Laporan Tugas NLP Minggu ke-2 / Language Model

Muhammad Asjad Adna Jihad 1301170242

- Deskripsi Masalah

Untuk topik artikel dipilih dengan topik ‘Sneakers Sampah Nike Space Hippie’ yang akan dirancang model Bahasa unigram serta bigram, penghitungan nilai probabilitas, pengimplementasian *Laplace Smoothing,* serta penghitungan nilai Perplexity. Untuk kalimat uji sendiri telah disediakan 4 kalimat pada file tester.txt, dengan sebagian dari jumlah kalimat tersebut sesuai dengan topik yang diambil dan sebagiannya lagi tidak sesuai dengan topik.

- Perancangan Sistem

i. Pembacaan file & melakukan penggabungan menjadi satu file .txt

Untuk tahap ini, digunakan library glob untuk membaca 20 file artikel, kemudian menghasilkan output file result.txt yang didalamnya merupakan gabungan 20 file tersebut. Dilakukan penggunaan fungsi lower() untuk mengubah huruf menjadi lowercase demi meminimalisir kesalahan.

ii. Tokenisasi

Tokenisasi menambahkan tag ‘<s>’ di awal kalimat dan </s> di akhir kalimat, kemudian membagi kalimat per-kata.

iii. Menghitung Frekuensi Unigram

Pada tahapan ini, apabila menemukan kata yang sama maka akan ditambahkan nilai 1 dalam penghitungan.

iv. Menampilkan 10 Unigram yang paling sering muncul

Melakukan sorting berdasarkan frekuensi tertinggi, output :

```
10 besar kata/simbol yang sering muncul (unigram) : {',': 333, '<s>': 300, '</s>': 300, '.': 277, 'nike': 181, 'yang': 177, 'space': 165, 'hippie': 150, 'dari': 144, 'dan': 121}
```

v. Menghitung Probabilitas Unigram

Didapatkan dengan membagi frekuensi tiap kata dengan jumlah semua kata.

vi. Menghitung Frekuensi Bigram

Pada tahap ini, apabila menemukan kombinasi kata yang sama maka akan ditambahkan nilai 1.

vii. Menghitung Probabilitas Bigram

Dilakukan dengan membagi frekuensi kombinasi 2 kara dengan frekuensi kata didepan kombinasi 2 kata.

viii. Menampilkan 10 Bigram dengan Probabilitas paling tinggi

Melakukan sorting berdasarkan probabilitas paling tinggi, output :

```
10 model bigram berdasarkan probabilitas tertinggi : {('berkenalan', 'dengan'): 1.0, ('mengeksplorasi', 'material'): 1.0, ('anyar', 'mereka'): 1.0, ('bertajuk', 'space'): 1.0, ('pasca-konsumsi', '.'): 1.0, ('junk—mungkin', 'terinspirasi'): 1.0, ('banyaknya', 'sampah'): 1.0, ('debris', ')'): 1.0, ('sebab', 'koleksi'): 1.0, ('memadukan', 'praktik'): 1.0}
```

ix. Pengecekan Bigram kalimat uji

Untuk kalimat uji, akan dilakukan penggunaan fungsi lower() kemudian melakukan tokenisasi. Setelahnya dimulai penghitungan probabilitas Bigram dan nilai Perplexity, pengimplementasian *Laplace Smoothing* juga terjadi pada tahap ini. Output dari tahap ini akan memunculkan probabilitas Bigram (sebelum & setelah *Laplace Smoothing*) dan nilai Perplexity (sebelum & setelah *Laplace Smoothing*).

- Analisis

Dari 4 kalimat uji, terdapat 2 kalimat yang sesuai dengan topik artikel dan 2 kalimat yang tidak sesuai yang kemudian ke-empat kalimat ini akan dibandingkan nilainya baik probabilitas dan perplexity.

+ Kalimat yang sesuai dengan topik

```
['<s>', 'sepatu', 'menggunakan', 'bahan', 'dasar', 'sampah', 'garment', '.', '</s>']
probabilitas (sebelum laplace smoothing) :  1.3639134085474753e-08
probabilitas (setelah laplace smoothing) :  1.169210611777334e-19
nilai perplexity (sebelum laplace smoothing) :  7.480189115190088
nilai perplexity (setelah laplace smoothing) :  126.93093302314425

['<s>', 'nike', 'akan', 'mengeluarkan', 'lini', 'sneakers', 'sustainable', 'terbarunya', 'yang', 'bernama', 'space', 'hippie', '.', '</s>']
probabilitas (sebelum laplace smoothing) :  2.0066322187408216e-20
probabilitas (setelah laplace smoothing) :  3.667163984980548e-49
nilai perplexity (sebelum laplace smoothing) :  25.52504915282127
nilai perplexity (setelah laplace smoothing) :  2881.9781952153444
```

+ Kalimat yang tidak sesuai dengan topik

```
['<s>', 'nike', 'mematok', 'harga', 'yang', 'cukup', 'tinggi', 'untuk', 'rilisan', 'kali', 'ini', '.', '</s>']
probabilitas (sebelum laplace smoothing) :  0.0
probabilitas (setelah laplace smoothing) :  8.693387608326684e-81
nilai perplexity (sebelum laplace smoothing) :  39.901964675867276
nilai perplexity (setelah laplace smoothing) :  1440535.3486715513

['<s>', 'sepatu', 'adidas', 'akan', 'berkolaborasi', 'apabila', 'rilisan', 'ini', 'terbilang', 'sukses', '</s>']
probabilitas (sebelum laplace smoothing) :  0.0
probabilitas (setelah laplace smoothing) :  4.865953034132471e-111
nilai perplexity (sebelum laplace smoothing) :  39.901964675867276
nilai perplexity (setelah laplace smoothing) :  10676755088.249254
```

- Kesimpulan

Berdasarkan hasil output diatas dapat dilihat, untuk kalimat yang sesuai dengan topik memiliki nilai probabilitas bigram (baik sebelum/setelah *Laplace Smoothing*) memiliki nilai yang lebih besar serta nilai perplexity yang jauh lebih kecil apabila disandingkan dengan hasil nilai dari kalimat yang tidak sesuai dengan topik. Hal tersebut disebabkan karena pada kalimat yang sesuai dengan topik memiliki beberapa kata yang juga terdapat dalam corpus sehingga nilai probabilitas bigramnya akan semakin besar serta untuk nilai perplexitynya lebih kecil.