Rancang Bangun Sistem Pencarian Teks dengan Menggunakan Model Continuous-Bag-of-Words model dan Continuous Skip-Gram Model pada Koleksi Dokumen

Preparation

1. download https://github.com/fathanq/web-crawler/
2. import skripsi.sql
3. install dependencies (python -m pip install -r requirements.txt)
4. change max_allowed_packet inside my.ini on section [mysqld] according to your data size (modified_crawling.py sekali run (default setting) bisa menghasilkan sekitar 150MB-an dictionary)
5. run modified_crawling.py
6. run onehotencode.py (memindahkan data dari data yang di crawl modified crawl (dbcrawl) ke database skripsi, menghapus stopwords, merapihkan data, lalu membuat one hot encode)

Training

Training data di batasi 3 jam

- run training.py untuk melatih data di database skripsi dengan metode Skip-Gram
- run trainingcbow.py untuk melatih data di database skripsi dengan metode CBOW dengan satu tetangga (dua kata)
- run trainingcbow4kata.py untuk melatih data di database skripsi dengan metode CBOW dengan dua tetangga (empat kata)

Searching

Searching akan dilakukan dengan menggabungkan kata input user dengan kata input paling relevan keluaran model masing-masing
Searching dibatasi lima hasil keluaran (jika hasil keluar kurang dari lima maka sisanya akan diisi data random dari database)
Data perlu di train terlebih dahulu dengan metode yang sama sebelum melakukan search
Search dilakukan dengan exact string match (regular expression)

- **run search.py untuk mencari dokumen relevan dengan menggunakan metode Skip-Gram**
  search.py menerima satu kata input
  pada search.py searching akan dilakukan dengan mencari kata:
  1. "(kata relevan 1) (kata input) (kata relevan 2)"
  2. "(kata relevan 2) (kata input) (kata relevan 1)"

- **run search2cbow.py untuk mencari dokumen relevan dengan menggunakan metode CBOW 2 kata**
  search2cbow.py menerima dua kata input
  pada search2cbow.py searching akan dilakukan dengan mencari kata:
  1. "(kata input 1) (kata relevan) (kata input 2)"
  2. "(kata input 2) (kata relevan) (kata input 1)"

- **run search4cbow.py untuk mencari dokumen relevan dengan menggunakan metode CBOW 4 kata**
  search4cbow.py menerima empat kata input
  pada search4cbow.py searching akan dilakukan dengan mencari kata:
  "(kata input 1) (kata input 2) (kata relevan) (kata input 3) (kata input 4)"