

Notebook

February 21, 2019

```
In [1]: # import modules & set up logging
import gensim, logging
import smart_open, os
logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.DEBUG)
import datetime
import pandas as pd
import multiprocessing

# fichier incltu dans le projet
import save_notebook
```

D:\Outil\Anaconda\envs\majeure-ml-env\lib\site-packages\gensim\utils.py:1197: UserWarning: detected Windows; aliasing chunkize to chunkize_serial)

1 D  claration donn  es

```
In [2]: now = str(datetime.datetime.now()).replace(" ", "")
```

```
In [3]: word_embedding = "word2vec"
```

2 Prepare data

```
In [4]: filenames = os.listdir("../wikipedia/data")
```

```
In [5]: #Cr       un fichier ou chaque ligne continent tout un fichier
# path="../wikipedia/data"
path="../wikipedia/data/"
with open('../data/wikipedia_informatic.txt', 'w+', encoding="utf8" ) as out_file:

    for fname in filenames:
#         print(fname)
        if "ipynb_checkpoints" in fname:
            continue
        try:
            with open(path + fname, encoding="utf8") as in_file:
                out_file.write(in_file.read().replace("\n",""))
```

```

except:
    continue

```

```

In [6]: # On lit et on tokenize le fichier
with open('./data/wikipedia_informatic.txt', 'r', encoding="utf8") as f:
    wiki_vocab = f.readlines()
    wiki_vocab = [x.strip() for x in wiki_vocab]

    wiki_vocab_tokenized = []
    # for line in wiki_vocab:
    #     print(gensim.utils.simple_preprocess(line))
    # wiki_vocab_tokenized.append(gensim.utils.simple_preprocess(str(wiki_vocab)))

In [7]: wiki_vocab_tokenized = gensim.utils.simple_preprocess(str(wiki_vocab))

```

3 Create model

```

In [8]: # build vocabulary and train model
model = gensim.models.Word2Vec(
    [wiki_vocab_tokenized],
    size=150,
    seed=1234,
    window=10,
    min_count=2,
    workers=multiprocessing.cpu_count())

date_before_learning = datetime.datetime.now()
model.train([wiki_vocab_tokenized], total_examples=len(wiki_vocab_tokenized), epochs=30)
time_training = datetime.datetime.now() - date_before_learning

```

```

2019-02-21 17:06:07,830 : WARNING : consider setting layer size to a multiple of 4 for greater
2019-02-21 17:06:07,831 : INFO : collecting all words and their counts
2019-02-21 17:06:07,833 : INFO : PROGRESS: at sentence #0, processed 0 words, keeping 0 word types
2019-02-21 17:06:09,208 : INFO : collected 133432 word types from a corpus of 5351366 raw words
2019-02-21 17:06:09,209 : INFO : Loading a fresh vocabulary
2019-02-21 17:06:09,603 : INFO : min_count=2 retains 62430 unique words (46% of original 133432)
2019-02-21 17:06:09,604 : INFO : min_count=2 leaves 5280364 word corpus (98% of original 5351366)
2019-02-21 17:06:09,839 : INFO : deleting the raw counts dictionary of 133432 items
2019-02-21 17:06:09,843 : INFO : sample=0.001 downsamples 28 most-common words
2019-02-21 17:06:09,844 : INFO : downsampling leaves estimated 4064182 word corpus (77.0% of previous)
2019-02-21 17:06:10,083 : INFO : estimated required memory for 62430 words and 150 dimensions: 150MB
2019-02-21 17:06:10,084 : INFO : resetting layer weights
2019-02-21 17:06:11,259 : INFO : training model with 4 workers on 62430 vocabulary and 150 features
2019-02-21 17:06:11,269 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:11,271 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:11,272 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:11,315 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:11,318 : INFO : EPOCH - 1 : training on 5351366 raw words (10000 effective words)

```

```

2019-02-21 17:06:11,327 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:11,327 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:11,329 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:11,364 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:11,364 : INFO : EPOCH - 2 : training on 5351366 raw words (10000 effective words)
2019-02-21 17:06:11,373 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:11,375 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:11,376 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:11,406 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:11,407 : INFO : EPOCH - 3 : training on 5351366 raw words (10000 effective words)
2019-02-21 17:06:11,416 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:11,417 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:11,417 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:11,447 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:11,448 : INFO : EPOCH - 4 : training on 5351366 raw words (10000 effective words)
2019-02-21 17:06:11,457 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:11,458 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:11,458 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:11,488 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:11,489 : INFO : EPOCH - 5 : training on 5351366 raw words (10000 effective words)
2019-02-21 17:06:11,490 : INFO : training on a 26756830 raw words (50000 effective words) took
2019-02-21 17:06:11,491 : WARNING : under 10 jobs per worker: consider setting a smaller `batch`
2019-02-21 17:06:11,493 : WARNING : Effective 'alpha' higher than previous training cycles
2019-02-21 17:06:11,493 : INFO : training model with 4 workers on 62430 vocabulary and 150 features
2019-02-21 17:06:11,509 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:11,510 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:11,510 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:11,540 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:11,541 : INFO : EPOCH - 1 : training on 5351366 raw words (10000 effective words)
2019-02-21 17:06:11,541 : WARNING : EPOCH - 1 : supplied example count (1) did not equal expected
2019-02-21 17:06:11,550 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:11,552 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:11,553 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:11,579 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:11,581 : INFO : EPOCH - 2 : training on 5351366 raw words (10000 effective words)
2019-02-21 17:06:11,581 : WARNING : EPOCH - 2 : supplied example count (1) did not equal expected
2019-02-21 17:06:11,590 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:11,591 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:11,591 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:11,620 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:11,621 : INFO : EPOCH - 3 : training on 5351366 raw words (10000 effective words)
2019-02-21 17:06:11,622 : WARNING : EPOCH - 3 : supplied example count (1) did not equal expected
2019-02-21 17:06:11,630 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:11,632 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:11,633 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:11,662 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:11,663 : INFO : EPOCH - 4 : training on 5351366 raw words (10000 effective words)
2019-02-21 17:06:11,663 : WARNING : EPOCH - 4 : supplied example count (1) did not equal expected

```

[illegible]

[illegible]

[illegible]

```

2019-02-21 17:06:12,795 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:12,795 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:12,797 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:12,826 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:12,827 : INFO : EPOCH - 29 : training on 5351366 raw words (10000 effective words)
2019-02-21 17:06:12,829 : WARNING : EPOCH - 29 : supplied example count (1) did not equal expected count (10000)
2019-02-21 17:06:12,839 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:06:12,839 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:06:12,840 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:06:12,869 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:06:12,870 : INFO : EPOCH - 30 : training on 5351366 raw words (10000 effective words)
2019-02-21 17:06:12,871 : WARNING : EPOCH - 30 : supplied example count (1) did not equal expected count (10000)
2019-02-21 17:06:12,872 : INFO : training on a 160540980 raw words (300000 effective words) total

```

```
In [9]: model.save(str("model/" + now.replace(".", "-").replace(":", "-") + ".model"))
```

```

2019-02-21 17:06:12,881 : INFO : saving Word2Vec object under model/2019-02-2117-05-54-119339.model
2019-02-21 17:06:12,883 : INFO : not storing attribute vectors_norm
2019-02-21 17:06:12,885 : INFO : not storing attribute cum_table
2019-02-21 17:06:14,815 : INFO : saved model/2019-02-2117-05-54-119339.model

```

4 Test

```
In [10]: result = model.wv.most_similar(positive="microsoft",topn=10)
```

```
2019-02-21 17:06:14,823 : INFO : precomputing L2-norms of word weight vectors
```

```
In [11]: print(result)
```

```
[('tel', 0.9972558617591858), ('pistes', 0.9972215890884399), ('audio', 0.9966777563095093), ('
```

5 Save

```
In [12]: import save_notebook
```

```
In [13]: name_notebook_exported = save_notebook.save_notebook("word2vec_with_gensim.ipynb")
```

```

def write_result(word_embedding, time_training,name_notebook_exported, fname ):
    if not os.path.isfile(fname):
        fname = "Nombre de mots;Model de word embedding;temps d'apprentissage;Notebook"
    f=open(fname, "a+")
    f.write(str(len(wiki_vocab_tokenized) + ";" +word_embedding + ";" + time_training + "\n"))
    f.close

```

```
In [14]: write_result(word_embedding, time_training,name_notebook_exported, "resultats.csv" )
```

```
-----  
OSError
```

```
Traceback (most recent call last)
```

```
<ipython-input-14-b3c8fb67f2f6> in <module>
```

```
----> 1 write_result(word_embedding, time_training,name_notebook_exported, "resultats.csv"
```

```
<ipython-input-13-4bc3e2481a3a> in write_result(word_embedding, time_training, name_notebook_exported, time_training)
```

```
4     if not os.path.isfile(fname):
```

```
5         fname = "Nombre de mots;Model de word embedding;temps d'apprentissage;Notebook"
```

```
----> 6     f=open(fname, "a+")
```

```
7     f.write(str(len(wiki_vocab_tokenized) + ";" +word_embedding + ";" + time_training))
```

```
8     f.close
```

```
OSError: [Errno 22] Invalid argument: "Nombre de mots;Model de word embedding;temps d'apprentissage;Notebook"
```