LAPORAN AKHIR PRAKTIKUM

Mata Praktikum	: Kecerdasan Buatan
Kelas	: 3IA13
Praktikum ke-	: 4
Tanggal	: 21 Januari 2023
Materi	: Natural Language Processing
NPM	: 51420249
Nama	: Tiara Puspita
Ketua Asisten	: David
Jumlah Lembar	: 4

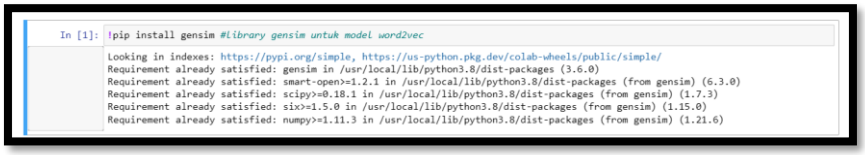
LABORATORIUM TEKNIK INFORMATIKA

UNIVERSITAS GUNADARMA

2022

Isi Laporan

1. Gensim adalah library Python untuk topic modelling, document indexing dan similarity retrieval dengan kumpulan data yang sangat besar. Install library gensim dengan kode berikut:



```
In [1]: !pip install gensim #library gensim untuk model word2vec

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: gensim in /usr/local/lib/python3.8/dist-packages (3.6.0)
Requirement already satisfied: smart-open>=1.2.1 in /usr/local/lib/python3.8/dist-packages (from gensim) (6.3.0)
Requirement already satisfied: scipy>=0.18.1 in /usr/local/lib/python3.8/dist-packages (from gensim) (1.7.3)
Requirement already satisfied: six>=1.5.0 in /usr/local/lib/python3.8/dist-packages (from gensim) (1.15.0)
Requirement already satisfied: numpy>=1.11.3 in /usr/local/lib/python3.8/dist-packages (from gensim) (1.21.6)
```

Lalu import library yang kan digunakan pada program

```
In [2]: import io
        import time
        from datetime import timedelta
        import gensim
```

2. Load dataset yang di download dengan menggunakan link yang disediakan pada forum.

```
!wget https://www.dropbox.com/s/exe4yv8rfi6hclf/ind_news_2020_1M.tar.gz?dl=1 -O corpus.tar.gz

--2023-01-14 14:44:56--  https://www.dropbox.com/s/exe4yv8rfi6hclf/ind_news_2020_1M.tar.gz?dl=1
Resolving www.dropbox.com (www.dropbox.com)... 162.125.65.18, 2620:100:6021:18::a27d:4112
Connecting to www.dropbox.com (www.dropbox.com)|162.125.65.18|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: /s/dl/exe4yv8rfi6hclf/ind_news_2020_1M.tar.gz [following]
--2023-01-14 14:44:56--  https://www.dropbox.com/s/dl/exe4yv8rfi6hclf/ind_news_2020_1M.tar.gz
Reusing existing connection to www.dropbox.com:443.
HTTP request sent, awaiting response... 302 Found
Location: https://uc6a2274f297145fc51e83a57684.dl.dropboxusercontent.com/cd/0/get/B0i9MdMFNBLuC4cJ-NRn1XOcwdEd81YfB8KV-PHDZBlZgid48MWP3waA6VeGu70pi0ZQQkv80Jj66dEf9m_p0BDmbbrpoPoKMuKLcgw_j776w8vGV45BCZhImFAw2whNtdEUVFv8nto4zva69bNt-6BFxKvSPDh5tVLzel_4FC5fFjNcYrRDu0QiNfdHabuiVzs/file?dl=1# [following]
--2023-01-14 14:44:57--  https://uc6a2274f297145fc51e83a57684.dl.dropboxusercontent.com/cd/0/get/B0i9MdMFNBLuC4cJ-NRn1XOcwdEd81YfB8KV-PHDZBlZgid48MWP3waA6VeGu70pi0ZQQkv80Jj66dEf9m_p0BDmbbrpoPoKMuKLcgw_j776w8vGV45BCZhImFAw2whNtdEUVFv8nto4zva69bNt-6BFxKvSPDh5tVLzel_4FC5fFjNcYrRDu0QiNfdHabuiVzs/file?dl=1
Resolving uc6a2274f297145fc51e83a57684.dl.dropboxusercontent.com (uc6a2274f297145fc51e83a57684.dl.dropboxusercontent.com)... 162.125.65.15, 2620:100:6027:15::a27d:480f
Connecting to uc6a2274f297145fc51e83a57684.dl.dropboxusercontent.com (uc6a2274f297145fc51e83a57684.dl.dropboxusercontent.com)|162.125.65.15|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 224376751 (214M) [application/binary]
Saving to: ‘corpus.tar.gz’

corpus.tar.gz       100%[===================>] 213.98M  22.1MB/s    in 10s

2023-01-14 14:45:08 (21.2 MB/s) - ‘corpus.tar.gz’ saved [224376751/224376751]
```

Setelah itu, data yang di download harus di ekstrak. Berikut merupakan syntax yang digunakan untuk mengekstrak data dan juga list dari dari setelah di ekstrak:

```
In [4]: !tar -xzvf corpus.tar.gz

ind_news_2020_1M/
ind_news_2020_1M/ind_news_2020_1M-sentences.txt
ind_news_2020_1M/ind_news_2020_1M-sources.txt
ind_news_2020_1M/ind_news_2020_1M-words.txt
ind_news_2020_1M/ind_news_2020_1M-inv_so.txt
ind_news_2020_1M/ind_news_2020_1M-inv_w.txt
ind_news_2020_1M/ind_news_2020_1M-import.sql
ind_news_2020_1M/ind_news_2020_1M-co_s.txt
ind_news_2020_1M/ind_news_2020_1M-co_n.txt
```

3. Selanjutnya adalah tahap data preprocessing dan cleansing. Pada tahap ini dataset akan dimasukkan kedalam satu file yang diberi nama corpus.txt dna diubah menjadi artikel yang akan digunakan sebagai kamus data pada program. Setelah itu, data akan dibersihkan dengan menghapus tanda baca sehingga hanya tersisa kata nya saja.

```
In [5]: corpora = gensim.corpora.TextCorpus('ind_news_2020_1M/ind_news_2020_1M-sentences.txt')
        article_count = 0
        with io.open('corpus.txt', 'a') as wiki_text:
          for text in corpora.get_texts():

            wiki_text.write(" ".join(map(str, text)) + '\n')
            article_count +=1

            if article_count % 10000 == 0:
              print('{} article processed'.format(article_count))

            print('total: {} articles'.format(article_count))

total: 995004 articles
total: 995005 articles
total: 995006 articles
total: 995007 articles
total: 995008 articles
total: 995009 articles
total: 995010 articles
total: 995011 articles
total: 995012 articles
total: 995013 articles
total: 995014 articles
total: 995015 articles
total: 995016 articles
total: 995017 articles
total: 995018 articles
total: 995019 articles
total: 995020 articles
total: 995021 articles
total: 995022 articles
```

4. Training model word2vec. Disini menggunakan fungsi LineSentence dimana merupakan sebuah fungsi yang digunakan untuk melakukan iterasi pada file yang memiliki banyak kalimat, sehingga menjadi satu baris satu kalimat. Setelah itu masukkan kedalam variable id_w2v dengan dimensinya itu 300 dan workers 16 dan save.

```
In [6]: import time
        import multiprocessing
        from datetime import timedelta

        from gensim.models import word2vec

        start_time = time.time()
        print('Training Word2Vec Model...')
        sentences = word2vec.LineSentence('corpus.txt')
        id_w2v = word2vec.Word2Vec(sentences, size=300, workers=16)
        id_w2v.save('model_word2vec_300_model')
        finish_time = time.time()

        print('Finished. Elapsed time : {}'
              .format(timedelta(seconds=finish_time-start_time)))

Training Word2Vec Model...
```

5. Setelah dilakukan model training, kitab isa menggunakan model. Disini model digunakan untuk mencari korelasi antar kata dengan ditampilkan juga persentase keakuratan nya. Selain itu, model juga digunakan untuk menemukan kata yang tidak memiliki korelasi dengan kumpulan kata yang dibuat menjadi string.

```
In [7]: id_w2v.wv.similarity('wanita', 'pria')

Out[7]: 0.8684156

In [8]: id_w2v.wv.most_similar('biru')

Out[8]: [('pink', 0.7926267385482788),
         ('berwarna', 0.7829585075378418),
         ('coklat', 0.7751073241233826),
         ('warni', 0.7713457345962524),
         ('hitam', 0.7712019681930542),
         ('metalik', 0.7120261192321777),
         ('nude', 0.7114129066467285),
         ('berkelir', 0.709497332572937),
         ('krem', 0.7094334959983826),
         ('dicat', 0.7065083384513855)]

In [10]: id_w2v.wv.most_similar(positive=['wanita', 'raja'], negative=['pria'])

Out[10]: [('ratu', 0.7381659746170044),
          ('kaisar', 0.6237672567367554),
          ('elizabeth', 0.6062129735946655),
          ('pewaris', 0.5993252992630005),
          ('permaisuri', 0.598286509513855),
          ('kesultanan', 0.5932478904724121),
          ('sultan', 0.5876309871673584),
          ('pangeran', 0.5850600004196167),
          ('vajiralongkorn', 0.5755350589752197),
          ('kristen', 0.5746060609817505)]
```