Tugas 1.
NPM : 1706005905
Nama : Rini Jannati

Bagian A.
Pada kondisi soal, yang dimaksud dengan kata adalah urutan huruf alphabet(tanpa kata dan karakter tanda baca lainnya) yang terpisah oleh whitespace. Berarti jika ada kata yang mengandung alfanumerik, kata tersebut tidak termasuk pada definisi “kata” yang dimaksud dalam soal. Jika ada kata majemuk ataupun kata yang berulang yang dipisah oleh tanda “-” maka kata tersebut dihitung secara terpisah. Setelah memahami pernyataan soal tersebut, maka jawaban yang didapat adalah sebagai berikut.
1. Jumlah distinct word yang dihasilkan.

```
Reneje:tugas1 rinijannati$ perl T1_1706005905_RiniJannatiA.pl
1. jumlah distinct word = 522699
```

Hasil didapat dengan alasan:

1. Kata dinormalisasikan menjadi lowercase.
2. Jika ada tanda “-” maka tanda tersebut diganti menjadi whitespace.
3. Kata yang diizinkan hanya kata yang mengandung alphabet. Jika ada yang mengandung alfanumerik atau tanda lainnya maka kata tersebut tidak diizinkan.

2. Screen Capture hasil keseluruhan.

```
2. hasil keseluruhan:
<ranking>       <frekuensi>     <kata>
1       915711  yang
2       879405  dan
3       792250  di
4       412504  pada
5       380929  dari
6       370081  dengan
7       340447  ini
8       323116  adalah
9       278845  dalam
10      277849  untuk
11      249973  tahun
12      238274  kategori
13      181152  oleh
14      179731  sebagai
15      150423  ke
16      147819  indonesia
17      138612  ia
18      126559  menjadi
19      126233  juga
20      123896  tidak
21      117280  itu
22      116124  atau
23      115676  merupakan
24      110315  class
25      109328  sebuah
26      105340  satu
27      96988   kota
28      96602   luar
29      95992   mereka
30      89793   orang
31      87758   memiliki
32      84980   desa
33      79298   pranala
34      77768   kecamatan
35      77158   tersebut
36      75159   karena
37      74820   dapat
38      70390   lebih
39      69867   akan
40      68522   salah
522659  1       ggo
522660  1       kabarya
522661  1       alkaloidnya
522662  1       lapiasan
522663  1       dimethylphenol
522664  1       gramatikus
522665  1       seperengkat
522666  1       nieslen
522667  1       elchaig
522668  1       variometer
522669  1       retropepsin
522670  1       ansouis
522671  1       mimesos
522672  1       inveigh
522673  1       utsugi
522674  1       samalewa
522675  1       hedrian
522676  1       gilippus
522677  1       teamone
522678  1       usenoides
522679  1       ghaill
522680  1       balmedie
522681  1       sutadipura
522682  1       supraptomo
522683  1       kiliks
522684  1       muhammadarasulullah
522685  1       shodanso
522686  1       civilizacija
522687  1       hamoed
522688  1       mkendaraan
522689  1       danamulya
522690  1       thumbsticks
522691  1       consilience
522692  1       sonan
522693  1       solehan
522694  1       acmecetca
522695  1       gondosaputro
522696  1       piedmontsetelah
522697  1       temeuan
522698  1       cacciamani
522699  1       montre
```

3. Kata Top 30:

```
3. Top 30 Kata sering muncul:
<ranking>       <frekuensi>     <kata>
1       915711  yang
2       879405  dan
3       792250  di
4       412504  pada
5       380929  dari
6       370081  dengan
7       340447  ini
8       323116  adalah
9       278845  dalam
10      277849  untuk
11      249973  tahun
12      238274  kategori
13      181152  oleh
14      179731  sebagai
15      150423  ke
16      147819  indonesia
17      138612  ia
18      126559  menjadi
19      126233  juga
20      123896  tidak
21      117280  itu
22      116124  atau
23      115676  merupakan
24      110315  class
25      109328  sebuah
26      105340  satu
27      96988   kota
28      96602   luar
29      95992   mereka
30      89793   orang
```

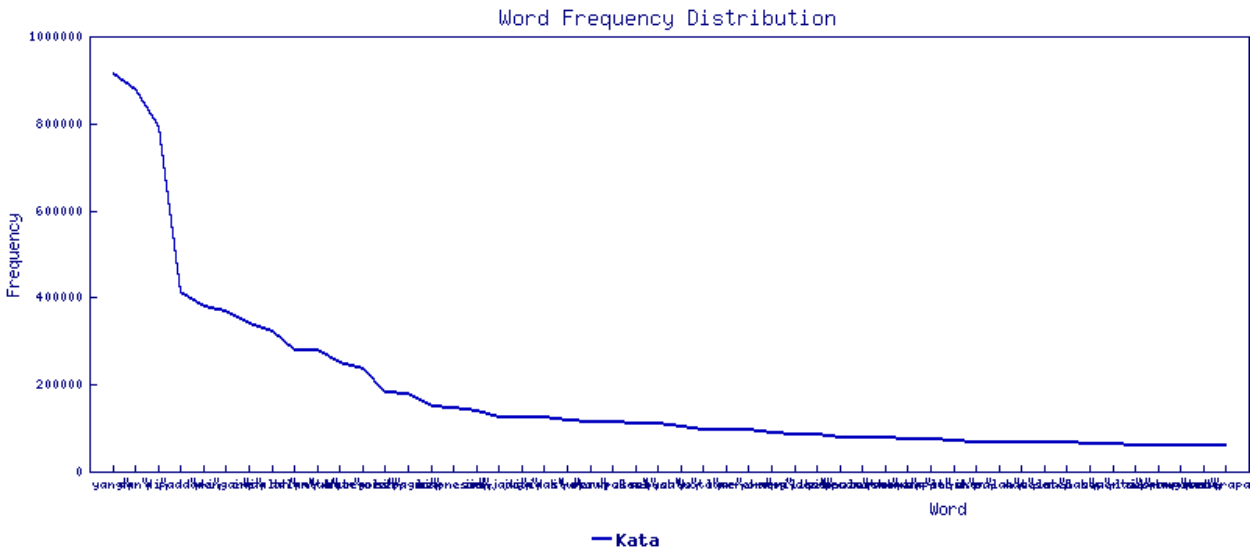
Dari hasil yang diperoleh dapat disimpulkan kata yang paling sering muncul merupakan kata seperti “yang”,”dan”,”di”,”pada”,”dari”,”dengan”,”adalah”,”dalam”,”untuk” dan “tahun” sebagai top 10. Kata-kata tersebut merupakan kata penghubung, kata depan ataupun kata yang kurang menunjukkan sebuah kata yang merujuk pada kata kunci sebuah topik atau dapat juga dikatakan sebagai kata-kata yang paling sering digunakan.

4. Top 20, Kata jarang muncul

```
<ranking>       <frekuensi>     <kata>
1       1       fnrs
2       1       qiuxia
3       1       neratta
4       1       minitaur
5       1       weht
6       1       penternak
7       1       hangbu
8       1       glenbranter
9       1       wechthari
10      1       gcais
11      1       peromaan
12      1       trichiuridae
13      1       finansija
14      1       totogong
15      1       djatibedrijf
16      1       mikrohidronya
17      1       monja
18      1       guixian
19      1       pammanakang
20      1       sukiswo
```

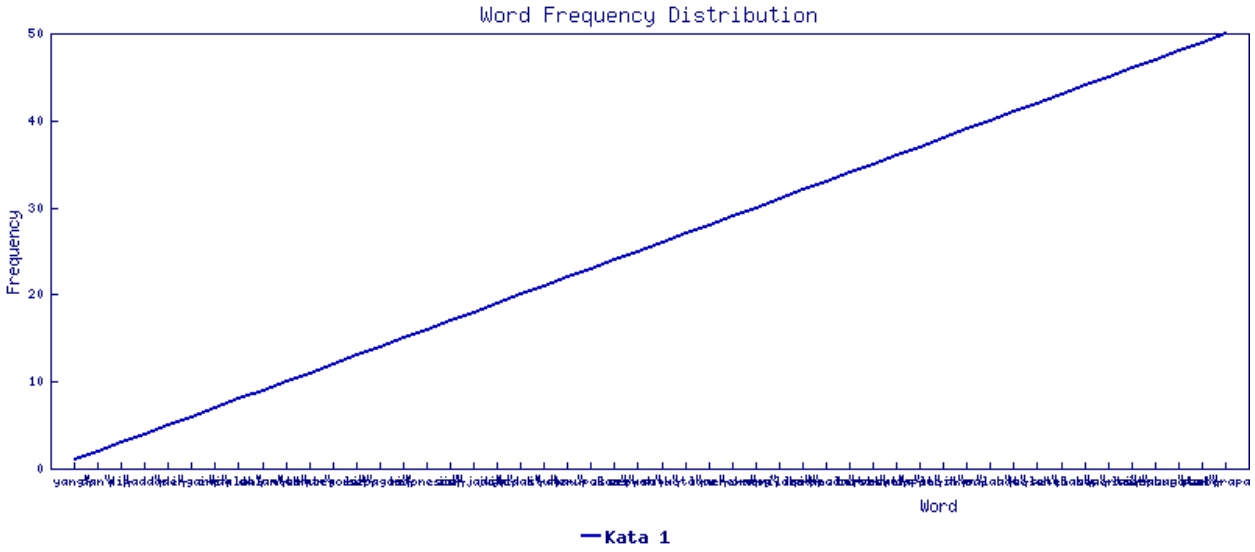
Dari hasil yang diperoleh pada Top 20 kata-kata yang paling jarang muncul adalah rata-rata kemunculan kata tersebut adalah sekali. Dapat dikatakan kemunculan kata tersebut tidak terlalu berkaitan dengan isi.

5. Distribusi frekuensi kata TOP 50



Grafik Word Frequency Distribution

6. Kaitan pada algoritma Zifp's Law:
Jika kata menara muncul sebanyak 4284 kali pada peringkat 1000, maka kata pemimpin pada peringkat 500 yaitu lebih kurang 8568 (kenyataannya muncul sebanyak 8645) kali.
Dari hasil program ini dapat dikatakan bahwa pernyataan zipf's law bahwa frekuensi kata berbanding terbalik dengan ranking kemunculan karena semakin kecil angka ranking berarti frekuensi sebakin besar.



Grafik Ranking

Jika dibandingkan antara grafik word frequency distribution dengan grafik ranking, maka dapat dilihat yang yang merupakan kata dengan ranking 1 jumlah frekuensinya lebih banyak daripada ranking 50.

7. Kata yang mengandung kalimat jakarta.

```
7. Banyak kalimat yang mengandung kata jakarta= 17991
```

Hasil didapat dengan cara baris yang mengandung jakarta dimasukkan kedalam array lalu jumlah array tersebut dihitung.

8. Pada soal no.7 sudah dihitung kalimat yang mengandung kata jakarta. Pada contoh soal pembuatan kata bigram, regular expression juga ikut dalam pembuatan bigram. Jika kondisi tersebut diikut sertakan, maka hasil dengan kondisi tersebut adalah

```
8. Top 40 kata bigram dari kalimat yang mengandung jakarta:
<ranking>       <frekuensi>     <kata>
1       14765   . <end>
2       4146    di jakarta
3       3539    , jakarta
4       3142    jakarta ,
5       1876    indonesia kategori
6       1514    kategori tokoh
7       1490    , dan
8       1402    , indonesia
9       1334    jakarta timur
10      1327    pranala luar
11      1267    pada tahun
12      1222    jakarta <end>
13      1052    luar kategori
14      1048    tokoh dari
15      963     timur ,
16      957     dki jakarta
17      917     ). <end>
18      879     pada tanggal
19      858     terletak di
20      839     <start> pada
21      825     ke jakarta
22      799     di indonesia
23      797     jakarta dan
24      749     dari jakarta
25      686     thumb |
26      678     jakarta kategori
27      677     jakarta selatan
28      657     sekolah dasar
29      561     jakarta pusat
30      556     indonesia ,
31      552     yang terletak
32      539     sma negeri
33      533     jakarta pada
34      532     salah satu
35      509     sekolah menengah
36      500     , yang
37      476     jakarta (
38      438     ) dan
39      397     dasar negeri
40      394     <start> di
DONE
```

Jika kondisi

B.

9. Pada program ini yang saya lakukan adalah mengecek line per line kata metro jaya sehingga didapat jumlah kalimat yang mengandung kata metro jaya adalah sebanyak 48.

```
9. Banyak kata metro dan jaya: 48
```

10. Pada program ini yang saya lakukan untuk mendapatkan banyak dokumen adalah dengan mengecek banyaknya jumlah <DOC> pada korpus yang telah saya gabungkan. Jumlahnya adalah 1288.

```
10. Banyak dokumen pada korpus: 1288
```

11.Pada program ini yang saya lakukan adalah menghapus semua tag-an dan menyimpan datanya pada sebuah array. Syarat untuk menjadi anggota array juga saya tambah agar tidak ada empty value didalam array dan juga memisahkan kalimat yang jika mereka tergabung dipisahkan dengan tanda titik(.), tanda tanya(?) dan tanda seru(!)

```
11. Jumlah rata-rata kalimat adalah 31.4883540372671
```

12. Jumlah Collocation:

1	13.2406405131954	salif diao
2	13.2406405131954	alain duclos
3	13.2406405131954	vedran runje
4	13.2406405131954	hedi yustaja
5	13.2406405131954	evvc vertical
6	13.2406405131954	naichi rainbow
7	13.2406405131954	belvoir castle
8	13.2406405131954	tunjukkanlah empati
9	13.2406405131954	tepekong lio
10	13.2406405131954	indro tjahjono
11	13.2406405131954	darius nggawa
12	13.2406405131954	cold storage
13	13.2406405131954	ninh binh
14	13.2406405131954	giuseppe macchion
15	13.2406405131954	degenerasi neuron
16	13.2406405131954	trabajadores venezolanos
17	13.2406405131954	wives where
18	13.2406405131954	hosea kogo
19	13.2406405131954	avocado bean
20	13.2406405131954	internalisasi konsitusi
21	13.2406405131954	evgenia koulikovskaya
22	13.2406405131954	isarangura na
23	13.2406405131954	coordinadora democratica
24	13.2406405131954	uut nita
25	13.2406405131954	santi wijayanti
26	13.2406405131954	kekonsistenan permaian
27	13.2406405131954	seat steering
28	13.2406405131954	iroda tulyaganova
29	13.2406405131954	bandera roja
30	13.2406405131954	muhammed abid

Dari jumlah collocation kebanyakan adalah nama orang. Kata ini berarti adalah kata bigram dan frekuensi kata-katanya juga termasuk tidak terlalu sering muncul.