Tutorial: Word2Vec + Gensim + Wikidump-id

Oleh : Abdiansah
Bidang : Pemrosesan Bahasa Alami
Tingkat : [Pemula / ~~Menengah~~ / ~~Ahli~~]
Alat : Python, Gensim

Tutorial ini akan membahas pembuatan model *word2vec*[1] menggunakan pustaka Gensim[2] dengan data set yang bersumber dari *wikipedia dump*[3] bahasa Indonesia. Tutorial ini bertujuan untuk memberikan langkah-langkah praktis sehingga mudah diduplikasi oleh pemula. Tutorial terbagi menjadi tiga tahapan, yaitu:

1. Mengambil dan membuat data set wikipedia berbahasa Indonesia.
2. Melatih dan menyimpan model *word2vec* menggunakan pustaka gensim.
3. Mengambil dan mengetes model *word2vec* menggunakan pustaka gensim.

Data dan kode program yang sudah jadi dapat diunduh di github[4].

Mengambil dan membuat data set wikipedia berbahasa Indonesia

1. Kunjungi alamat URL wikipedia dump bahasa Indonesia di sini:
   - *https://dumps.wikimedia.org/idwiki/latest/* (akses tanggal 20 April 2020)
2. Cari dan unduh file ***idwiki-latest-pages-articles.xml.bz2*** (file kompresi ~ 558 MB)
3. Ekstrak file *idwiki-latest-pages-articles.xml.bz2* sehingga diperoleh file ***idwiki-latest-pages-articles.xml***
4. File *idwiki-latest-pages-articles.xml* bertipe xml yang masih mengandung tag-tag xml. Untuk mengambil teks artikel kita menggunakan bantuan program **WikiExtractor**[5]. Berikut langkah-langkahnya:
   - Unduh file *WikiExtractor.py* kemudian letakkan satu direktori dengan file *idwiki-latest-pages-articles.xml*
   - Format perintah: **python** WikiExtractor.py -o [*nama direktori*] [*file wiki dump xml*]
   - Buka konsol dan jalankan **Python** (**versi 2**)
   - Eksekusi: **python** WikiExtractor.py -o **wiki_dump** *idwiki-latest-pages-articles.xml*
   - Hasil eksekusi di atas adalah sebuah direktori bernama **wiki_dump** yang berisi lagi 6 (enam) sub-direktori (AA, AB, AC, AD, AE, AF) dengan setiap sub-direktori berisi file-file teks berukuran mirip (~ 1 MB).
5. Antar sub-direktori memiliki nama file yang sama sehingga membutuhkan waktu jika nama file diubah satu-persatu. Oleh karena itu nama file akan diubah secara otomatis menggunakan program **RenameFiles** (filenya ada di github). Berikut langkah-langkahnya:

1 https://en.wikipedia.org/wiki/Word2vec
2 https://radimrehurek.com/gensim/models/word2vec.html
3 https://dumps.wikimedia.org/idwiki/latest/
4 https://github.com/abdiansah/tutorial-word2vec
5 https://github.com/attardi/wikiextractor

- Unduh file *RenameFiles.py* kemudian letakkan satu direktori dengan wiki_dump
- Format perintah: **python** RenameFiles.py [*path direktori*] [*nama inisial file*]
- Buka konsol dan jalankan **Python**
- Eksekusi: **python** RenameFiles.py ./wiki_dump/AA/ **a**
- Eksekusi: **python** RenameFiles.py ./wiki_dump/AB/ **b**
- Eksekusi: **python** RenameFiles.py ./wiki_dump/AC/ **c**
- Eksekusi: **python** RenameFiles.py ./wiki_dump/AD/ **d**
- Eksekusi: **python** RenameFiles.py ./wiki_dump/AE/ **e**
- Eksekusi: **python** RenameFiles.py ./wiki_dump/AF/ **f**
- Hasilnya adalah nama file setiap sub-direktori sudah berbeda-beda. Selanjutnya buat direktori dengan nama **dataset_wiki**, kemudian pindahkan seluruh file teks dalam direktori wiki_dump ke dataset_wiki.

6. Direktori **dataset_wiki** (592 file - 595 MB) akan menjadi data latih untuk model *word2vec*.

Melatih dan menyimpan model word2vec menggunakan pustaka gensim

1. Program **TrainW2V** digunakan untuk melatih dan menyimpan model word2vec. Program tersebut dibuat menggunakan pustaka gensim dan dapat diunduh di github.
2. Berikut langkah-langkah melatih model word2vec menggunakan data set wikipedia yang dibuat pada tahap sebelumnya:
   - Unduh file *TrainW2V.py* kemudian letakkan satu direktori dengan dataset_wiki
   - Format perintah: **python** TrainW2V.py [*path dataset*] [*nama model word2vec*]
   - Buka konsol dan jalankan **Python**
   - Buat **virtual environment** (venv) kemudian install **gensim**
   - Eksekusi: **python** TrainW2V.py ./dataset_wiki/ **model_w2v.bin**
3. Hasil pelatihan akan diperoleh file model word2vec bernama **model_w2v.bin**

Mengambil dan mengetes model word2vec menggunakan pustaka gensim

1. Program **TesW2V** digunakan untuk mengambil dan mengetes model word2vec. Program tersebut dibuat menggunakan pustaka gensim dan dapat diunduh di github.
2. Berikut langkah-langkah mengetes model word2vec:
   - Unduh file *TesW2V.py* kemudian letakkan satu direktori dengan file **model_w2v.bin**
   - Format perintah: **python** TesW2V.py [*nama model word2vec*]
   - Isi kata yang akan dicari kemiripannya
   - Isi ‘qqq’ untuk keluar program
3. Selesai.