

# 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 17:14:51,773 : WARNING : consider setting layer size to a multiple of 4 for greater
2019-02-21 17:14:51,775 : INFO : collecting all words and their counts
2019-02-21 17:14:51,775 : INFO : PROGRESS: at sentence #0, processed 0 words, keeping 0 word ty
2019-02-21 17:14:53,206 : INFO : collected 134060 word types from a corpus of 5386246 raw words
2019-02-21 17:14:53,208 : INFO : Loading a fresh vocabulary
2019-02-21 17:14:53,600 : INFO : min_count=2 retains 62658 unique words (46% of original 134060)
2019-02-21 17:14:53,601 : INFO : min_count=2 leaves 5314844 word corpus (98% of original 5386246)
2019-02-21 17:14:53,838 : INFO : deleting the raw counts dictionary of 134060 items
2019-02-21 17:14:53,842 : INFO : sample=0.001 downsamples 28 most-common words
2019-02-21 17:14:53,844 : INFO : downsampling leaves estimated 4091074 word corpus (77.0% of pr
2019-02-21 17:14:54,078 : INFO : estimated required memory for 62658 words and 150 dimensions:
2019-02-21 17:14:54,079 : INFO : resetting layer weights
2019-02-21 17:14:55,254 : INFO : training model with 4 workers on 62658 vocabulary and 150 fea
2019-02-21 17:14:55,263 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,264 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,264 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,308 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,310 : INFO : EPOCH - 1 : training on 5386246 raw words (10000 effective wor

```

```

2019-02-21 17:14:55,318 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,320 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,321 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,354 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,355 : INFO : EPOCH - 2 : training on 5386246 raw words (10000 effective words)
2019-02-21 17:14:55,364 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,366 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,367 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,397 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,398 : INFO : EPOCH - 3 : training on 5386246 raw words (10000 effective words)
2019-02-21 17:14:55,407 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,409 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,410 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,439 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,441 : INFO : EPOCH - 4 : training on 5386246 raw words (10000 effective words)
2019-02-21 17:14:55,448 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,450 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,451 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,480 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,480 : INFO : EPOCH - 5 : training on 5386246 raw words (10000 effective words)
2019-02-21 17:14:55,481 : INFO : training on a 26931230 raw words (50000 effective words) took 0.000000 seconds
2019-02-21 17:14:55,481 : WARNING : under 10 jobs per worker: consider setting a smaller `batch_size`
2019-02-21 17:14:55,483 : WARNING : Effective 'alpha' higher than previous training cycles
2019-02-21 17:14:55,484 : INFO : training model with 4 workers on 62658 vocabulary and 150 features
2019-02-21 17:14:55,495 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,500 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,501 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,527 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,528 : INFO : EPOCH - 1 : training on 5386246 raw words (10000 effective words)
2019-02-21 17:14:55,529 : WARNING : EPOCH - 1 : supplied example count (1) did not equal expected count (5386246)
2019-02-21 17:14:55,542 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,542 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,543 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,571 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,572 : INFO : EPOCH - 2 : training on 5386246 raw words (10000 effective words)
2019-02-21 17:14:55,573 : WARNING : EPOCH - 2 : supplied example count (1) did not equal expected count (5386246)
2019-02-21 17:14:55,582 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,583 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,584 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,612 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,613 : INFO : EPOCH - 3 : training on 5386246 raw words (10000 effective words)
2019-02-21 17:14:55,614 : WARNING : EPOCH - 3 : supplied example count (1) did not equal expected count (5386246)
2019-02-21 17:14:55,623 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,624 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,625 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,652 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,653 : INFO : EPOCH - 4 : training on 5386246 raw words (10000 effective words)
2019-02-21 17:14:55,654 : WARNING : EPOCH - 4 : supplied example count (1) did not equal expected count (5386246)

```

```

2019-02-21 17:14:55,663 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,664 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,665 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,695 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,696 : INFO : EPOCH - 5 : training on 5386246 raw words (10000 effective wo
2019-02-21 17:14:55,697 : WARNING : EPOCH - 5 : supplied example count (1) did not equal expect
2019-02-21 17:14:55,707 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,708 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,709 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,741 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,742 : INFO : EPOCH - 6 : training on 5386246 raw words (10000 effective wo
2019-02-21 17:14:55,743 : WARNING : EPOCH - 6 : supplied example count (1) did not equal expect
2019-02-21 17:14:55,754 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,762 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,763 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,788 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,789 : INFO : EPOCH - 7 : training on 5386246 raw words (10000 effective wo
2019-02-21 17:14:55,789 : WARNING : EPOCH - 7 : supplied example count (1) did not equal expect
2019-02-21 17:14:55,801 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,802 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,803 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,833 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,834 : INFO : EPOCH - 8 : training on 5386246 raw words (10000 effective wo
2019-02-21 17:14:55,835 : WARNING : EPOCH - 8 : supplied example count (1) did not equal expect
2019-02-21 17:14:55,846 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,847 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,848 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,880 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,881 : INFO : EPOCH - 9 : training on 5386246 raw words (10000 effective wo
2019-02-21 17:14:55,881 : WARNING : EPOCH - 9 : supplied example count (1) did not equal expect
2019-02-21 17:14:55,891 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,893 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,894 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,925 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,926 : INFO : EPOCH - 10 : training on 5386246 raw words (10000 effective w
2019-02-21 17:14:55,926 : WARNING : EPOCH - 10 : supplied example count (1) did not equal expe
2019-02-21 17:14:55,937 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,938 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,939 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:55,970 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:55,971 : INFO : EPOCH - 11 : training on 5386246 raw words (10000 effective w
2019-02-21 17:14:55,972 : WARNING : EPOCH - 11 : supplied example count (1) did not equal expe
2019-02-21 17:14:55,985 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:14:55,987 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:14:55,988 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:14:56,019 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:14:56,028 : INFO : EPOCH - 12 : training on 5386246 raw words (10000 effective w
2019-02-21 17:14:56,029 : WARNING : EPOCH - 12 : supplied example count (1) did not equal exp

```

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



```
2019-02-21 17:15:08,870 : INFO : training on a 1615873800 raw words (3000000 effective words) +
```

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

```
2019-02-21 17:15:08,881 : INFO : saving Word2Vec object under model/2019-02-2117-14-38-361318.model
```

```
2019-02-21 17:15:08,882 : INFO : not storing attribute vectors_norm
```

```
2019-02-21 17:15:08,884 : INFO : not storing attribute cum_table
```

```
2019-02-21 17:15:10,897 : INFO : saved model/2019-02-2117-14-38-361318.model
```

## 4 Test

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

```
2019-02-21 17:15:10,905 : INFO : precomputing L2-norms of word weight vectors
```

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

```
[('dos', 0.8956843614578247), ('système', 0.8914103507995605), ('assimiler', 0.83317112922605)]
```

## 5 Save

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

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

```
In [14]: def write_result(word_embedding, time_training,name_notebook_exported, fname ):
    if not os.path.isfile(fname):
        f=open(fname, "a+")
        f.write("Nombre de mots;Model de word embedding;temps d'apprentissage;Notebook")
        f.close
    f=open(fname, "a+")
    f.write("\n" + str( str(len(wiki_vocab_tokenized)) + ";" +word_embedding + ";" + str(time_training)))
    f.close
```

```
In [15]: write_result(word_embedding, time_training,name_notebook_exported, "resultats.csv" )
```

```
In [16]: name_notebook_exported
```

```
Out[16]: 'resultats/word2vec_with_gensim2019-02-2117-15-11-106036.pdf'
```