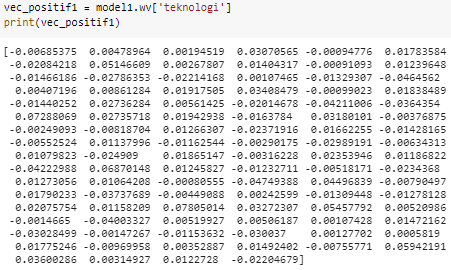
1. **Deskripsi Masalah**

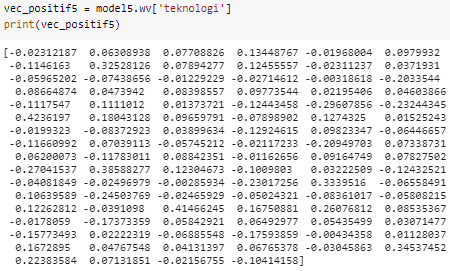
Membuat model Word2Vec Skip-Gram menggunakan library Gensim dari 100 artikel dengan topik ‘teknologi, ‘ekonomi’, dan ‘politik’ yang diambil dari ‘www.jawapos.com’. Akan dilakukan 2 eksperimen dengan jumlah minimal kemunculan kata = 1 dan 5 dengan panjang vector/embedding = 100. Kemudian analisis dari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

1. **Perancangan Sistem & Analisis**

* Akses representasi vektor / embedding kata ‘teknologi’ pada model 1 (min\_count = 1)

****

* Akses representasi vektor / embedding kata ‘teknologi’ pada model 2 (min\_count = 5)

****

* Similarity antar kata pada model 1 (min\_count=1)
  + Similarity > 0,5

****

* + 0 < Similarity < 0,5

****

* + -1 < Similarity < -0,5

****

* Similarity antar kata pada model 2 (min\_count=5)
  + Similarity > 0,5

****

* + 0 < Similarity < 0,5

****

* + -1 < Similarity < -0,5

****

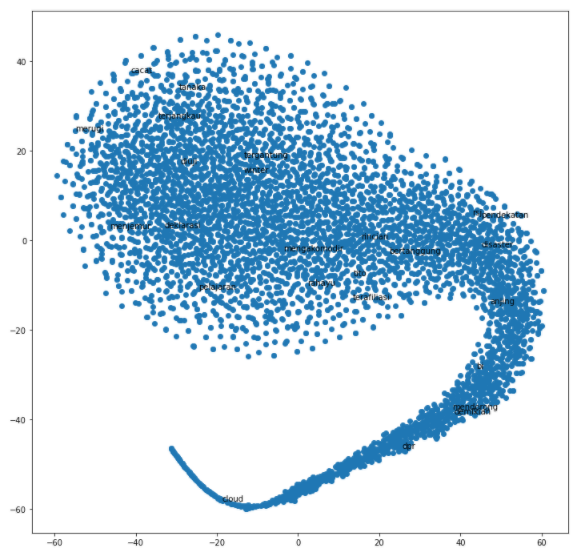
* Top 5 kata ‘teknologi’ yang similar dengan sebuah kata tertentu pada model 1 (min\_count = 1)

****

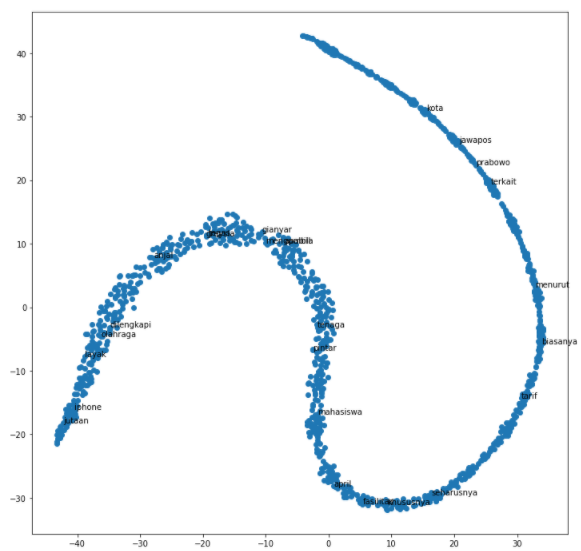
* Top 5 kata ‘teknologi’ yang similar dengan sebuah kata tertentu pada model 2 (min\_count = 5)

****

* Visualisasi embedding pada model 1 (min\_count = 1)

****

* Visualisasi embedding pada model 2 (min\_count = 5)

****

1. **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.