Jovita Nurvania (11301174215)

LAPORAN TUGAS NLP PEKAN 4 – WORD2VEC

I. Deskripsi Masalah

Word2vec adalah sebuah metode yang dapat merepresentasikan setiap kata yang ada dalam suatu kalimat sebagai vector dengan N dimensi. Dengan word2vec kita dapat menghitung *semantic similarity* dari setiap kata yang membentuk *one-hot encoding.* Hasil dari hitungan ini dapat merepresentasikan relasi suatu kata dengan kata lainnya. Model word2vec yang digunakan adalah skip-gram.

II. Penyelesaian Masalah

Pada penelitian ini, dilakukan pelatihan pada 100 artikel berbahasa Indonesia yang memiliki topik ekonomi. Kemudian 100 artikel tersebut digabung menjadi satu teks untuk diolah. Baru setelah digabung menjadi satu corpus file tersebut diolah dengan gensim dan word2vec, untuk proses pertama dilakukan dua eksperimen yaitu pertama jumlah minimal kemunculan katanya adalah satu yang kedua jumlah minimal kemunculan katanya adalah lima.

Setelah itu, dicari vector positif dari suatu kata dengan model pertama minimal jumlah katanya 1 dan model kedua minimal kemunculan katanya 5. Setelah itu dicari nilai similarity antar dua kata dan dicari kata dengan similarity terbesar sebanyak lima kata. Tahap terakhir yang dilakukan adalah visualisasi data menjadi bentuk grafik seperti pada code.

III. Analisis Hasil

Nomor 1 :

```
#min_count = 1
vec_positif1 = model1.wv['ekonomi']
print(vec_positif1)
```

```
[-0.21821296  0.3505536  -0.30365995  0.03594282 -0.39777398 -0.18146995
 -0.1009039  -0.01999632  0.22065525 -0.23756807 -0.08133017 -0.07669984
 -0.13474132 -0.5978026  -0.5445283  -0.23687656  0.5387245   0.57738703
 -0.15306844  0.149465    0.1597305  -0.12001058  0.41217947 -0.4590664
  0.13140973  0.6599007  -0.18539003 -0.02951195  0.12551582  0.04167398
 -0.4205641   0.1777522  -0.1548502   0.21892    -0.54290384  0.40446258
 -0.24227934 -0.30001634 -0.68408173  0.29322794  0.00553682  0.21306384
  0.26753575  0.49201834 -0.5085635  -0.08168729  0.3227569  -0.30992228
 -0.22441635 -0.10493813 -0.15239745 -0.65633565  0.12441855  0.10252198
  0.2521341   0.21551757 -0.5675029   0.16329741 -0.28190848 -0.4118242
 -0.14842317 -0.07635441 -0.04465564  0.35404068  0.2066088   0.09666917
 -0.2259598   0.17698808  0.11196317 -0.02173929  0.170759   -0.35852432
 -0.6427907  -0.00838329 -0.19760807 -0.29160914 -0.05231457  0.07038242
  0.04887882 -0.17168775 -0.129588   -0.07256512  0.05192376  0.05731605
 -0.17475553 -0.04288153 -0.31087035  0.27118182  0.49028823  0.27754012
  0.07100399 -0.06835267  0.10937974 -0.14680715 -0.22384177 -0.10398419
  0.38796678 -0.1262463  -0.3141286  -0.31537792]
```

```
#min_count = 5
vec_positif5 = model5.wv['ekonomi']
print(vec_positif5)
```

```
[-0.10401885  0.25808173 -0.27564526  0.07879519 -0.31633368 -0.19018741
 -0.10621651  0.17689762  0.29895383 -0.20306955 -0.1964491  -0.0264108
 -0.17129791 -0.5054012  -0.42924115 -0.17439957  0.485274    0.5296832
 -0.2322301   0.00982699  0.09204915 -0.02622337  0.33420202 -0.46430928
  0.05036088  0.6067716  -0.22873603  0.0325788   0.03561233  0.0594497
 -0.36142978  0.1248097  -0.11495671  0.1617228  -0.44043076  0.38620922
 -0.26495126 -0.30872956 -0.6361669   0.2063882   0.02627122  0.20369332
  0.10335957  0.34955385 -0.437807   -0.0523992   0.27133983 -0.2079585
 -0.18070498 -0.05151795 -0.06554385 -0.6073037   0.18912539  0.0461653
  0.32630682  0.14798234 -0.53048295  0.06352374 -0.26112863 -0.2681246
 -0.12731785 -0.11330143  0.03055774  0.40362754  0.1799301   0.15578309
 -0.18316017  0.17481469  0.12797113 -0.05557889  0.16753331 -0.30586836
 -0.58461463  0.08360689 -0.13684866 -0.23593971  0.00138913  0.10663519
 -0.00235818 -0.1280295  -0.02371704  0.0509578   0.06737821 -0.01169784
 -0.11827686 -0.05551163 -0.13319728  0.24962416  0.5513489   0.24993922
 -0.09496875 -0.06747966  0.10600298 -0.09230137 -0.1602607  -0.14978781
  0.2270841  -0.27208647 -0.14960162 -0.3487938 ]
```

Terdapat vector dari kata ‘ekonomi’ yang pertama adalah hasil dari perhitungan dengan minimal kemunculan kata sekali dan yang kedua dengan minimal kemunculan kada

lima kali. Yang berbeda di antara kedua model hanyalah nilai vektornya, untuk bentuk dan positif negatifnya sama saja.

Nomor 2 :

min_count = 1

```
[ ]  1  #Similarity > 0,5
     2  print(model1.wv.similarity('kebijakan', 'ekonomi'))
```

```
0.9994686
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Fu
  if np.issubdtype(vec.dtype, np.int):
```

```
[ ]  1  # 0 < Similarity < 0,5
     2  print(model1.wv.similarity('keuangan', 'dipertimbangkan'))
```

```
0.33667386
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Fu
  if np.issubdtype(vec.dtype, np.int):
```

```
[ ]  1  # -1 < Similarity < -0,5
     2  print(model1.wv.similarity('kompetisi', 'berkala'))
```

```
-0.7613075
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: Fu
  if np.issubdtype(vec.dtype, np.int):
```

min_count = 5

```
[ ]  1  #Similarity > 0,5
     2  print(model5.wv.similarity('uang', 'tunai'))
```

```
0.995443
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:7
  if np.issubdtype(vec.dtype, np.int):
```

```
[ ]  1  # 0 < Similarity < 0,5
     2  print(model5.wv.similarity('peringatan', 'ramadhan'))
```

```
0.9713444
/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:7
  if np.issubdtype(vec.dtype, np.int):
```

```
[ ]  1  # -1 < Similarity < -0,5
     2  print(model1.wv.similarity('asuransi', 'kotor'))
```

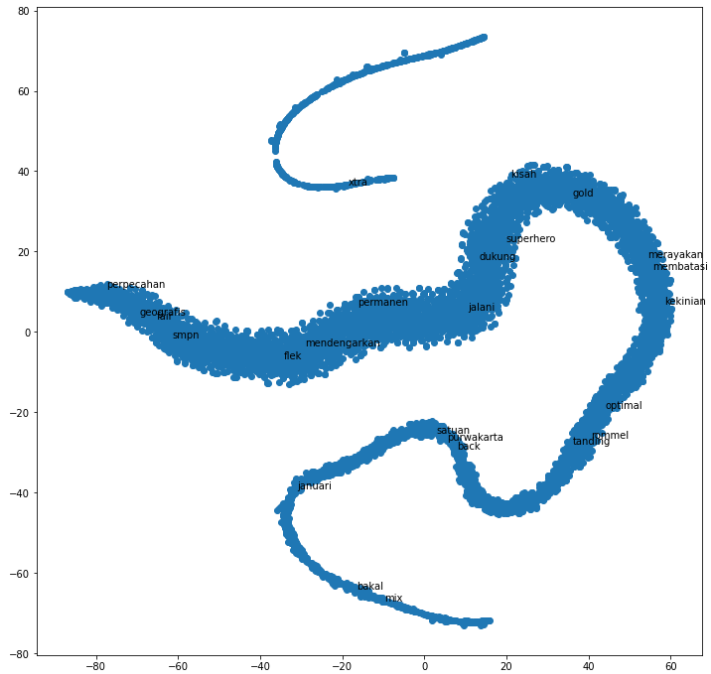
```
0.9839549
```

Nilai similarity model 1 sudah sesuai dengan perkiraan di awal sedangkan untuk model 2 tidak. Pada model 2 similarity antar katanya bernilai 0,9 semua karena pada model ini menghitung minimal kemunculan kata lima kali, jadi lebih besar kemungkinannya kedua kata bertemu dan memiliki nilai similarity yang tinggi dibandingkan dengan yang minimal kemunculannya satu.

Nomor 3 :

Lima kata teratas dengan hitungan tertinggi untuk minimal kemunculan satu adalah kata ; untuk, dan, seperti, sebagai dan belum. Untuk minimal kemunculan lima adalah kata; untuk, namun, seperti, bisa dan jika.

Nomor 4 :



Penyebaran data sesuai dengan kata dan kemunculan kata juga nilai similaritynya jika divisualisasikan berbentuk seperti gambar di atas. Persebaran datanya tidak merata, banyak kata yang nilainya sama oleh karena itu penyebaran data berdempetan seperti gambar.