**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)

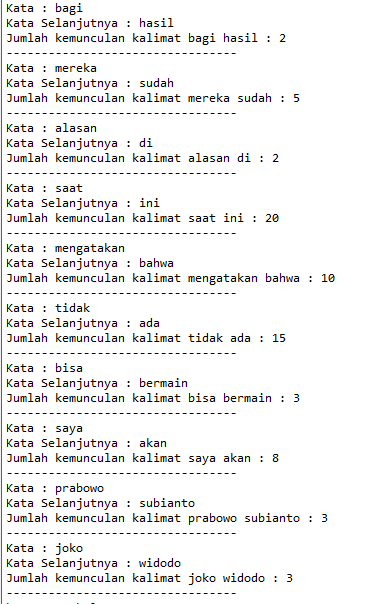
Bisa 🡪 bermain (jumlah kemunculan kalimat :3)

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 :



* 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 :

