

LAPORAN AKHIR PRAKTIKUM

Mata Praktikum : Kecerdasan Buatan
Kelas : 3IA02
Praktikum ke- : 4
Tanggal : 11/1/23
Materi : Natural Language Processing
NPM : 50420562
Nama : Ibrahim Bramullah
Ketua Asisten : David
Paraf Asisten :
Nama Asisten :
Jumlah Lembar : 6 Lembar

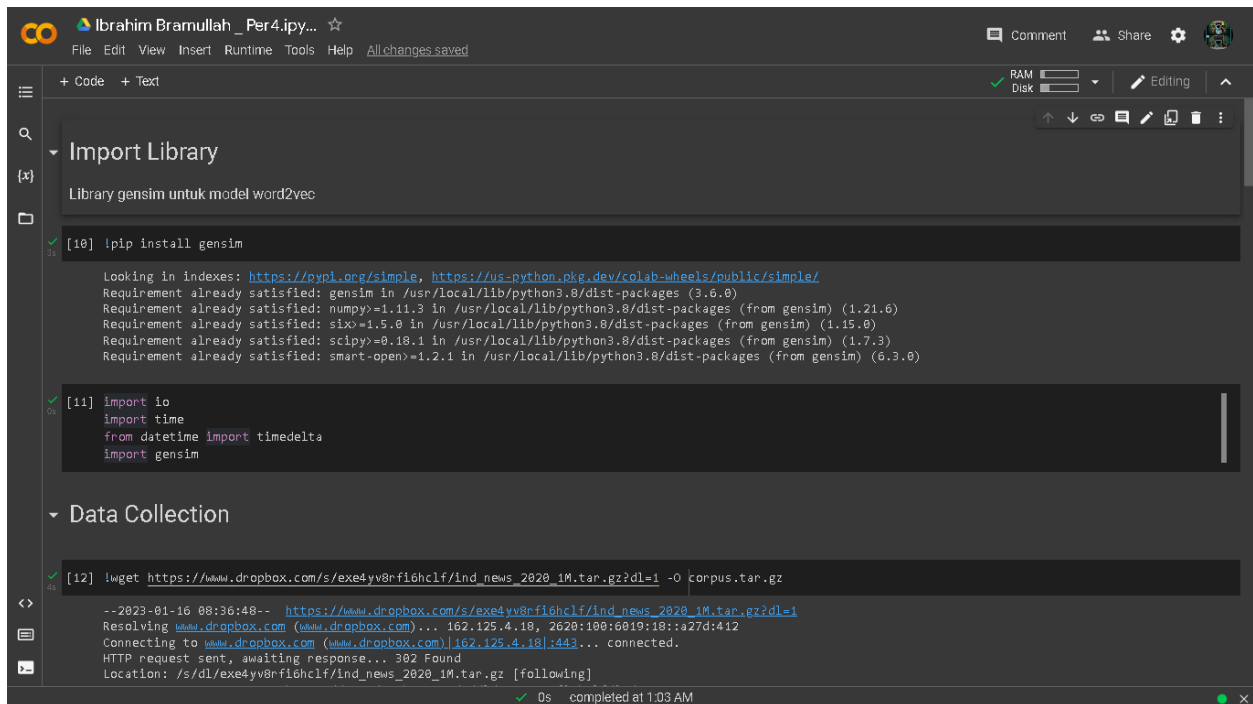
LABORATORIUM TEKNIK INFORMATIKA

UNIVERSITAS GUNADARMA

2022

--1

Mengimport library



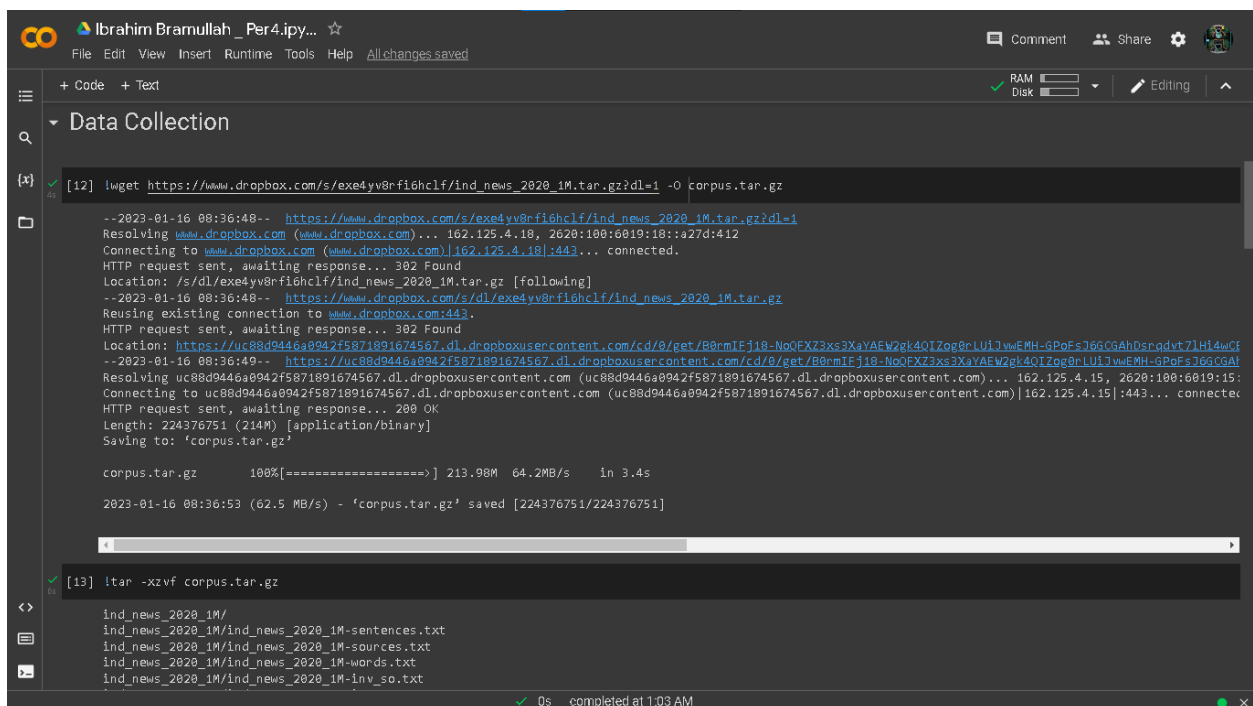
The screenshot shows a Jupyter Notebook interface with the following content:

- Import Library**
 - Library gensim untuk model word2vec
 - [10] `!pip install gensim`
 - 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: numpy>=1.11.3 in /usr/local/lib/python3.8/dist-packages (from gensim) (1.21.6)
 - Requirement already satisfied: six>=1.5.0 in /usr/local/lib/python3.8/dist-packages (from gensim) (1.15.0)
 - Requirement already satisfied: scipy>=0.18.1 in /usr/local/lib/python3.8/dist-packages (from gensim) (1.7.3)
 - Requirement already satisfied: smart-open>=1.2.1 in /usr/local/lib/python3.8/dist-packages (from gensim) (6.3.0)
 - [11] `import io`
`import time`
`from datetime import timedelta`
`import gensim`
- Data Collection**
 - [12] `!wget https://www.dropbox.com/s/exe4yv8rfi6hclf/ind_news_2020_1M.tar.gz?dl=1 -O corpus.tar.gz`
 - 2023-01-16 08:36:48-- https://www.dropbox.com/s/exe4yv8rfi6hclf/ind_news_2020_1M.tar.gz?dl=1
 - Resolving www.dropbox.com (www.dropbox.com)... 162.125.4.18, 2620:100:6019:18::a27d:412
 - Connecting to www.dropbox.com (www.dropbox.com)|162.125.4.18|:443... connected.
 - HTTP request sent, awaiting response... 302 Found
 - Location: /s/dl/exe4yv8rfi6hclf/ind_news_2020_1M.tar.gz [following]

The notebook status bar at the bottom indicates "0s completed at 1:03 AM".

--2

Masukkan dataset



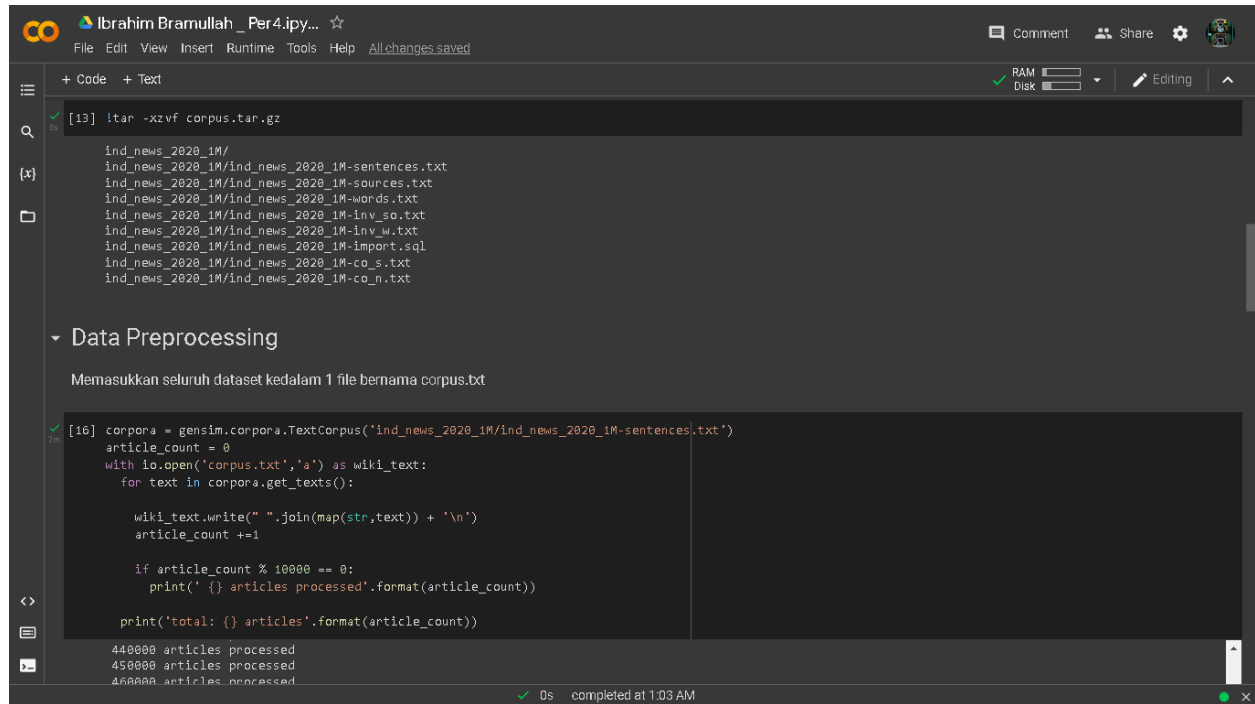
The screenshot shows a Jupyter Notebook interface with the following content:

- Data Collection**
 - [12] `!wget https://www.dropbox.com/s/exe4yv8rfi6hclf/ind_news_2020_1M.tar.gz?dl=1 -O corpus.tar.gz`
 - 2023-01-16 08:36:48-- https://www.dropbox.com/s/exe4yv8rfi6hclf/ind_news_2020_1M.tar.gz?dl=1
 - Resolving www.dropbox.com (www.dropbox.com)... 162.125.4.18, 2620:100:6019:18::a27d:412
 - Connecting to www.dropbox.com (www.dropbox.com)|162.125.4.18|:443... connected.
 - HTTP request sent, awaiting response... 302 Found
 - Location: /s/dl/exe4yv8rfi6hclf/ind_news_2020_1M.tar.gz [following]
 - 2023-01-16 08:36:48-- 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://uc88d9446a0942f5871891674567.dl.dropboxusercontent.com/cd/0/get/B0nmIEj18-NoQFX73xs3XaYAEW2gk4QIZqg8cUj1wJF#H-GPoFs366CGAh0sngdvt7LH4wCf>
 - 2023-01-16 08:36:49-- <https://uc88d9446a0942f5871891674567.dl.dropboxusercontent.com/cd/0/get/B0nmIEj18-NoQFX73xs3XaYAEW2gk4QIZqg8cUj1wJF#H-GPoFs366CGAh0sngdvt7LH4wCf>
 - Resolving uc88d9446a0942f5871891674567.dl.dropboxusercontent.com (uc88d9446a0942f5871891674567.dl.dropboxusercontent.com)... 162.125.4.15, 2620:100:6019:15::
 - Connecting to uc88d9446a0942f5871891674567.dl.dropboxusercontent.com (uc88d9446a0942f5871891674567.dl.dropboxusercontent.com)|162.125.4.15|: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 64.2MB/s in 3.4s
 - 2023-01-16 08:36:53 (62.5 MB/s) - 'corpus.tar.gz' saved [224376751/224376751]
- [13] `!tar -xvzf 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

The notebook status bar at the bottom indicates "0s completed at 1:03 AM".

--3

Ekstrak file



The screenshot shows a Jupyter Notebook titled "Ibrahim Bramullah_Per4.ipynb". The interface includes a top menu bar with options like File, Edit, View, Insert, Runtime, Tools, and Help. Below the menu, there's a toolbar with icons for file operations and a status bar showing RAM and Disk usage. The notebook content is divided into two cells. The first cell, labeled [13], contains a terminal command: `ltar -xzvf corpus.tar.gz`. Below the command, a list of files is displayed, including `ind_news_2020_1M/` and various sub-files like `ind_news_2020_1M/ind_news_2020_1M-sentences.txt`. The second cell, labeled [16], is titled "Data Preprocessing" and contains a description: "Memasukkan seluruh dataset kedalam 1 file bernama corpus.txt". Below the description, there's a code block that initializes a `gensim.corpora.TextCorpus` object, iterates over the corpus to write each text entry to a file named `corpus.txt`, and prints progress updates. The output of the code shows the progress of processing 440,000, 450,000, and 460,000 articles. The notebook status bar at the bottom indicates that the execution completed at 1:03 AM.

```
[13] ltar -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

- Data Preprocessing

Memasukkan seluruh dataset kedalam 1 file bernama corpus.txt

[16] 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('{} articles processed'.format(article_count))

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

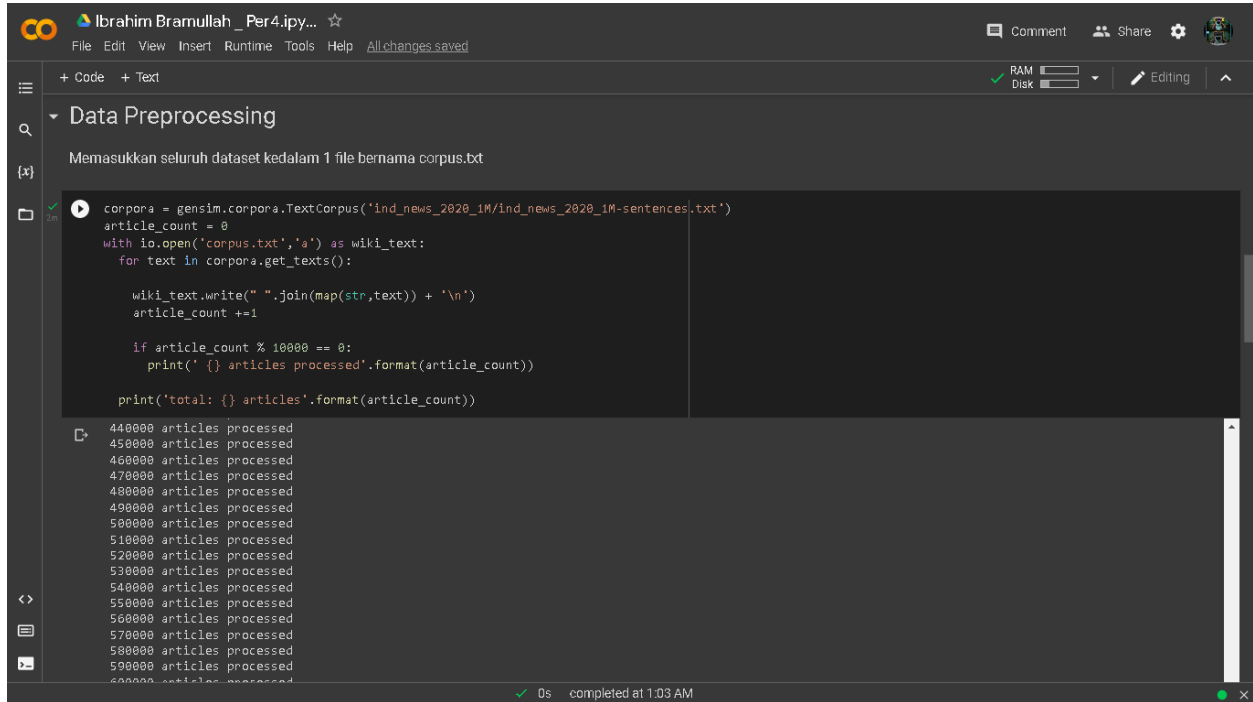
440000 articles processed
450000 articles processed
460000 articles processed

✓ 0s completed at 1:03 AM
```

--4

Penamaan file (corpus.txt)

Dan meminta 10.000 article



The screenshot shows a Jupyter Notebook interface. The top bar indicates the file name is 'Ibrahim Bramullah_Per4.ipynb'. The notebook is titled 'Data Preprocessing' and contains a single code cell. The code in the cell is as follows:

```
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

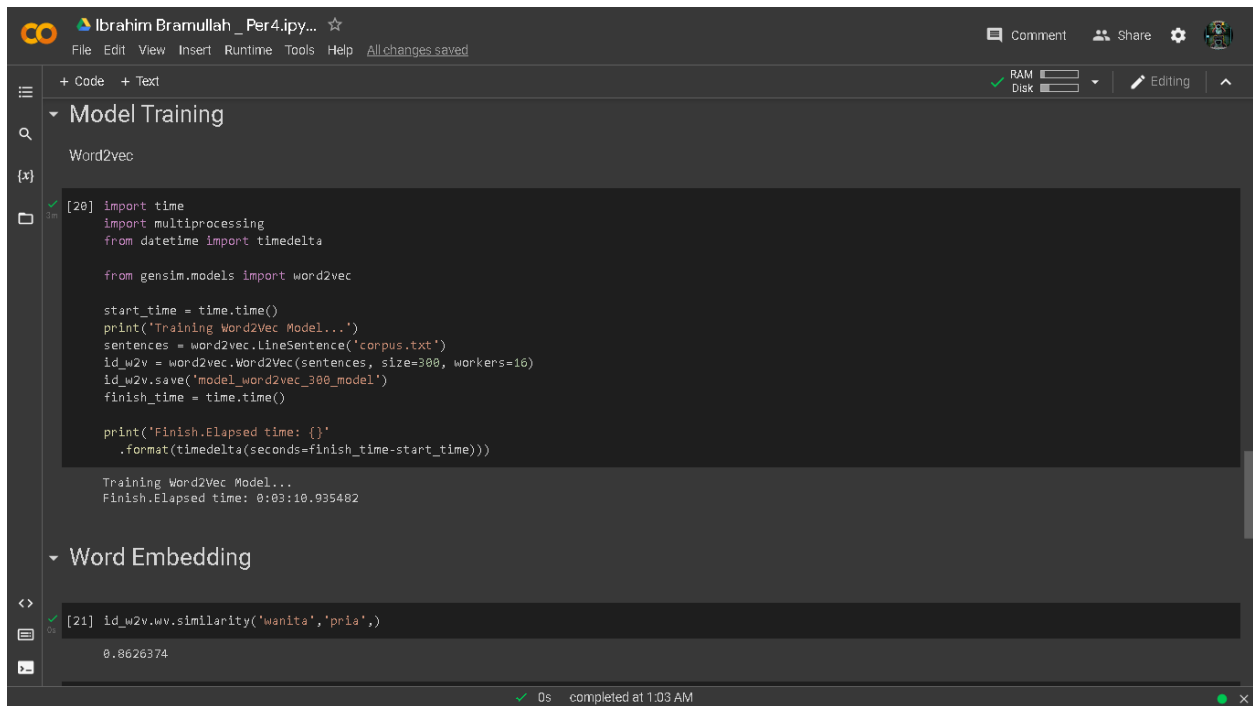
    if article_count % 10000 == 0:
        print('{} articles processed'.format(article_count))

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

The output of the code cell shows a series of status messages: '440000 articles processed' through '590000 articles processed'. The bottom status bar indicates the notebook is 'completed at 1:03 AM'.

--5

Lamanya pemodelan dalam waktu



The screenshot shows a Jupyter Notebook titled 'Ibrahim Bramullah_Per4.ipynb'. The interface includes a top menu bar with options like File, Edit, View, Insert, Runtime, Tools, and Help. Below the menu, there's a toolbar with icons for RAM, Disk, and Editing. The notebook is divided into two main sections: 'Model Training' and 'Word Embedding'.

Model Training

```
[20] 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('Finish.Elapsed time: {}'.format(timedelta(seconds=finish_time-start_time)))
```

Training Word2Vec Model...
Finish.Elapsed time: 0:03:10.935482

Word Embedding

```
[21] id_w2v.wv.similarity('wanita','pria',)
```

0.8626374

The bottom status bar indicates the notebook is 'completed at 1:03 AM'.

--6

- Mencari kesamaan dalam data
- Relasi dari data dan mengecualikan pria
- Mencari yang tidak sama sama sekali

```
ibrahim Bramullah _ Per4.ipyn... ☆
File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text
RAM
Disk
Editing

Word Embedding

[21] id_w2v.wv.similarity('wanita','pria',)
0.8626374

[27] id_w2v.wv.most_similar('permainan')
[('permainannya', 0.7268516209457397),
 ('performa', 0.6823062806728516),
 ('kemampuannya', 0.6209434866906212),
 ('penampilan', 0.6109982132911682),
 ('ritme', 0.5970042943954468),
 ('keunggulan', 0.5905264616812573),
 ('penampilannya', 0.5905006527900696),
 ('sentuhan', 0.5903928875923157),
 ('performanya', 0.5836355686187744),
 ('game', 0.5806973576545715)]

[28] id_w2v.wv.most_similar(positives=['wanita','video'], negative=['pria'])
[('videonya', 0.612392783164978),
 ('foto', 0.5570433139801025),
 ('meme', 0.526890754699707),
 ('postingan', 0.5174978971481323),
 ('konten', 0.5087568759916213),
 ('fotonya', 0.5039861868171692),
 ('percakapan', 0.49844485116004944),
 ('rekaman', 0.4933618009004236),
 ('unggahan', 0.49102503061294556),
 ('caption', 0.48867833614349365)]

id_w2v.wv.doesnt_match("baju kuning tumbuhan hijau".split())
'baju'
```

0s completed at 1:42 AM