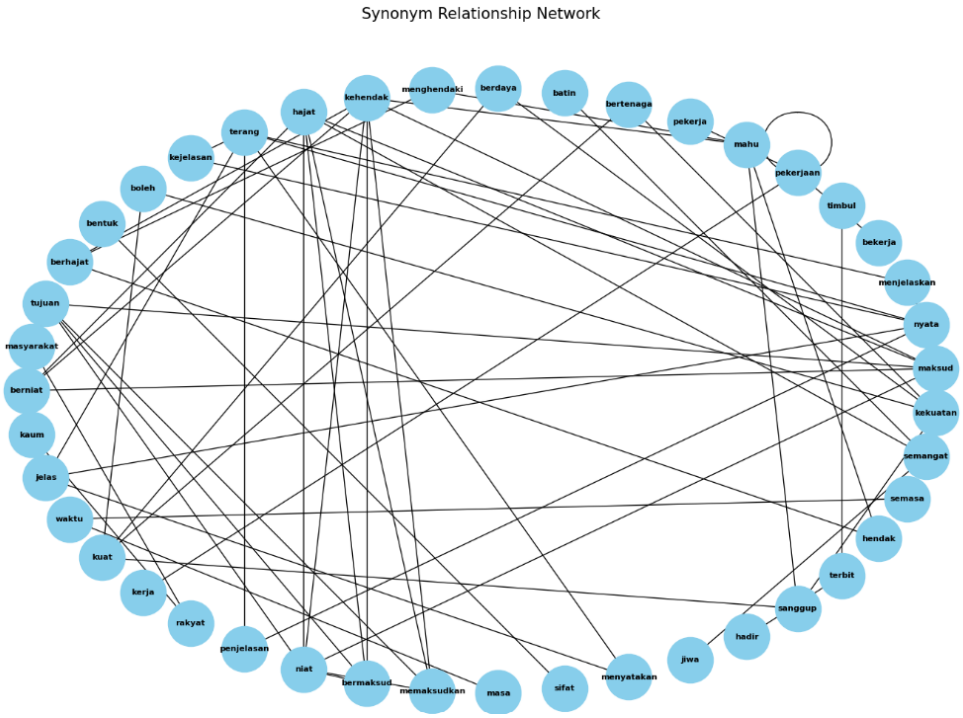


Fig. 4.6 *Synonyms Relationship Network*

In this network diagram, the line connected between nodes implies that there is a synonym relationship between them. For example, the word ‘hajat’ is connected to ‘niat’, ‘bermaksud’, ‘bermaksudkan’, etc. This means that ‘hajat’ is a synonym of ‘niat’. With this network diagram, we can easily conclude the synonym relationship between different words, as well as identifying whether a word has multiple synonyms.

4.1.5 Polarity Analysis

Polarity is categorized into threes, *positive*, *none* (neutral), and *negative*. Polarity of a word represents the tendency of the word in providing a definition, either positively, negatively, or neutral. Since it is categorical, we plot a bar chart to visualize the distribution of words’ polarity.

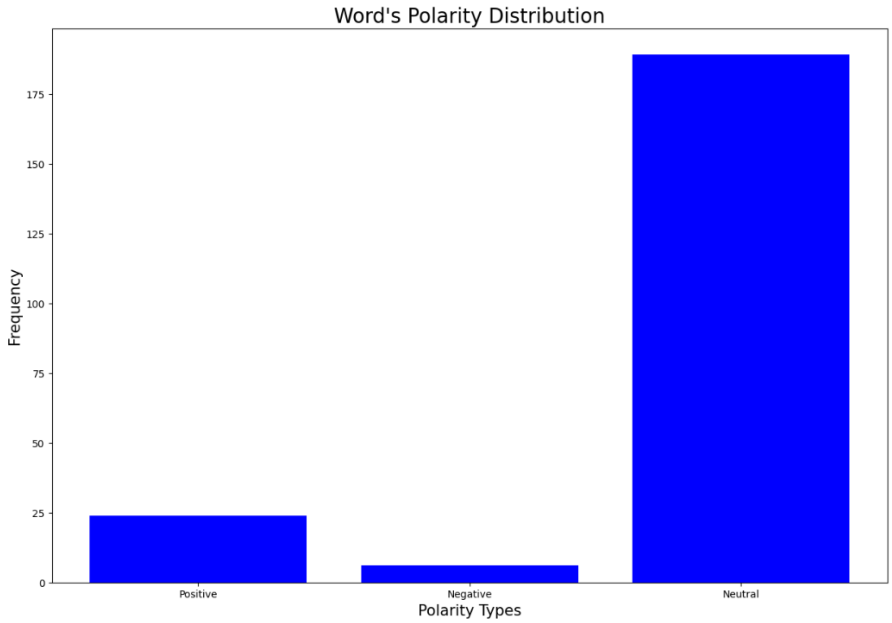


Fig. 4.7 *Bar Chart of Polarity Distribution*

From the bar chart, it is obvious that most of the words have the tendency neither in positive or negative side. In other words, words are mostly neutral in providing their meaning on emotional values. We believe that this may be caused by the part of speech of the words since adjectives may be the words that have the most obvious emotional values, making their polarity differentiable.

If we ignore the bar representing neutral, we still can see the difference between *positive* and *negative* polarities distribution. It is basically easy to observe and conclude that the frequency of positive vocabulary words is higher than the negative ones.

4.1.6 Distribution of Part of Speech

In our data set, we define the value of part of speech with 6 values, namely *kata nama* (nouns), *kata kerja* (verbs), *kata tugas* (function words), *adverba* (adverbs), *adjektif* (adjectives), and null. Without considering the entries with null part of speech, the following distribution of part of speech is plotted based on 219 available data.

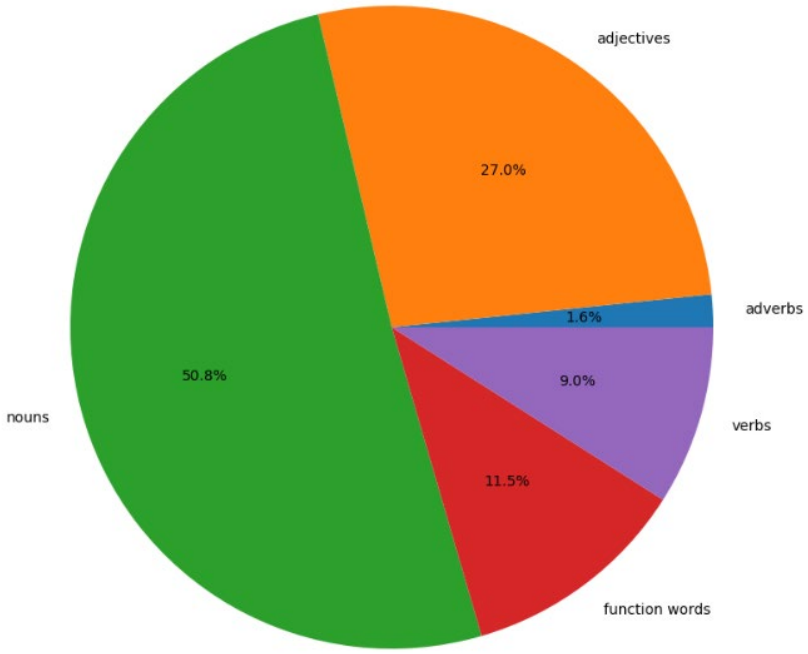


Fig. 4.8 *Pie Chart of Part of Speech Distribution*

From the pie chart, there are a few insights worth further elaboration. First, we can see that the part of speech 'nouns' is the most opaque in the distribution. The percentage is up to 50.8%, implying that at least half of the words in the available data are categorized as nouns. This is followed by adjectives, function words, verbs, and lastly adverbs.

Furthermore, since the category of adverbs is extremely minimal, we can somehow ignore the contribution of this part of speech in building the lexicon. With 219 data, there is only 1.6% words considered as adverbs, which equivalently means that only about 3 to 4 of the words are adverbs. To be honest, in the further study where data is scraped from Google Books whose data amount is about one fourth the current studied dataset, there does not even exist 1 adverb in the dataset.

Lastly, the contribution of function words is about the same as that of verbs. However, function words can be treated as a collection of multiple parts of speech. To be more precise, articles (a, an, the), prepositions, pronouns, auxiliaries, and conjunctions are all considered as categories of function words. Hence, averagely the contribution of one category of function words is only about 2.3% (= 11.5% total contribution ÷ 5 number of categories), which is comparatively smaller than that of verbs but gets closer to the contribution of adverbs.

4.1.7 Correlation between Word Lengths and Frequency of Usage

Last but not least, the correlation between word lengths and frequency of its usage is one of the interesting points that we could study about. Given the preface that people usually use short instead of long words to express one identical meaning, we are concerned about whether shorter words have a higher frequency of being used in daily life speech or writing. For example, people prefer using 'special' instead of 'extraordinary' to express the same meaning: something is unique. The same concept applied in Malay, as well as other languages. This is mentioned by Zipf's law, an empirical law, which states that the word frequency is inversely proportional to the word rank. (Wikipedia contributors, 2024e)

$$\text{word frequency} \propto \frac{1}{\text{word rank}}.$$

To proceed with this analysis, we plot the scatter plot of word lengths against the frequency of respective usage, and we fit the data into a regression line. The following is the result of graphing.

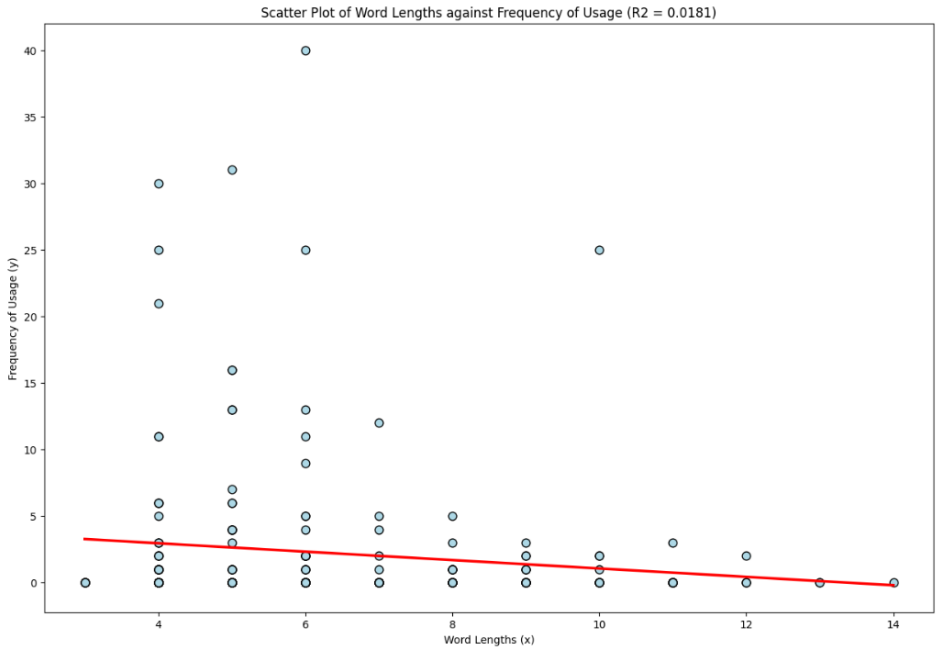


Fig. 4.9 *Scatter Plot and Regression Line of Word Lengths against Frequency of Usage.*

From the plot, although we can see there is a tendency where most of the data points are concentrated in the word lengths from 3 to 8 (especially 4 to 7), the regression line denies the conclusion. Statistically, if we consider this to be a dot plot, then it may show a strong relationship between these two variables. However, the regression line had shown that the relationship between them is extremely weak, provided the coefficient of determination of 0.0181. In

other words, only 1.81% of the variation in frequency of usage can be explained by the word lengths.

As a short conclusion, this study has proven that there may exist an association between word lengths and frequency of usage theoretically, but mathematically not.

4.2. Results of Intrinsic Evaluation (if any)

4.2.1 Lexicon Coverage

Lexicon coverage uses the base lexicon (scraped from DBP) as the comparison to calculate the coverage of a specific version of data scraped from Google Books. It calculates the percentage of existence of vocabularies in data scraped from Google Books as compared to the larger data set. For example, one Google Books data named gb_data_0001.csv has a lexicon coverage of about 4.11%. This means that this particular dataset has 4.11% of the vocabulary words found in the DBP dataset.

```
from python_script.metrics_calc import *

coverage_mark = coverage(base_words, gb_words)
print(coverage_mark)
print('or in percentage,', f'{coverage_mark*100:.2f}%')
```

```
0.0410958904109589
or in percentage, 4.11%
```

Fig. 4.10 *Google Books Data Coverage Measure*

4.2.2 Lexicon Accuracy

Lexicon accuracy measures how accurate a scraped dataset can be, as compared to the DBP dataset. By doing so, we implement two similarity measures, *exact similarity* and *Levenshtein Distance similarity*.

Exact similarity calculates the percentage where the data are exactly matched as the data in DBP dataset. However, this usually brings us unwanted results as it is hard to achieve exact similarity within two scraped strings.

To resolve such, we introduce the Levenshtein distance to calculate the similarity between two strings. Formally, Levenshtein distance is a string metric that is used to measure the difference between two sequences. The formal definition of Levenshtein in mathematics is given as follows. (Wikipedia contributors, 2024a)

$$\operatorname{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \operatorname{lev}\left(\operatorname{tail}(a), \operatorname{tail}(b)\right) & \text{if } \operatorname{head}(a) = \operatorname{head}(b), \\ 1 + \min \begin{cases} \operatorname{lev}\left(\operatorname{tail}(a), b\right) \\ \operatorname{lev}\left(a, \operatorname{tail}(b)\right) \\ \operatorname{lev}\left(\operatorname{tail}(a), \operatorname{tail}(b)\right) \end{cases} & \text{otherwise} \end{cases}$$

Surprisingly, the similarity measures are improved from the result of exact similarity calculation. For instance, the exact similarity measures of data file gb_data_0001.csv is 0.00%; while by using the Levenshtein distance method, the similarity is enhanced to about 9.32%.

Fig. 4.11 *Exact Annotation Accuracy VS Levenshtein Distance Similarity*

4.3. Results of Extrinsic Evaluation (if any)

For the part of extrinsic evaluation, we study the relationship between polarity and subjectivity of a word. We are interested in whether there is any influence to subjectivity due to the polarity of the word. In other words, we are studying the question where the emotional value of a word identifies or implies the tendency of it being driven by fact or opinion.

To do so, we apply regression analysis on polarity and subjectivity scores as they are both numerical, as well as recorded in floating numbers. The regression method that we apply is **polynomial regression**. We fit the data into polynomial functions of degree 1, 2, 3, and 4, and the result is shown as follows.

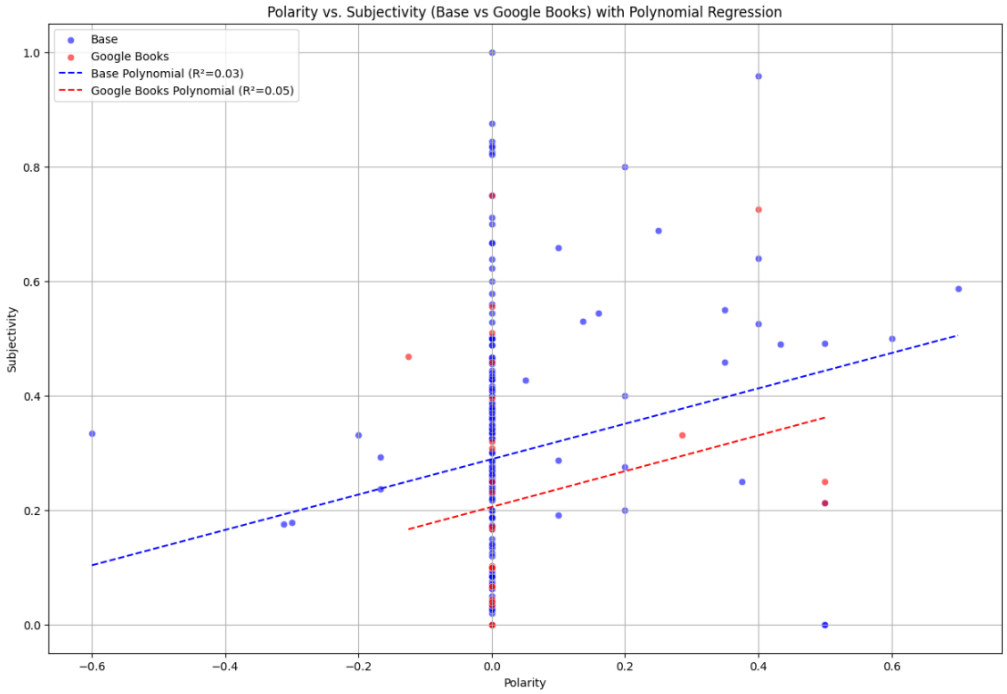


Fig. 4.12 *Polynomial Regression (degree = 1)*

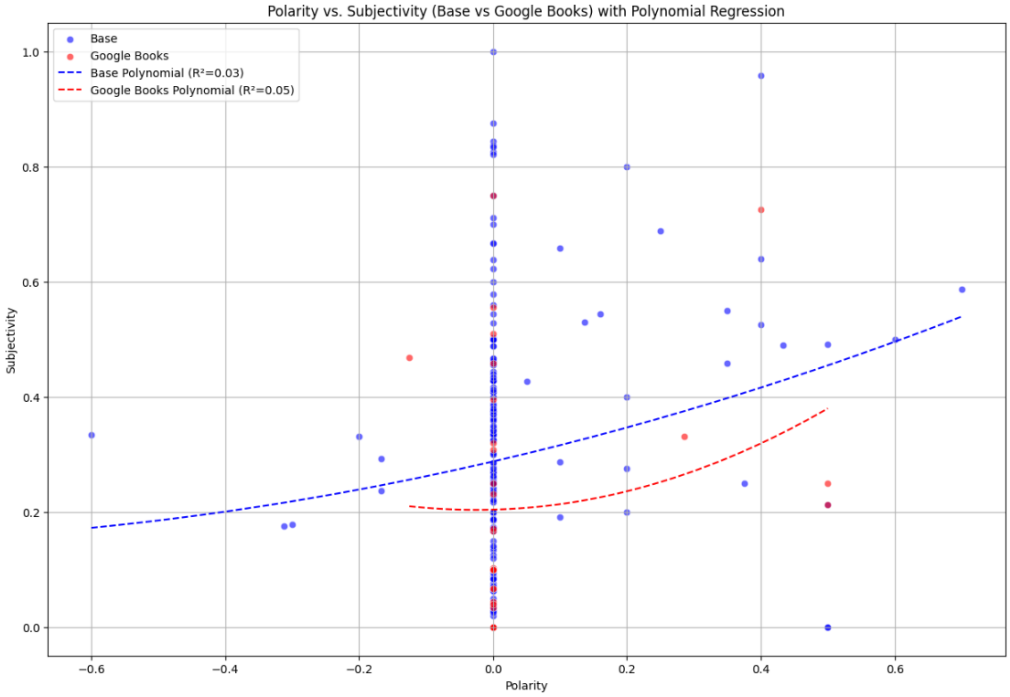


Fig. 4.13 *Polynomial Regression (degree = 2)*

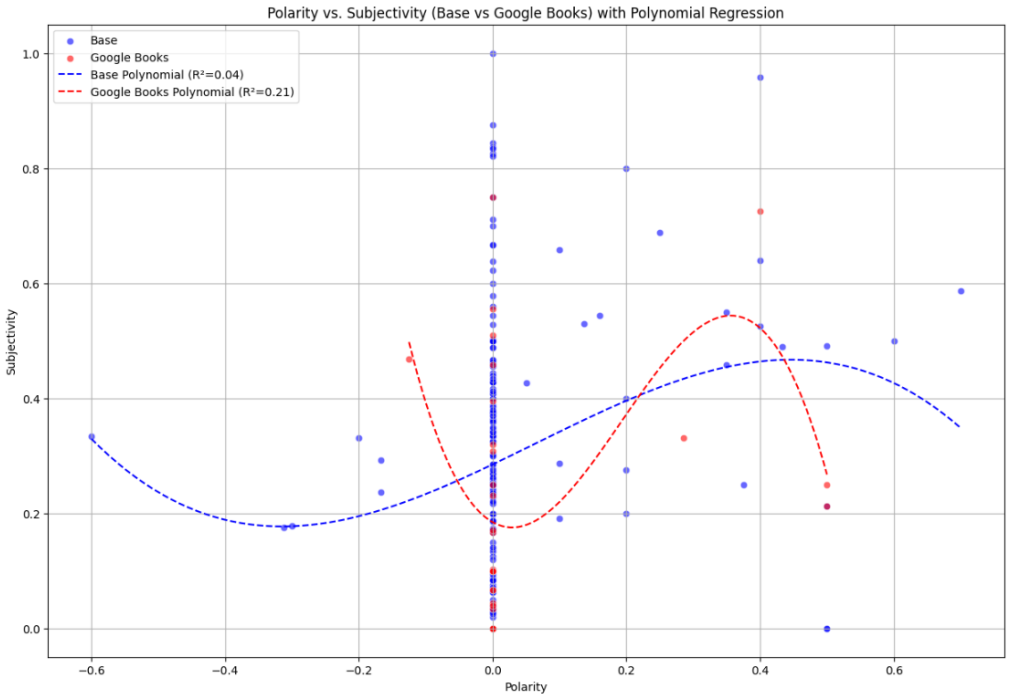


Fig. 4.14 *Polynomial Regression (degree = 3)*

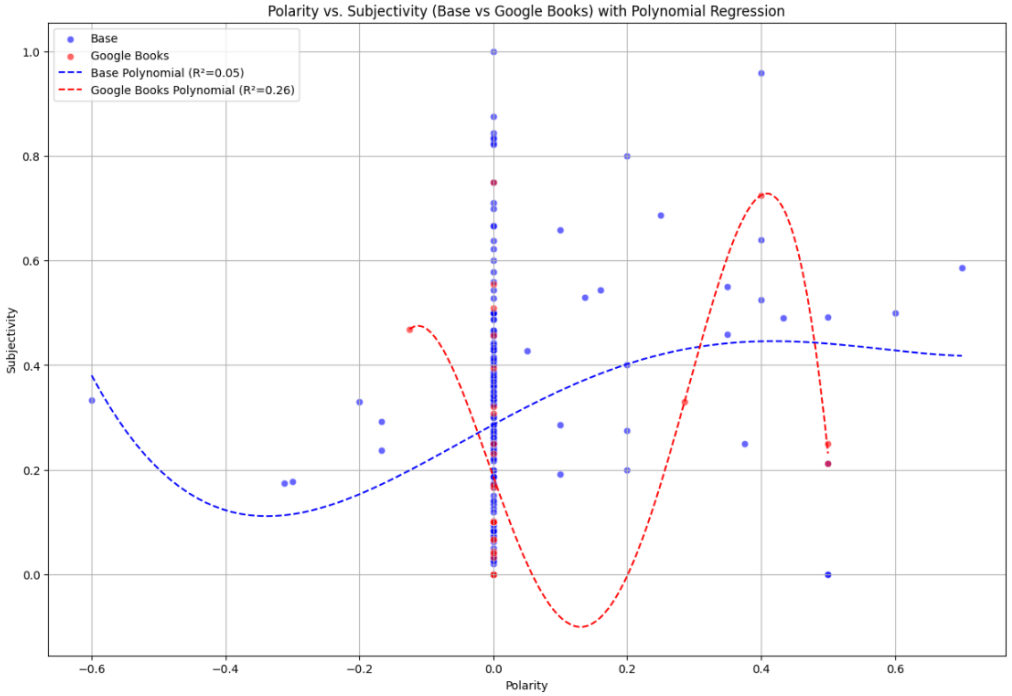


Fig. 4.15 *Polynomial Regression (degree = 4)*

From these four diagrams, it is clear that the regression method is not suitable to represent the relationship between polarity and subjectivity of a word. The reason is that all the regression does not provide good statistics, which can be observed especially from the **coefficient of determination**. It just can simply conclude that the relationship between polarity and subjectivity is extremely weak, especially when the polynomial degree is small in value (1 and 2).

Moreover, although the coefficient of determination gets better as the degree of polynomial regression gets larger, it is not suitable to fit the data into polynomial regression with high degree, as specific as greater than 3. This is because it can cause improper and unreliable predictions. Obverse the graph in figure of polynomial regression degree 4. It is obvious that within the interval of polarity in somewhere 0.05 to 0.2, there is an upward concave in the regression line. In this interval, we can see that the subjectivity is invalid as the value is less than 0, which does not rely on the range of subjectivity score of 0 to 1. Hence, there will exist the issue when predicting subjectivity by using polarity, by fitting the data into polynomial regression of high degree.

Despite the weak statistical measure, the polynomial regression with degree 3 seems to be the best polynomial regression to represent the association between polarity and subjectivity due to its moderate fitting performance and stability in prediction.

4.4. List of Python classes and Jupyter notebook(s) for lexicon enrichment:

Name of Python class(es) / notebooks	Author
lexicon_analysis.ipynb	Chiam Zi Wei
statistics_calc.py	Chiam Zi Wei
metrics_calc.py	Chiam Zi Wei

4.5. Code for Python Classes

4.5.1 lexicon_analysis.ipynb

Lexicon Analysis

Create Spark Session

```
[1]: from pyspark.sql import SparkSession
     spark = SparkSession.builder.getOrCreate()
```

```
24/12/15 11:48:47 WARN Utils: Your hostname, JORDAN. resolves to a loopback address: 127.0.1.1; using 10.255.255.254 instead (on interface lo)
24/12/15 11:48:47 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/12/15 11:48:48 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

[2]:
```
df = spark.read.csv('lexicon_data/enriched_lexicon.csv', header=True)
df.show()
```

```
+-----------+----------+-------------+--------------------+--------------------+--------------------+--------------------+--------------------+------
-------+--------+--------------------+------------------+
|      vocab|word_length|part_of_speech|         definitions|        sentences_eg|       derived_words|            synonyms|            antonyms|freq_o
f_usage|polarity|        subjectivity| subjectivity_score|
+-----------+----------+-------------+--------------------+--------------------+--------------------+--------------------+--------------------+------
-------+--------+--------------------+------------------+
|    berniat|         7|     kata nama|1. bertujuan (aka...|1. aku pun tidak ...|berniat;berniat-n...|maksud;tujuan;ing...|                NULL|
0|    none|         fact-driven|                0.0|
|     mundur|         6|      adjektif|1. (berjalan, ber...|1. dia mundur set...|memundurkan;kemun...|merosot;meleset;m...|           maju;maju|
0|    none|partially fact-dr...|0.30083333333333334|
|       kaum|         4|     kata nama|1. puak, suku ban...|1. banyak kaum or...|   berkaum;perkauman|suku;kabilah;taef...|                NULL|
5|    none|         fact-driven|                0.1|
|       wang|         4|     kata nama|1. alat pertukara...|1. satuan wang ne...|  berwang;mengewangi|duit;fulus;pitis;...|                NULL|
6|    none|partially opinion...| 0.5777777777777777|
|      jelas|         5|      adjektif|1. terang (perkat...|1. suara itu tida...|berjelas-jelas;me...|terang;nyata;keta...|               kabur|
0|positive|         fact-driven|0.19166666666666665|
|      waktu|         5|     kata nama|1. rangkaian atau...|1. secara rohania...|sewaktu;sewaktu-w...|masa;jangka masa;...|                NULL|
0|    none|partially fact-dr...| 0.3712962962962963|
|      ramai|         5|      adjektif|1. riuh-rendah, h...|1. orang menyabun...|seramai;beramai-r...|bising;riuh;hiruk...|sunyi;sugul;lemba...|
1|positive|partially fact-dr...|0.49166666666666664|
|       kuat|         4|      adjektif|1. banyak tenagan...|1. walaupun ia su...|sekuat-kuatnya;be...|gagah;berdaya;ber...|lemah;rapuh;lemah...|
0|positive|partially fact-dr...|0.4897530864197307|
|     sahaja|         6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|bersahaja;bersaha...|satu-satunya;cuma...|               jarang|
13|    none|partially fact-dr...|0.45833333333333337|
|      kerja|         5|     kata nama|1. usaha (kegiata...|1. beberapa mingg...|sekerja;bekerja;m...|khidmat;pekerjaan...|                NULL|
0|    none|partially fact-dr...|0.27769360269360266|
|     rakyat|         6|     kata nama|1. seluruh pendud...|1. rakyat Malaysi...|          kerakyatan|warganegara;pendu...|    pemerintah;ketua|
25|    none|partially fact-dr...|                0.26|
|      penuh|         5|      adjektif|1. seluruhnya ber...|1. tong itu penuh...|sepenuhnya;memenu...|berisi;tepu;padat...|kosong;sedikit;ku...|
4|positive|partially opinion...|                0.55|
|  merupakan|         9|     kata nama|1. membentuk (mem...|1. peti rahsia it...|rupanya;rupa-rupa...|paras;muka;wajah;...|                NULL|
3|    none|      opinion-driven|                0.75|
| penjelasan|        10|      adjektif|1. keterangan (yg...|1. Ali meminta pe...|berjelas-jelas;me...|terang;nyata;keta...|               kabur|
0|    none|         fact-driven|                0.0|
|       niat|         4|     kata nama|1. maksud atau tu...|1. mematikan oran...|berniat;berniat-n...|maksud;tujuan;ing...|                NULL|
0|    none|partially fact-dr...| 0.2857142857142857|
|  bermaksud|         9|     kata nama|1. bertujuan, ber...|1. kami bermaksud...|bermaksud;memaksu...|tujuan;hajat;hasr...|                NULL|
1|    none|      opinion-driven|             0.84375|
|       tahu|         4|    kata kerja|1. maklum akan ke...|1. entah kapal pe...|tahu-tahu;mengeta...|maklum;faham;meng...|    jahil;membiarkan|
0|    none|         fact-driven|0.18571428571428572|
|memaksudkan|        11|     kata nama|1. menghendaki, m...|1. dialah yg dima...|bermaksud;memaksu...|tujuan;hajat;hasr...|                NULL|
0|negative|         fact-driven|               0.175|
|       masa|         4|     kata nama|1. waktu, ketika;...|1. masa dan tempa...|              semasa|waktu;ketika;deti...|                NULL|
3|    none|         fact-driven|0.08333333333333333|
|       jauh|         4|      adjektif|1. besar jaraknya...|1. rumahnya agak ...|berjauh;menjauhi;...|tidak dekat;besar...|               dekat|
2|positive|partially opinion...| 0.6583333333333334|
+-----------+----------+-------------+--------------------+--------------------+--------------------+--------------------+--------------------+------
-------+--------+--------------------+------------------+
only showing top 20 rows
```

Descriptive Analysis

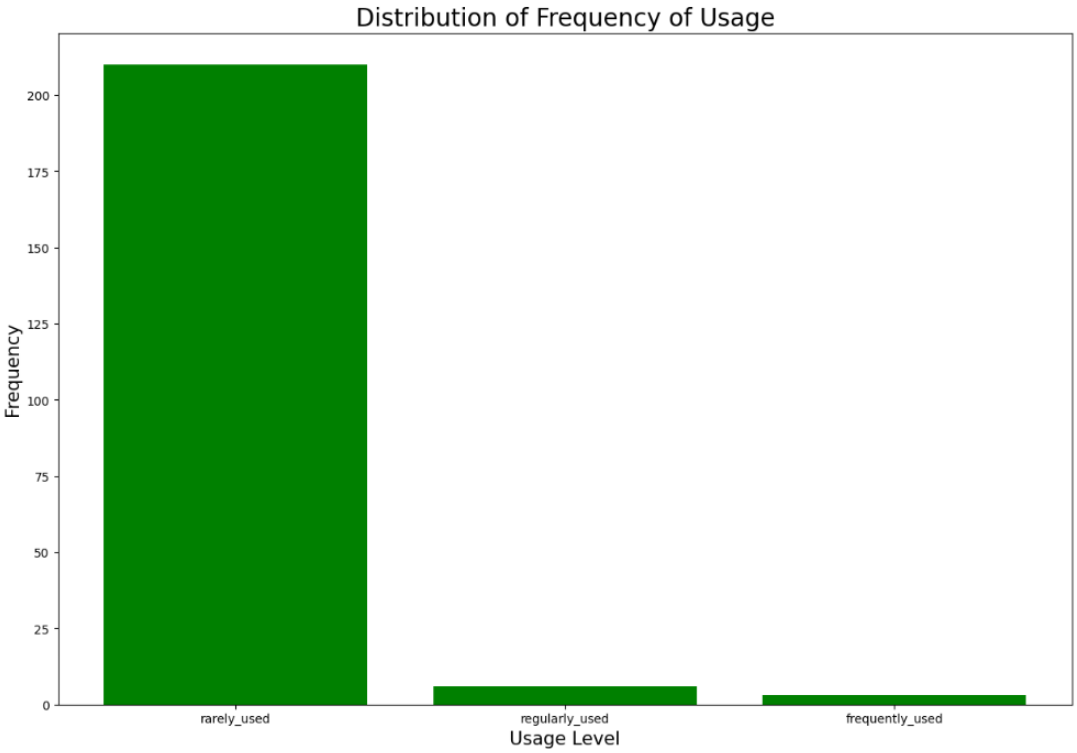
Distribution of Frequency of Usage

[3]:
```
df_rows = df.collect()

freq_of_use_level = {
    "rarely_used"    : 0,
    "regularly_used" : 0,
    "frequently_used": 0
}

min_freq = 99999
max_freq = -99999
for row in df_rows:
    freq_of_use = int(row[8])
    if freq_of_use < min_freq:
        min_freq = freq_of_use
    if freq_of_use > max_freq:
        max_freq = freq_of_use

    if freq_of_use >= 0 and freq_of_use < 15:
        freq_of_use_level["rarely_used"] = freq_of_use_level["rarely_used"] + 1
    elif freq_of_use >= 15 and freq_of_use < 30:
        freq_of_use_level["regularly_used"] = freq_of_use_level["regularly_used"] + 1
    else:
        freq_of_use_level["frequently_used"] = freq_of_use_level["frequently_used"] + 1

print(freq_of_use_level)
```

```
{'rarely_used': 210, 'regularly_used': 6, 'frequently_used': 3}
```

[4]:
```
import numpy as np
import matplotlib.pyplot as plt

plt.figure(figsize = (15, 10))
plt.title('Distribution of Frequency of Usage', fontsize = 20)
plt.xlabel('Usage Level', fontsize = 15)
plt.ylabel('Frequency', fontsize = 15)
plt.bar(freq_of_use_level.keys(), freq_of_use_level.values(), color='g')
```

[4]: <BarContainer object of 3 artists>

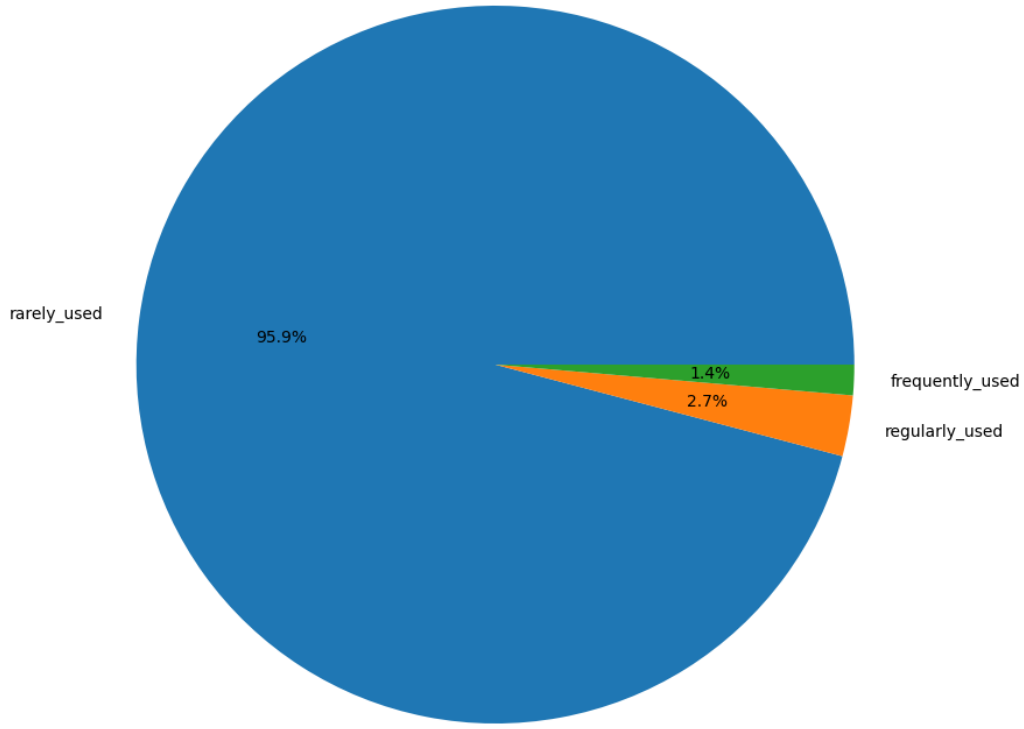


From the histogram above, the following insights can be concluded:

- most of the words in lexicon are rarely being used
- only a few of the words are regularly and frequently used

Pie Chart of Frequency of Usage Distribution

[5]:
```
fig, ax = plt.subplots(figsize = (10, 10))
ax.pie(freq_of_use_level.values(), labels = freq_of_use_level.keys(), autopct='%1.1f%%')

plt.title('Pie Chart of Frequency of Usage Distribution', fontsize = 20)
plt.show()
```



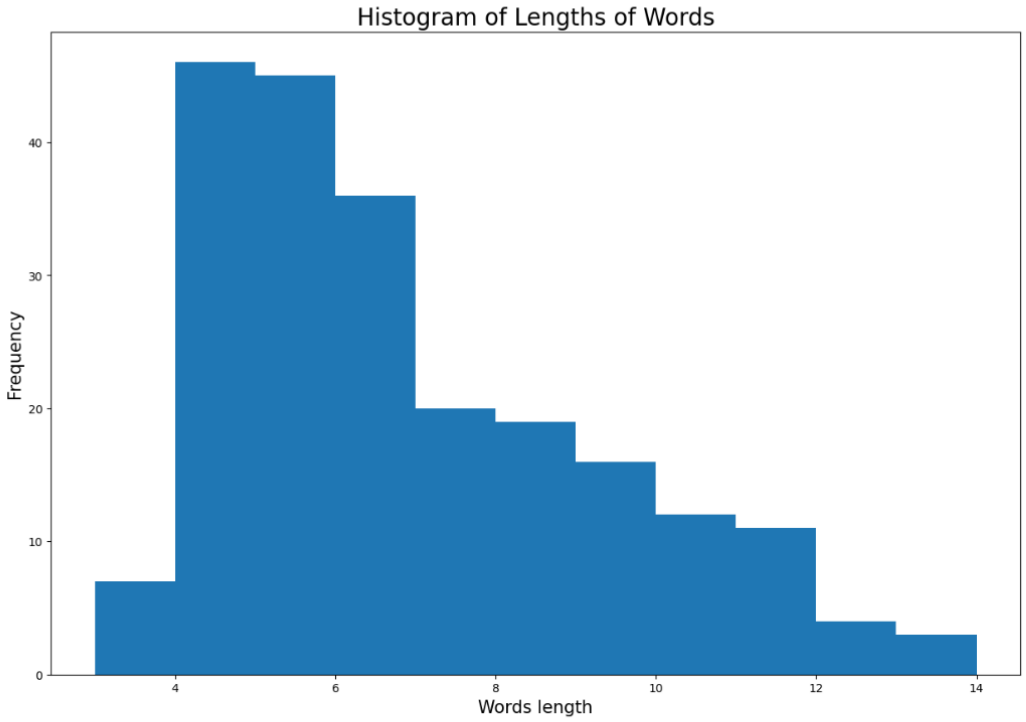
[6]:
```
print("Least Frequent Used Vocab:", df.where(df.freq_of_usage == min_freq).select('vocab', 'freq_of_usage').count())

print("Most Frequent Used Vocab:",df.where(df.freq_of_usage == max_freq).select('vocab', 'freq_of_usage').count())
```

```
Least Frequent Used Vocab: 141
Most Frequent Used Vocab: 1
```

Distribution of Lengths of Words

[7]:
```
word_lengths = []

for row in df_rows:
    word_lengths.append(int(row[1]))

plt.figure(figsize=(15, 10))
plt.title('Histogram of Lengths of Words', fontsize = 20)
plt.xlabel('Words length', fontsize = 15)
plt.ylabel('Frequency', fontsize=15)
plt.hist(word_lengths, bins=11)
```

[7]:
```
(array([ 7., 46., 45., 36., 20., 19., 16., 12., 11.,  4.,  3.]),
 array([ 3.,  4.,  5.,  6.,  7.,  8.,  9., 10., 11., 12., 13., 14.]),
 <BarContainer object of 11 artists>)
```



The following insights are obtained from the histogram above:

- the distribution in length of words skewed to the right.
- the distribution in length of words is concentrated at 4 - 6.
- the distribution in length of words is approximately uniform in the centre of distribution (7 - 12).

Prefixes and Suffixes Analysis

[8]:
```
prefixes = []
suffixes = []

for row in df_rows:
    vocab = row[0]

    if row[5] != None:
        derived_words_list = row[5].split(';')

        for i in derived_words_list:
            try:
                pref_end_index = i.index(vocab)
                prefix = i[:pref_end_index]
                suffix = i[pref_end_index + len(vocab):]

                if prefix.find('-') == -1 and prefix != None and prefix not in prefixes:
                    prefixes.append(prefix)
                if suffix.find('-') == -1 and suffix != None and suffix not in suffixes:
                    suffixes.append(suffix)
            except:
                pass

while '' in prefixes:
    prefixes.remove('')

while '' in suffixes:
    suffixes.remove('')

prefixes = [i.replace(' ', '') for i in prefixes]
suffixes = [i.replace(' ', '') for i in suffixes]
print(prefixes, suffixes)
```

```
['me', 'ke', 'ber', 'per', 'menge', 'men', 'pen', 'se', 'be', 'memper', 'ter', 'pe', 'sepe', 'berpe', 'berke', 'penge', 'berpenge', 'meng', 'peng', 'seke', 'di', 'ke', 'bers
e', 'berper', 'menye'] ['kan', 'an', 'i', 'nya', 'in', 'hati', 'takmahu', 'pun', '2', '3']
```

[9]:
```
malay_prefix = ['me', 'mem', 'meng', 'memper', 'mempe', 'be', 'ber', 'ke', 'pe', 'per', 'peng', 'pen', 'se', 'di', 'anti', 'juru']
malay_suffix = ['an', 'kan', 'i', 'nya', 'pun']

def intersection(l1, l2):
    return [value for value in l1 if value in l2]

prefixes = intersection(malay_prefix, prefixes)
suffixes = intersection(malay_suffix, suffixes)
print("Prefixes Existed: " + str(prefixes) + "\nSuffixes Existed: " + str(suffixes))
```

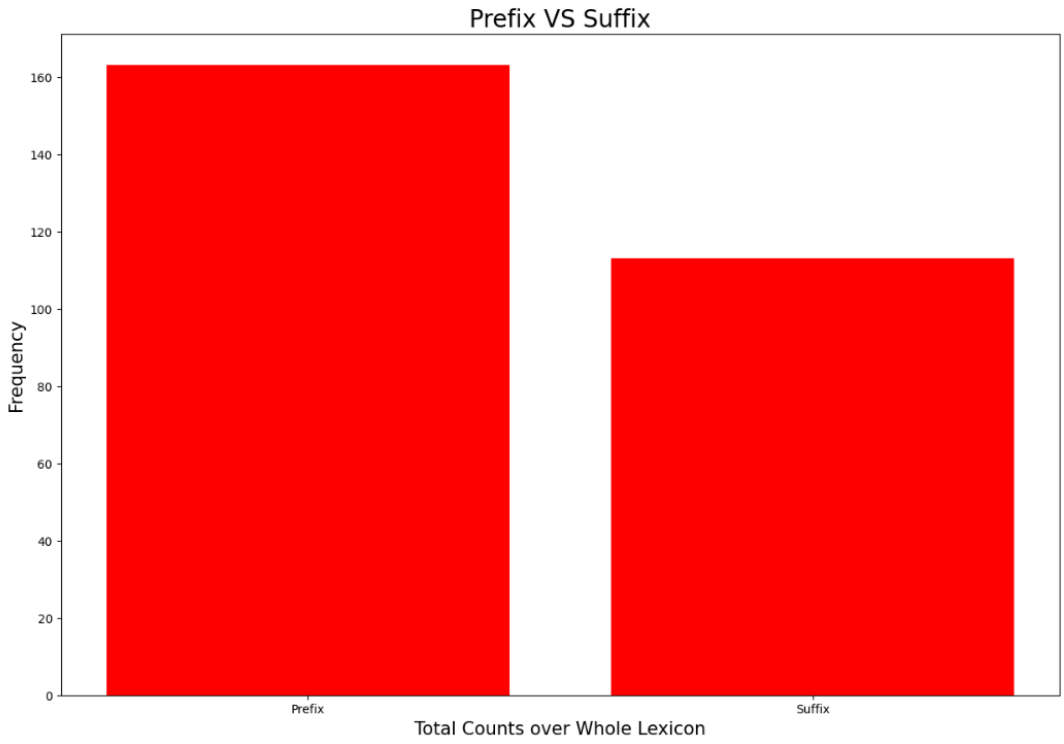
```
Prefixes Existed: ['me', 'meng', 'memper', 'be', 'ber', 'ke', 'pe', 'per', 'peng', 'pen', 'se', 'di']
Suffixes Existed: ['an', 'kan', 'i', 'nya', 'pun']
```

[10]:
```
total_prefix_count = 0
total_suffix_count = 0

for row in df_rows:
    vocab = row[0]

    if row[5] != None:
        derived_words_list = row[5].split(';')

    for i in derived_words_list:
        try:
            pref_end_index = i.index(vocab)
            prefix = i[:pref_end_index]
            suffix = i[pref_end_index + len(vocab):]

            if prefix in prefixes:
                total_prefix_count += 1
            if suffix in suffixes:
                total_suffix_count += 1
        except:
            pass

print(
    "Total Prefix Count: " + str(total_prefix_count) +
    "\nTotal Suffix Count: " + str(total_suffix_count)
)
```

```
Total Prefix Count: 163
Total Suffix Count: 113
```

[11]:
```
plt.figure(figsize=(15, 10))
plt.title('Prefix VS Suffix', fontsize=20)
plt.xlabel('Total Counts over Whole Lexicon', fontsize=15)
plt.ylabel('Frequency', fontsize=15)
plt.bar(['Prefix', 'Suffix'], [total_prefix_count, total_suffix_count], color='r')
```

[11]:
```
<BarContainer object of 2 artists>
```



In here, we can simply conclude that **the usage of prefix is more frequent than the usage of suffix**.

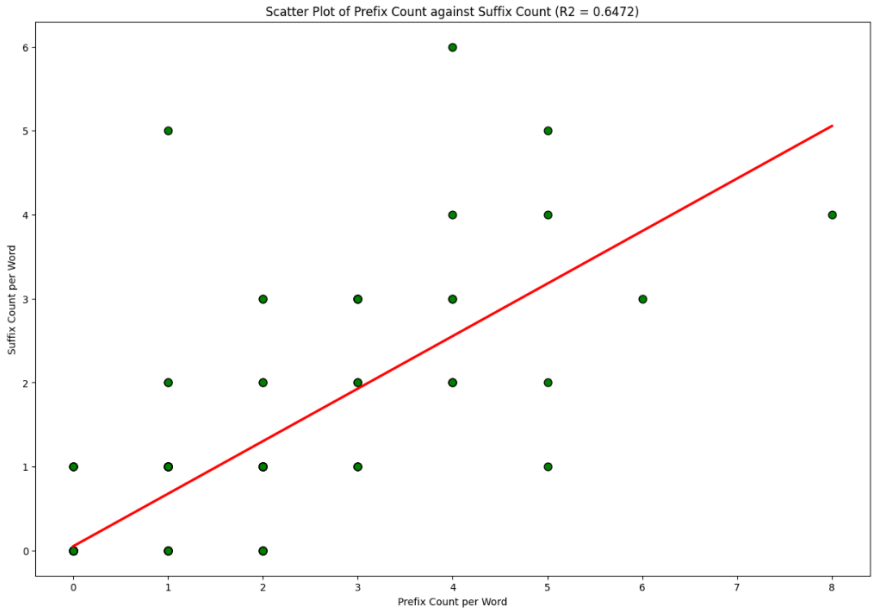
Relationship between Prefixes and Suffixes

[12]:
```
prefix_per_word = []
suffix_per_word = []

for row in df_rows:
    vocab = row[0]
    temp_pref_quant = 0
    temp_suff_quant = 0

    if row[5] != None:
        derived_words_list = row[5].split(';')

    for i in derived_words_list:
        try:
            pref_end_index = i.index(vocab)
            prefix = i[:pref_end_index]
            suffix = i[pref_end_index + len(vocab):]

            if prefix in prefixes:
                temp_pref_quant += 1
            if suffix in suffixes:
                temp_suff_quant += 1
        except:
            pass

    prefix_per_word.append(temp_pref_quant)
    suffix_per_word.append(temp_suff_quant)

print(
    "Prefix Count Per Word: " + str(prefix_per_word) +
    "\nSuffix Count Per Word: " + str(suffix_per_word)
)
```

```
Prefix Count Per Word: [0, 2, 2, 1, 3, 2, 4, 5, 2, 4, 1, 1, 0, 0, 3, 0, 1, 0, 1, 1, 2, 2, 1, 1, 0, 0, 1, 0, 2, 0, 1, 3, 2, 0, 0, 1, 1, 4, 0, 2, 1, 0, 1, 0, 2, 1, 5, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 3, 2,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 2, 0, 0, 2, 4, 0, 0, 1, 0, 3, 0, 0, 1, 0, 0, 0, 0, 4, 0, 2, 0, 0, 2, 0, 2, 0, 2, 0, 0, 0, 0, 0, 4, 0, 0, 1, 0,
3, 0, 0, 0, 0, 0, 3, 1, 0, 0, 0, 0, 0, 0, 1, 4, 0, 0, 0, 2, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 1, 0, 3, 0, 2, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 8, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 6, 0, 0, 0, 0, 0, 0, 3, 2, 0]
Suffix Count Per Word: [0, 2, 1, 1, 3, 0, 2, 1, 0, 4, 1, 1, 0, 0, 1, 0, 5, 0, 0, 2, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 3, 1, 0, 0, 2, 0, 3, 1, 1, 1, 0, 0, 0, 1, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 2, 6, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 2, 0, 1, 0, 0, 1, 0, 3, 0, 1, 0, 0, 0, 0, 0, 3, 0, 0, 1, 0,
3, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 1, 0, 0, 2, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 1, 0, 3, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 2, 1, 0]
```

[13]:
```
import numpy as np
import scipy

r = scipy.stats.linregress(prefix_per_word, suffix_per_word)[2]

fig, ax = plt.subplots(figsize=(15, 10))
plt.title(f'Scatter Plot of Prefix Count against Suffix Count (R2 = {r**2:.4f})')
plt.xlabel('Prefix Count per Word')
plt.ylabel('Suffix Count per Word')
ax.scatter(prefix_per_word, suffix_per_word, s=60, color='g', edgecolors='k')

b1, b0 = np.polyfit(prefix_per_word, suffix_per_word, deg=1)
xseq = np.linspace(0, 8, num = 100)


ax.plot(xseq, b0 + b1 * xseq, color='r', lw=2.5)
```

[13]:
```
[<matplotlib.lines.Line2D at 0x7fc228d73d30>]
```



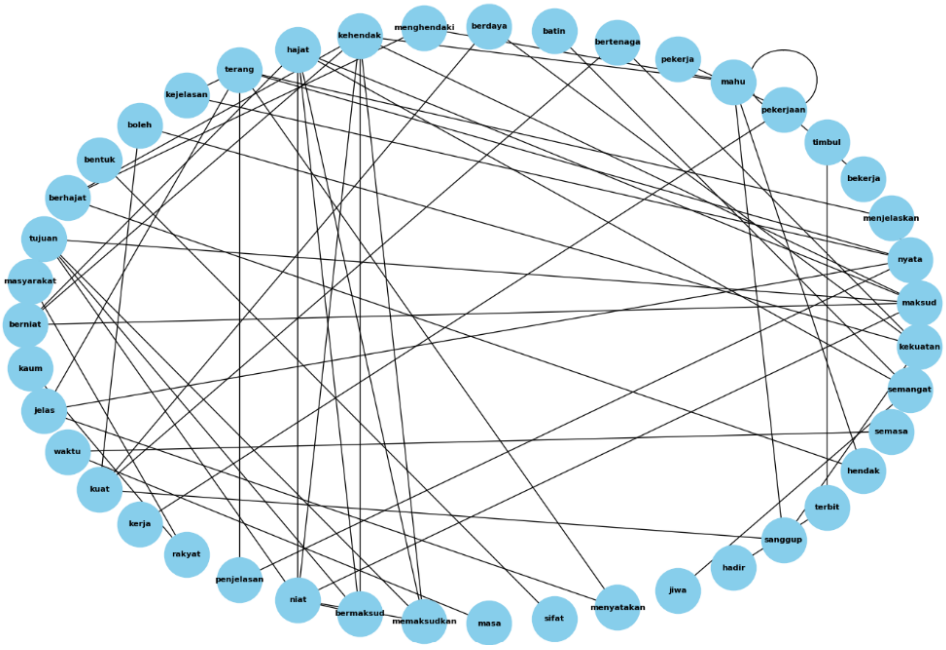
The regression line by plotting prefix count per word againt suffix count per word gives a coefficient of determination of 0.6479 ($r^2 = 0.6479$). This indicates that the appearance of possible prefix and suffix in one single word has a **relatively strong** association.

Synonyms Relationship Network

[14]:
```
import networkx as nx

ng = nx.Graph()

vocabs = []
for row in df_rows:
    ng.add_node(row[0])
    vocabs.append(row[0])

for k in range(len(df_rows)):
    if df_rows[k][6] != None:
        synonyms = df_rows[k][6].split(';')

        for i in range(len(synonyms)):
            for j in range(len(vocabs)):
                if synonyms[i] == vocabs[j] and i != j:
                    ng.add_edge(df_rows[k][0], vocabs[j])
                    break
```

[15]:
```
isolated_nodes = [node for node, degree in dict(ng.degree()).items() if degree == 0]
ng.remove_nodes_from(isolated_nodes)

plt.figure(figsize=(15,10))
pos = nx.shell_layout(ng)
nx.draw(
    ng, pos, with_labels=True, node_color="skyblue", node_size=2000, font_size=8, font_color="black", font_weight="bold"
)
plt.title("Synonym Relationship Network", fontsize=15)
plt.show()
```

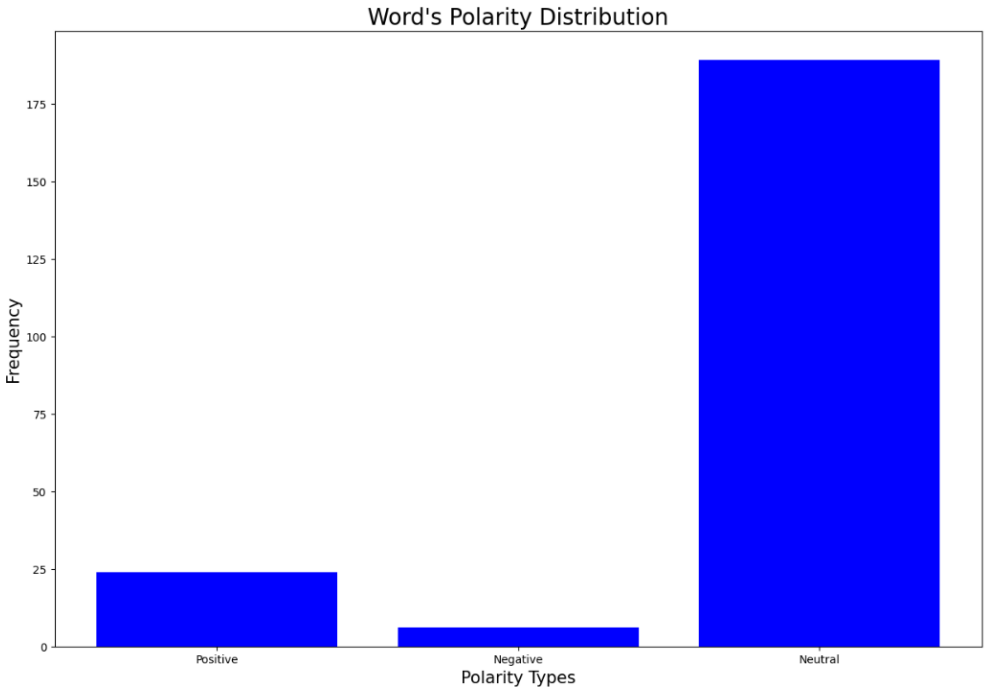
Synonym Relationship Network



Words' Polarity Analysis

[16]:
```
total_positive_count = 0
total_negative_count = 0
total_neutral_count = 0

for row in df_rows:
    polarity = row[9]

    if polarity == 'positive':
        total_positive_count += 1
    elif polarity == 'negative':
        total_negative_count += 1
    elif polarity == 'none':
        total_neutral_count += 1

plt.figure(figsize=(15, 10))
plt.title("Word's Polarity Distribution", fontsize=20)
plt.xlabel('Polarity Types', fontsize=15)
plt.ylabel('Frequency', fontsize=15)
plt.bar(['Positive', 'Negative', 'Neutral'], [total_positive_count, total_negative_count, total_neutral_count], color='b')
```

[16]: <BarContainer object of 3 artists>



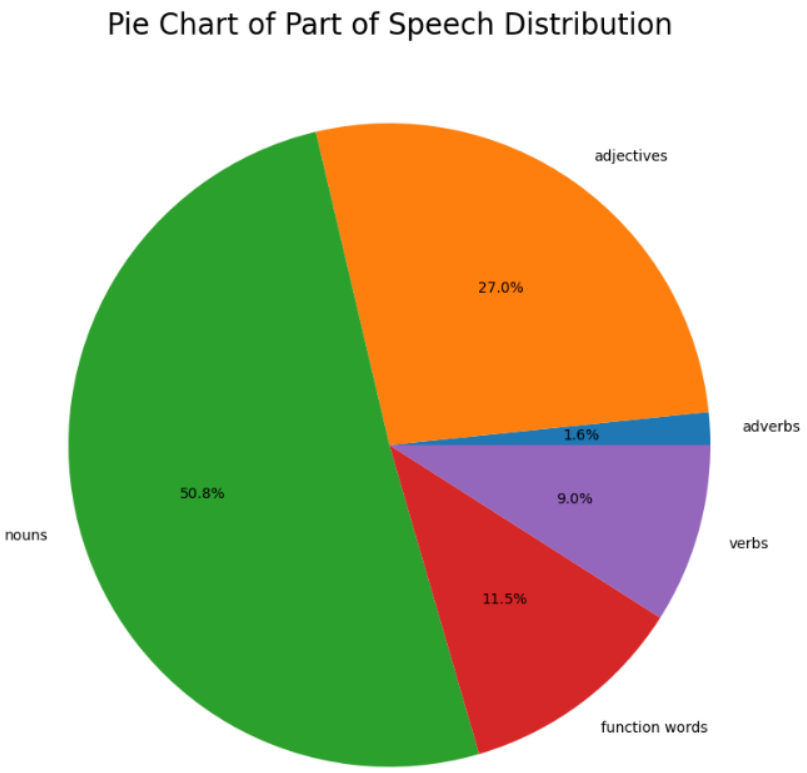
From the data scrapped, it is found that most of the words are in neutral polarity. This may be caused by the reason where **most of the words are not adjectives** (adjectives have more tendency of being positive or negative).

Distribution of Part of Speech

[17]:
```
df.select('part_of_speech').distinct().show()
```

```
+--------------+
|part_of_speech|
+--------------+
|       adverba|
|      adjektif|
|     kata nama|
|    kata tugas|
|    kata kerja|
|          NULL|
+--------------+
```

[18]:
```
total_adverb_count = 0
total_adjective_count = 0
total_nouns_count = 0
total_func_words_count = 0
total_verbs_count = 0

for row in df_rows:
    part_of_speech = row[2]

    if part_of_speech == 'adverba':
        total_adverb_count += 1
    elif part_of_speech == 'adjektif':
        total_adjective_count += 1
    elif part_of_speech == 'kata nama':
        total_nouns_count += 1
    elif part_of_speech == 'kata tugas':
        total_func_words_count += 1
    elif part_of_speech == 'kata kerja':
        total_verbs_count += 1

fig, ax = plt.subplots(figsize = (10, 10))
ax.pie([total_adverb_count, total_adjective_count, total_nouns_count, total_func_words_count, total_verbs_count], labels = ['adverbs', 'adjectives', 'nouns', 'function words', 'verbs'], autopct='%1.1f%%')

plt.title('Pie Chart of Part of Speech Distribution', fontsize = 20)
plt.show()
```



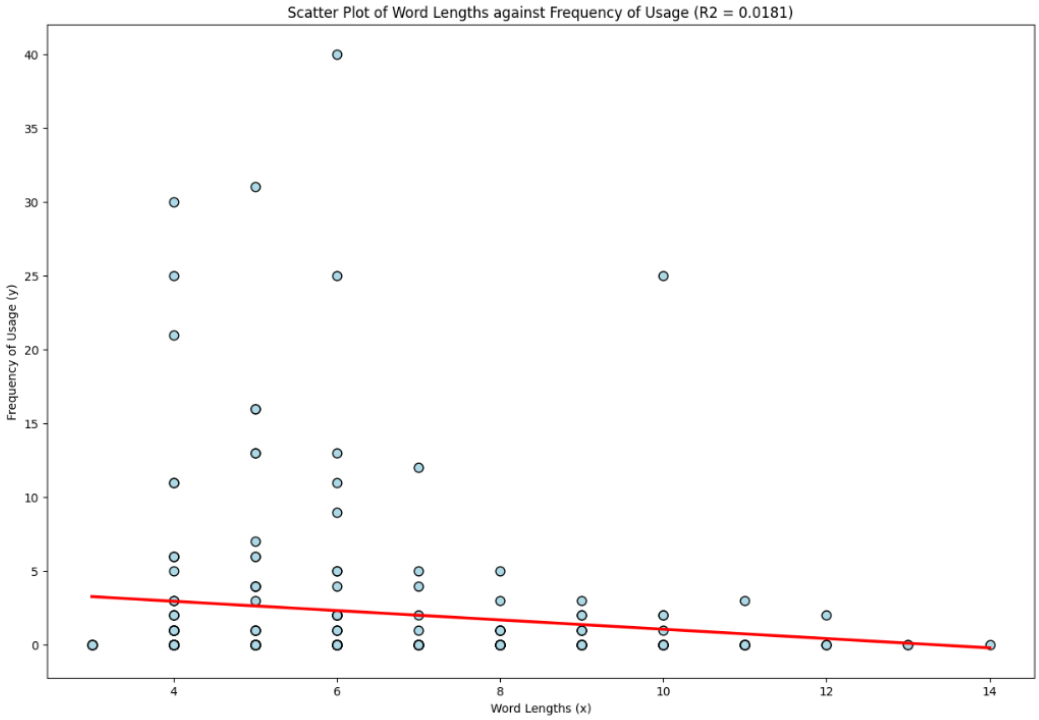
In here, we found that

1. the most part of speech: **nouns**.
2. the least part of speech: **adverbs**.
3. **function words** and **verbs** has similar contribution.

Study about Correlation of Word Lengths and Frequency of Usage

[19]:
```
freq_of_use = []

for row in df_rows:
    freq_of_use.append(int(row[8]))

r = scipy.stats.linregress(word_lengths, freq_of_use)[2]

fig, ax = plt.subplots(figsize=(15, 10))
plt.title(f'Scatter Plot of Word Lengths against Frequency of Usage (R2 = {r**2:.4f})')
plt.xlabel('Word Lengths (x)')
plt.ylabel('Frequency of Usage (y)')
ax.scatter(word_lengths, freq_of_use, s=60, color='lightblue', edgecolors='k')

b1, b0 = np.polyfit(word_lengths, freq_of_use, deg=1)
xseq = np.linspace(3, 14, num = 100)

ax.plot(xseq, b0 + b1 * xseq, color='r', lw=2.5)
```

[19]: [<matplotlib.lines.Line2D at 0x7fc2285f35b0>]



Statistics about Subjectiivty Scores

[20]:
```
from python_script.statistics_calc import *
from pyspark.sql import Row
import pandas as pd

subjectivity_scores = []

for row in df_rows:
    subjectivity_scores.append(float(row[-1]))

label_list = ['N', 'mean', 'variance', 'standard deviation', 'standard error', 'range', 'coefficient of range', 'confidence interval']
value_list = [len(subjectivity_scores), mean(subjectivity_scores), var(subjectivity_scores), stddev(subjectivity_scores), stderr(subjectivity_scores),
              val_range(subjectivity_scores), cor(subjectivity_scores), conf_int(subjectivity_scores, 0.01)]

pd.DataFrame({
    'Statistic': label_list,
    'Value': value_list
})

for i in range(len(label_list)):
    if i != len(label_list) - 1:
        print(f'{label_list[i]:>20} : {value_list[i]:.4f}')
    else:
        print(f'{label_list[i]:>20} : {value_list[i]}')
```

```
                   N : 219.0000
                mean : 0.2972
            variance : 0.0504
  standard deviation : 0.2246
      standard error : 0.0152
               range : 1.0000
coefficient of range : 1.0000
 confidence interval : (0.2619, 0.3325)
```

Intrinsic Evaluation

Lexicon Coverage

[24]:
```
data_file = f'google_books_data/gb_data_{int(input()):04d}.csv'

gb_df = spark.read.csv(data_file, header=True)
gb_df.show()
```

```
 1
+-----------+----------+--------------+--------------------+--------------------+--------------------+--------------------+--------+-------------+--------+--------------------+--------------------+
|      vocab|word_length|part_of_speech|         definitions|        sentences_eg|       derived_words|            synonyms|antonyms|freq_of_usage|polarity|        subjectivity| subjectivity_score|
+-----------+----------+--------------+--------------------+--------------------+--------------------+--------------------+--------+-------------+--------+--------------------+--------------------+
|       asing|         5|          NULL|1. berlainan, ter...|1. banyak bangsa ...|                NULL|                NULL|    NULL|            2|negative|partially fact-dr...|             0.46875|
|        atas|         4|          NULL|1 tempat atau bah...|1. atas bumbung. ...|                NULL|                NULL|    NULL|            1|positive|partially fact-dr...|                0.25|
|       badan|         5|          NULL|1 semua bahagian ...|1. badan orang it...|                NULL|                NULL|    NULL|            0|    none|         fact-driven|0.06607142857142857|
|      berisi|         6|     kata nama|1. mempunyai isi,...|1. bintit-bintit ...|seisi;berisi;meng...|kandungan;muatan;...|    NULL|            0|    none|partially opinion...| 0.7250000000000001|
|berkedudukan|        12|          NULL|1. mempunyai kedu...|1. jarum penimban...|                NULL|                NULL|    NULL|            0|    none|partially fact-dr...|0.39499999999999996|
|       besar|         5|          NULL|1 tidak kecil uku...|1. Rumahnya besar...|                NULL|                NULL|    NULL|            7|    none|         fact-driven|0.16666666666666669|
|        buku|         4|          NULL|1 bahagian yg ker...|1. Ahmad menghanc...|                NULL|                NULL|    NULL|            0|    none|         fact-driven|0.03333333333333333|
|       dalam|         5|          NULL|1. jarak (jauhnya...|1. dalam perigi i...|                NULL|                NULL|    NULL|           56|    none|         fact-driven|                0.1|
|       dapat|         5|          NULL|1 boleh (berupaya...|1. Saya dapat men...|                NULL|                NULL|    NULL|           13|    none|partially opinion...| 0.5548611111111111|
|       dasar|         5|          NULL|1. bahagian yg di...|1. mutiara didapa...|                NULL|                NULL|    NULL|            0|    none|         fact-driven|0.10250000000000001|
|      dengan|         6|          NULL|1. (ber)serta, be...|1. saya pergi men...|                NULL|                NULL|    NULL|           28|    none|         fact-driven|             0.04375|
|         hak|         3|          NULL|1. (yg) benar, ke...|1. katakanlah yg ...|              berhak|                NULL|    NULL|            0|positive|partially fact-dr...| 0.3303571428571428|
|       hukum|         5|     kata nama|1. peraturan yg d...|1. hukum mengerja...|hukuman;menghukum...|                NULL|    NULL|            0|    none|         fact-driven|                0.1|
|       lebih|         5|          NULL|1 melampaui had y...|1. Panjangnya leb...|lebih-lebih lagi;...|                NULL|  kurang|           16|positive|         fact-driven|0.21208333333333332|
|   merupakan|         9|     kata nama|1. membentuk (mem...|1. peti rahsia it...|rupanya;rupa-rupa...|paras;muka;wajah;...|    NULL|            3|    none|      opinion-driven|                0.75|
|       milik|         5|     kata nama|1. kepunyaan, hak...|1. semua tanah ad...|    memiliki;pemilik|                NULL|    NULL|            0|    none|partially opinion...| 0.5089285714285714|
|        oleh|         4|          NULL|1 kata sendi nama...|1. Daerah itu tel...|                NULL|                NULL|    NULL|           11|    none|         fact-driven|             0.23125|
|   peraturan|         9|          NULL|1. segala yg tela...|1. setiap pekerja...|                NULL|                NULL|    NULL|            0|    none|         fact-driven|                0.0|
| persembahan|        11|          NULL|1. hadiah (kpd ra...|1. segala negeri ...|                NULL|                NULL|    NULL|            0|    none|         fact-driven|                0.0|
|        pokok|         5|     kata nama|1 segala jenis tu...|1. pokok pisang;2...|            berpokok|                NULL|    NULL|            0|    none|partially fact-dr...|0.45746464183964186|
+-----------+----------+--------------+--------------------+--------------------+--------------------+--------------------+--------+-------------+--------+--------------------+--------------------+
only showing top 20 rows
```

[25]:
```
from sklearn.metrics import *

base_words = df.select('vocab', 'definitions')
gb_words = gb_df.select('vocab', 'definitions')
```

[26]:
```
from python_script.metrics_calc import *

coverage_mark = coverage(base_words, gb_words)
print(coverage_mark)
print('or in percentage,', f'{coverage_mark*100:.2f}%')
```

```
0.0410958904109589
or in percentage, 4.11%
```

Lexicon Accuracy

exact annotation accuracy

[27]:
```
anno_acc = annotation_accuracy_exact(base_words, gb_words)
print(anno_acc)
print('or in percentage,', f'{anno_acc*100:.2f}%')
```

```
0.0
or in percentage, 0.00%
```

levenshtein distance similarity

[28]:
```
lev_simi = similarity_levenshtein(base_words, gb_words)
print(lev_simi)
print('or in percentage,', f'{lev_simi:.2f}%')
```

```
9.322321935754918
or in percentage, 9.32%
```

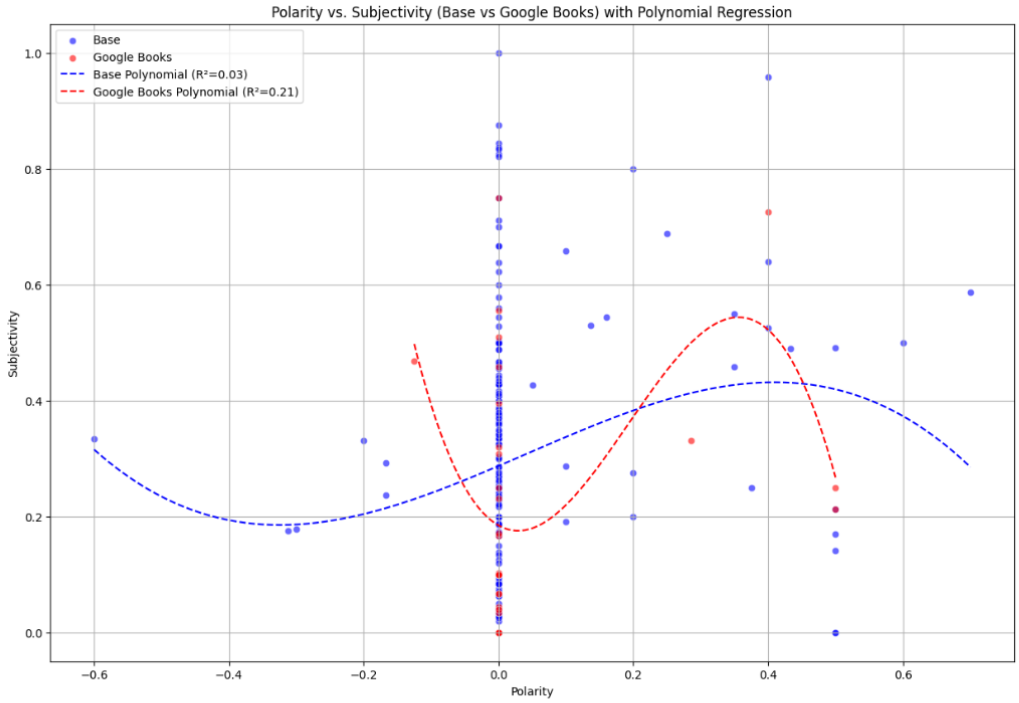
Extrinsic Evaluation

Polarity VS Subjectivity ***with polynomial regression***

[29]:
```
from python_script.sentiment import *

base_words = df.select('vocab', 'subjectivity_score')
gb_words = gb_df.select('vocab', 'subjectivity_score')

base_rows = base_words.collect()
gb_rows = gb_words.collect()

base_polarity = []
base_subjectivity = []
for row in base_rows:
    vocab = row[0]
    subjectivity = row[1]

    trans_vocab = translateFromMyToEn(vocab)

    polarity = getPolarityValue(trans_vocab)

    base_polarity.append(polarity)
    base_subjectivity.append(float(subjectivity))

gb_polarity = []
gb_subjectivity = []
for row in gb_rows:
    vocab = row[0]
    subjectivity = row[1]

    trans_vocab = translateFromMyToEn(vocab)

    polarity = getPolarityValue(trans_vocab)

    gb_polarity.append(polarity)
    gb_subjectivity.append(float(subjectivity))
```

[30]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
from scipy.stats import linregress

max_len = max(len(base_polarity), len(gb_polarity))

base_polarity += [None] * (max_len - len(base_polarity))
base_subjectivity += [None] * (max_len - len(base_subjectivity))
gb_polarity += [None] * (max_len - len(gb_polarity))
gb_subjectivity += [None] * (max_len - len(gb_subjectivity))

data = {
    'base_polarity': base_polarity,
    'base_subjectivity': base_subjectivity,
    'gb_polarity': gb_polarity,
    'gb_subjectivity': gb_subjectivity
}

df = pd.DataFrame(data)

plt.figure(figsize=(15, 10))
sns.scatterplot(x='base_polarity', y='base_subjectivity', data=df, color='blue', label='Base', alpha=0.6)
sns.scatterplot(x='gb_polarity', y='gb_subjectivity', data=df, color='red', label='Google Books', alpha=0.6)

degree = int(input('Enter the degree of polynomial regression (positive integers) >> '))
base_non_nan = df.dropna(subset=['base_polarity', 'base_subjectivity'])
if not base_non_nan.empty:
    poly_base = np.polyfit(base_non_nan['base_polarity'], base_non_nan['base_subjectivity'], degree)
    poly_base_func = np.poly1d(poly_base)

    x_base_poly = np.linspace(base_non_nan['base_polarity'].min(), base_non_nan['base_polarity'].max(), 100)
    y_base_poly = poly_base_func(x_base_poly)

    y_base_pred = poly_base_func(base_non_nan['base_polarity'])
    r2_base = r2_score(base_non_nan['base_subjectivity'], y_base_pred)

    plt.plot(x_base_poly, y_base_poly, color='blue', linestyle='--', label=f'Base Polynomial (R²={r2_base:.2f})')

gb_non_nan = df.dropna(subset=['gb_polarity', 'gb_subjectivity'])
if not gb_non_nan.empty:
    poly_gb = np.polyfit(gb_non_nan['gb_polarity'], gb_non_nan['gb_subjectivity'], degree)
    poly_gb_func = np.poly1d(poly_gb)

    x_gb_poly = np.linspace(gb_non_nan['gb_polarity'].min(), gb_non_nan['gb_polarity'].max(), 100)
    y_gb_poly = poly_gb_func(x_gb_poly)

    y_gb_pred = poly_gb_func(gb_non_nan['gb_polarity'])
    r2_gb = r2_score(gb_non_nan['gb_subjectivity'], y_gb_pred)

    plt.plot(x_gb_poly, y_gb_poly, color='red', linestyle='--', label=f'Google Books Polynomial (R²={r2_gb:.2f})')

plt.title('Polarity vs. Subjectivity (Base vs Google Books) with Polynomial Regression')
plt.xlabel('Polarity')
plt.ylabel('Subjectivity')
plt.legend()
plt.grid(True)

plt.show()
```

```
Enter the degree of polynomial regression (positive integers) >>  3
```

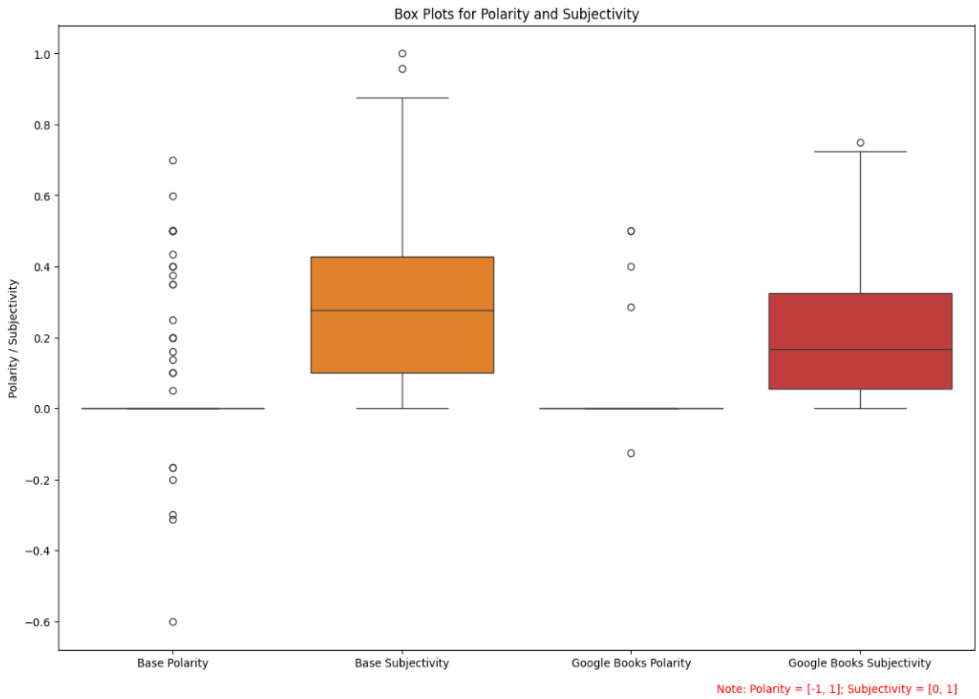


In here, we can conclude the following insights:

1. Regression model may not be suitable to represent the relationship between polarity and subjectivity.
2. In other word, there is no polynomial relationship between polarity and subjectivity.
3. The most suitable polynomial degree is 3.
4. As the polynomial degree get larger than 3, i.e. $n > 3, \forall n \in \mathbb{Z}$, it will cause the problem of invalid prediction. (The value predicted will exceed the acceptable range of subjectivity ($0 \leqslant$ subjectivity $\leqslant 1$))

Polarity VS Subjectivity ***with box plots***

[47]:
```
plt.figure(figsize=(15, 10))
sns.boxplot(data=df)

plt.title('Box Plots for Polarity and Subjectivity')
plt.ylabel('Polarity / Subjectivity')
plt.xticks([0, 1, 2, 3], ['Base Polarity', 'Base Subjectivity', 'Google Books Polarity', 'Google Books Subjectivity'])
plt.text(2.9, -0.8, 'Note: Polarity = [-1, 1]; Subjectivity = [0, 1]', fontsize=10, ha='center', color='red')

plt.show()
```



[32]:
```
spark.stop()
```

4.5.2 statistics_calc.py

```
import scipy.stats

def mean(l: list):
    return sum(l) / len(l)

def var(l: list):
    ssx = sum([x ** 2 for x in l])
    sx = sum(l)
    n = len(l)

    return (ssx - sx ** 2/n)/(n-1)

def stddev(l: list):
    return var(l) ** 0.5

def stderr(l: list):
    return stddev(l) / (len(l)**0.5)

def val_range(l: list):
    return (max(l) - min(l))

def cor(l: list):
    return val_range(l)/(max(l) + min(l))

def conf_int(l: list, alpha: float):
    if len(l) >= 30:
        crit_v = scipy.stats.norm.ppf(1 - alpha)
        std_err = stderr(l)
    else:
        crit_v = scipy.stats.t.ppf(1 - alpha, df=len(l) - 1)
        std_err = stderr(l)

    return (round(float(mean(l) - crit_v * std_err), 4), round(float(mean(l)
             + crit_v * std_err), 4))
```

4.5.3 metrics_calc.py

```
import pandas as pd
from pyspark.sql import DataFrame
from rapidfuzz import fuzz
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

def merge_dfs(df_base, df_com):
    if isinstance(df_base, DataFrame):
        df_base = df_base.toPandas()

    if isinstance(df_com, DataFrame):
        df_com = df_com.toPandas()

    merged_df = pd.merge(df_base, df_com, left_on=df_base.columns[0],
       right_on=df_com.columns[0], how='left', suffixes=('_base', '_com'))
    return merged_df, df_base, df_com

def coverage(df_base, df_com):
    merged_df, df_base, df_com = merge_dfs(df_base, df_com)

    total_base_words = len(df_base)
    overlapping_words = merged_df['definitions_com'].notna().sum()
    return overlapping_words / total_base_words

def annotation_accuracy_exact(df_base, df_com):
    merged_df, df_base, df_com = merge_dfs(df_base, df_com)

    merged_df['exact_match'] = merged_df.apply(
        lambda row: 1 if row['definitions_base'] == row['definitions_com']
         else 0, axis=1
    )

    exact_accuracy = merged_df['exact_match'].mean()
    return exact_accuracy

def similarity_levenshtein(df_base, df_com):
    merged_df, df_base, df_com = merge_dfs(df_base, df_com)

    merged_df['similarity'] = merged_df.apply(
        lambda row: fuzz.ratio(str(row['definitions_base']),
         str(row['definitions_com'])), axis=1
    )

    average_similarity = merged_df['similarity'].mean()
    return average_similarity
```

5. Lexicon Maintenance and Update (if any)

5.1. List of lexicon maintenance and update tasks carried out

5.1.1 Data Scraping

With one randomly selected word from our DBP data, we scrape the description of a book, from Google Books, that its title contains the randomly drawn word. The selection of books is iterative until a non-null description is scrapped from Google Books. Once found, the data is cleansed by removing non-alphabetical characters such as punctuations, spaces, etc., followed by tokenization.

The tokenized data is then explored and produced with length of word, part of speech, definitions, example sentences, derived words and synonyms, and temporarily stored as a dataframe.

5.1.2 Data Cleansing

Followed after simple attributes capturing, data cleansing is performed to remove vocabulary with no definitions, drop duplicate records, and standardize data.

5.1.3 Data Enrichment

After data cleansing, data is enriched with a few new columns, namely antonyms, frequency of usage, polarity, subjectivity, and subjectivity scores.

5.1.4 Data Loading

Lastly, the enriched data is stored into two different storages, namely HDFS and MongoDB. In HDFS, the dataframe is stored into a CSV file named with the prefix of gb_data_ followed by a 4-digit sequence number. Similarly in MongoDB, a new collection named the same as the CSV file in HDFS is created to store the data inside the collection in document form.

5.2. List of Python classes and Jupyter notebook(s) for lexicon maintenance and update

Name of Python class(es) / notebooks	Author
lexicon_maintenance.ipynb	Lau Kit Lim
scrape_google_books.py	Lau Kit Lim

5.3. Code for Python Classes

5.4.1 lexicon_maintenance.ipynb

Lexicon Maintenance

Read Data from Enriched Lexicons

[1]:
```
from python_script.scrape_google_books import *
from pyspark.sql import SparkSession

spark = SparkSession.builder.getOrCreate()

df = spark.read.csv('lexicon_data/enriched_lexicon.csv', header=True)
df.show()
```

```
24/12/15 18:23:23 WARN Utils: Your hostname, JORDAN. resolves to a loopback address: 127.0.1.1; using 10.255.255.254 instead (on interface lo)
24/12/15 18:23:23 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/12/15 18:23:23 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

vocab	word_length	part_of_speech	definitions	sentences_eg	derived_words	synonyms	antonyms	freq_of_usage	polarity	subjectivity	subjectivity_score
berniat	7	kata nama	1. bertujuan (aka...	1. aku pun tidak ...	berniat;berniat-n...	maksud;tujuan;ing...	NULL	0	none	fact-driven	0.0
mundur	6	adjektif	1. (berjalan, ber...	1. dia mundur set...	memundurkan;kemun...	merosot;meleset;m...	maju;maju	0	none	partially fact-dr...	0.3008333333333334
kaum	4	kata nama	1. puak, suku ban...	1. banyak kaum or...	berkaum;perkauman	suku;kabilah;taef...	NULL	5	none	fact-driven	0.1
wang	4	kata nama	1. alat pertukara...	1. satuan wang ne...	berwang;mengewangi	duit;fulus;pitis;...	NULL	6	none	partially opinion...	0.5777777777777777
jelas	5	adjektif	1. terang (perkat...	1. suara itu tida...	berjelas-jelas;me...	terang;nyata;keta...	kabur	0	positive	fact-driven	0.1916666666666665
waktu	5	kata nama	1. rangkaian atau...	1. secara rohania...	sewaktu;sewaktu-w...	masa;jangka masa;...	NULL	0	none	partially fact-dr...	0.3712962962962963
ramai	5	adjektif	1. riuh-rendah, h...	1. orang menyabun...	seramai;beramai-r...	bising;riuh;hiruk...	sunyi;sugul;lemba...	1	positive	partially fact-dr...	0.4916666666666664
kuat	4	adjektif	1. banyak tenagan...	1. walaupun ia su...	sekuat-kuatnya;be...	gagah;berdaya;ber...	lemah;rapuh;lemah...	0	positive	partially fact-dr...	0.4897530864197531
sahaja	6	kata tugas	1. yg satu-satuny...	1. tuan sahaja yg...	bersahaja;bersaha...	satu-satunya;cuma...	jarang	13	none	partially fact-dr...	0.4583333333333337
kerja	5	kata nama	1. usaha (kegiata...	1. beberapa mingg...	sekerja;bekerja;m...	khidmat;pekerjaan...	NULL	0	none	partially fact-dr...	0.2776936026936266
rakyat	6	kata nama	1. seluruh pendud...	1. rakyat Malaysi...	kerakyatan	warganegara;pendu...	pemerintah;ketua	25	none	partially fact-dr...	0.26
penuh	5	adjektif	1. seluruhnya ber...	1. tong itu penuh...	sepenuhnya;memenu...	berisi;tepu;padat...	kosong;sedikit;ku...	4	positive	partially opinion...	0.55
merupakan	9	kata nama	1. membentuk (mem...	1. peti rahsia it...	rupanya;rupa-rupa...	paras;muka;wajah;...	NULL	3	none	opinion-driven	0.75
penjelasan	10	adjektif	1. keterangan (yg...	1. Ali meminta pe...	berjelas-jelas;me...	terang;nyata;keta...	kabur	0	none	fact-driven	0.0
niat	4	kata nama	1. maksud atau tu...	1. mematikan oran...	berniat;berniat-n...	maksud;tujuan;ing...	NULL	0	none	partially fact-dr...	0.2857142857142857
bermaksud	9	kata nama	1. bertujuan, ber...	1. kami bermaksud...	bermaksud;memaksu...	tujuan;hajat;hasr...	NULL	1	none	opinion-driven	0.84375
tahu	4	kata kerja	1. maklum akan ke...	1. entah kapal pe...	tahu-tahu;mengeta...	maklum;faham;meng...	jahil;membiarkan	0	none	fact-driven	0.1857142857148572
memaksudkan	11	kata nama	1. menghendaki, m...	1. dialah yg dima...	bermaksud;memaksu...	tujuan;hajat;hasr...	NULL	0	negative	fact-driven	0.175
masa	4	kata nama	1. waktu, ketika;...	1. masa dan tempa...	semasa	waktu;ketika;deti...	NULL	3	none	fact-driven	0.08333333333333333
jauh	4	adjektif	1. besar jaraknya...	1. rumahnya agak ...	berjauh;menjauhi;...	tidak dekat;besar...	dekat	2	positive	partially opinion...	0.6583333333333334

only showing top 20 rows

Randomly Select ONE (1) Word from Enriched Lexicon

[10]:
```
import random

df_rows = df.collect()

random_vocab = df_rows[random.randint(0, len(df_rows) - 1)][0]
print(random_vocab)

gb_data = scrapeWith(random_vocab)
```

```
perkataan
```

Scrape the Words in Google Books

[11]:
```
while gb_data == []:
    random_vocab = df_rows[random.randint(0, len(df_rows) - 1)][0]
    print('Reassigned Random Vocab:', random_vocab)

    gb_data = scrapeWith(random_vocab)

for i in gb_data:
    print(f'{"Book Title :": >20} {i[0]}')
    print(f'{"Author(s) :": >20} {i[1]}')
    print(f'{"Publisher :": >20} {i[2]}')
    print(f'{"Description :": >20} {i[3]}')
    print(f'{"Link :": >20} {i[4]}')
    print(f'{"Published Date :": >20} {i[5]}')
    print('-' * 180, end='\n'*2)
```

```
        Book Title : PERKATAAN ALQURAN MUDAH
         Author(s) : ['Syed Nasir Bin Omar']
         Publisher : Syed Nasir Bin Omar
       Description : Menuntut ilmu sangat ditekankan dalam Islam kerana kita disuruh berbuat amal dan ibadah berdasarkan pengetahuan dan bukan sekadar
ikut-ikutan sahaja. Al-Quran yang kita baca setiap hari akan lebih memberi manfaat kepada kita jika isi, maksud dan tujuannya dapat kita fahami dan pra
ktikkan dalam segala hal seharian kita. Hanya dalam mendalami ilmu kita tentang Al-Quran barulah kita akan dapat lebih menjiwai dan menjalani hidup ini
sebagai Muslim sejati. Buku ini diterbitkan khas untuk mengajak anda memulakan jelajah ilmu ke arah mendalami isi maksud Al-Quran dengan cara penyampai
an yang mudah dan senang diikuti. Insya Allah, setelah anda membaca, anda akan faham dan mengenali Al-Quran yang kita baca setiap hari. Amin
             Link : https://play.google.com/store/books/details?id=6YwgEAAAQBAJ&source=gbs_api
    Published Date : 2020-07-29
------------------------------------------------------------------------------------------------------------------------------------------------
------------------------------
```

Obtain, Cleanse, and Tokenize the Description

description is selected as the one to be tokenized because it has higher probability to be written in Malay language.

[12]:
```
desc = gb_data[0][3]

import re
import string

desc = desc.replace('\xad', ' ').replace('\xa0', ' ').replace('/', ' ').lower()
desc = desc.translate(str.maketrans('', '', string.punctuation))
desc = re.sub(r"\d+", "", desc)

desc = desc.split(' ')
while '' in desc:
    desc.remove('')

print(desc)
```

```
['menuntut', 'ilmu', 'sangat', 'ditekankan', 'dalam', 'islam', 'kerana', 'kita', 'disuruh', 'berbuat', 'amal', 'dan', 'ibadah', 'berdasarkan', 'pengeta
huan', 'dan', 'bukan', 'sekadar', 'ikutikutan', 'sahaja', 'alquran', 'yang', 'kita', 'baca', 'setiap', 'hari', 'akan', 'lebih', 'memberi', 'manfaat',
'kepada', 'kita', 'jika', 'isi', 'maksud', 'dan', 'tujuannya', 'dapat', 'kita', 'fahami', 'dan', 'praktikkan', 'dalam', 'segala', 'hal', 'seharian', 'k
ita', 'hanya', 'dalam', 'mendalami', 'ilmu', 'kita', 'tentang', 'alquran', 'barulah', 'kita', 'akan', 'dapat', 'lebih', 'menjiwai', 'dan', 'menjalani',
'hidup', 'ini', 'sebagai', 'muslim', 'sejati', 'buku', 'ini', 'diterbitkan', 'khas', 'untuk', 'mengajak', 'anda', 'memulakan', 'jelajah', 'ilmu', 'ke',
'arah', 'mendalami', 'isi', 'maksud', 'alquran', 'dengan', 'cara', 'penyampaian', 'yang', 'mudah', 'dan', 'senang', 'diikuti', 'insya', 'allah', 'setel
ah', 'anda', 'membaca', 'anda', 'akan', 'faham', 'dan', 'mengenali', 'alquran', 'yang', 'kita', 'baca', 'setiap', 'hari', 'amin']
```

Extract the Attributes from each Tokenized Word

[13]:
```
from python_script.web_scrapping import *
lexicons = []

for k in desc:
    w = Word(k)
    w.extractLexicon()

    lexicons.append(w.getLexicon())

print(lexicons)
```

Create a Temporary DataFrame for the Lexicon

[14]:
```
columns = ['vocab', 'word_length', 'part_of_speech', 'definitions', 'sentences_eg', 'synonyms', 'derived_words']
inc_df = spark.createDataFrame(lexicons, columns)
```

[15]:
```
from pyspark.sql.functions import col, concat_ws

for cols in ['definitions', 'sentences_eg', 'synonyms', 'derived_words']:
    inc_df = inc_df.withColumn(cols, concat_ws(";", col(cols)))

inc_df.show()
```

```
+-----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|      vocab|word_length|part_of_speech|         definitions|        sentences_eg|            synonyms|       derived_words|
+-----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|   menuntut|          8|              |1. meminta dgn me...|1. pekerja-pekerj...|                    |                    |
|       ilmu|          4|    kata nama,|1. pengetahuan dl...|1. ilmu bumi;2. i...|                    |    berilmu;keilmuan|
|     sangat|          6|              |1. terlalu, terla...|1. bilangan murid...|                    |                    |
| ditekankan|         10|              |                    |                    |                    |                    |
|      dalam|          5|              |1. jarak (jauhnya...|1. dalam perigi i...|                    |                    |
|      islam|          5|              |1. memasukkan ses...|1. orang Perlak i...|                    |                    |
|     kerana|          6|    kata tugas|                    |                    |                    |                    |
|       kita|          4|     kata nama|1 saya dan kamu s...|1. Mari kita perg...|                    |            kekitaan|
|    disuruh|          7|              |                    |                    |                    |                    |
|    berbuat|          7|              |1. mengerjakan (m...|1. berbuat baik (...|                    |                    |
|       amal|          4|              |1. apa sahaja yg ...|1. banyak amal, s...|                    |                    |
|        dan|          3|              |                    |                    |                    |                    |
|     ibadah|          6|              |                    |                    |                    |                    |
|berdasarkan|         11|              |                    |                    |                    |                    |
|pengetahuan|         11|    kata kerja|1. perihal menget...|1. kisah serong y...|maklum;faham;meng...|tahu-tahu;mengeta...|
|        dan|          3|              |                    |                    |                    |                    |
|      bukan|          5|              |1 berlain drpd yg...|1. Dia bukan saud...|                    |                    |
|    sekadar|          7|     kata nama|1. selaras dgn (k...|1. sekadar yg dip...|daya;kemampuan;ke...|             sekadar|
| ikutikutan|         10|              |                    |                    |                    |                    |
|      sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|satu-satunya;cuma...|bersahaja;bersaha...|
+-----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Data Cleansing

The following actions are performed in this stage.

1. drop duplicated values
2. limit the value accepted in `part_of_speech`
3. replace empty string entries with `NULL`
4. keep those words which are derived words of other

[16]:
```
from pyspark.sql.functions import *

inc_df = inc_df.dropDuplicates()

inc_df = inc_df.withColumn(
    'part_of_speech',
    regexp_replace(col('part_of_speech'), "[^\w\s]", "")
)

allowed_pos = ['kata nama', 'kata tugas', 'kata kerja', 'adjektif', 'adverba']

# remove the data that is not considered part of speech
inc_df = inc_df.withColumn(
    'part_of_speech',
    when(col('part_of_speech').isin(allowed_pos), col('part_of_speech')).otherwise(None)
)
```

[17]:
```
for row in ['definitions', 'sentences_eg', 'synonyms', 'derived_words']:
    inc_df = inc_df.withColumn(
        row,
        when(col(row) == '', None).otherwise(col(row))
    )
```

[18]:
```
# check the records with no definition scrapped
inc_df = inc_df.filter(col('definitions').isNotNull())

inc_df.show()
print("Number of Unique Record with No Definition:", inc_df.count())
```

```
+-----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|      vocab|word_length|part_of_speech|         definitions|        sentences_eg|            synonyms|       derived_words|
+-----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
|       kita|          4|     kata nama|1 saya dan kamu s...|1. Mari kita perg...|                NULL|            kekitaan|
|      dalam|          5|          NULL|1. jarak (jauhnya...|1. dalam perigi i...|                NULL|                NULL|
|     sangat|          6|          NULL|1. terlalu, terla...|1. bilangan murid...|                NULL|                NULL|
|       ilmu|          4|     kata nama|1. pengetahuan dl...|1. ilmu bumi;2. i...|                NULL|    berilmu;keilmuan|
|   menuntut|          8|          NULL|1. meminta dgn me...|1. pekerja-pekerj...|                NULL|                NULL|
|      islam|          5|          NULL|1. memasukkan ses...|1. orang Perlak i...|                NULL|                NULL|
|pengetahuan|         11|    kata kerja|1. perihal menget...|1. kisah serong y...|maklum;faham;meng...|tahu-tahu;mengeta...|
|      bukan|          5|          NULL|1 berlain drpd yg...|1. Dia bukan saud...|                NULL|                NULL|
|    sekadar|          7|     kata nama|1. selaras dgn (k...|1. sekadar yg dip...|daya;kemampuan;ke...|             sekadar|
|    berbuat|          7|          NULL|1. mengerjakan (m...|1. berbuat baik (...|                NULL|                NULL|
|       amal|          4|          NULL|1. apa sahaja yg ...|1. banyak amal, s...|                NULL|                NULL|
|     sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|satu-satunya;cuma...|bersahaja;bersaha...|
|       baca|          4|          NULL|1. memerhatikan i...|1. baca akhbar (b...|                NULL|                NULL|
|       yang|          4|          NULL|1. (Lin) kata pen...|1. rumah yang bes...|                NULL|                NULL|
|       hari|          4|          NULL|1 masa dua puluh ...|1. Dlm seminggu, ...|                NULL|                NULL|
|     setiap|          6|          NULL|1. satu-satu drpd...|1. setiap orang y...|                NULL|                NULL|
|        isi|          3|     kata nama|1. sesuatu yg ter...|1. apakah isi bak...|kandungan;muatan;...|seisi;berisi;meng...|
|     kepada|          6|    kata tugas|1. kata utk menya...|1. mengirim surat...|                NULL|                NULL|
|     maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|    manfaat|          7|     kata nama|1. guna, faedah; ...|1. apa yg di-laku...|                NULL|bermanfaat;memanf...|
+-----------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows

Number of Unique Record with No Definition: 42
```

[19]:
```
# Split the DataFrame into two parts: with definition and without definition
null_df = inc_df.filter(col("definitions").isNull())
not_null_df = inc_df.filter(col("definitions").isNotNull())

# If vocab is a derived_words of other vocab even if it does not have definition, keep.
# Else (no definition and is not derived_words of others), delete.
matched_df = null_df.alias("null_df").join(
    not_null_df.alias("not_null_df"),
    expr("instr(null_df.derived_words, not_null_df.vocab) > 0"),
    "inner"
).select("null_df.*")

inc_df = not_null_df.union(matched_df)
inc_df.show()
print("Number of Records after Definitions Cleansing:", inc_df.count())
```

```
+----------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+
|      vocab|word_length|part_of_speech|         definitions|        sentences_eg|            synonyms|       derived_words|
+----------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+
|       kita|          4|     kata nama|1 saya dan kamu s...|1. Mari kita perg...|                NULL|            kekitaan|
|      dalam|          5|         NULL|1. jarak (jauhnya...|1. dalam perigi i...|                NULL|                NULL|
|     sangat|          6|         NULL|1. terlalu, terla...|1. bilangan murid...|                NULL|                NULL|
|       ilmu|          4|     kata nama|1. pengetahuan dl...|1. ilmu bumi;2. i...|                NULL|    berilmu;keilmuan|
|   menuntut|          8|         NULL|1. meminta dgn me...|1. pekerja-pekerj...|                NULL|                NULL|
|      islam|          5|         NULL|1. memasukkan ses...|1. orang Perlak i...|                NULL|                NULL|
|pengetahuan|         11|    kata kerja|1. perihal menget...|1. kisah serong y...|maklum;faham;meng...|tahu-tahu;mengeta...|
|      bukan|          5|         NULL|1 berlain drpd yg...|1. Dia bukan saud...|                NULL|                NULL|
|    sekadar|          7|     kata nama|1. selaras dgn (k...|1. sekadar yg dip...|daya;kemampuan;ke...|            sekadar|
|    berbuat|          7|         NULL|1. mengerjakan (m...|1. berbuat baik (...|                NULL|                NULL|
|       amal|          4|         NULL|1. apa sahaja yg ...|1. banyak amal, s...|                NULL|                NULL|
|     sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|satu-satunya;cuma...|bersahaja;bersaha...|
|       baca|          4|         NULL|1. memerhatikan i...|1. baca akhbar (b...|                NULL|                NULL|
|       yang|          4|         NULL|1. (Lin) kata pen...|1. rumah yang bes...|                NULL|                NULL|
|       hari|          4|         NULL|1 masa dua puluh ...|1. Dlm seminggu, ...|                NULL|                NULL|
|     setiap|          6|         NULL|1. satu-satu drpd...|1. setiap orang y...|                NULL|                NULL|
|        isi|          3|     kata nama|1. sesuatu yg ter...|1. apakah isi bak...|kandungan;muatan;...|seisi;berisi;meng...|
|     kepada|          6|    kata tugas|1. kata utk menya...|1. mengirim surat...|                NULL|                NULL|
|     maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|tujuan;hajat;hasr...|bermaksud;memaksu...|
|    manfaat|          7|     kata nama|1. guna, faedah; ...|1. apa yg di-laku...|                NULL|bermanfaat;memanf...|
+----------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows

Number of Records after Definitions Cleansing: 42
```

Data Enrichment

The following actions are performed in this stage:

1. append antonym (if any) for each word
2. append frequency of usage from 5 Malay essays
3. append polarity and subjectivity mark for each word

[20]:
```
from python_script.antonym_finder import findAntonym
```

[21]:
```
inc_df = inc_df.select('vocab', 'word_length', 'part_of_speech', 'definitions', 'sentences_eg', 'derived_words', 'synonyms')

append_antonym_udf = udf(findAntonym, StringType())

inc_df = inc_df.withColumn(
    'antonyms', append_antonym_udf(col('vocab'))
)

inc_df = inc_df.withColumn(
    'antonyms',
    when(col('antonyms') == '', None).otherwise(col('antonyms'))
)

inc_df.show()
```

```
[Stage 28:>                                                         (0 + 1) / 1]
+----------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+---------------+
|      vocab|word_length|part_of_speech|         definitions|        sentences_eg|       derived_words|            synonyms|       antonyms|
+----------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+---------------+
|       kita|          4|     kata nama|1 saya dan kamu s...|1. Mari kita perg...|            kekitaan|                NULL|           NULL|
|      dalam|          5|         NULL|1. jarak (jauhnya...|1. dalam perigi i...|                NULL|                NULL|           NULL|
|     sangat|          6|         NULL|1. terlalu, terla...|1. bilangan murid...|                NULL|                NULL|           NULL|
|       ilmu|          4|     kata nama|1. pengetahuan dl...|1. ilmu bumi;2. i...|    berilmu;keilmuan|                NULL|           NULL|
|   menuntut|          8|         NULL|1. meminta dgn me...|1. pekerja-pekerj...|                NULL|                NULL|           NULL|
|      islam|          5|         NULL|1. memasukkan ses...|1. orang Perlak i...|                NULL|                NULL|           NULL|
|pengetahuan|         11|    kata kerja|1. perihal menget...|1. kisah serong y...|tahu-tahu;mengeta...|maklum;faham;meng...|jahil;membiarkan|
|      bukan|          5|         NULL|1 berlain drpd yg...|1. Dia bukan saud...|                NULL|                NULL|           NULL|
|    sekadar|          7|     kata nama|1. selaras dgn (k...|1. sekadar yg dip...|            sekadar|daya;kemampuan;ke...|           NULL|
|    berbuat|          7|         NULL|1. mengerjakan (m...|1. berbuat baik (...|                NULL|                NULL|           NULL|
|       amal|          4|         NULL|1. apa sahaja yg ...|1. banyak amal, s...|                NULL|                NULL|           NULL|
|     sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|bersahaja;bersaha...|satu-satunya;cuma...|         jarang|
|       baca|          4|         NULL|1. memerhatikan i...|1. baca akhbar (b...|                NULL|                NULL|           NULL|
|       yang|          4|         NULL|1. (Lin) kata pen...|1. rumah yang bes...|                NULL|                NULL|           NULL|
|       hari|          4|         NULL|1 masa dua puluh ...|1. Dlm seminggu, ...|                NULL|                NULL|           NULL|
|     setiap|          6|         NULL|1. satu-satu drpd...|1. setiap orang y...|                NULL|                NULL|           NULL|
|        isi|          3|     kata nama|1. sesuatu yg ter...|1. apakah isi bak...|seisi;berisi;meng...|kandungan;muatan;...|           NULL|
|     kepada|          6|    kata tugas|1. kata utk menya...|1. mengirim surat...|                NULL|                NULL|           NULL|
|     maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|bermaksud;memaksu...|tujuan;hajat;hasr...|           NULL|
|    manfaat|          7|     kata nama|1. guna, faedah; ...|1. apa yg di-laku...|bermanfaat;memanf...|                NULL|           NULL|
+----------+-----------+-------------+--------------------+--------------------+--------------------+--------------------+---------------+
only showing top 20 rows
```

[22]:
```
import requests
from bs4 import BeautifulSoup as bs
import string

urls = [
    "https://infopelajar.com.my/karangan-buli-siber-kesan-kepada-emosi/",
    "https://infopelajar.com.my/karangan-jenayah-siber/",
    "https://infopelajar.com.my/karangan-rumah-terbuka-di-malaysia/",
    "https://infopelajar.com.my/karangan-rasuah-ancaman-kepada-negara/",
    "https://infopelajar.com.my/contoh-karangan-pidato/"
]

url_contents = []
for url in urls:
    r = requests.get(url)
    soup = bs(r.content, 'html.parser')

    entry_content = soup.find('div', class_='entry-content')
    entry_content.find('div', class_='cta-box').decompose()
    essay_str = entry_content.get_text().replace('\n', ' ')
    essay_str = essay_str.translate(str.maketrans('', '', string.punctuation))

    url_contents.append(essay_str)

full_content_str = ' '.join(url_contents)
words_in_content = full_content_str.lower().split(' ')

while '' in words_in_content:
    words_in_content.remove('')
```

[23]:
```
df_rows = inc_df.collect()

freq_of_usage = []

for row in df_rows:
    count = 0
    vocab = row[0]

    for item in words_in_content:
        if item == vocab:
            count = count + 1

    freq_of_usage.append([vocab, count])
```

[24]:
```
freq_df = spark.createDataFrame(freq_of_usage, ["vocab", "freq_of_usage"])

inc_df = inc_df.join(freq_df, on='vocab')
inc_df.show()
```

```
[Stage 35:>                                                 (0 + 1) / 1]
+----------+-----------+---------------+--------------------+--------------------+--------------------+--------------------+--------------+------------+
|     vocab|word_length|part_of_speech|         definitions|        sentences_eg|       derived_words|            synonyms|      antonyms|freq_of_usage|
+----------+-----------+---------------+--------------------+--------------------+--------------------+--------------------+--------------+------------+
|      kita|          4|     kata nama|1 saya dan kamu s...|1. Mari kita perg...|            kekitaan|                NULL|          NULL|          30|
|     dalam|          5|           NULL|1. jarak (jauhnya...|1. dalam perigi i...|                NULL|                NULL|          NULL|          56|
|    sangat|          6|           NULL|1. terlalu, terla...|1. bilangan murid...|                NULL|                NULL|          NULL|           7|
|      ilmu|          4|     kata nama|1. pengetahuan dl...|1. ilmu bumi;2. i...|    berilmu;keilmuan|                NULL|          NULL|           6|
|  menuntut|          8|           NULL|1. meminta dgn me...|1. pekerja-pekerj...|                NULL|                NULL|          NULL|           0|
|     islam|          5|           NULL|1. memasukkan ses...|1. orang Perlak i...|                NULL|                NULL|          NULL|           0|
|     bukan|          5|           NULL|1 berlain drpd yg...|1. Dia bukan saud...|                NULL|                NULL|          NULL|          13|
|pengetahuan|         11|    kata kerja|1. perihal menget...|1. kisah serong y...|tahu-tahu;mengeta...|maklum;faham;meng...|jahil;membiarkan|           3|
|    sekadar|          7|     kata nama|1. selaras dgn (k...|1. sekadar yg dip...|            sekadar|daya;kemampuan;ke...|          NULL|           0|
|      amal|          4|           NULL|1. apa sahaja yg ...|1. banyak amal, s...|                NULL|                NULL|          NULL|           0|
|   berbuat|          7|           NULL|1. mengerjakan (m...|1. berbuat baik (...|                NULL|                NULL|          NULL|           0|
|    sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|bersahaja;bersaha...|satu-satunya;cuma...|         jarang|          13|
|      yang|          4|           NULL|1. (Lin) kata pen...|1. rumah yang bes...|                NULL|                NULL|          NULL|          93|
|      hari|          4|           NULL|1 masa dua puluh ...|1. Dlm seminggu, ...|                NULL|                NULL|          NULL|           6|
|      baca|          4|           NULL|1. memerhatikan i...|1. baca akhbar (b...|                NULL|                NULL|          NULL|           0|
|     lebih|          5|           NULL|1 melampaui had y...|1. Panjangnya leb...|    lebih-lebih lagi;...|                NULL|         kurang|          16|
|       isi|          3|     kata nama|1. sesuatu yg ter...|1. apakah isi bak...|seisi;berisi;meng...|kandungan;muatan;...|          NULL|           0|
|   manfaat|          7|     kata nama|1. guna, faedah; ...|1. apa yg di-laku...|bermanfaat;memanf...|                NULL|          NULL|           4|
|    kepada|          6|    kata tugas|1. kata utk menya...|1. mengirim surat...|                NULL|                NULL|          NULL|          42|
|    setiap|          6|           NULL|1. satu-satu drpd...|1. setiap orang y...|                NULL|                NULL|          NULL|           2|
+----------+-----------+---------------+--------------------+--------------------+--------------------+--------------------+--------------+------------+
only showing top 20 rows
```

[25]:
```
from python_script.sentiment import *

import re

df_rows = inc_df.collect()

sentiments = []

for rows in df_rows:
    vocab = rows[0]
    sentences = rows[4]

    trans_vocab = translateFromMyToEn(vocab)

    polarity = getPolarityOf(trans_vocab)
    if sentences != None:
        trans_sentences = []
        clean_sentences = re.sub('\d.', '', sentences)
        splitted_sentences = clean_sentences.split(';')

        for sente in splitted_sentences:
            trans_sentences.append(translateFromMyToEn(sente))

        total_subjectivity = 0
        sentences_count = 0
        for sente in trans_sentences:
            total_subjectivity = total_subjectivity + getSubjectivityValue(sente)
            sentences_count = sentences_count + 1

        subjectivity_score = total_subjectivity / sentences_count
        subjectivity = getSubjectivityFrom(subjectivity_score)
    else:
        subjectivity_score = getSubjectivityValue(trans_vocab)
        subjectivity = getSubjectivityOf(trans_vocab)

    sentiments.append([vocab, polarity, subjectivity, subjectivity_score])
```

[26]:
```
sentiment_df = spark.createDataFrame(sentiments, ["vocab", "polarity", "subjectivity", "subjectivity_score"])

inc_df = inc_df.join(sentiment_df, on='vocab')
inc_df = inc_df.dropDuplicates()
inc_df.show()
```

```
+----------+-----------+--------------+--------------------+--------------------+--------------+--------------------+--------------+-------+--------------------+--------------------+
|     vocab|word_length|part_of_speech|         definitions|        sentences_eg|  derived_words|            synonyms|      antonyms|freq_of_usage|polarity|        subjectivity| subjectivity_score|
+----------+-----------+--------------+--------------------+--------------------+--------------+--------------------+--------------+-------------+--------+--------------------+--------------------+
|pengetahuan|         11|    kata kerja|1. perihal menget...|1. kisah serong y...|tahu-tahu;mengeta...|maklum;faham;meng...|jahil;membiarkan|            3|    none|partially opinion...| 0.6638888888888889|
|    sahaja|          6|    kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|bersahaja;bersaha...|satu-satunya;cuma...|        jarang|           13|    none|partially fact-dr...|0.45833333333333337|
|    senang|          6|      adjektif|1. (serba) mudah,...|1. kerja mereka s...|bersenang;bersena...|mudah;gampang;rin...|susah;benci;sesak...|            0|positive|partially opinion...| 0.7183333333333334|
|      amal|          4|          NULL|1. apa sahaja yg ...|1. banyak amal, s...|          NULL|                NULL|          NULL|            0|    none|partially fact-dr...|0.45909090909090916|
|      amin|          4|          NULL|1. perkenankanlah...|1. doa itu diakhi...|          NULL|                NULL|          NULL|            0|    none|         fact-driven|                 0.0|
|      arah|          4|          NULL|1. hala, jurusan;...|1. ia mencari ara...|          NULL|                NULL|          NULL|            0|    none|         fact-driven| 0.19537037037037037|
|      baca|          4|          NULL|1. memerhatikan i...|1. baca akhbar (b...|          NULL|                NULL|          NULL|            0|    none|         fact-driven|0.08571428571428573|
|   berbuat|          7|          NULL|1. mengerjakan (m...|1. berbuat baik (...|          NULL|                NULL|          NULL|            0|    none|partially fact-dr...| 0.3611111111111111|
|     bukan|          5|          NULL|1 berlain drpd yg...|1. Dia bukan saud...|          NULL|                NULL|          NULL|           13|    none|partially opinion...| 0.5928571428571429|
|      buku|          4|          NULL|1 bahagian yg ker...|1. Ahmad menghanc...|          NULL|                NULL|          NULL|            0|    none|         fact-driven|0.03333333333333333|
|      cara|          4|          NULL|1 peraturan membu...|1. Bagaimanakah c...|          NULL|                NULL|          NULL|            1|    none|partially fact-dr...|0.33888888888888885|
|     dalam|          5|          NULL|1. jarak (jauhnya...|1. dalam perigi i...|          NULL|                NULL|          NULL|           56|    none|         fact-driven|                 0.1|
|     dapat|          5|          NULL|1 boleh (berupaya...|1. Saya dapat men...|          NULL|                NULL|          NULL|           13|    none|partially opinion...| 0.5548611111111111|
|    dengan|          6|          NULL|1. (ber)serta, be...|1. saya pergi men...|          NULL|                NULL|          NULL|           28|    none|         fact-driven|             0.05625|
|     faham|          5|          NULL|1. = fahaman peng...|1. fahamnya menge...|          NULL|                NULL|          NULL|            0|    none|         fact-driven|              0.1675|
|     hanya|          5|    kata tugas|1. tidak lebih dr...|1. kemanakan hamb...|          NULL|                NULL|          NULL|            4|    none|      opinion-driven| 0.8166666666666667|
|      hari|          4|          NULL|1 masa dua puluh ...|1. Dlm seminggu, ...|          NULL|                NULL|          NULL|            6|    none|partially fact-dr...| 0.4089143990929705|
|     hidup|          5|          NULL|1 bernyawa dan be...|1. Kedua-dua ibu ...|hidup-hidup;hidup...|                NULL|  mati;palsu;mati|            4|positive|partially opinion...| 0.5295153061224489|
|      ilmu|          4|     kata nama|1. pengetahuan di...|1. ilmu bumi;2. i...|    berilmu;keilmuan|                NULL|          NULL|            6|    none|partially fact-dr...|0.25297619047619047|
|       isi|          3|     kata nama|1. sesuatu yg ter...|1. apakah isi bak...|seisi;berisi;meng...|kandungan;muatan;...|          NULL|            0|    none|partially fact-dr...|           0.453125|
+----------+-----------+--------------+--------------------+--------------------+--------------+--------------------+--------------+-------------+--------+--------------------+--------------------+
only showing top 20 rows
```

Data Loading

Data are loaded into two different storages:

1. MongoDB
2. HDFS

[27]:
```
from pymongo import MongoClient
import pprint

uri = "mongodb+srv://jordan:jordan0551@cluster0.nzsyn.mongodb.net/?retryWrites=true&w=majority&appName=Cluster0"
client = MongoClient(uri)

db = client["lexicondb"]
collections = db.list_collection_names()
collection_name = f'gb_data_{len(collections):04d}'
```

[28]:
```
current_collection = db[collection_name]
```

[29]:
```
df_rows = inc_df.collect()

data_list = []
for i in range(len(df_rows)):
    data_dict = {
        'seq_no'            : f'{i+1:02d}',
        'vocab'             : df_rows[i][0],
        'word_length'       : df_rows[i][1],
        'part_of_speech'    : df_rows[i][2],
        'definitions'       : df_rows[i][3],
        'sentences_eg'      : df_rows[i][4],
        'derived_words'     : df_rows[i][5],
        'synonyms'          : df_rows[i][6],
        'antonyms'          : df_rows[i][7],
        'freq_of_usage'     : df_rows[i][8],
        'polarity'          : df_rows[i][9],
        'subjectivity'      : df_rows[i][10],
        'subjectivity_score': df_rows[i][11]
    }

    if data_dict not in data_list:
        data_list.append(data_dict)

    if not current_collection.find_one({'vocab': data_dict['vocab']}):
        current_collection.insert_one(data_dict)
    else:
        print('Document with Vocab ' + data_dict['vocab'] + ' Existed!')
```

[30]:
```
from python_script.hdfs_commander import *

inc_df.coalesce(1).write.csv(collection_name, header=True, mode='overwrite')

run_hdfs_command(f'cd DE-Assignment')
run_hdfs_command(f'mkdir google_books_data')
run_hdfs_command(f'hdfs dfs -mkdir google_books_data')
run_hdfs_command(f'hdfs dfs -mv {collection_name}/part-00000-*.csv google_books_data/{collection_name}.csv')
run_hdfs_command(f'hdfs dfs -rm -R {collection_name}')
run_hdfs_command(f'hdfs dfs -get google_books_data/{collection_name}.csv google_books_data/{collection_name}.csv')
```

```
Error occurred: Command 'cd DE-Assignment' returned non-zero exit status 2.
Error occurred: Command 'mkdir google_books_data' returned non-zero exit status 1.
Error occurred: Command 'hdfs dfs -mkdir google_books_data' returned non-zero exit status 1.
Output: Deleted gb_data_0002
```

[31]:
```
spark.stop()
```

5.4.2 scrape_google_books.py

```
import requests
from datetime import datetime

def scrapeWith(query, max_result=1, year=None):
    if len(str(year)) != 4:
        raise ValueError('Please Provide year in 4 digits (e.g.: 2024)')

    url = "https://www.googleapis.com/books/v1/volumes"

    params = {
        'q': query,
        'langRestrict': 'my'
    }

    response = requests.get(url, params=params)

    book_info = []
    if response.status_code == 200:
        data = response.json()

        if 'items' in data:
            for i, book in enumerate(data['items']):
                # Stop scraping if we've reached the max_result
                if i >= max_result:
                    break

                description = book['volumeInfo']
                    .get('description', 'No description available')
                if description != 'No description available':
                    title = book['volumeInfo']
                           .get('title', 'No title available')
                    authors = book['volumeInfo']
                           .get('authors', ['No authors available'])
                    publisher = book['volumeInfo']
                           .get('publisher', 'No publisher available')
                    link = book['volumeInfo']
                           .get('infoLink', 'No link available')
                    published_date = book['volumeInfo']
                           .get('publishedDate', None)

                    if year:
                        if published_date and
                           published_date.startswith(str(year)):
                            book_info.append([title, authors, publisher,
                                  description, link, published_date])
                    else:
                        book_info.append([title, authors, publisher,
                                  description, link, published_date])

        else:
            print("No books found.")
    else:
        print("Failed to retrieve data:", response.status_code)

    return book_info
```

6. Data Streaming

6.1. Kafka Streaming

Use case	Real-time Natural Language Processing (NLP)
Kafka Topic	‘kata nama’, ‘kata kerja’, ‘kata tugas’, ‘adverba’, ‘adjektif’
End user(s)	Book authors, translators

Screenshot of message content:

On running producers_starter.sh:

```
jordan@JORDAN:~/DE-Assignment/shell_script$ ./producers_starter.sh
Streaming Topic: nouns >>>

Sent: Row 18: Vocab: ilmu, POS: kata nama
Sent: Row 19: Vocab: isi, POS: kata nama
Sent: Row 23: Vocab: kita, POS: kata nama
Sent: Row 25: Vocab: maksud, POS: kata nama
Sent: Row 26: Vocab: manfaat, POS: kata nama
Sent: Row 33: Vocab: muslim, POS: kata nama
Sent: Row 37: Vocab: sekadar, POS: kata nama
Sent: Row 3: Vocab: berisi, POS: kata nama
Sent: Row 12: Vocab: hukum, POS: kata nama
Sent: Row 14: Vocab: merupakan, POS: kata nama
Sent: Row 15: Vocab: milik, POS: kata nama
Sent: Row 19: Vocab: pokok, POS: kata nama
Sent: Row 21: Vocab: rumah, POS: kata nama
Data Topic nouns Streamed Successfully!


Streaming Topic: verbs >>>

Sent: Row 0: Vocab: pengetahuan, POS: kata kerja
Data Topic verbs Streamed Successfully!


Streaming Topic: functionWords >>>

Sent: Row 1: Vocab: sahaja, POS: kata tugas
Sent: Row 15: Vocab: hanya, POS: kata tugas
Sent: Row 22: Vocab: kepada, POS: kata tugas
Sent: Row 39: Vocab: tentang, POS: kata tugas
Sent: Row 40: Vocab: untuk, POS: kata tugas
Sent: Row 26: Vocab: tentang, POS: kata tugas
Data Topic functionWords Streamed Successfully!


Streaming Topic: adverbs >>>

Data Topic adverbs Streamed Successfully!


Streaming Topic: adjectives >>>

Sent: Row 2: Vocab: senang, POS: adjektif
Sent: Row 32: Vocab: mudah, POS: adjektif
Data Topic adjectives Streamed Successfully!
```

On running consumers_starter.sh:

```
jordan@JORDAN:~/DE-Assignment/shell_script$ ./consumers_starter.sh
ConsumerRecord(topic='nouns', partition=0, offset=0, timestamp=1734501235387, timestamp_type=0, key=None, value=b'Row 18: V
ocab: ilmu, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=35, serialized_header
_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=1, timestamp=1734501236390, timestamp_type=0, key=None, value=b'Row 19: V
ocab: isi, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=34, serialized_header_
size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=2, timestamp=1734501237391, timestamp_type=0, key=None, value=b'Row 23: V
ocab: kita, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=35, serialized_header
_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=3, timestamp=1734501238394, timestamp_type=0, key=None, value=b'Row 25: V
ocab: maksud, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=37, serialized_head
er_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=4, timestamp=1734501239396, timestamp_type=0, key=None, value=b'Row 26: V
ocab: manfaat, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=38, serialized_hea
der_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=5, timestamp=1734501240398, timestamp_type=0, key=None, value=b'Row 33: V
ocab: muslim, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=37, serialized_head
er_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=6, timestamp=1734501241401, timestamp_type=0, key=None, value=b'Row 37: V
ocab: sekadar, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=38, serialized_hea
der_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=7, timestamp=1734501242408, timestamp_type=0, key=None, value=b'Row 3: Vo
cab: berisi, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=36, serialized_heade
r_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=8, timestamp=1734501243410, timestamp_type=0, key=None, value=b'Row 12: V
ocab: hukum, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=36, serialized_heade
r_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=9, timestamp=1734501244413, timestamp_type=0, key=None, value=b'Row 14: V
ocab: merupakan, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=40, serialized_h
eader_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=10, timestamp=1734501245415, timestamp_type=0, key=None, value=b'Row 15:
Vocab: milik, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=36, serialized_head
er_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=11, timestamp=1734501246417, timestamp_type=0, key=None, value=b'Row 19:
Vocab: pokok, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=36, serialized_head
er_size=-1)
ConsumerRecord(topic='nouns', partition=0, offset=12, timestamp=1734501247420, timestamp_type=0, key=None, value=b'Row 21:
Vocab: rumah, POS: kata nama', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=36, serialized_head
er_size=-1)
Received shutdown signal. Terminating consumer...
Consumer nouns Terminated Successfully!


ConsumerRecord(topic='verbs', partition=0, offset=0, timestamp=1734501248856, timestamp_type=0, key=None, value=b'Row 0: Vo
cab: pengetahuan, POS: kata kerja', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=42, serialized
_header_size=-1)
Received shutdown signal. Terminating consumer...
Consumer verbs Terminated Successfully!
ConsumerRecord(topic='functionWords', partition=0, offset=0, timestamp=1734501250321, timestamp_type=0, key=None, value=b'R
ow 1: Vocab: sahaja, POS: kata tugas', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=37, seriali
zed_header_size=-1)
ConsumerRecord(topic='functionWords', partition=0, offset=1, timestamp=1734501251324, timestamp_type=0, key=None, value=b'R
ow 15: Vocab: hanya, POS: kata tugas', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=37, seriali
zed_header_size=-1)
ConsumerRecord(topic='functionWords', partition=0, offset=2, timestamp=1734501252326, timestamp_type=0, key=None, value=b'R
ow 22: Vocab: kepada, POS: kata tugas', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=38, serial
ized_header_size=-1)
ConsumerRecord(topic='functionWords', partition=0, offset=3, timestamp=1734501253328, timestamp_type=0, key=None, value=b'R
ow 39: Vocab: tentang, POS: kata tugas', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=39, seria
lized_header_size=-1)
ConsumerRecord(topic='functionWords', partition=0, offset=4, timestamp=1734501254331, timestamp_type=0, key=None, value=b'R
ow 40: Vocab: untuk, POS: kata tugas', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=37, seriali
zed_header_size=-1)
ConsumerRecord(topic='functionWords', partition=0, offset=5, timestamp=1734501255337, timestamp_type=0, key=None, value=b'R
ow 26: Vocab: tentang, POS: kata tugas', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=39, seria
lized_header_size=-1)
Received shutdown signal. Terminating consumer...
Consumer functionWords Terminated Successfully!


Received shutdown signal. Terminating consumer...
Consumer adverbs Terminated Successfully!


ConsumerRecord(topic='adjectives', partition=0, offset=0, timestamp=1734501257270, timestamp_type=0, key=None, value=b'Row
2: Vocab: senang, POS: adjektif', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=35, serialized_h
eader_size=-1)
ConsumerRecord(topic='adjectives', partition=0, offset=1, timestamp=1734501258272, timestamp_type=0, key=None, value=b'Row
32: Vocab: mudah, POS: adjektif', headers=[], checksum=None, serialized_key_size=-1, serialized_value_size=35, serialized_h
eader_size=-1)
Received shutdown signal. Terminating consumer...
Consumer adjectives Terminated Successfully!
```

List of Python classes for Kafka producer and consumer:

Name of Python class(es) / notebooks	Author
data_streaming.ipynb	Chiam Zi Wei
Consumer Python scripts (for all topics)	Lau Kit Lim
Producer Python scripts (for all topics)	Ang Yu Wen
consumers_starter.sh	Chiam Zi Wei
producers_starter.sh	Chiam Zi Wei

6.2. Python Code

6.2.1 data_streaming.ipynb

Data Streaming

Kafka Streaming

Topic Creation

```
from python_script.hdfs_commander import run_hdfs_command as hdfs

topics = ['nouns', 'verbs', 'functionWords', 'adverbs', 'adjectives']

for t in topics:
    hdfs(f'kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic {t}')
```

```
Output: Created topic nouns.

Output: Created topic verbs.

Output: Created topic functionWords.

Output: Created topic adverbs.

Output: Created topic adjectives.
```

Topic Listing and Describing

```
print("List of Topics:")
hdfs('kafka-topics.sh --list --bootstrap-server localhost:9092')

print('-' * 150)
print("Description of Topics:")
for t in topics:
    hdfs(f'kafka-topics.sh --describe --bootstrap-server localhost:9092 --topic {t}')
```

```
List of Topics:
Output: adjectives
adverbs
functionWords
nouns
verbs

------------------------------------------------------------------------------------------------------------------------------------------------------
Description of Topics:
Output: Topic: nouns    TopicId: nqOFl_57RqOeDxelyiM3Aw PartitionCount: 1       ReplicationFactor: 1    Configs:
        Topic: nouns    Partition: 0    Leader: 0       Replicas: 0     Isr: 0

Output: Topic: verbs    TopicId: QiBj1dgURTWd0xjFCL1ShA PartitionCount: 1       ReplicationFactor: 1    Configs:
        Topic: verbs    Partition: 0    Leader: 0       Replicas: 0     Isr: 0

Output: Topic: functionWords    TopicId: m_xnnO3bSsatqT0wnOwoGQ PartitionCount: 1       ReplicationFactor: 1    Configs:
        Topic: functionWords    Partition: 0    Leader: 0       Replicas: 0     Isr: 0

Output: Topic: adverbs  TopicId: Csco1UO1Q9S31DoM1MEQXw PartitionCount: 1       ReplicationFactor: 1    Configs:
        Topic: adverbs  Partition: 0    Leader: 0       Replicas: 0     Isr: 0

Output: Topic: adjectives       TopicId: vjCeEEO-QDqj2-UPxa-9ew PartitionCount: 1       ReplicationFactor: 1    Configs:
        Topic: adjectives       Partition: 0    Leader: 0       Replicas: 0     Isr: 0
```

Topic Deletion (run when necessary)

```
for t in topics:
    hdfs(f'kafka-topics.sh --delete --bootstrap-server localhost:9092 --topic {t}')
```

After topic creation, please proceed to the following steps:

1. open two different terminals.
2. change the directory to `~/DE-Assignment/shell_script` using `cd` command.
3. in one of the terminal, run the command `./consumers_starter.sh`.
4. in the another, run the command `./producers_starter.sh`.

6.2.2 consumer python script (e.g.: consumer_nouns.py)

```
from kafka import KafkaConsumer
import time
import socket

bootstrap_servers = "localhost:9092"
topic = 'nouns'

consumer = KafkaConsumer(topic,bootstrap_servers=bootstrap_servers)

for msg in consumer:
    if msg.value == b'SHUTDOWN':
        print("Received shutdown signal. Terminating consumer...")
        consumer.close()
        break

    print(msg)

consumer.close()
print(f"Consumer {topic} Terminated Successfully!\n\n")
```

6.2.3 producer python script (e.g.: producer_nouns.py)

```
from kafka import KafkaProducer
import time
import socket
import os
import pandas as pd

bootstrap_servers = "localhost:9092"
topic = 'nouns'
time_interval = 1

directory_path = "google_books_data"
all_files = os.listdir(directory_path)
file_names = [f for f in all_files if
      os.path.isfile(os.path.join(directory_path, f))]

producer = KafkaProducer(bootstrap_servers=bootstrap_servers)

print(f"Streaming Topic: {topic} >>> \n")
for file in file_names:
    df = pd.read_csv(directory_path + '/' + file)
    df = df[df['part_of_speech'] == 'kata nama']

    for idx, row in df.iterrows():
        message = f"Row {idx}: Vocab: {row['vocab']},
                    POS: {row['part_of_speech']}"
        producer.send(topic, value=message.encode('utf-8'))
        print(f"Sent: {message}")
        time.sleep(time_interval)

shutdown_signal = "SHUTDOWN"
producer.send(topic, value=shutdown_signal.encode('utf-8'))

producer.flush()
print(f"Data Topic {topic} Streamed Successfully!\n\n")
```

6.2.4 consumers_starter.sh

```
#!/bin/sh

cd ~/DE-Assignment

python streaming_script/consumer_nouns.py
echo -n
python streaming_script/consumer_verbs.py
echo -n
python streaming_script/consumer_function_words.py
echo -n
python streaming_script/consumer_adverbs.py
echo -n
python streaming_script/consumer_adjectives.py
echo -n
```

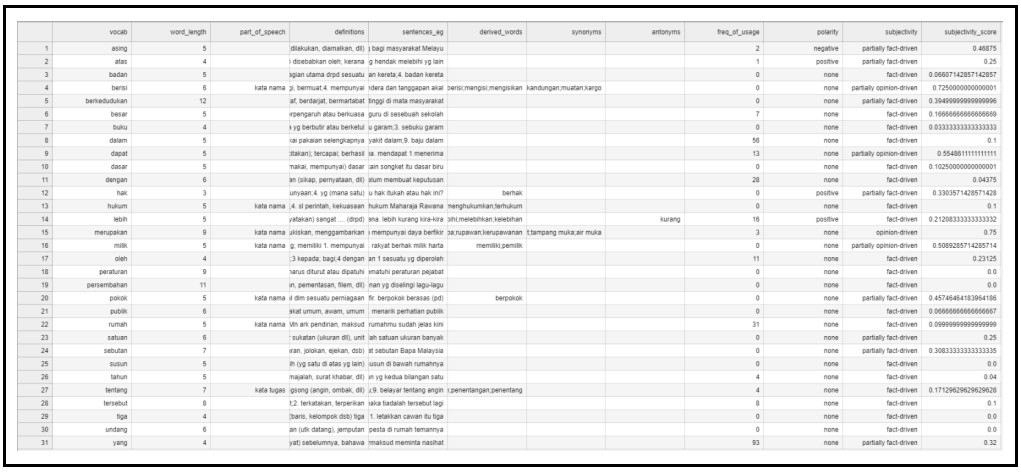
6.2.5 producers_starter.sh

```
#!/bin/sh

cd ~/DE-Assignment

python streaming_script/producer_nouns.py
echo -n
python streaming_script/producer_verbs.py
echo -n
python streaming_script/producer_function_words.py
echo -n
python streaming_script/producer_adverbs.py
echo -n
python streaming_script/producer_adjectives.py
echo -n
```

6.3. Spark Structured Streaming

Use case	Real-time Natural Language Processing (NLP)
Source	Socket Source
Sink	Memory Sink
End user(s)	Book authors, translators

Screenshot of aggregated data that will be streamed:



	vocab	word_length	part_of_speech	definitions	sentences_eg	derived_words	synonyms	antonyms	freq_of_usage	polarity	subjectivity	subjectivity_score
1	asing	5		dilakukan, diamalkan, dll)	bagi masyarakat Melayu				2	negative	partially fact-driven	0.46875
2	atas	4		disebabkan oleh; kerana	g hendak melebihi yg lain				1	positive	partially fact-driven	0.25
3	badan	5		agian utama dr pd sesuatu	an kereta;4. badan kereta				0	none	fact-driven	0.06607142857142857
4	berisi	6	kata nama	i, bermuat;4. mempunyai	ndera dan tanggapan akal	berisi;mengisi;mengisikan	kandungan;muatan;kargo		0	none	partially opinion-driven	0.7250000000000001
5	berkedudukan	12		af, berdarjat, bermartabat	tinggi di mata masyarakat				0	none	partially fact-driven	0.3949999999999996
6	besar	5		rpengaruh atau berkuasa	guru di sesebuah sekolah				7	none	fact-driven	0.16666666666666669
7	buku	4		yg berbutir atau berkelul	u garam;3. sebuku garam				0	none	fact-driven	0.03333333333333333
8	dalam	5		kai pakaian selengkapnya	yakit dalam;9. baju dalam				56	none	fact-driven	0.1
9	dapat	5		ditakan); tercapai; berhasil	sa. mendapat 1 menerima				13	none	partially opinion-driven	0.5548611111111111
10	dasar	5		makai, mempunyai) dasar	ain songket itu dasar biru				0	none	fact-driven	0.10250000000000001
11	dengan	6		an (sikap, pernyataan, dll)	alum membuat keputusan				28	none	fact-driven	0.04375
12	hak	3		unyaan;4. yg (mana satu)	u hak itukah atau hak ini?	berhak			0	positive	partially fact-driven	0.3303571428571428
13	hukum	5	kata nama	4. sl perintah, kekuasaan	hukum Maharaja Ravana	menghukumkan;terhukum			0	none	fact-driven	0.1
14	lebih	5		yatakan) sangat ... (dpd)	sna. lebih kurang kira-kira	ihi;melebihkan;kelebihan		kurang	16	positive	fact-driven	0.21208333333333332
15	merupakan	9	kata nama	ukiskan, menggambarkan	i mempunyai daya berfikir	sa;rupawan;kerupawanan	t;tampang muka;air muka		3	none	opinion-driven	0.75
16	milik	5	kata nama	g, memiliki 1. mempunyai	rakyat berhak milik harta	memiliki;pemilik			0	none	partially opinion-driven	0.5089285714285714
17	oleh	4		:3 kepada; bagi;4 dengan	an 1 sesuatu yg diperoleh				11	none	fact-driven	0.23125
18	peraturan	9		harus diturut atau dipatuhi	matuhi peraturan pejabat				0	none	fact-driven	0.0
19	persembahan	11		an, pementasan, filem, dll)	nan yg disellingi lagu-lagu				0	none	fact-driven	0.0
20	pokok	5	kata nama	d dim sesuatu perniagaan	fir. berpokok berasas (pd)	berpokok			0	none	partially fact-driven	0.45746464183964186
21	publik	6		akat umum, awam, umum	menarik perhatian publik				0	none	fact-driven	0.06666666666666667
22	rumah	5	kata nama	Vn ark pendirian, maksud	rumahmu sudah jelas kini				31	none	fact-driven	0.09999999999999999
23	satuan	6		sukatan (ukuran dll), unit	ah satuan ukuran banyak				0	none	partially fact-driven	0.25
24	sebutan	7		iran, jolokan, ejekan, dsb)	at sebutan Bapa Malaysia				0	none	partially fact-driven	0.30833333333333335
25	susun	5		ih (yg satu di atas yg lain)	usun di bawah rumahnya				0	none	fact-driven	0.0
26	tahun	5		majalah, surat khabar, dll)	in yg kedua bilangan satu				4	none	fact-driven	0.04
27	tentang	7	kata tugas	gsiong (angin, ombak, dll)	;9. belayar tentang angin	penentangan;penentang			4	none	fact-driven	0.17129629629629628
28	tersebut	8		t;2. terkatakan, terperikan	aka tiadalah tersebut lagi				8	none	fact-driven	0.1
29	tiga	4		bara, kelompok dsb) tiga	1. letakkan cawan itu tiga				0	none	fact-driven	0.0
30	undang	6		an (utk datang), jemputan	pesta di rumah temannya				0	none	fact-driven	0.0
31	yang	4		yat) sebelumnya, bahawa	maksud meminta nasihat				93	none	partially fact-driven	0.32

List of Python classes / Jupyter notebook(s) for Structured Streaming:

Name of Python class(es) / notebooks	Author
data_streaming.ipynb	Chiam Zi Wei

6.4. Python Code

6.4.1 data_streaming.ipynb

Spark Structured Streaming

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *

spark = SparkSession.builder.getOrCreate()
```

```
24/12/18 21:14:26 WARN Utils: Your hostname, JORDAN. resolves to a loopback address: 127.0.1.1; using 10.255.255.254 instead (on interface lo)
24/12/18 21:14:26 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
24/12/18 21:14:27 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

```
directory_path = "google_books_data/"

import os
if not os.path.isdir(directory_path):
    raise ValueError(f"The path {directory_path} is not a valid directory.")

csv_stream_df = spark.readStream \
    .format("csv") \
    .option("header", "true") \
    .option("inferSchema", "true") \
    .schema("vocab STRING, word_length INT, part_of_speech STRING, \
            definitions STRING, sentences_eg STRING, derived_words STRING, synonyms STRING, \
            antonyms STRING, freq_of_usage INT, polarity STRING, subjectivity STRING, subjectivity_score STRING") \
    .option("basePath", directory_path) \
    .load(directory_path)
```

```
print("Streaming Topic: kata nama")
filtered_df = csv_stream_df.filter(csv_stream_df['part_of_speech'] == 'kata nama')

query = filtered_df.writeStream \
    .outputMode("append") \
    .format("console") \
    .start()
```

```
Streaming Topic: kata nama
24/12/18 21:14:31 WARN ResolveWriteToStream: Temporary checkpoint location created which is deleted normally when the query didn't fail: /tmp/temporary-33efe804-7426-4925-93d1-9264b31db098. If it's required to delete it under any circ
umstances, please set spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know deleting temp checkpoint folder is best effort.
24/12/18 21:14:31 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
-------------------------------------------
Batch: 0
-------------------------------------------
+---------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+--------+-----------+--------+--------------------+--------------------+--------------------+
|    vocab|word_length|part_of_speech|         definitions|        sentences_eg|        derived_words|            synonyms|antonyms|freq_of_usage|polarity|        subjectivity| subjectivity_score|
+---------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+--------+-----------+--------+--------------------+--------------------+
|     ilmu|          4|     kata nama|1. pengetahuan dl...|1. ilmu bumi;2. i...|    berilmu;keilmuan|                NULL|    NULL|          6|    none|partially fact-dr...|0.25297619047619047|
|      isi|          3|     kata nama|1. sesuatu yg ter...|1. apakah isi bak...|seisi;berisi;meng...|kandungan;muatan;...|    NULL|          0|    none|partially fact-dr...|               0.453125|
|     kita|          4|     kata nama|1 saya dan kamu s...|1. Mari kita perg...|            kekitaan|                NULL|    NULL|         30|    none|        fact-driven|                 0.0|
|   maksud|          6|     kata nama|1. tujuan, hajat,...|1. maksud aku ke ...|bermaksud;memaksu...|tujuan;hajat;hasr...|    NULL|          0|    none|partially fact-dr...|              0.34375|
| manfaat|          7|     kata nama|1. guna, faedah; ...|1. apa yg di-laku...|bermanfaat;memanf...|                NULL|    NULL|          4|    none|        fact-driven|                 0.0|
|  muslim|          6|     kata nama|1. menjadikan ses...|1. serba-serbi me...|                NULL|                NULL|    NULL|          0|    none|        fact-driven|                 0.0|
| sekadar|          7|     kata nama|1. selaras dgn (k...|1. sekadar yg dip...|            sekadar|daya;kemampuan;ke...|    NULL|          0|    none|        fact-driven|                 0.0|
|  berisi|          6|     kata nama|1. mempunyai isi,...|1. bintit-bintit ...|seisi;berisi;meng...|kandungan;muatan;...|    NULL|          0|    none|partially opinion...| 0.7250000000000001|
|   hukum|          5|     kata nama|1. peraturan yg d...|1. hukum mengerja...|hukuman;menghukum...|                NULL|    NULL|          0|    none|        fact-driven|                 0.1|
|merupakan|          9|     kata nama|1. membentuk (mem...|1. peti rahsia it...|rupanya;rupa-rupa...|paras;muka;wajah;...|    NULL|          3|    none|     opinion-driven|                0.75|
|   milik|          5|     kata nama|1. kepunyaan, hak...|1. semua tanah ad...|   memiliki;pemilik|                NULL|    NULL|          0|    none|partially opinion...| 0.5089285714285714|
|   pokok|          5|     kata nama|1 segala jenis tu...|1. pokok pisang;2...|           berpokok|                NULL|    NULL|          0|    none|partially fact-dr...| 0.45746464183964186|
|    rumah|          5|     kata nama|1. binaan utk tem...|1. Rumah Perseku...|                NULL|                NULL|    NULL|         31|    none|        fact-driven| 0.09999999999999999|
+---------+-----------+--------------+--------------------+--------------------+--------------------+--------------------+--------+-----------+--------+--------------------+--------------------+
```

```
print("Streaming Topic: kata kerja")
filtered_df = csv_stream_df.filter(csv_stream_df['part_of_speech'] == 'kata kerja')

query = filtered_df.writeStream \
    .outputMode("append") \
    .format("console") \
    .start()
```

```
Streaming Topic: kata kerja
24/12/18 21:14:35 WARN ResolveWriteToStream: Temporary checkpoint location created which is deleted normally when the query didn't fail: /tmp/temporary-51aae749-649e-48b5-a2f4-8f3c6eac347f. If it's required to delete it under any circ
umstances, please set spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know deleting temp checkpoint folder is best effort.
24/12/18 21:14:35 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
-------------------------------------------
Batch: 0
-------------------------------------------
+-----------+-----------+---------------+--------------------+--------------+--------------------+--------------------+-----------+-------------+--------+--------------------+--------------------+
|      vocab|word_length|part_of_speech|         definitions|  sentences_eg|       derived_words|            synonyms|   antonyms|freq_of_usage|polarity|        subjectivity|subjectivity_score|
+-----------+-----------+---------------+--------------------+--------------+--------------------+--------------------+-----------+-------------+--------+--------------------+--------------------+
|pengetahuan|         11|      kata kerja|1. perihal menget...|1. kisah serong y...|tahu-tahu;mengeta...|maklum;faham;meng...|jahil;membiarkan|            3|    none|partially opinion...|0.6638888888888889|
+-----------+-----------+---------------+--------------------+--------------+--------------------+--------------------+-----------+-------------+--------+--------------------+--------------------+
```

```
print("Streaming Topic: kata tugas")
filtered_df = csv_stream_df.filter(csv_stream_df['part_of_speech'] == 'kata tugas')

query = filtered_df.writeStream \
    .outputMode("append") \
    .format("console") \
    .start()
```

```
Streaming Topic: kata tugas
24/12/18 21:14:36 WARN ResolveWriteToStream: Temporary checkpoint location created which is deleted normally when the query didn't fail: /tmp/temporary-417b49bd-b73b-498b-b5ca-3b99c569a451. If it's required to delete it under any circ
umstances, please set spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know deleting temp checkpoint folder is best effort.
24/12/18 21:14:36 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
-------------------------------------------
Batch: 0
-------------------------------------------
+-------+-----------+---------------+--------------------+--------------+--------------------+--------------------+-----------+-------------+--------+--------------------+--------------------+
|  vocab|word_length|part_of_speech|         definitions|  sentences_eg|       derived_words|            synonyms|   antonyms|freq_of_usage|polarity|        subjectivity| subjectivity_score|
+-------+-----------+---------------+--------------------+--------------+--------------------+--------------------+-----------+-------------+--------+--------------------+--------------------+
| sahaja|          6|     kata tugas|1. yg satu-satuny...|1. tuan sahaja yg...|bersahaja;bersaha...|satu-satunya;cuma...|  jarang|           13|    none|partially fact-dr...|0.4583333333333337|
|  hanya|          5|     kata tugas|1. tidak lebih dr...|1. kemanakan hamb...|                NULL|                NULL|   NULL|            4|    none|       opinion-driven| 0.8166666666666667|
| kepada|          6|     kata tugas|1. kata utk menya...|1. mengirim surat...|                NULL|                NULL|   NULL|           42|    none|          fact-driven| 0.2222222222222222|
|tentang|          7|     kata tugas|1. = di tentang b...|1. tentang sebuah...|bersetentang;bers...|                NULL|   NULL|            4|    none|          fact-driven|0.17129629629629628|
|  untuk|          5|     kata tugas|1. ditentukan (di...|1. tanah getah yg...|memperuntukkan;pe...|bagi;bahagian;hab...|   NULL|           42|    none|          fact-driven|0.14305555555555555|
|tentang|          7|     kata tugas|1. = di tentang b...|1. tentang sebuah...|bersetentang;bers...|                NULL|   NULL|            4|    none|          fact-driven|0.17129629629629628|
+-------+-----------+---------------+--------------------+--------------+--------------------+--------------------+-----------+-------------+--------+--------------------+--------------------+
```

```
print("Streaming Topic: adverba")
filtered_df = csv_stream_df.filter(csv_stream_df['part_of_speech'] == 'adverba')

query = filtered_df.writeStream \
    .outputMode("append") \
    .format("console") \
    .start()
```

```
Streaming Topic: adverba
24/12/18 21:14:37 WARN ResolveWriteToStream: Temporary checkpoint location created which is deleted normally when the query didn't fail: /tmp/temporary-76c9cc62-99db-4f7a-bba6-e19ed1ca8e1e. If it's required to delete it under any circ
umstances, please set spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know deleting temp checkpoint folder is best effort.
24/12/18 21:14:37 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
-------------------------------------------
Batch: 0
-------------------------------------------
+-----+-----------+---------------+-----------+--------------+--------------+--------+---------+-------------+--------+------------+------------------+
|vocab|word_length|part_of_speech|definitions|sentences_eg|derived_words|synonyms|antonyms|freq_of_usage|polarity|subjectivity|subjectivity_score|
+-----+-----------+---------------+-----------+--------------+--------------+--------+---------+-------------+--------+------------+------------------+
+-----+-----------+---------------+-----------+--------------+--------------+--------+---------+-------------+--------+------------+------------------+
```

```
print("Streaming Topic: adjektif")
filtered_df = csv_stream_df.filter(csv_stream_df['part_of_speech'] == 'adjektif')

query = filtered_df.writeStream \
    .outputMode("append") \
    .format("console") \
    .start()
```

```
Streaming Topic: adjektif
24/12/18 21:14:37 WARN ResolveWriteToStream: Temporary checkpoint location created which is deleted normally when the query didn't fail: /tmp/temporary-3901ce55-6140-4d63-826d-207fa37c1ba4. If it's required to delete it under any circ
umstances, please set spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know deleting temp checkpoint folder is best effort.
24/12/18 21:14:37 WARN ResolveWriteToStream: spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and will be disabled.
-------------------------------------------
Batch: 0
-------------------------------------------
+------+-----------+---------------+--------------------+--------------+--------------------+--------------------+--------------------+-------------+--------+--------------------+--------------------+
| vocab|word_length|part_of_speech|         definitions|  sentences_eg|       derived_words|            synonyms|            antonyms|freq_of_usage|polarity|        subjectivity|subjectivity_score|
+------+-----------+---------------+--------------------+--------------+--------------------+--------------------+--------------------+-------------+--------+--------------------+--------------------+
|senang|          6|      adjektif|1. (serba) mudah,...|1. kerja mereka s...|bersenang;bersena...|mudah;gampang;rin...|susah;benci;sesak...|            0|positive|partially opinion...|0.7183333333333334|
| mudah|          5|      adjektif|1. tidak susah at...|1. sajak baru, bu...|bermudah-mudah;me...|                NULL|                NULL|            2|positive|partially opinion...|0.6441077441077442|
+------+-----------+---------------+--------------------+--------------+--------------------+--------------------+--------------------+-------------+--------+--------------------+--------------------+
```

```
spark.stop()
```

7. References

7.1. Reference Links

Pusat Rujukan Persuratan Melayu @PRPM. (n.d.). http://prpm.dbp.gov.my/

Wikipedia contributors. (2024, November 26). *Dewan Bahasa dan Pustaka*. Wikipedia. https://en.wikipedia.org/wiki/Dewan_Bahasa_dan_Pustaka

Wikipedia contributors. (2024, December 17). *Google Books*. Wikipedia. https://en.wikipedia.org/wiki/Google_Books

What is Hadoop Distributed File System (HDFS). (n.d.). Databricks. https://www.databricks.com/glossary/hadoop-distributed-file-system-hdfs#:~:text=HDFS%20(Hadoop%20Distributed%20File%20System,rapidly%20transferring%20data%20between%20nodes.

Wikipedia contributors. (2024b, December 2). *MongoDB*. Wikipedia. https://en.wikipedia.org/wiki/MongoDB

Wikipedia contributors. (2024c, December 11). *Neo4J*. Wikipedia. https://en.wikipedia.org/wiki/Neo4j

Wikipedia contributors. (2024a, August 28). *Levenshtein distance*. Wikipedia. https://en.wikipedia.org/wiki/Levenshtein_distance

Wikipedia contributors. (2024e, December 16). *Zipf's law*. Wikipedia. https://en.wikipedia.org/wiki/Zipf%27s_law