LAPORAN TUGAS PEKAN 3 NLP : SEMANTIK VEKTOR I (TF-IDF & PPMI)

CLARISA HASYA YUTIKA | 1301174256 | IF 41 GAB01

1. Deskripsi Masalah

Terdapat 20 file artikel berbahasa indonesia dengan topik ‘Cyber Crime Whatsapp’ dan ‘Pembelajaran Jarak Jauh’ yang diambil dari www.detik.com, kemudian akan dibuat matriks TF-IDF dan PPMI. Lalu akan dilakukan eksperimen menghitung nilai cosine similarity antar kalimat dan antar kata berdasakan matriks TF-IDF, menghitung nilai cosine similarity antar kata berdasarkan matriks co-occurrence term-context, dan menghitung nilai PPMI antar kata.

2. Perancangan Sistem

- Tokenisasi

  Tokenisasi bertujuan untuk memisahkan kata menjadi per token dari sebuah kalimat. Kemudian dilakukan case folding (mengubah semua kalimat menjadi lowercase, punctuation removal (penghilangan tanda baca).

- IDF

  IDF adalah menghitung frekuensi kata muncul dalam semua dokumen. Dalam eksperimen ini, sebuah dokumen adalah 1 kalimat. Semakin sedikit frekuensi kata muncul dalam semua dokumen, maka makin besar nilainya

- TF

  TF adalah menghitung frekuensi kata muncul dalam sebuah dokumen. Semakin banyak frekuensi kata muncul dalam sebuah dokumen, maka semakin besar nilainya.

- TF-IDF

  TF-IDF didapatkan dari perkalian nilai TF dan nilai IDF. Ukuran matriks TF-IDF yang didapat adalah 1544 x 344 (jum kata x jum kalimat).

```
Ukuran Matriks TF-IDF (jum kata x jum kalimat) :  (1544, 334)
Ukuran Matriks TF-IDF (jum kalimat x jum kata) :  (334, 1544)
```

  Persentase TF-IDF yang tidak bernilai 0 adalah 1.1%.

```
Persentase TF-IDF yang tidak bernilai 0 :  1.1 %
```

- Cosine Similarity

  Dihitung cosine similarity untuk mengukur similarity antar dua kalimat dan dua kata dari matriks TF-IDF.

- Co-Occurrence Term-Context

  Akan dihitung kemunculan term dengan term sebelum dan sesudah dengan jarak window sizenya yang sebesar 2. Ukuran matriks co-occurent term-context adalah 1544x1544.

```
Ukuran Matriks co-occurence term-context : [ 1544 , 1544 ]
```

- Cosine Similarity

  Dihitung cosine similarity antar dua kata dari matriks co-occurrence term-context.

- PPMI

  Akan menghitung dari dua kata yang muncul lebih sering secara bersamaan dibanding muncul sendiri-sendiri. Ukuran matriks PPMI adalah 1544x1544.

```
Ukuran Matriks PPMI : [ 1544 , 1544 ]
```

  Persentase PPMI yang tidak bernilai 0 adalah 0.68%.

```
Persentase PPMI yang tidak bernilai 0 :  0.68 %
```

LAPORAN TUGAS PEKAN 3 NLP : SEMANTIK VEKTOR I (TF-IDF & PPMI)

CLARISA HASYA YUTIKA | 1301174256 | IF 41 GAB01

3. Analisis

Akan dilakukan eskperimen menghitung nilai cosine similarity antar kalimat dan antar kata berdasarkan matriks TF-IDF, menghitung nilai cosine similarity antar kata berdasarkan matriks co-occurrence term-context, dan menghitung nilai PPMI antar kata.

- **Cosine Similarity Antar Kalimat Topik yang Sama (TF-IDF)**
  Didapatkan nilai cosine similarity sebesar 0.00597.

```
Antar Kalimat Topik yang Sama (TF-IDF) :  0.005967038675015867
```

- **Cosine Similarity Antar Kalimat Topik yang Berbeda (TF-IDF)**
  Didapatkan nilai cosine similarity sebesar 0.00176 dan 0.00079.

```
Antar Kalimat Topik yang Berbeda (TF-IDF) :  0.001755506663504508
Antar Kalimat Topik yang Berbeda (TF-IDF) :  0.0007868741002916099
```

- **Cosine Similarity Antar Kata Topik yang Sama (TF-IDF)**
  Didapatkan nilai cosine similarity sebesar 0.05984 (‘yang’& ‘ whatsapp’) dan 0.01002 (‘whatsapp’ & ‘menjadi’).

```
Antar Kata (yang & whatsapp) Topik yang Sama (TF-IDF) :  0.059836885406521
Antar Kata (whatsapp & menjadi) Topik yang Sama (TF-IDF) :  0.010012256604451828
```

- **Cosine Similarity Antar Kata Topik yang Berbeda (TF-IDF)**
  Didapatkan nilai cosine similarity sebesar 0.00089 (‘whatsapp’ & ‘pembelajaran’) dan 0.00663 (‘pembelajaran’ & ‘keamanan’).

```
Antar Kata (whatsapp & pembelajaran) Topik yang Berbeda (TF-IDF) :  0.000891772525581406
Antar Kata (pembelajaran & keamanan) Topik yang Berbeda (TF-IDF) :  0.006623689752983703
```

- **Cosine Similarity Antar Kata Topik yang Sama (Co-Occurrence Term-Context)**
  Didapatkan nilai cosine similarity sebesar 0.45004 (‘yang’ & ‘whatsapp’) dan 0.21676 (‘whatsapp’ & ‘menjadi’).

```
(yang & whatsapp) Topik yang Sama (co-occurrence term-context) :  0.450037090412966
(whatsapp & menjadi) Topik yang Sama (co-occurrence term-context) :  0.21676341625018133
```

- **Cosine Similarity Antar Kata Topik yang Berbeda (Co-Occurrence Term-Context)**
  Didapatkan nilai cosine similarity sebesar 0.1448 (‘whatsapp’ & ‘pembelajaran’) dan 0.09599 (‘pembelajaran’ & ‘keamanan’).

```
(whatsapp & pembelajaran) Topik yang Berbeda (co-occurrence term-context) :  0.1447996403889752
(pembelajaran & keamanan) Topik yang Berbeda (co-occurrence term-context) :  0.09599583352216755
```

- **PPMI Antar Kata Topik yang Sama**
  Didapatkan nilai PPMI sebesar 0 (‘yang’ & ‘whatsapp’) dan 0 (‘whatsapp’ & ‘menjadi’).

```
PPMI antar kata (yang & whatsapp) :  0
PPMI antar kata (whatsapp & menjadi) :  0
```

- **PPMI Antar Kata Topik yang Berbeda**
  Didapatkan nilai PPMI sebesar None (‘whatsapp’ & ‘pembelajaran’) dan None (‘pembelajaran’ & ‘keamanan’).

```
PPMI antar kata (whatsapp & pembelajaran) :  None
PPMI antar kata (pembelajaran & keamanan) :  None
```

4. Kesimpulan

Dari beberapa eksperimen yang dilakukan, nilai cosine similarity antar kalimat TF-IDF nilai tertinggi sebesar 0.00597, antar kata TF-IDF nilai tertinggi sebesar 0.05984, antar kata co-occurrence term-context nilai tertinggi sebesar 0.45004, maka artinya kalimat/kata dengan topik yang sama tingkat kemiripannya lebih tinggi. Nilai cosine similarity antar kata berdasarkan matriks co-occurrence term-context lebih tinggi nilainya. Nilai PPMI antar kata topik yang sama mendapat nilai sebesar 0, dan antar kata topik yang berbeda mendapat nilai None, maka kemungkinan dua kata muncul secara bersamaan sangat kecil dan tidak ada.