Laporan Analisis Semantik Vektor

Untuk topik A disini saya menggunakan topik tentang kesehatan yaitu virus covid dan untuk topik B tentang bencana alam yaitu banjir khusus di Jakarta. Untuk sumber artikel topik A dari https://www.suara.com/tag/vaksin-virus-corona dan untuk topik B bersumber dari https://www.kompasiana.com/tag/banjir-awal-tahun-2020. Alasan pemilihan topik karena topik A dan Topik B tidak berkaitan dan cenderung berbanding terbalik.

1. **Ukuran matriks Tf-IDF, *co-occurrence term-context*, dan PPMI**

a. Ukuran matriks Tf-IDF

```
Ukuran Matriks TF-IDF (jummlah kata x jumlah kalimat) :  (200, 180)
Ukuran Matriks TF-IDF (jumlah kalimat x jumlah kata) :  (180, 200)
```

b. Ukuran matriks *co-occurrence term-context*

```
Ukuran Matriks co-occurence term-context : [ 1896 , 1896 ]
```

c. Ukuran matriks PPMI

```
Ukuran Matriks PPMI : [ 1896 , 1896 ]
```

2. Hasil Analisis Eksperimen

a.-Hitung berapa persen elemen matriks Tf-IDF yang bernilai tidak sama dengan 0

```
Persentase TF-IDF yang tidak bernilai 0 :  0.06863888888888889
```

Hasil persentasi 0.06863888888888889 atau sekitar 6.86 %

-Hitung berapa persen elemen matriks PPMI yang bernilai tidak sama dengan 0

```
Persentase PPMI yang tidak bernilai 0 :  0.54 %
```

b.- Cosine similarity antar dokumen dengan topik yang sama berdasarkan matriks TF-IDF

```
cos_sim_1 = np.dot(tf_idf_model[0],tf_idf_model[1])/np.linalg.norm(tf_idf_model[0])*np.linalg.norm(tf_idf_model[1])
print('Cosine Similarity Antar Dokumen (yang & banjir) Topik yang Sama (TF-IDF) :', str(cos_sim_1))
```

```
Cosine Similarity Antar Dokumen (yang & banjir) Topik yang Sama (TF-IDF) : 0.09656445576905319
```

- Cosine similarity antar dokumen dengan topik yang berbeda berdasarkan matriks TF-IDF

```
cos_sim_2 = np.dot(tf_idf_model[0],tf_idf_model[174])/np.linalg.norm(tf_idf_model[0])*np.linalg.norm(tf_idf_model[174])
print('Cosine Similarity Antar Dokumen (yang & baru) Topik yang Berbeda (TF-IDF) : ',str(cos_sim_2))
```

```
Cosine Similarity Antar Dokumen (yang & baru) Topik yang Berbeda (TF-IDF) :  0.002873172546588043
```

Berdasarkan analisis hasil dari cosine similarity, nilai dari cosine similarity antar dokumen dengan topik yang sama cenderung lebih tinggi yaitu sekitar 9.65 % dibandingkan dengan nilai cosine similarity yang berbeda topik yaitu sekitar 0.28 % karena keterkaitan makna kata dengan topik yang sama jauh lebih mirip dibandingkan dengan topik yang berbeda . Hasil nilai cosine similarity yang dipereloeh sudah sesuai karena seharusnya hasil antar dokumen dengan topik yang sama lebih baik karena memiliki similarity yang mirip.

c. -Cosine similarity antar kata dengan topik yang sama berdasarkan matriks TF-IDF

```
[64] cos_simm_01 = np.dot(tf_idf_modell[6],tf_idf_modell[19])/np.linalg.norm(tf_idf_modell[6])*np.linalg.norm(tf_idf_modell[19])
     print('Cosine Similarity Antar Kata (vaksin dan covid) Topik yang Sama (TF-IDF) : ', str(cos_simm_01))

     Cosine Similarity Antar Kata (vaksin dan covid) Topik yang Sama (TF-IDF) :  0.09496833009217388
```

```
[65] cos_simm_01 = np.dot(tf_idf_modell[3],tf_idf_modell[23])/np.linalg.norm(tf_idf_modell[3])*np.linalg.norm(tf_idf_modell[23])
     print('Cosine Similarity Antar Kata (banjir dan hujan) Topik yang Sama (TF-IDF) : ', str(cos_simm_01))

     Cosine Similarity Antar Kata (banjir dan hujan) Topik yang Sama (TF-IDF) :  0.05655249880692815
```

- Cosine similarity antar kata dengan topik yang beda berdasarkan matriks TF-IDF

```
[67] cos_simm_03 = np.dot(tf_idf_modell[3],tf_idf_modell[19])/np.linalg.norm(tf_idf_modell[3])*np.linalg.norm(tf_idf_modell[19])
     print('Cosine Similarity Antar Kata (banjir dan covid) Topik yang Berbeda (TF-IDF) : ', str(cos_simm_03))

     Cosine Similarity Antar Kata (banjir dan covid) Topik yang Berbeda (TF-IDF) :  0.00023438473517001452
```

```
[68] cos_simm_04 = np.dot(tf_idf_modell[22],tf_idf_modell[19])/np.linalg.norm(tf_idf_modell[22])*np.linalg.norm(tf_idf_modell[19])
     print('Cosine Similarity Antar Kata (hujan dan covid) Topik yang Berbeda (TF-IDF) : ', str(cos_simm_04))

     Cosine Similarity Antar Kata (hujan dan covid) Topik yang Berbeda (TF-IDF) :  0.0
```

Berdasarkan analisis hasil dari cosine similarity, nilai dari 2 percobaan perhitungan cosine similarity antar kata “vaksin” dengan “covid” dan “banjir” dengan “hujan” dengan topik yang sama cenderung lebih tinggi yaitu sekitar 9.49% dan 5.6% dibandingkan dengan nilai cosine similarity antar kata “banjir” dengan “covid” dan “hujan” dengan “covid” dengan topik yang berbeda yaitu sekitar 0.023% dan 0.0% karena keterkaitan makna kata dengan topik yang sama tinggi artinya lebih baik dan jauh lebih mirip dibandingkan dengan topik yang berbeda .

d. -Cosine similarity antar kata dengan topik yang sama berdasarkan matriks *co-occuraence* term-*context*

```
Cosine Similarity Antar Kata (vaksin dan covid) Topik yang Sama (co-occurrence term-context) :  0.570360356379574
Cosine Similarity Antar Kata (banjir & hujan) Topik yang Sama (co-occurrence term-context) :  0.36813638000392596
```

- Cosine similarity antar kata dengan topik yang beda berdasarkan matriks *co-occuraenc*e term-*context*

```
Cosine Similarity Antar Kata (banjir & covid) Topik yang Berbeda (co-occurrence term-context) :  0.06516504689359356
Cosine Similarity Antar Kata (hujan & covid) Topik yang Berbeda (co-occurrence term-context) :  0.06168274970199546
```

Berdasarkan hasil yang diperoleh dengan menggunakan *co-occurrence* term-*context*, nilai dari 2 percobaan perhitungan cosine similarity antar kata “vaksin” dengan “covid” dan “banjir” dengan “hujan” dengan topik yang sama cenderung lebih tinggi yaitu sekitar 57% dan 36% dibandingkan dengan nilai cosine similarity antar kata “banjir” dengan “covid” dan “hujan” dengan “covid” dengan topik yang berbeda yaitu sekitar 6.5% dan 6.1% karena keterkaitan makna kata dengan topik yang sama tinggi artinya lebih baik dan jauh lebih mirip dibandingkan dengan topik yang berbeda . Jika dibandingkan dengan cosine similarity dengan matriks TF-IDF cenderung lebih tinggi nilai cosine similarity dengan menggunakan matriks *co-occuraenc*e term-*context* artinya dengan menggunakan matriks *co-occuraenc*e term-*context lebih baik dibanding dengan matriks TF-IDF*

e. -Cosine similarity antar kata dengan topik yang sama berdasarkan matriks PPMI

```
PPMI antar kata (vaksin dan covid) :  3.6556
PPMI antar kata (banjir & hujan) :  None
```

-Cosine similarity antar kata dengan topik yang beda berdasarkan matriks PPMI

```
PPMI antar kata (banjir & covid) :  None
PPMI antar kata (hujan & covid) :  None
```

Berdasarkan hasil yang diperoleh dengan menggunakan PPMI, nilai dari 2 percobaan perhitungan cosine similarity antar kata “vaksin” dengan “covid” dan “banjir” dengan “hujan” dengan topik yang sama cenderung lebih tinggi yaitu sekitar 3.65 dan none dibandingkan dengan nilai cosine similarity antar kata “banjir” dengan “covid” dan “hujan” dengan “covid” dengan topik yang berbeda yaitu none untuk keduanya karena keterkaitan makna kata dengan topik yang sama tinggi artinya lebih baik dan jauh lebih mirip dibandingkan dengan topik yang berbeda . Jika dibandingkan dengan cosine similarity dengan PPMI cenderung lebih tinggi nilai cosine similarity dengan menggunakan matriks *co-occuraenc*e term-*context* artinya dengan menggunakan matriks *co-occuraenc*e term-*context lebih baik dibanding dengan matriks PPMI*