LAPORAN TUGAS PEKAN 2 NLP : LANGUAGE MODEL UNIGRAM BIGRAM

CLARISA HASYA YUTIKA | 1301174256 | IF 41 02

1. Deskripsi Masalah

Terdapat 20 file artikel berbahasa indonesia dengan topik ‘Cyber Crime Whatsapp’ akan dibuat sebuah model bahasa bigram. Kemudian akan diimplementasikan laplace smoothing dan akan dihitung nilai probabilitas bigram dan nilai perplexity. Terdapat 6 kalimat uji, 3 kalimat sesuai dengan topik, 3 kalimat tidak sesuai topik. Dari 6 kalimat uji, akan dibandingkan nilai perplexitynya, lebih bagus kalimat yang sesuai topik atau kalimat yang tidak sesuai topik.

2. Perancangan Sistem

- Membaca 20 file artikel .txt

```
#Read file txt
data = []
list_of_files = glob('a*.txt')
for file_name in list_of_files:
  f = open(file_name, 'r')
  data.append(f.read())
  f.close()
```

- Case Folding

  Dilakukannya case folding untuk mengubah semua huruf menjadi lowercase.

- Tokenisasi

  Tokenisasi bertujuan untuk menambah tag ‘<s> pada awal kalimat, dan tag </s> pada akhir kalimat. Kemudian membagi kalimat menjadi per-kata.

- Frekuensi Unigram

  Menghitung frekuensi unigram dengan menambahkan nilai 1 setiap menemukan kata yang sama.

- 10 Unigram yang palung sering muncul

```
10 Unigram yang paling sering muncul
<s> 330
</s> 330
, 329
. 318
yang 195
whatsapp 185
dan 115
di 99
untuk 90
ini 84
```

- Probabilitas Unigram

  Nilai probabilitas unigram didapat dengan cara frekuensi kata dibagi dengan jumlah semua kata

- Frekuensi Bigram

  Menghitung frekuensi bigram dengan menambahkan nilai 1 setiap menemukan kombinasi kata yang sama

- Probabilitas Bigram

  Nilai probabilitas bigram didapat dengan cara frekuensi kombinasi 2 kata dibagi dengan frekuensi 1 kata diawal kombinasi 2 kata.

- 10 Probabilitas Bigram tertinggi

```
10 Bigram dengan probability paling tinggi
('?', '</s>') 1.0
('apac', 'communications') 1.0
('communications', 'director') 1.0
('director', 'sravanthi') 1.0
('sravanthi', 'dev') 1.0
('bersama', 'sejumlah') 1.0
('tahapan-tahapan', 'yang') 1.0
('percobaan', 'untuk') 1.0
('one', 'time') 1.0
('time', 'password') 1.0
```

- Cek Bigram untuk kalimat uji

  Untuk cek probabilitas bigram pada kalimat uji, akan dilakukan case folding dan tokenisasi. Setelah itu dihitung nilai probabilitas bigram dengan mencari kombinasi 2 kata yang sama dengan probabilitas bigram dari kalimat latih. Jika sama, nilai probabilitas akan dikalikan

dengan probabilitas bigram. Jika tidak sama, nilai probabilitas akan dikalikan dengan 0.

- Laplace Smoothing

  Supaya nilai probabilitas bigram kalimat uji tidak 0, diperlukannya laplace smoothing. Dengan menambahkan nilai 1 di setiap frekuensi bigramnya. Kemudian untuk mendapatkan nilai probabilitas bigram kalimat uji adalah dengan frekuensi bigram (Ci) ditambah 1 kemudian dibagi dengan frekuensi unigram (N) kata sebelumnya ditambah dengan jumlah kata dari kalimat uji (V).

  awalnya: $P(w_i) = \frac{c_i}{N}$ menjadi: $P_{\text{Laplace}}(w_i) = \frac{c_i + 1}{N + V}$

- Perplexity

  Nilai perplexity merupakan invers dari nilai probabilitas bigram. Semakin kecil nilai perplexity, maka semakin bagus kualitas language modelnya.

3. Analisis

Terdapat 6 kalimat uji, 3 kalimat sesuai dengan topik, 3 kalimat tidak sesuai topik. Akan dibandingkan nilai perplexitynya

- Kalimat yang sesuai dengan topik

```
Non Smoothing
Probabilitas Bigram :  0.0
Perplexity :  0
Smoothing
Probabilitas Bigram :  4.241620669689574e-38
Perplexity :  467.15608736452066
```

```
Non Smoothing
Probabilitas Bigram :  0.0
Perplexity :  0
Smoothing
Probabilitas Bigram :  1.4704524757738434e-31
Perplexity :  371.00442022697644
```

```
Non Smoothing
Probabilitas Bigram :  0.0
Perplexity :  0
Smoothing
Probabilitas Bigram :  5.5253887129916376e-76
Perplexity :  612.8016062147624
```

- Kalimat yang tidak sesuai dengan topik

```
Non Smoothing
Probabilitas Bigram :  0.0
Perplexity :  0
Smoothing
Probabilitas Bigram :  7.237226820639794e-78
Perplexity :  719.5365873852458
```

```
Non Smoothing
Probabilitas Bigram :  0.0
Perplexity :  0
Smoothing
Probabilitas Bigram :  2.611342230283748e-69
Perplexity :  720.4945089828541
```

```
Non Smoothing
Probabilitas Bigram :  0.0
Perplexity :  0
Smoothing
Probabilitas Bigram :  3.2346483592810946e-64
Perplexity :  768.9833757727514
```

4. Kesimpulan

Dari ketiga jenis kalimat yang berbeda dapat disimpulkan bahwa kalimat yang sesuai dengan topik, nilai probabilitas bigram yang sudah di smoothing lebih besar dan nilai perplexitynya lebih kecil daripada kalimat uji tidak sesuai topik. Hal ini dapat terjadi karena ada beberapa kata dalam kalimat sesuai topik yang sudah terdapat dalam corpus. Sehingga probabilitasnya akan semakin besar, dan nilai perplexity akan lebih kecil.