

CONTOH SOLUSI TL-4

Nama : Muhammad Faisal Adi Soesatyo

NPM : 1906293184

1. Dua skenario model

a. Lowercasing (uncased)

- i. Lowercase: True
- ii. $n = 1$, Perplexity: 12.8372 (unigram)
- iii. $n = 2$, Perplexity: 10.9014 (bigram)
- iv. $n = 3$, Perplexity: 12.4165 (trigram)

b. No lowercasing (cased)

- i. Lowercase: False
- ii. $n = 1$, Perplexity: 12.9860 (unigram)
- iii. $n = 2$, Perplexity: 10.9727 (bigram)
- iv. $n = 3$, Perplexity: 12.4940 (trigram)

Apabila disajikan dalam bentuk tabel, berikut adalah ringkasannya,

| No. | Lowercasing (<i>uncased</i>) | No Lowercasing (<i>cased</i>) |
|-----|--------------------------------|---------------------------------|
| 1. | 12.8372 | 12.9860 |
| 2. | 10.9014 | 10.9727 |
| 3. | 12.4165 | 12.4940 |

Semakin rendah perplexity menandakan model tersebut bekerja semakin baik. Secara keseluruhan terlihat bahwa model yang dilakukan lowercasing sedikit lebih baik dibandingkan dengan tanpa lowercasing.

2. Proses yang dilakukan sebelum menggunakan *library* Aksara.

- a. Install foma.
- b. Mengingat OS yang digunakan saat ini adalah windows, maka yang saya lakukan adalah install WSL-Ubuntu untuk dapat menjalankan environment linux di atas windows.
- c. Dengan demikian install foma cukup menggunakan command `apt-get install foma-bin.`
- d. Membuat python virtual environment dengan command `python -m venv env.`

- e. Terakhir install *library* Aksara menggunakan command `pip install aksara`.
 - f. Pada link [user guide Aksara](#) terdapat tatacara menggunakan Aksara untuk melakukan tokenization. Terdapat dua jenis class, yakni BaseTokenizer serta MultiwordTokenizer. Pada langkah ini saya menggunakan MultiwordTokenizer untuk menangani kata-kata MWE di dalam bahasa Indonesia.
3. Generate sentence
 - a. Unigram
 - i. ...
 - b. Bigram
 - i. ...
4. Saya mendapatkan empat kata yang tidak terkandung di dalam *vocabulary*, yakni **peron**, **lincah**, **ustad**, **pekerti**. Contoh kalimat yang dapat dibangun dari OOV tersebut adalah
 - a. Unigram: <s> saya sedang menunggu di **peron** 5 stasiun tersebut </s>
 - b. Unigram: <s> para pekerja terlihat **lincah** saat membersihkan lokomotif tersebut </s>
 - c. Bigram: <s> <s> pak **ustad** berceramah di atas mimbar masjid </s>
 - d. Bigram: <s> <s> para murid diajarkan budi **pekerti** di sekolah </s>

Performa perplexity untuk masing-masing kalimat tersebut adalah,

 - ❖ n = 1, Perplexity: 1.5431
 - ❖ n = 1, Perplexity: 1.5431
 - ❖ n = 2, Perplexity: 10.2320
 - ❖ n = 2, Perplexity: 10.3767
5. Ukuran masing-masing vocab,
 - a. Unigram: 8,227
 - b. Bigram: 111,405
6. Terkait aspek analisis disesuaikan dengan hasil serta step yang dilakukan oleh masing-masing peserta. Hasil perplexity (menggunakan pendekatan logaritma) untuk masing-masing kalimat tersebut,
 - a. ...