Jovita Nurvania (1301174215)

LAPORAN TUGAS NLP PEKAN 2 - MODEL BAHASA

I. Deskripsi Masalah

Model Bahasa atau yang biasa disebut dengan *language model* merupakan model yang memberikan nilai probabilitas pada rangkaian kata. Rangkaian kata tersebut biasa disebut dengan n-gram, dimana n merupakan jumlah dari kata contoh unigram (satu kata), bigram (dua kata) dan trigram (tiga kata). Tujuan dari pemodelan Bahasa ini adalah untuk mengetahui frekuensi munculnya sebuah kata dari serangkaian kalimat yang banyak. Selain itu, juga menghitung probabilitas kata tersebut muncul pada kalimat baru yang diberikan.

II. Penyelesaian Masalah

Pada penelitian ini, dilakukan pelatihan pada 20 file yang masing-masing filenya berasal dari artikel-artikel yang membahas tentang 'Pendidikan di masa pandemic covid-19'. Kemudian semua file tersebut digabung menjadi 1 file yang akan dijadikan sebagai data latih untuk mengetahui frekuensi unigram, probabilitas unigram, frekuensi bigram dan probabilitas bigram. Penghitungan tersebut dilakukan menggunakan bahasa algoritma python dan library nltk dan operator.

Setelah semua file digabung menjadi satu, lalu dilakukan *lowercasing* atau pengubahan huruf kapital menjadi huruf kecil semua pada isi file tersebut. Setelah itu, untuk memudahkan pengolahan, file ditokenisasi atau diberi pembatas antar kalimat menggunakan tag <s> di awal kalimat dan tag </s> di akhir kalimat. Kemudian baru dilakukan penghitungan frekuensi dan probabilitas.

Untuk menguji kalimat dari data latih tersebut, diberikan sebuah file yang berisi 6 kalimat, yaitu 3 kalimat yang sesuai dengan topik dan 3 kalimat yang tidak sesuai dengan topik. Kalimat yang dipilih adalah :

- Pada masa pandemi covid -19 seperti pada saat ini banyak pihak yang terdampak, salah satunya yaitu dunia pendidikan.
- Pendidikan jarak jauh tentu juga memiliki kekurangan terutama dalam hal pemahaman materi oleh peserta didik.
- Dan tentunya dengan adanya perubahan system pembelajaran membutuhkan fasilitas tambahan.
- Perusahaan minyak mengalami penurunan saham.
- Pilkada tahun ini tidak jadi diadakan.
- Harga bawang bombay naik tinggi.

Alasan dari pemilihan kalimat uji tersebut adalah untuk kalimat pertama dan kedua diambil kalimat yang persis dari artikel agar probabilitasnya tidak 0, kalimat ketiga diambil acak dari kata-kata yang telah ada di artikel dan kalimat keempat, kelima maupun keenam diambil acak kata yang berada di luar topik.

III. Analisis Hasil

Setelah semua nilai ada, dilakukan pemrosesan untuk mencari 10 unigram dengan frekuensi tertinggi dan 10 probabilitas bigram tertinggi. Dari hasil pengkodean seperti pada gambar, didapatkan 10 unigram yang paling sering muncul seperti di bawah ini.

```
[ ]  1  sorted_u = dict(sorted(freq_tab.items(), key=operator.itemgetter(1),reverse=True)[:10])
     2  print(sorted_u)
```

```
{'<s>': 817, '</s>': 817, '.': 813, 'yang': 718, ',': 699, 'dan': 455, 'di': 284, 'untuk': 283, 'dengan': 261, 'ini': 254}
```

Kemudian dilakukan pengkodean untuk mencari 10 nilai probabilitas bigram yang tertinggi,

```
1  sorted_b = dict(sorted(bigram_prob_tab.items(), key=operator.itemgetter(1),reverse=True)[:10])
2  print(sorted_b)
```

yang memiliki hasil : {('maraknya', 'pendidikan'): 1.0, ('musim', 'pandemi'): 1.0, ('satu-satunya', 'solusi'): 1.0, ('diliburkannya', 'intansi'): 1.0, ('intansi', 'pendidikan'): 1.0, ('tatap', 'muka'): 1.0, ('resiko', 'penyebaran'): 1.0, ('penjelasan', 'materi'): 1.0, ('ditekankan', 'demi'): 1.0, ('memudahkan', 'siswa'): 1.0}.

Dapat dilihat pada hasil running kode, bahwa dilakukan sebuah fungsi bernama *laplace smoothing*. Fungsi ini berguna untuk menambahkan satu nilai pada frekuensi bigram agar tidak adanya nilai 0 pada saat penghitungan probabilitas. Hasil running menampilkan jika sebelum dilakukan *laplace smoothing*, nilai probabilitas kalimat uji ketiga hingga keenam bernilai 0. Hal tersebut dikarenakan jika ada setidaknya satu saja ataupun lebih bigram yang tidak ada pada data latih. Oleh karena itu, diperlukan penambahan nilai satu pada setiap frekuensi bigram agar tidak adanya probabilitas yang bernilai 0.

Selain itu, analisis yang dilakukan pada laporan ini adalah perbandingan nilai perplexity dari kalimat uji yang sesuai topik dengan kalimat uji yang tidak sesuai dengan topik. Dapat dilihat pada hasil running program, bahwa kalimat pertama memiliki nilai perplexity terkecil dengan probabilitas terbesar karena kalimat tersebut berisi kata-kata yang sama dengan data latih. Kemudian pada kalimat kedua dan ketiga juga perplexity nya lebih kecil dari perplexity kalimat keempat, kelima dan keenam. Hal ini menyatakan bahwa, kalimat uji dengan topik yang sesuai lebih baik karena semakin kecil nilai perplexitynya suatu kalimat uji akan semakin bagus.

Pada program terdapat tambahan penghitungan frekuensi trigram, yang dimaksudkan untuk melihat keterkaitan antar tiga kata dan frekuensi nya tersebut dari serangkaian file yang ada.