TUGAS NATURAL LANGUAGE PROCESSING

Fero Resyanto

1301154318

IF-39-GAB

JURUSAN S1 TEKNIK INFORMATIKA

FAKULTAS INFORMATIKA

UNIVERSITAS TELKOM

BANDUNG

2018

- Sumber Artikel

Sumber artikel yang digunakan pada tugas ini berasal dari website Viva Indonesia. Artikel diambil dari 8 kategori artikel yaitu kategori berita, bola, digital, gaya hidup, otomotif, showbiz, sport dan other yang berjumlah 100 artikel. Pengambilan artikel dilakukan dengan metode crawling menggunakan bahasa pemrograman python.

- Analisis terhadap hasil pengujian prediksi kemunculan kata.

10 kata digunakan untuk melakukan prediksi kemunculan kata. 10 kata yang digunakan/diambil berdasarkan kata yang memiliki frekuensi yang cukup banyak dan dan cukup sedikit. 5 kata diambil yang memiliki frekuensi kata cukup banyak dann 5 kata lainnya diambil yg memiliki frekuensi kata cukup sedikit. Kata yang diambil bukan kata yang memiliki frekuensi paling banyak karena kata-kata yang memiliki frekuensi paling banyak merupakan kata penghubung dan sebagainya. Jumlah kemunculan kata untuk frekuensi kata cukup sedikit berjumlah kurang dari sama dengan 50, sedangkan jumlah kata untuk frekuensi kata cukup banyak lebih dari 50.

Kata-kata uji yang digunakan adalah sebagai berikut :

1. Bagi = 29 kata
2. Mereka = 83 kata
3. Alasan = 11 kata
4. Saat = 88 kata
5. Mengatakan = 38 kata
6. Tidak = 108 kata
7. Bisa = 153 kata
8. Saya = 136 kata
9. Prabowo = 14 kata
10. Joko = 3 kata

Setelah melakukan running pada program, berikut hasil kata yang muncul setelah kata uji :

Bagi → hasil (jumlah kemunculan kalimat : 2)

Mereka → sudah (jumlah kemunculan kalimat : 5)

Alasan → di (jumlah kemunculan kalimat :2)

Saat → ini (jumlah kemunculan kalimat :20)

Mengatakan → bahwa (jumlah kemunculan kalimat :10)

Tidak → ada (jumlah kemunculan kalimat : 15)

Bisa → bermain (jumlah kemunculan kalimat :3)

Saya → akan (jumlah kemunculan kalimat :8)

Prabowo → subianto (jumlah kemunculan kalimat : 3)

Joko → widodo (jumlah kemunculan kalimat : 3)

Berdasarkan hasil dari percobaan 10 kata yang diambil secara acak, kata “saat” diikuti kata “ini” memiliki frekuensi kemunculan paling banyak yaitu 20 kalimat. Sedangkan kata “bagi” diikuti kata “hasil” memiliki frekuensi kemunculan paling sedikit 2 kalimat, begitu juga dengan kata “alasan” diikuti kata “di” memiliki frekuensi kemunculan paling sedikit 2 kalimat. Berikut hasil runningan program :

```
Kata : bagi
Kata Selanjutnya : hasil
Jumlah kemunculan kalimat bagi hasil : 2
---------------------------------
Kata : mereka
Kata Selanjutnya : sudah
Jumlah kemunculan kalimat mereka sudah : 5
---------------------------------
Kata : alasan
Kata Selanjutnya : di
Jumlah kemunculan kalimat alasan di : 2
---------------------------------
Kata : saat
Kata Selanjutnya : ini
Jumlah kemunculan kalimat saat ini : 20
---------------------------------
Kata : mengatakan
Kata Selanjutnya : bahwa
Jumlah kemunculan kalimat mengatakan bahwa : 10
---------------------------------
Kata : tidak
Kata Selanjutnya : ada
Jumlah kemunculan kalimat tidak ada : 15
---------------------------------
Kata : bisa
Kata Selanjutnya : bermain
Jumlah kemunculan kalimat bisa bermain : 3
---------------------------------
Kata : saya
Kata Selanjutnya : akan
Jumlah kemunculan kalimat saya akan : 8
---------------------------------
Kata : prabowo
Kata Selanjutnya : subianto
Jumlah kemunculan kalimat prabowo subianto : 3
---------------------------------
Kata : joko
Kata Selanjutnya : widodo
Jumlah kemunculan kalimat joko widodo : 3
---------------------------------
```

- Analisis terhadap evaluasi perplexity

5 kalimat digunakan untuk melakukan pengujian perplexity. Kalimat yang digunakan merupakan kalimat judul/highlight dari beberapa artikel. Berikut ini kalimat yang digunakan :

1. Kalau Deddy Mizwar Keluar Kita Ucapkan Selamat Jalan.
2. Sandiaga Ajak Presiden Jokowi Tukarkan Dolar AS ke Rupiah.
3. Erick Thohir Masuk Top List Calon Ketua Timses Jokowi.
4. Pertamina Mulai Kelola Blok SES.
5. Luhut Pastikan Proyek Kelistrikan Ditunda.

Kalimat 1 memiliki nilai probabilitas 1.529 dengan nilai perplexity 6.77

Kalimat 2 memiliki nilai probabilitas 4.168 dengan nilai perplexity 8.36

Kalimat 3 memiliki nilai probabilitas 2.546 dengan nilai perplexity 3.75

Kalimat 4 memiliki nilai probabilitas 0.001 dengan nilai perplexity 5.09

Kalimat 5 memiliki nilai probabilitas 0.001 dengan nilai perplexity 5.44

Berdasarkan evaluasi perplexity diatas, kalimat 5 memiliki nilai perplexity paling tinggi yaitu 5.44. sedangkan kalimat 3 memiliki nilai perplexity paling kecil yaitu 3.75, hal ini menandakan bahwa model bahasa yang digunakan cukup baik pada kalimat 3. Karena semakin kecil nilai perplexity, maka semakin baik suatu model bahasa tersebut. Berikut hasil runningan program :

```
kata 1 :  kalau
kata 2 :  deddy
Probabilitas :  0.022222222222222223
kata 1 :  deddy
kata 2 :  mizwar
Probabilitas :  0.5
kata 1 :  mizwar
kata 2 :  keluar
Probabilitas :  0.2
kata 1 :  keluar
kata 2 :  kita
Probabilitas :  0.1
kata 1 :  kita
kata 2 :  ucapkan
Probabilitas :  0.024096385542168676
kata 1 :  ucapkan
kata 2 :  selamat
Probabilitas :  0.6666666666666666
kata 1 :  selamat
kata 2 :  jalan
Probabilitas :  0.42857142857142855
total probabilitas kalimat :  1.5299292407726143e-06
Preplexity:  6.7726906066166785
```

```
kata 1 :  sandiaga
kata 2 :  ajak
Probabilitas :  0.1
kata 1 :  ajak
kata 2 :  presiden
Probabilitas :  0.5
kata 1 :  presiden
kata 2 :  jokowi
Probabilitas :  0.05555555555555555
kata 1 :  jokowi
kata 2 :  tukarkan
Probabilitas :  0.058823529411764705
kata 1 :  tukarkan
kata 2 :  dolar
Probabilitas :  1.0
kata 1 :  dolar
kata 2 :  as
Probabilitas :  0.1
kata 1 :  as
kata 2 :  ke
Probabilitas :  0.25
kata 1 :  ke
kata 2 :  rupiah
Probabilitas :  0.01020408163265306
total probabilitas kalimat :  4.1683340002667736e-08
Preplexity:  8.36574622555889
```

```
kata 1 :  erick
kata 2 :  thohir
Probabilitas :  0.4
kata 1 :  thohir
kata 2 :  masuk
Probabilitas :  0.25
kata 1 :  masuk
kata 2 :  top
Probabilitas :  0.045454545454545456
kata 1 :  top
kata 2 :  list
Probabilitas :  1.0
kata 1 :  list
kata 2 :  calon
Probabilitas :  0.25
kata 1 :  calon
kata 2 :  ketua
Probabilitas :  0.2857142857142857
kata 1 :  ketua
kata 2 :  timses
Probabilitas :  0.23529411764705882
kata 1 :  timses
kata 2 :  jokowi
Probabilitas :  0.3333333333333333
total probabilitas kalimat :  2.546473134708429e-05
Preplexity:  3.7519549286298024

kata 1 :  pertamina
kata 2 :  mulai
Probabilitas :  0.25
kata 1 :  mulai
kata 2 :  kelola
Probabilitas :  0.043478260869565216
kata 1 :  kelola
kata 2 :  blok
Probabilitas :  0.5
kata 1 :  blok
kata 2 :  ses
Probabilitas :  0.2727272727272727
total probabilitas kalimat :  0.0014822134387351778
Preplexity:  5.0965033423290205
```

```
kata 1 :  luhut
kata 2 :  pastikan
Probabilitas :  0.1
kata 1 :  pastikan
kata 2 :  proyek
Probabilitas :  0.5
kata 1 :  proyek
kata 2 :  kelistrikan
Probabilitas :  0.09090909090909091
kata 1 :  kelistrikan
kata 2 :  ditunda
Probabilitas :  0.25
total probabilitas kalimat :  0.0011363636363636365
Preplexity:  5.4465396306630005
```