

**LAPORAN**  
**RENCANA TUGAS MANDIRI (RTM) Ke-3**  
**MATA KULIAH BIG DATA KELAS B**  
**“MENDISTRIBUSIKAN DATA TEKS BERITA KE DALAM**  
**HDFS & MEMBUAT MAPREDUCE UNTUK PROGRAM**  
**WORDCOUNT”**



**DOSEN PENGAMPU**

Tresna Maulana Fahrudin S.ST., M.T. (NIP. 199305012022031007)

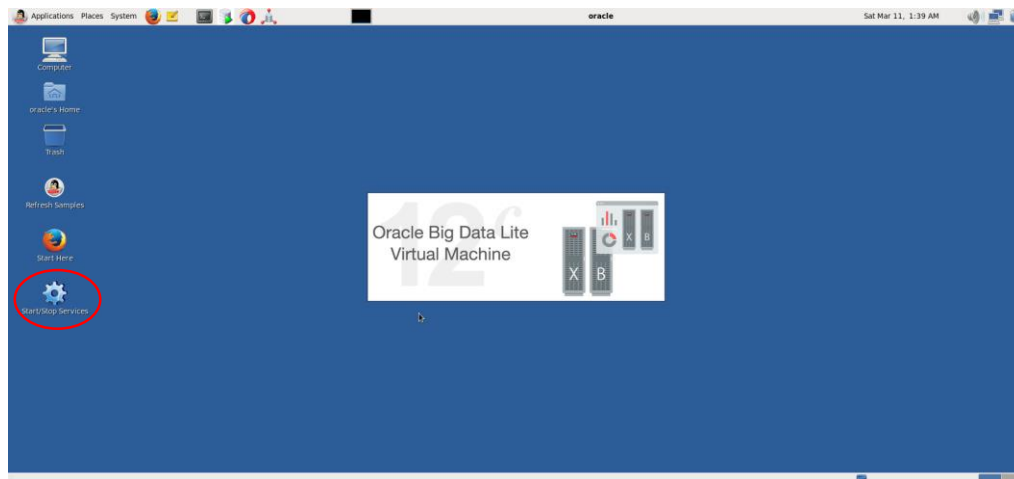
**NAMA PENYUSUN**

Fadlila Agustina (21083010050)

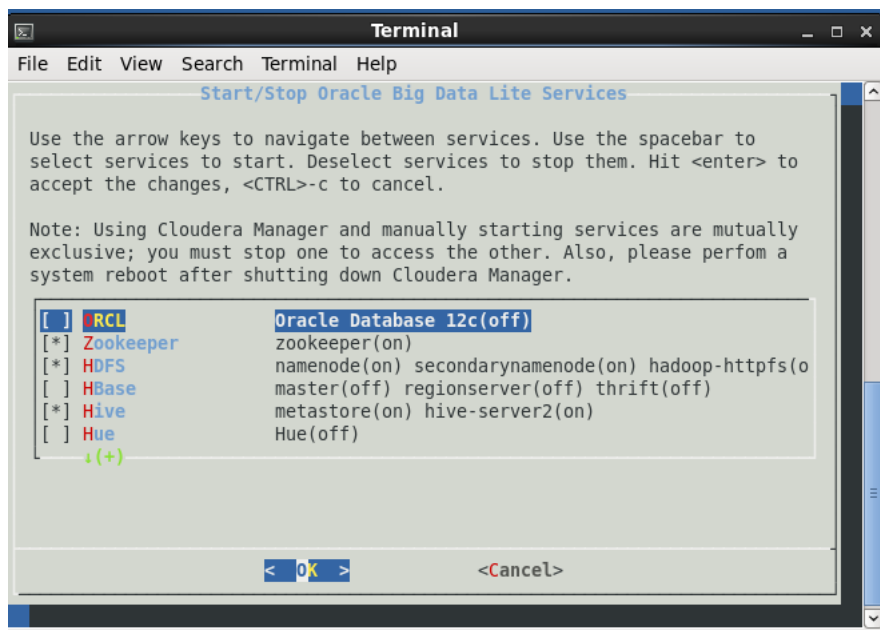
**UNIVERSITAS PEMBANGUNAN NASIONAL**  
**“VETERAN” JAWA TIMUR**  
**TAHUN 2023**

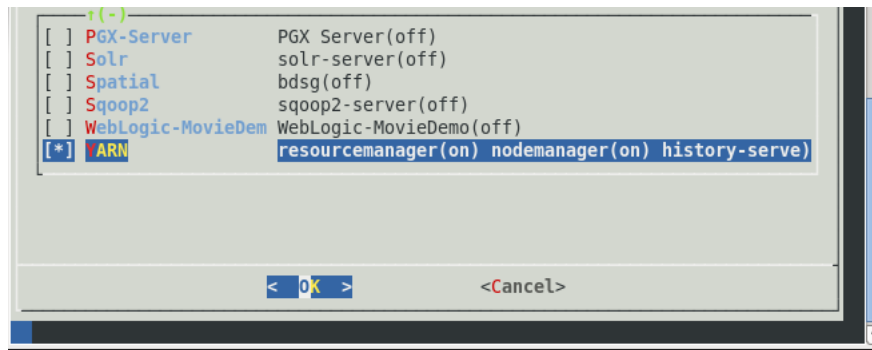
## Latihan

1. Langkah pertama yaitu membuka Oracle VM VirtualBox dan Start pada bagian BigDataLite.
2. Terdapat akun oracle dan memasukkan passwordnya.
3. Pada tampilan desktop sudah tersedia Service pada Oracle Big data.

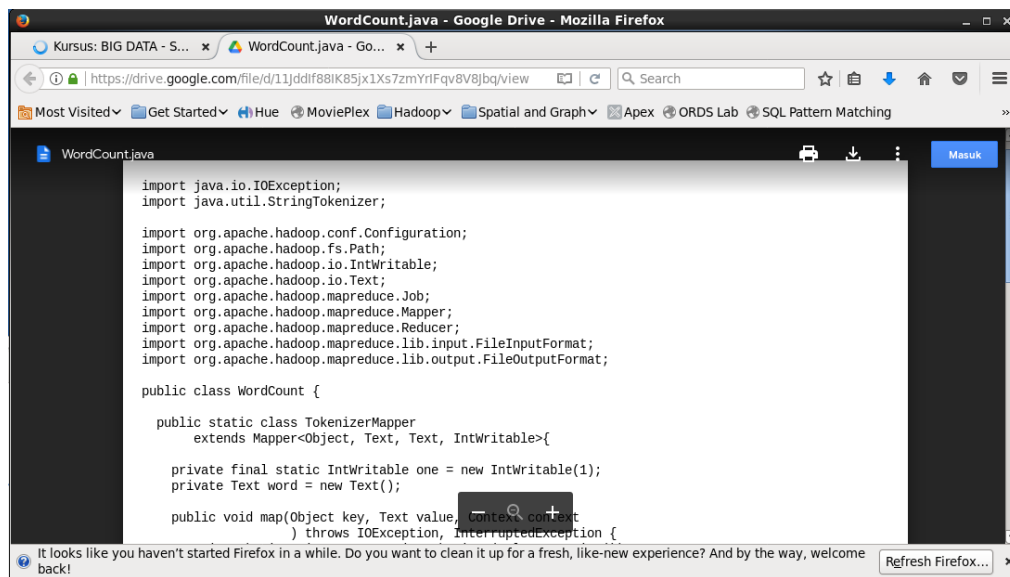


4. Klik **Service** dan akan muncul tampilan service yang aktif secara default, yaitu **Zookeeper**, **HDFS**, **Hive**, dan **YARN**.

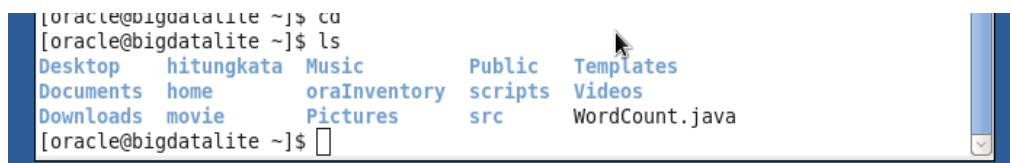




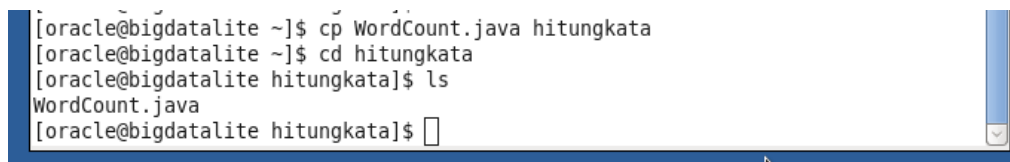
5. Buka web ilmu.upnjatim.ac.id untuk download file WordCount.java



6. Setelah terdownload cek pada terminal apakah sudah ada file nya.



7. Jika sudah ada, maka copy filenya ke dalam direktori hitungkata.



8. Setelah itu, compile program java dan membuat JAR file dalam direktori hitungkata. Pastikan hadoop sudah berjalan dengan baik (sudo jps). Jika berhasil, maka akan muncul output beberapa file seperti di bawah ini.

```

WordCount.java
[oracle@bigdatalite hitungkata]$ HADOOP_CLASSPATH=/usr/java/jdk1.8.0_151/lib/too
ls.jar
[oracle@bigdatalite hitungkata]$ hadoop com.sun.tools.javac.Main WordCount.java
[oracle@bigdatalite hitungkata]$ jar cf wc.jar WordCount*.class
[oracle@bigdatalite hitungkata]$ ls -al
total 32
drwxr-xr-x. 2 oracle oinstall 4096 Mar 11 03:16 .
drwxr-xr-x. 44 oracle oracle 4096 Mar 11 01:54 ..
-rw-r--r--. 1 oracle oinstall 3075 Mar 11 03:16 wc.jar
-rw-r--r--. 1 oracle oinstall 1491 Mar 11 03:14 WordCount.class
-rw-r--r--. 1 oracle oinstall 1739 Mar 11 03:14 WordCount$IntSumReducer.class
-rw-r--r--. 1 oracle oinstall 2089 Mar 11 01:55 WordCount.java
-rw-r--r--. 1 oracle oinstall 1736 Mar 11 03:14 WordCount$TokenizerMapper.class
[oracle@bigdatalite hitungkata]$

```

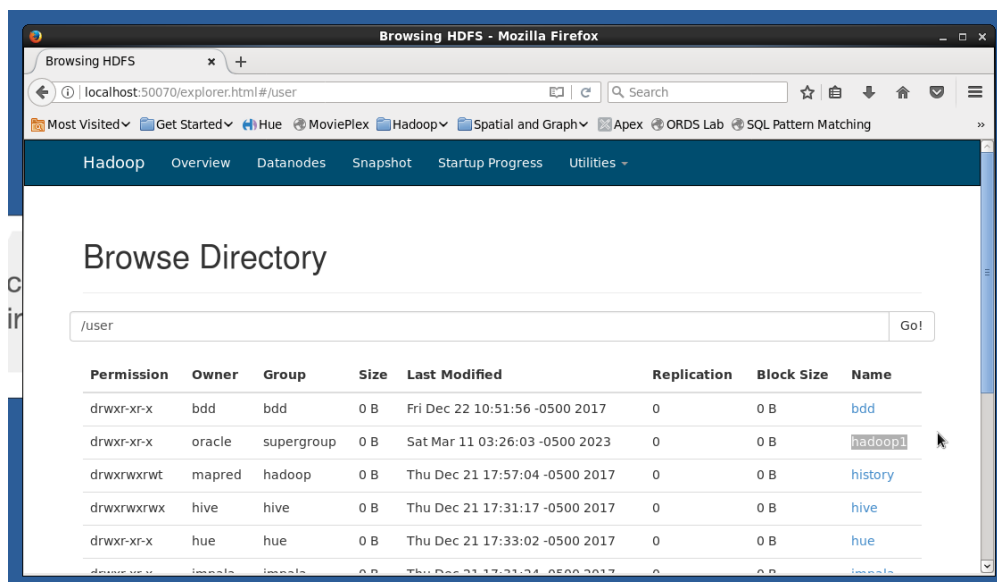
9. Membuat direktori sesuai dengan user di HDFS, contohnya hadoop1.

```

[oracle@bigdatalite hitungkata]$ hadoop fs -mkdir hdfs:///user/hadoop1

```

10. Setelah membuat direktori, cek pada browsing HDFS apakah direktori sudah ada di Hadoop file system atau belum.



11. Kemudian, membuat direktori baru di hadoop untuk menyimpan program WordCount beserta input kata yang akan diproses.

```

[oracle@bigdatalite hitungkata]$ hadoop fs -ls /user/hadoop1
[oracle@bigdatalite hitungkata]$ hadoop fs -mkdir /user/hadoop1/hitungkata
[oracle@bigdatalite hitungkata]$

```

12. Membuat file input kalimat apapun di sebuah file.

```

[oracle@bigdatalite hitungkata]$ hadoop fs -ls /user/hadoop1/hitungkata
[oracle@bigdatalite hitungkata]$ echo "hello world bye world" > file01
[oracle@bigdatalite hitungkata]$ echo "hello hadoop goodbye hadoop" > file02
[oracle@bigdatalite hitungkata]$ ls
file01 WordCount.class WordCount$TokenizerMapper.class
file02 WordCount$IntSumReducer.class
wc.jar WordCount.java
[oracle@bigdatalite hitungkata]$

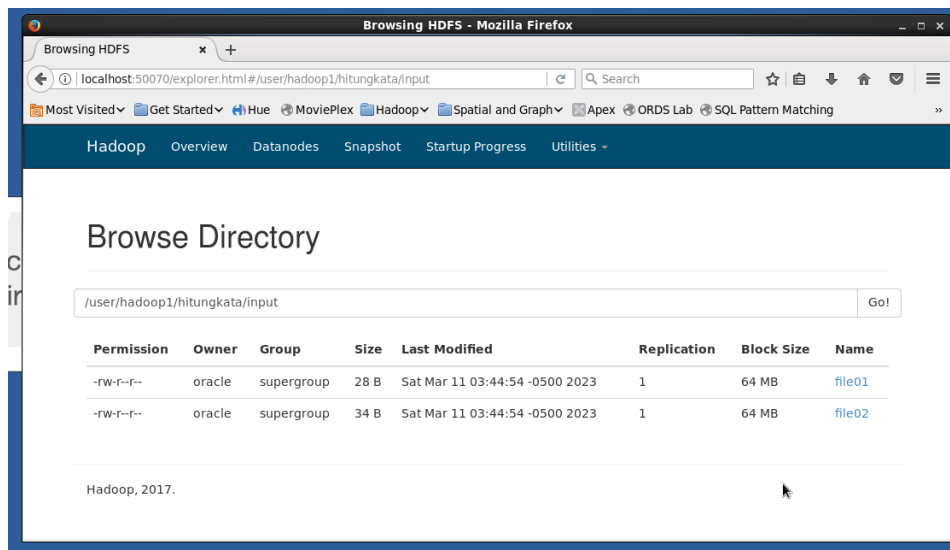
```

13. Membuat direktori input di hadoop.

```
[oracle@bigdatalite hitungkata]$ hadoop fs -mkdir hdfs:///user/hadoop1/hitungkata/input
[oracle@bigdatalite hitungkata]$ hadoop fs -mkdir /user/hadoop1/hitungkata/input
mkdir: '/user/hadoop1/hitungkata/input': File exists
[oracle@bigdatalite hitungkata]$
```

14. Kedua file tersebut berada di local direktori, maka dari itu pindahkan kedua file tersebut ke hadoop. Lalu cek file pada browsing HDFS sudah ada atau belum.

```
[oracle@bigdatalite hitungkata]$ hadoop fs -copyFromLocal file0* /user/hadoop1/hitungkata/input
```



15. Untuk memastikan apakah file sudah tersimpan dalam hadoop dan membaca isi file nya dapat melakukan perintah seperti di bawah ini.

```
[oracle@bigdatalite hitungkata]$ hadoop fs -ls /user/hadoop1/hitungkata/input
Found 2 items
-rw-r--r-- 1 oracle supergroup 28 2023-03-11 03:44 /user/hadoop1/hitungkata/input/file01
-rw-r--r-- 1 oracle supergroup 34 2023-03-11 03:44 /user/hadoop1/hitungkata/input/file02
[oracle@bigdatalite hitungkata]$ hadoop fs -cat /user/hadoop1/hitungkata/input/file01
"hello world bye world"
[oracle@bigdatalite hitungkata]$ hadoop fs -cat /user/hadoop1/hitungkata/input/file02
"hello hadoop goodbye hadoop"
[oracle@bigdatalite hitungkata]$
```

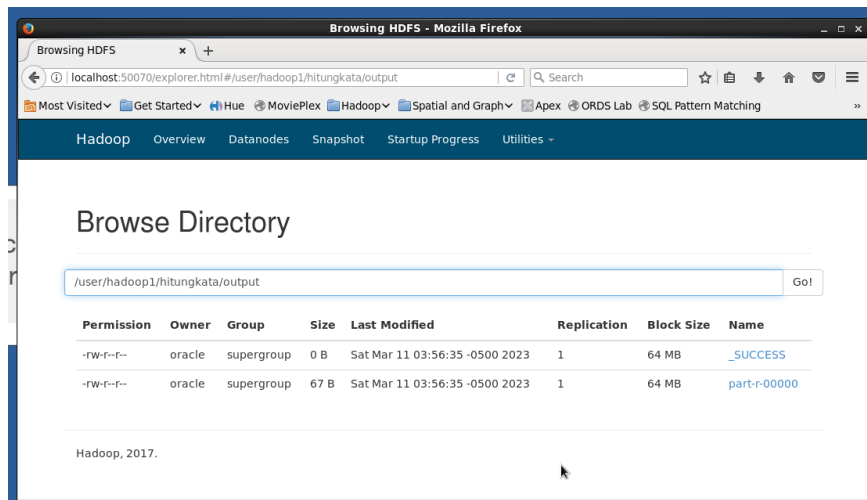
16. Sebelum menjalankan JAR, pastikan tetap berada pada direktori hitungkata.

```

[oracle@bigdatalite hitungkata]$ hadoop jar wc.jar WordCount /user/hadoop1/hitungkata/input /user/hadoop1/hitungkata/output
23/03/11 03:55:59 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/03/11 03:56:00 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
23/03/11 03:56:01 INFO input.FileInputFormat: Total input paths to process : 2
23/03/11 03:56:01 INFO mapreduce.JobSubmitter: number of splits:2
23/03/11 03:56:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1678519786910_0001
23/03/11 03:56:03 INFO impl.YarnClientImpl: Submitted application application_1678519786910_0001
23/03/11 03:56:03 INFO mapreduce.Job: The url to track the job: http://bigdatalite.localdomain:8088/proxy/application_1678519786910_0001/
23/03/11 03:56:03 INFO mapreduce.Job: Running job: job_1678519786910_0001
23/03/11 03:56:17 INFO mapreduce.Job: Job job_1678519786910_0001 running in uber mode : false
23/03/11 03:56:17 INFO mapreduce.Job: map 0% reduce 0%
23/03/11 03:56:29 INFO mapreduce.Job: map 50% reduce 0%
23/03/11 03:56:30 INFO mapreduce.Job: map 100% reduce 0%
23/03/11 03:56:37 INFO mapreduce.Job: map 100% reduce 100%
23/03/11 03:56:38 INFO mapreduce.Job: Job job_1678519786910_0001 completed successfully
23/03/11 03:56:38 INFO mapreduce.Job: Counters: 49
    File System Counters
      FILE: Number of bytes read=116
      FILE: Number of bytes written=435453
      FILE: Number of read operations=0
      FILE: Number of large read operations=0
      FILE: Number of write operations=0
      HDFS: Number of bytes read=336
      HDFS: Number of bytes written=67
      HDFS: Number of read operations=9
      HDFS: Number of large read operations=0
      HDFS: Number of write operations=2
    Job Counters
      Launched map tasks=2
      Launched reduce tasks=1
      Data-local map tasks=2
      Total time spent by all maps in occupied slots (ms)=17207
      Total time spent by all reduces in occupied slots (ms)=5955
      Total time spent by all map tasks (ms)=17207
      Total time spent by all reduce tasks (ms)=5955
      Total vcore-milliseconds taken by all map tasks=17207
      Total vcore-milliseconds taken by all reduce tasks=5955
      Total megabyte-milliseconds taken by all map tasks=17619968
      Total megabyte-milliseconds taken by all reduce tasks=6097920
    Map-Reduce Framework
      Map input records=2
      Map output records=8
      Map output bytes=94
      Map output materialized bytes=122
      Input split bytes=274
      Combine input records=8
      Combine output records=8
      Reduce input groups=7
      Reduce shuffle bytes=122
      Reduce input records=8
      Reduce output records=7
      Spilled Records=16
      Shuffled Maps =2
      Failed Shuffles=0
      Merged Map outputs=2
      GC time elapsed (ms)=286
      CPU time spent (ms)=3000
      Physical memory (bytes) snapshot=732479488
      Virtual memory (bytes) snapshot=5697773568
      Total committed heap usage (bytes)=484966400
    Shuffle Errors
      BAD_ID=0
      CONNECTION=0
      IO_ERROR=0
      WRONG_LENGTH=0
      WRONG_MAP=0
      WRONG_REDUCE=0
    File Input Format Counters
      Bytes Read=62
    File Output Format Counters
      Bytes Written=67

```

17. Perintah di atas membuat direktori baru /user/hadoop1/hitungkata/output dan terbentuk 2 file baru viz. \_SUCCESS dan part-r-00000.



18. Untuk melihat hasilnya, jalankan perintah di bawah ini.

```
[oracle@bigdatalite hitungkata]$ hadoop fs -cat /user/hadoop1/hitungkata/output/part*
bye      1
goodbye 1
hadoop   1
hadoop"  1
world    1
world"   1
"hello   2
[oracle@bigdatalite hitungkata]$
```

# Tugas 1

Coba terapkan program WordCount menggunakan data teks berita dari 2 sumber yang berbeda namun membahas topik berita yang sama. Simpan kedua berita tersebut pada 2 file txt lalu jalankan program WordCount seperti eksperimen di atas!

1. Langkah pertama yang dilakukan adalah mencari 2 link berita yang sedang trending. Pada penugasan ini, saya menggunakan topik Selena Gomez dan Hailey Bieber. Berikut link yang saya gunakan:
  - <https://lifestyle.kompas.com/read/2023/03/04/082659920/kronologi-drama-terbaru-selena-gomez-vs-hailey-bieber-libatkan-kylie>
  - <https://hot.detik.com/celeb/d-6592610/kronologi-ribut-ribut-selena-gomez-versus-hailey-bieber-dan-kylie-jenner>
2. Setelah menemukan 2 link yang berbeda, lakukan scraping di jupyterlab anaconda dan simpan hasil teksnya dalam bentuk .txt

Teks 1:

```
import re
from newspaper import Article

# mendefinisikan artikel dan menyalin link yang akan discraping
article = Article('https://hot.detik.com/celeb/d-6592610/kronologi-ribut-ribut-selena-gomez-versus-hailey-bieber-dan-kylie-jenner', 'id')

# download artikel dan mengurai halaman web
article.download()
article.parse()

# membersihkan artikel dari tanda baca
def remove_punctuation(text):
    # menghapus tanda baca pada teks
    return re.sub(r'^\w\s', '', text)

# menyimpan teks artikel yang telah diolah (tanpa tanda baca) ke dalam variabel
text = article.text
clean_text = remove_punctuation(text)

# menyimpan teks artikel yang telah diolah ke dalam file .txt
with open("selena1.txt", "w", encoding="utf-8") as file_output:
    file_output.write(clean_text)
```

Teks 2:

```
import re
from newspaper import Article

# mendefinisikan artikel dan menyalin link yang akan discraping
article = Article('https://www.liputan6.com/showbiz/read/5218870/kilas-balik-perjalanan-cinta-hailey-bieber-dan-selena-gomez-dengan-justin-bieber', 'id')

# download artikel dan mengurai halaman web
article.download()
article.parse()



# membersihkan artikel dari tanda baca
def remove_punctuation(text):
    # menghapus tanda baca pada teks
    return re.sub(r'^\w\s', '', text)

# menyimpan teks artikel yang telah diolah (tanpa tanda baca) ke dalam variabel
text = article.text
clean_text = remove_punctuation(text)

# menyimpan teks artikel yang telah diolah ke dalam file .txt
with open("selena2.txt", "w", encoding="utf-8") as file_output:
    file_output.write(clean_text)
```



File .txt:

 selenal	3/11/2023 4:51 PM	Text Document	5 KB
 selenal2	3/11/2023 4:51 PM	Text Document	2 KB

3. Kemudian proses file hasil scraping tersebut untuk menghitung wordcount pada hadoop. Sebelumnya cek dulu apakah file nya sudah ada atau belum menggunakan perintah ls.

```
oracle@bigdatalite:~  
File Edit View Search Terminal Help  
File Input Format Counters  
Bytes Read=62  
File Output Format Counters  
Bytes Written=67  
[oracle@bigdatalite hitungkata]$ hadoop fs -cat /user/hadoop1/hitungkata/output/  
part*  
bye 1  
goodbye 1  
hadoop 1  
hadoop" 1  
world 1  
world" 1  
"hello 2  
[oracle@bigdatalite hitungkata]$ cd  
[oracle@bigdatalite ~]$ cd  
[oracle@bigdatalite ~]$ ls  
Desktop hitungkata Music Public Templates  
Documents home oraInventory scripts Videos  
Downloads movie Pictures src WordCount.java  
[oracle@bigdatalite ~]$ ls  
Desktop hitungkata Music Public selenal2.txt Videos  
Documents home oraInventory scripts src WordCount.java  
Downloads movie Pictures Pictures selenal.txt Templates  
[oracle@bigdatalite ~]$
```

4. Setelah itu membuat direktori RTM3 dan copy file selenal.txt, selenal2.txt, dan WordCount.java ke dalam direktori RTM3.

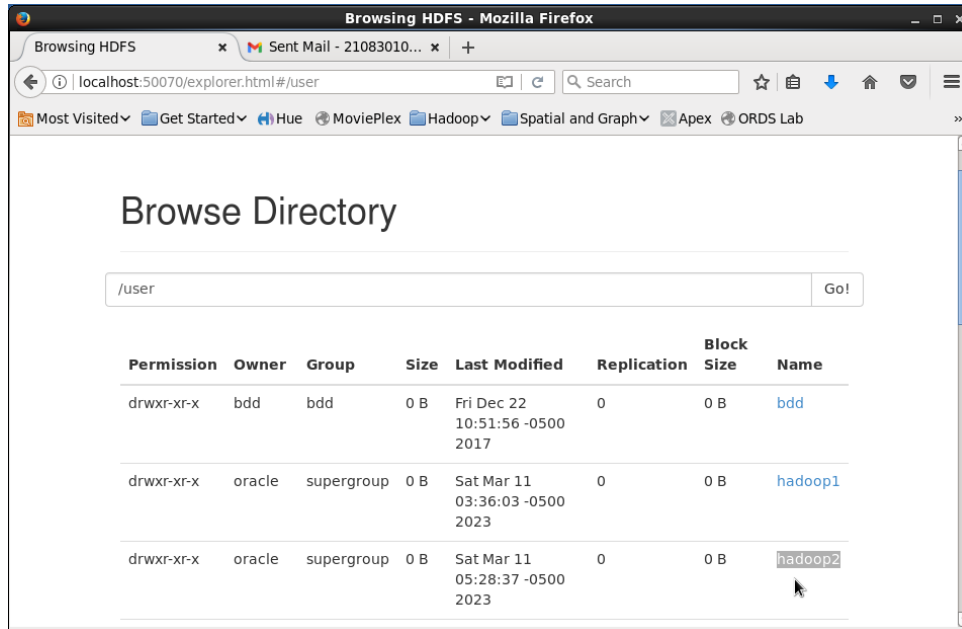
```
[oracle@bigdatalite ~]$ cp selenal.txt RTM3  
[oracle@bigdatalite ~]$ cp selenal2.txt RTM3  
[oracle@bigdatalite ~]$ cp WordCount.java RTM3  
[oracle@bigdatalite ~]$ cd RTM3  
[oracle@bigdatalite RTM3]$ ls  
selenal.txt selenal2.txt WordCount.java  
[oracle@bigdatalite RTM3]$
```

5. Compile program java dan buat JAR pada direktori RTM3.

```
[oracle@bigdatalite RTM3]$ export HADOOP_CLASSPATH=/usr/java/jdk1.8.0_151/lib/to  
ols.jar  
[oracle@bigdatalite RTM3]$ hadoop com.sun.tools.javac.Main WordCount.java  
[oracle@bigdatalite RTM3]$ jar cf wc.jar WordCount*.class  
[oracle@bigdatalite RTM3]$ ls -al  
total 44  
drwxr-xr-x. 2 oracle oinstall 4096 Mar 11 05:25 .  
drwxr-xr-x. 45 oracle oracle 4096 Mar 11 05:16 ..  
-rw-r--r--. 1 oracle oinstall 4274 Mar 11 05:17 selenal.txt  
-rw-r--r--. 1 oracle oinstall 1077 Mar 11 05:17 selenal2.txt  
-rw-r--r--. 1 oracle oinstall 3075 Mar 11 05:25 wc.jar  
-rw-r--r--. 1 oracle oinstall 1491 Mar 11 05:25 WordCount.class  
-rw-r--r--. 1 oracle oinstall 1739 Mar 11 05:25 WordCount$IntSumReducer.class  
-rw-r--r--. 1 oracle oinstall 2089 Mar 11 05:17 WordCount.java  
-rw-r--r--. 1 oracle oinstall 1736 Mar 11 05:25 WordCount$TokenizerMapper.class  
[oracle@bigdatalite RTM3]$
```

6. Buat direktori baru sesuai dalam HDFS yang bernama hadoop2.

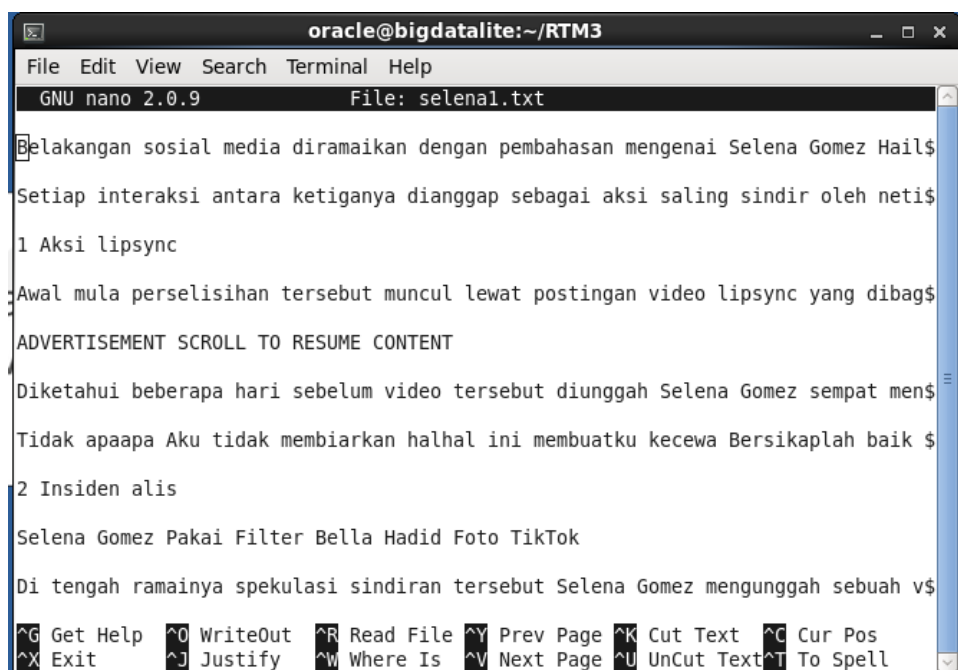
```
[oracle@bigdatalite RTM3]$ hadoop fs -mkdir hdfs:///user/hadoop2
[oracle@bigdatalite RTM3]$
```



7. Buat direktori baru bernama rtm3 yang ada di dalam user hadoop2. File tersebut dibuat di hadoop untuk menyimpan program WordCount beserta input kata yang akan diproses.

```
[oracle@bigdatalite RTM3]$ hadoop fs -ls hdfs:///user/hadoop2
[oracle@bigdatalite RTM3]$ hadoop fs -mkdir /user/hadoop2/rtm3
[oracle@bigdatalite RTM3]$
```

8. Berikut merupakan isi file selena1.txt dan selena2.txt menggunakan perintah nano.



```
oracle@bigdatalite:~/RTM3
File Edit View Search Terminal Help
GNU nano 2.0.9 File: selena2.txt

Liputan6com Jakarta Hailey Bieber dan Selena Gomez memiliki keterkaitan satu s$
Ternyata keterkaitan keduanya bukan hanya terjadi barubaru ini Jauh sebelum Hai$
Mari menengok beberapa tahun sebelumnya ketika Selena dan Justin mengonfirmasi $
Baca Juga Viral Perseteruan Hailey Bieber dan Kylie Jenner ke Selena Gomez
Hailey Bieber Tampil di Sampul Majalah Vogue Australia Diwawancarai Justin Bieb$
Hailey Bieber Potong Rambut Gaya Bob Bakal Populer di 2023
Beberapa bulan setelahnya yakni pada Mei 2011 Hailey menuliskan dukungan untuk $

[ Read 13 lines (Converted from DOS format) ]
^G Get Help ^O WriteOut ^R Read File ^Y Prev Page ^K Cut Text ^C Cur Pos
^X Exit ^J Justify ^W Where Is ^V Next Page ^U UnCut Text ^T To Spell
```

9. Buat direktori baru input di hadoop dan copy file txt ke dalam file input.

```
[oracle@bigdatalite RTM3]$ hadoop fs -mkdir /user/hadoop2/rtm3/input
[oracle@bigdatalite RTM3]$ hadoop fs -copyFromLocal selena* /user/hadoop2/rtm3/i
nput
[oracle@bigdatalite RTM3]$
```

Browsing HDFS - Mozilla Firefox

Browsing HDFS x Sent Mail - 21083010... x +

localhost:50070/explorer.html#/user/hadoop2/rtm3/input

Browse Directory

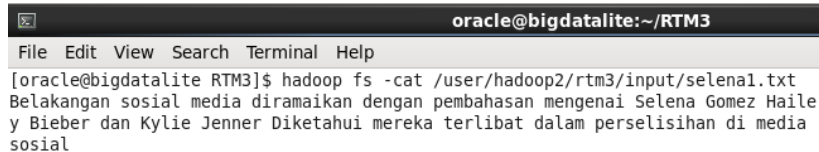
/user/hadoop2/rtm3/input Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	oracle	supergroup	4.17 KB	Sat Mar 11 05:44:11 -0500 2023	1	64 MB	<a href="#">selena1.txt</a>
-rw-r--r--	oracle	supergroup	1.05 KB	Sat Mar 11 05:44:11 -0500 2023	1	64 MB	<a href="#">selena2.txt</a>

Hadoop,

10. Untuk memastikan bahwa file sudah berada di dalam hadoop dapat menggunakan perintah seperti di bawah ini.

```
[oracle@bigdatalite RTM3]$ hadoop fs -ls /user/hadoop2/rtm3/input
Found 2 items
-rw-r--r-- 1 oracle supergroup      4274 2023-03-11 05:44 /user/hadoop2/rtm3/
input/selena1.txt
-rw-r--r-- 1 oracle supergroup      1077 2023-03-11 05:44 /user/hadoop2/rtm3/
input/selena2.txt
```



```
oracle@bigdatalite:~/RTM3
File Edit View Search Terminal Help
[oracle@bigdatalite RTM3]$ hadoop fs -cat /user/hadoop2/rtm3/input/selena1.txt
Belakangan sosial media diramaikan dengan pembahasan mengenai Selena Gomez Haile
y Bieber dan Kylie Jenner Diketahui mereka terlibat dalam perselisihan di media
sosial
```

Setiap interaksi antara ketiganya dianggap sebagai aksi saling sindir oleh netiz en Sebenarnya apa yang terjadi antara mereka bertiga Berikut kronologi dari dram a sosial media para bintang Amerika Serikat ini

#### 1 Aksi lipsync

Awal mula perselisihan tersebut muncul lewat postingan video lipsync yang dibagi kan oleh Hailey Bieber melalui akun TikTok pada Januari 2023 Lirik lipsync terse but berbunyi Im not saying she deserves it but Gods timing is always right Tak d isangka video ini justru membuat banyak orang berspekulasi dan mengklaim bahwa a ksi lipsync itu digunakan untuk membuat Selena Gomez malu

#### ADVERTISEMENT SCROLL TO RESUME CONTENT

Diketahui beberapa hari sebelum video tersebut diunggah Selena Gomez sempat meng alami body shaming akibat kenaikan berat badanya Tak berlangsung lama seorang pe ngguna TikTok ellenacuario memposting video yang bereaksi terhadap spekulasi ter sebut Secara mengejutkan Selena justru memberikan tanggapannya melalui kolom kom entar

Tidak apaapa Aku tidak membiarkan halhal ini membuatku kecewa Bersikaplah baik k epada semua orang X tulisnya

#### 2 Insiden alis

Selena Gomez Pakai Filter Bella Hadid Foto TikTok

Di tengah ramainya spekulasi sindiran tersebut Selena Gomez mengunggah sebuah vi deo lewat Instagram Story pada Februari 2023 Dia mengungkapkan secara tak sengaj a melakukan riasan laminated eyebrow Di samping itu Selena juga berharap dirinya bisa secantik Bella Hadid Namun banyak penggemar justru membandingkan keduanya meski pada akhirnya banyak yang setuju kecantikan Selena Gomez natural

Beberapa jam setelahnya Kylie Jenner juga mengunggah sebuah foto melalui Instagr am Story Foto tersebut menampilkan selfie dirinya yang ditambahkan sebuah teks i ni kecelakaan

Setelah itu Kylie juga mengunggah tangkapan layar antara dirinya dan Hailey Bieb er melalui FaceTime dengan menunjukan alis mereka secara closeup Ini membuat par a penggemar langsung bereaksi lagi bahwa unggahan tersebut merupakan balasan unt

```
[oracle@bigdatalite RTM3]$ hadoop fs -cat /user/hadoop2/rtm3/input/selena2.txt
Liputan6com Jakarta Hailey Bieber dan Selena Gomez memiliki keterkaitan satu sa
ma lain Seperti yang kita ketahui Hailey yang saat ini menikah dengan Justin Bie
ber dan Selena yang merupakan mantan dari pelantun lagu Love Yourself tersebut
```

Ternyata keterkaitan keduanya bukan hanya terjadi barubaru ini Jauh sebelum Hail ey dan Justin Bieber menikah diamdiam pada 2018 banyak drama yang terjadi antara mereka

Mari menengok beberapa tahun sebelumnya ketika Selena dan Justin mengonfirmasi h ubungan mereka pada Februari 2011 Keduanya tampil di red carpet dalam acara pest a Vanity Fair Oscar

Baca Juga Viral Perseteruan Hailey Bieber dan Kylie Jenner ke Selena Gomez

Hailey Bieber Tampil di Sampul Majalah Vogue Australia Diwawancarai Justin Biebe r

Hailey Bieber Potong Rambut Gaya Bob Bakal Populer di 2023

Beberapa bulan setelahnya yakni pada Mei 2011 Hailey menuliskan dukungan untuk m ereka Dilansir dari Cosmopolitan cuitan Hailey tersebut sudah ia hapus berisikan Aku yakin 100 tim Jelena Jelena sendiri adalah sebutan untuk fans dari couple S [oracle@bigdatalite RTM3]\$ □

## 11. Setelah itu, jalankan JAR dan pastikan untuk tetap berada di dalam direktori RTM3.

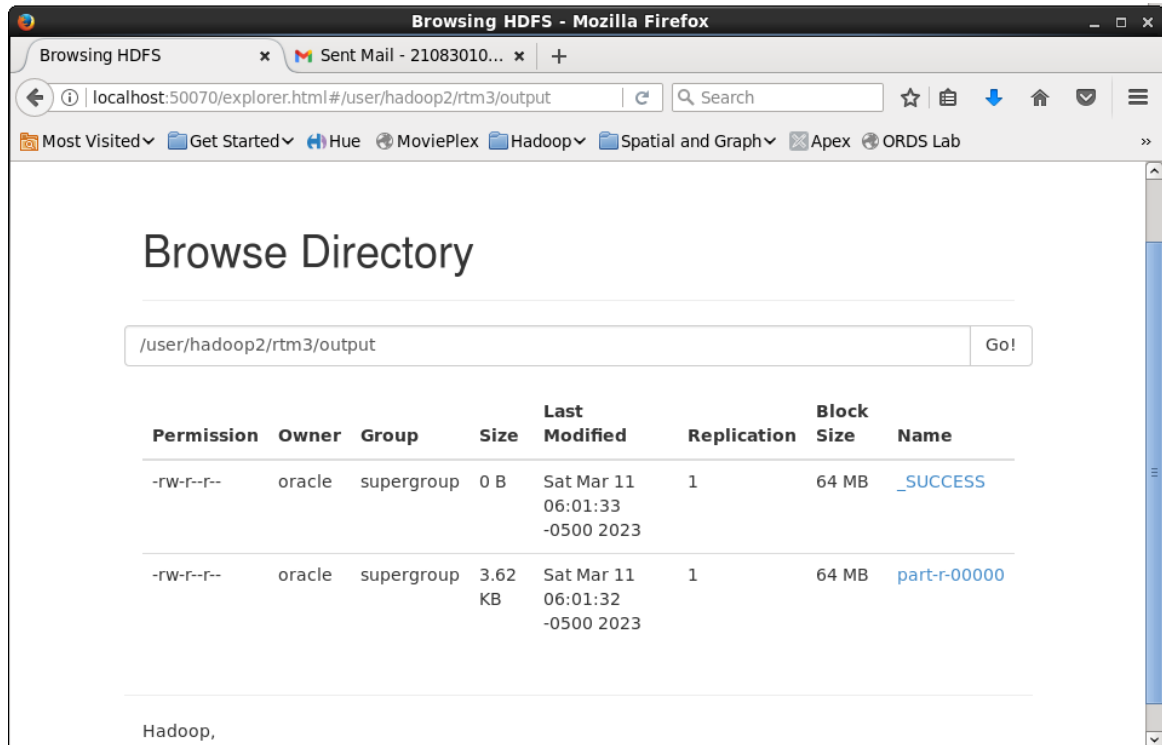
```
oracle@bigdatalite:~/RTM3
File Edit View Search Terminal Help
[oracle@bigdatalite RTM3]$ hadoop jar wc.jar WordCount /user/hadoop2/rtm3/input
/user/hadoop2/rtm3/output
23/03/11 06:00:54 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0
:8032
23/03/11 06:00:56 WARN mapreduce.JobResourceUploader: Hadoop command-line option
parsing not performed. Implement the Tool interface and execute your applicatio
n with ToolRunner to remedy this.
23/03/11 06:00:56 INFO input.FileInputFormat: Total input paths to process : 2
23/03/11 06:00:56 INFO mapreduce.JobSubmitter: number of splits:2
23/03/11 06:00:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
78519786910_0002
23/03/11 06:00:57 INFO impl.YarnClientImpl: Submitted application application_16
78519786910_0002
23/03/11 06:00:57 INFO mapreduce.Job: The url to track the job: http://bigdatali
te.localdomain:8088/proxy/application_1678519786910_0002/
23/03/11 06:00:57 INFO mapreduce.Job: Running job: job_1678519786910_0002
23/03/11 06:01:10 INFO mapreduce.Job: Job job_1678519786910_0002 running in uber
mode : false
23/03/11 06:01:10 INFO mapreduce.Job: map 0% reduce 0%
23/03/11 06:01:22 INFO mapreduce.Job: map 50% reduce 0%
23/03/11 06:01:23 INFO mapreduce.Job: map 100% reduce 0%
23/03/11 06:01:34 INFO mapreduce.Job: map 100% reduce 100%
23/03/11 06:01:34 INFO mapreduce.Job: Job job_1678519786910_0002 completed succe
ssfully
23/03/11 06:01:35 INFO mapreduce.Job: Counters: 49
  File System Counters
    FILE: Number of bytes read=5747
    FILE: Number of bytes written=446679
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=5623
    HDFS: Number of bytes written=3709
    HDFS: Number of read operations=9
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
  Job Counters
    Launched map tasks=2
    Launched reduce tasks=1
    Data-local map tasks=2
    Total time spent by all maps in occupied slots (ms)=18976
    Total time spent by all reduces in occupied slots (ms)=8609
    Total time spent by all map tasks (ms)=18976
    Total time spent by all reduce tasks (ms)=8609
    Total vcore-milliseconds taken by all map tasks=18976
```

```

Total vcore-milliseconds taken by all reduce tasks=8609
Total megabyte-milliseconds taken by all map tasks=19431424
Total megabyte-milliseconds taken by all reduce tasks=8815616
Map-Reduce Framework
  Map input records=60
  Map output records=750
  Map output bytes=8265
  Map output materialized bytes=5753
  Input split bytes=272
  Combine input records=750
  Combine output records=428
  Reduce input groups=390
  Reduce shuffle bytes=5753
  Reduce input records=428
  Reduce output records=390
  Spilled Records=856
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=296
  CPU time spent (ms)=4060
  Physical memory (bytes) snapshot=719646720
  Virtual memory (bytes) snapshot=5691052032
  Total committed heap usage (bytes)=490733568
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=5351
File Output Format Counters
  Bytes Written=3709
[oracle@bigdatalite RTM3]$ [ ]

```

12. Jika berhasil muncul output seperti di atas, maka akan muncul direktori baru seperti di bawah ini.



Browse Directory

/user/hadoop2/rtm3/output Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	oracle	supergroup	0 B	Sat Mar 11 06:01:33 -0500 2023	1	64 MB	<a href="#">_SUCCESS</a>
-rw-r--r--	oracle	supergroup	3.62 KB	Sat Mar 11 06:01:32 -0500 2023	1	64 MB	<a href="#">part-r-00000</a>

Hadoop,

13. Untuk melihat hasilnya, jalankan perintah seperti di bawah ini.

```
oracle@bigdatalite:~/RTM3
File Edit View Search Terminal Help
[oracle@bigdatalite RTM3]$ hadoop fs -cat /user/hadoop2/rtm3/output/part*
1 1
100 1
2 1
2011 2
2017 1
2018 1
2023 3
3 1
30 1
4 1
5 1
ADVERTISEMENT 1
Akhirnya 1
Akibatnya 1
Aksi 1
Aku 4
Amerika 1
Atas 1
Australia 1
Awal 1
Baca 1
Bakal 1
Beauty 1
Beberapa 3
Belakangan 1
Bella 2
Berikut 1
Bersikaplah 1
Bieber 13
Bob 1
CONTENT 1
Cosmopolitan 1
Dan 1
Di 2
Dia 4
Diketahui 2
Dilansir 1
Diwawancarai 1
Drama 1
Drop 1
FaceTime 2
Fair 1
Februari 2
Filter 1
Foto 3
Gaya 1
Gods 1
Gomez 17
Hadid 2
Hailey 15
Im 1
Ini 3
Insiden 1
Instagram 1
Instagramselenagomez 1
Instagram 1
Jakarta 1
Januari 1
Jauh 1
Jelena 2
Jenner 7
Juga 1
Justin 5
Kalian 1
Keduanya 1
Kylie 13
Lalu 1
Liputan6com 1
Lirik 1
Love 1
Maaf 1
Majalah 1
Mari 1
Mei 1
Menonaktifkan 1
Mic 1
Namun 1
Oscar 1
Pakai 1
Perseteruan 1
Populer 1
Potong 1
RESUME 1
Rambut 1
Rekaman 1
Respons 1
SCROLL 1
Sampul 1
Saya 1
```

Sebenarnya	1
Secara	1
Selena	24
Seorang	1
Seperti	1
Serikat	1
Setelah	2
Setiap	1
Setuju	1
Story	2
Swift	2
TO	1
Tak	2
Tampil	1
Taylor	2
Terdapat	1
Ternyata	1
The	1
Tidak	2
TikTok	7
TikToknya	1
Vanity	1
Video	1
Viral	1
Vogue	1
X	1
Yourself	1
acara	2
ada	2
adalah	1
akan	2
akhirnya	2
akibat	1
aksi	2
aku	4
akun	2
akunnya	1
alis	3
alisnya	1
always	1
antara	6
apa	1
apaapa	1
badanya	1
bagi	1

Dan masih banyak lagi kata yang ada di bawahnya.



## Tugas 2

Tugas selanjutnya adalah melakukan WordCount pada python, code scriptnya adalah sebagai berikut:

```
RTM 3.ipynb
[2]: # mengimport modul Regular Expression yg digunakan utk menyocokkan, mengganti, dan memproses string dgn menggunakan polanya
import re

# mendefinisikan fungsi baru yg menerima 1 argumen
def remove_punctuation(teks):
    # kembali dgn teks yg sudah dihapus tanda bacanya
    # fungsi sub digunakan utk mengganti tiap pola yg cocok dgn string kosong
    # pola [^\w\s] digunakan utk menyocokkan tiap karakter yg bkn huruf, angka, spasi dan diganti dgn string kosong
    return re.sub(r'[^\w\s]', '', teks)

# membuka file dgn mode read dan menempatkannya dlm var file1
with open("selena1.txt", "r") as file1:
    # membaca isi file dan disimpan ke dlm var file1
    teks1 = file1.read()

# membuka file dgn mode read dan menempatkannya dlm var file2
with open("selena2.txt", "r") as file2:
    # membaca isi file dan disimpan ke dlm var file2
    teks2 = file2.read()

# menghapus tanda baca dr var teks1 menggunakan fungsi remove_punctuation dan menyimpan hasilnya ke var clean_teks1
clean_teks1 = remove_punctuation(teks1)
# menghapus tanda baca dr var teks2 menggunakan fungsi remove_punctuation dan menyimpan hasilnya ke var clean_teks2
clean_teks2 = remove_punctuation(teks2)

# menggabungkan 2 teks yg sudah dihapus tanda bacanya
# tanda spasi di antara kedua var digunakan utk memisahkan teks agar gabungan teks tsb menjadi teks yg utuh dan disimpan ke dlm var kedua_teks
kedua_teks = clean_teks1 + " " + clean_teks2
```

```
RTM 3.ipynb
# tanda spasi di antara kedua var digunakan utk memisahkan teks agar gabungan teks tsb menjadi teks yg utuh dan disimpan ke dlm var kedua_teks
kedua_teks = clean_teks1 + " " + clean_teks2

# memisahkan tiap kata dlm teks yg sudah digabung dan disimpan ke var pisah_kata
pisah_kata = kedua_teks.split()

# membuat dict kosong yg akan diisi dgn kata-kata beserta frek nya
hitungkata = {}
# melakukan iterasi pd tiap kata dlm list pisah_kata
for kata in pisah_kata:
    # mengecek apakah kata saat ini blm ada dlm dict hitungkata
    if kata not in hitungkata:
        # jika blm ada, maka membuat key dgn nama kata tsb dan memberi nilai awal 1
        hitungkata[kata] = 1
    # jika sudah ada dlm hitung kata, maka
    else:
        # menambahkan nilai pada key kata tsb dgn 1 (menambah frek kemunculan kata tsb)
        hitungkata[kata] += 1

# membuka file dlm mode write sebagai hasil
with open("WordCountSelena.txt", "w") as hasil:
    # melakukan iterasi tiap item dlm dict
    # kata akan berisi kata-kata yg dihitung
    # hitung akan berisi jumlah kemunculan kata dlm teks
    for kata, hitung in hitungkata.items():
        # menulis tiap kata dan jumlah munculnya ke dlm file txt
        # str(hitung) digunakan utk mengubah nilai dr tipe int mjd str
        hasil.write(kata + " " + str(hitung) + "\n")
```

Lalu untuk melihat apakah sudah berhasil, dapat membuka file WordCountSelena.txt

```
WordCountSelena
File Edit View
Belakangan 1
sosial 7
media 7
diramaikan 1
dengan 6
pembahasan 1
mengensi 1
Selena 24
Gomez 17
Hailey 15
Bieber 13
dan 15
Kylie 13
Jenner 7
Diketahui 2
mereka 6
terlibat 1
dalam 4
perselisihan 2
di 8
Setiap 1
Ln 1, Col 1 100% Windows (CRLF) UTF-8
```