LAPORAN TUGAS PEKAN 4 NLP : SEMANTIK VEKTOR II (Word2Vec)

CLARISA HASYA YUTIKA | 1301174256 | IF 41 GAB01

1. Deskripsi Masalah

Membuat model Word2Vec Skip-Gram menggunakan library Gensim dari 100 artikel dengan topik ‘teknologi, ‘ekonomi’, dan ‘politik’ yang diambil dari ‘www.jawapos.com’. Akan dilakukan 2 eksperimen dengan jumlah minimal kemunculan kata = 1 dan 5 dengan panjang vector/embedding = 100. Kemudian analisis dari hasil similarity kata dan visualisasi embedding dari 2 eksperimen.

2. Perancangan Sistem & Analisis

- Akses representasi vektor / embedding kata ‘teknologi’ pada model 1 (min_count = 1)

```
vec_positif1 = model1.wv['teknologi']
print(vec_positif1)
```

```
[-0.00685375  0.00478964  0.00194519  0.03070565 -0.00094776  0.01783584
 -0.02084218  0.05146609  0.00267807  0.01404317 -0.00091093  0.01239648
 -0.01466186 -0.02786353 -0.02214168  0.00107465 -0.01329307 -0.0464562
  0.00407196  0.00861284  0.01917505  0.03408479 -0.00099023  0.01838489
 -0.01440252  0.02736284  0.00561425 -0.02014678 -0.04211006 -0.0364354
  0.07288069  0.02735718  0.01942938 -0.0163784   0.03180101 -0.00376875
 -0.00249093 -0.00818704  0.01266307 -0.02371916  0.01662255 -0.01428165
 -0.00552524  0.01137996 -0.01162544 -0.00290175 -0.02989191 -0.00634313
  0.01079823 -0.024909    0.01865147 -0.00316228  0.02353946  0.01186822
 -0.04222988  0.06870148  0.01245827 -0.01232711 -0.00518171 -0.0234368
  0.01273056  0.01064208 -0.00080555 -0.04749388  0.04496839 -0.00790497
  0.01790233 -0.03737689 -0.00449088  0.00242599 -0.01309448 -0.01278128
  0.02075754  0.01158209  0.07805014  0.03272307  0.05457792  0.00520986
 -0.0014665  -0.04003327  0.00519927  0.00506187  0.00107428  0.01472162
 -0.03028499 -0.00147267 -0.01153632 -0.030037    0.00127702  0.0005819
  0.01775246 -0.00969958  0.00352887  0.01492402 -0.00755771  0.05942191
  0.03600286  0.00314927  0.0122728  -0.02204679]
```

- Akses representasi vektor / embedding kata ‘teknologi’ pada model 2 (min_count = 5)

```
vec_positif5 = model5.wv['teknologi']
print(vec_positif5)
```

```
[-0.02312187  0.06308938  0.07708826  0.13448767 -0.01968004  0.0979932
 -0.1146163   0.32528126  0.07894277  0.12455557 -0.02311237  0.0371931
 -0.05965202 -0.07438656 -0.01229229 -0.02714612 -0.00318618 -0.2033544
  0.08664874  0.0473942   0.08398557  0.09773544  0.02195406  0.04603866
 -0.1117547   0.1111012   0.01373721 -0.12443458 -0.29607856 -0.23244345
  0.4236197   0.18043128  0.09659791 -0.07898902  0.1274325   0.01525243
 -0.0199323  -0.08372923  0.03899634 -0.12924615  0.09823347 -0.06446657
 -0.11660992  0.07039113 -0.05745212 -0.02117233 -0.20949703  0.07338731
  0.06200073 -0.11783011  0.08842351 -0.01162656  0.09164749  0.07827502
 -0.27041537  0.38588277  0.12304673 -0.1009803   0.03222509 -0.12432521
 -0.04081849 -0.02496979 -0.00285934 -0.23017256  0.3339516  -0.06558491
  0.10639589 -0.24503769 -0.02465929 -0.05024321 -0.08361017 -0.05808215
  0.12262812 -0.0391098   0.41466245  0.16750881  0.26076812  0.08535367
 -0.0178059  -0.17373359  0.05842921  0.06492977  0.05435499  0.03071477
 -0.15773493  0.02222319 -0.06885548 -0.17593859 -0.00434358  0.01128037
  0.1672895   0.04767548  0.04131397  0.06765378 -0.03045863  0.34537452
  0.22383584  0.07131851 -0.02156755 -0.10414158]
```

- Similarity antar kata pada model 1 (min_count=1)
  - Similarity > 0,5

```
print(model1.wv.similarity('pemerintah', 'politik'))
```

```
0.9806969
```

  - 0 < Similarity < 0,5

```
print(model1.wv.similarity('mengatur', 'harga'))
```

```
0.4358218
```

  - -1 < Similarity < -0,5

```
print(model1.wv.similarity('mengoptimalkan', 'news'))
```

```
0.16660695
```

- Similarity antar kata pada model 2 (min_count=5)
  - Similarity > 0,5

```
print(model5.wv.similarity('pembangunan', 'kerja'))
```

```
0.99971503
```

  - 0 < Similarity < 0,5

```
print(model5.wv.similarity('teknologi', 'telkomsel'))
```

```
0.9958548
```

  - -1 < Similarity < -0,5

```
print(model5.wv.similarity('memanfaatkan', 'singapura'))
```

```
0.99026674
```

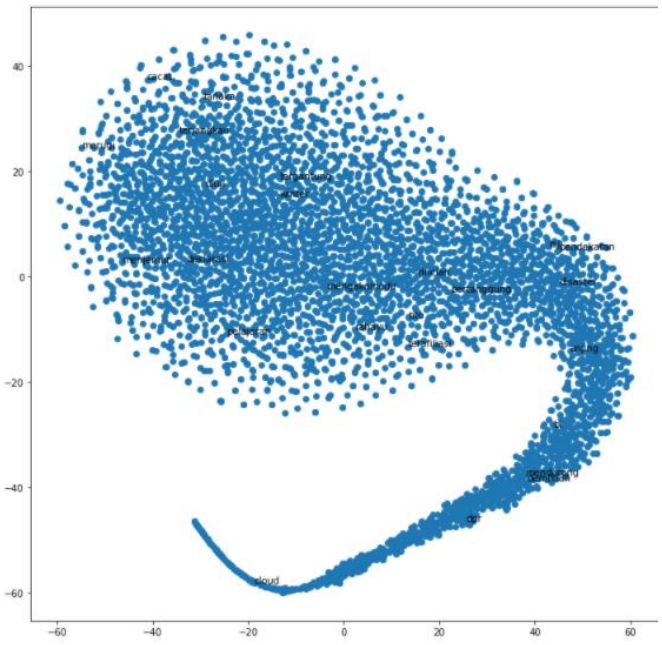
- Top 5 kata ‘teknologi’ yang similar dengan sebuah kata tertentu pada model 1 (min_count = 1)

```
[('bahwa', 0.9936176538467407), ('bisa', 0.9935950040817261)
('tidak', 0.9935855269432068), ('ke', 0.9935356378555298)
('di', 0.9934906959533691)]
```

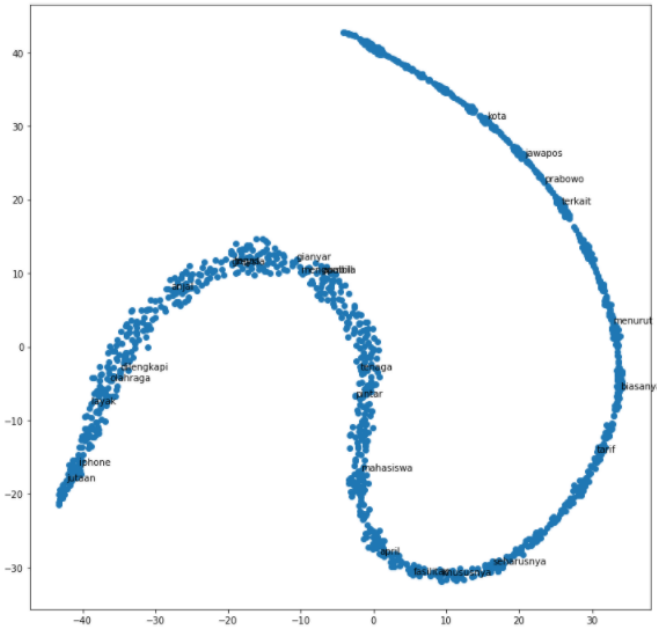
- Top 5 kata ‘teknologi’ yang similar dengan sebuah kata tertentu pada model 2 (min_count = 5)

```
[('di', 0.9997941255569458), ('bisa', 0.9997861385345459)
('tidak', 0.9997852444648743), ('bahwa', 0.9997849464416504)
('yang', 0.9997818470001221)]
```

- Visualisasi embedding pada model 1 (min_count = 1)



- Visualisasi embedding pada model 2 (min_count = 5)



3. Kesimpulan

- Dari percobaan mencari similarity antar kata, pada model 1 didapatkan pasangan kata yang nilai similarity nya > 0,5 dan 0 < similarity < 0,5 , tetapi tidak didapatkan pasangan kata yang nilai similaritynya antara -0,5 dan -1. Kemudian pada model 2 hanya didapatkan nilai similarity > 0,5 , hal ini dikarenakan pada model 2 setting kemunculan kata minimal 5, sehingga kemungkinan kemunculan sebuah kata dengan kata yang lain akan lebih tinggi dan nilai similaritynya pun lebih tinggi.
- Dari percobaan mencari 5 nilai similarity tertinggi dari kata ‘teknologi’, terdapat beberapa perbedaan. Pada model 1 kata tertinggi similaritynya adalah ‘bahwa’, sedangkan pada model 2 adalah ‘di’. Hal ini terjadi karena perbedaan kemunculan kata, pada model 1 lebih bebas dalam menghitung nilai similarity karena hanya sedikit kata yang di drop pada saat pemodelan karna hanya memliki jumlah minimum kemunculan kata 1.
- Dari hasil visualisasi embedding pada model 1 dan 2. Dapat dilihat bahwa sebaran kata pada model 1 lebih tersebar dibandingkan model 2. Hal ini dapat disimpulkan bahwa semakin besar kemunculan kata, maka nilai similarity akan semakin besar juga karena jarak antara sebaran data kecil.