

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 [ ]: # 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=20)
time_training = datetime.datetime.now() - date_before_learning

```

2019-02-21 17:16:40,363 : WARNING : consider setting layer size to a multiple of 4 for greater
 2019-02-21 17:16:40,364 : INFO : collecting all words and their counts
 2019-02-21 17:16:40,365 : INFO : PROGRESS: at sentence #0, processed 0 words, keeping 0 word types
 2019-02-21 17:16:41,604 : INFO : collected 134060 word types from a corpus of 5386246 raw words
 2019-02-21 17:16:41,606 : INFO : Loading a fresh vocabulary
 2019-02-21 17:16:41,995 : INFO : min_count=2 retains 62658 unique words (46% of original 134060)
 2019-02-21 17:16:41,997 : INFO : min_count=2 leaves 5314844 word corpus (98% of original 5386246)
 2019-02-21 17:16:42,233 : INFO : deleting the raw counts dictionary of 134060 items
 2019-02-21 17:16:42,237 : INFO : sample=0.001 downsamples 28 most-common words
 2019-02-21 17:16:42,238 : INFO : downsampling leaves estimated 4091074 word corpus (77.0% of previous)
 2019-02-21 17:16:42,477 : INFO : estimated required memory for 62658 words and 150 dimensions: 1.5 MB
 2019-02-21 17:16:42,479 : INFO : resetting layer weights
 2019-02-21 17:16:43,719 : INFO : training model with 4 workers on 62658 vocabulary and 150 features
 2019-02-21 17:16:43,728 : INFO : worker thread finished; awaiting finish of 3 more threads
 2019-02-21 17:16:43,729 : INFO : worker thread finished; awaiting finish of 2 more threads
 2019-02-21 17:16:43,730 : INFO : worker thread finished; awaiting finish of 1 more threads
 2019-02-21 17:16:43,775 : INFO : worker thread finished; awaiting finish of 0 more threads
 2019-02-21 17:16:43,776 : INFO : EPOCH - 1 : training on 5386246 raw words (10000 effective words)

2019-02-21 17:16:43,786 : INFO : worker thread finished; awaiting finish of 3 more threads
 2019-02-21 17:16:43,787 : INFO : worker thread finished; awaiting finish of 2 more threads
 2019-02-21 17:16:43,788 : INFO : worker thread finished; awaiting finish of 1 more threads
 2019-02-21 17:16:43,821 : INFO : worker thread finished; awaiting finish of 0 more threads
 2019-02-21 17:16:43,823 : INFO : EPOCH - 2 : training on 5386246 raw words (10000 effective words)
 2019-02-21 17:16:43,831 : INFO : worker thread finished; awaiting finish of 3 more threads
 2019-02-21 17:16:43,833 : INFO : worker thread finished; awaiting finish of 2 more threads
 2019-02-21 17:16:43,833 : INFO : worker thread finished; awaiting finish of 1 more threads
 2019-02-21 17:16:43,864 : INFO : worker thread finished; awaiting finish of 0 more threads
 2019-02-21 17:16:43,866 : INFO : EPOCH - 3 : training on 5386246 raw words (10000 effective words)
 2019-02-21 17:16:43,875 : INFO : worker thread finished; awaiting finish of 3 more threads
 2019-02-21 17:16:43,876 : INFO : worker thread finished; awaiting finish of 2 more threads
 2019-02-21 17:16:43,877 : INFO : worker thread finished; awaiting finish of 1 more threads
 2019-02-21 17:16:43,906 : INFO : worker thread finished; awaiting finish of 0 more threads
 2019-02-21 17:16:43,907 : INFO : EPOCH - 4 : training on 5386246 raw words (10000 effective words)
 2019-02-21 17:16:43,915 : INFO : worker thread finished; awaiting finish of 3 more threads
 2019-02-21 17:16:43,917 : INFO : worker thread finished; awaiting finish of 2 more threads
 2019-02-21 17:16:43,918 : INFO : worker thread finished; awaiting finish of 1 more threads
 2019-02-21 17:16:43,946 : INFO : worker thread finished; awaiting finish of 0 more threads
 2019-02-21 17:16:43,947 : INFO : EPOCH - 5 : training on 5386246 raw words (10000 effective words)
 2019-02-21 17:16:43,948 : INFO : training on a 26931230 raw words (50000 effective words) took
 2019-02-21 17:16:43,949 : WARNING : under 10 jobs per worker: consider setting a smaller `batch`
 2019-02-21 17:16:43,977 : WARNING : Effective 'alpha' higher than previous training cycles
 2019-02-21 17:16:43,978 : INFO : training model with 4 workers on 62658 vocabulary and 150 features
 2019-02-21 17:16:43,989 : INFO : worker thread finished; awaiting finish of 3 more threads
 2019-02-21 17:16:43,990 : INFO : worker thread finished; awaiting finish of 2 more threads
 2019-02-21 17:16:43,991 : INFO : worker thread finished; awaiting finish of 1 more threads
 2019-02-21 17:16:44,020 : INFO : worker thread finished; awaiting finish of 0 more threads
 2019-02-21 17:16:44,021 : INFO : EPOCH - 1 : training on 5386246 raw words (10000 effective words)
 2019-02-21 17:16:44,022 : WARNING : EPOCH - 1 : supplied example count (1) did not equal expected
 2019-02-21 17:16:44,034 : INFO : worker thread finished; awaiting finish of 3 more threads
 2019-02-21 17:16:44,035 : INFO : worker thread finished; awaiting finish of 2 more threads
 2019-02-21 17:16:44,036 : INFO : worker thread finished; awaiting finish of 1 more threads
 2019-02-21 17:16:44,066 : INFO : worker thread finished; awaiting finish of 0 more threads
 2019-02-21 17:16:44,067 : INFO : EPOCH - 2 : training on 5386246 raw words (10000 effective words)
 2019-02-21 17:16:44,067 : WARNING : EPOCH - 2 : supplied example count (1) did not equal expected
 2019-02-21 17:16:44,077 : INFO : worker thread finished; awaiting finish of 3 more threads
 2019-02-21 17:16:44,079 : INFO : worker thread finished; awaiting finish of 2 more threads
 2019-02-21 17:16:44,080 : INFO : worker thread finished; awaiting finish of 1 more threads
 2019-02-21 17:16:44,108 : INFO : worker thread finished; awaiting finish of 0 more threads
 2019-02-21 17:16:44,109 : INFO : EPOCH - 3 : training on 5386246 raw words (10000 effective words)
 2019-02-21 17:16:44,109 : WARNING : EPOCH - 3 : supplied example count (1) did not equal expected
 2019-02-21 17:16:44,121 : INFO : worker thread finished; awaiting finish of 3 more threads
 2019-02-21 17:16:44,122 : INFO : worker thread finished; awaiting finish of 2 more threads
 2019-02-21 17:16:44,123 : INFO : worker thread finished; awaiting finish of 1 more threads
 2019-02-21 17:16:44,153 : INFO : worker thread finished; awaiting finish of 0 more threads
 2019-02-21 17:16:44,153 : INFO : EPOCH - 4 : training on 5386246 raw words (10000 effective words)
 2019-02-21 17:16:44,154 : WARNING : EPOCH - 4 : supplied example count (1) did not equal expected

[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]

[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]

[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]

[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]

[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]

[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]

[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 17:18:04,783 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:18:04,784 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:18:04,785 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:18:04,810 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:18:04,811 : INFO : EPOCH - 2037 : training on 5386246 raw words (10000 effective
2019-02-21 17:18:04,811 : WARNING : EPOCH - 2037 : supplied example count (1) did not equal ex
2019-02-21 17:18:04,820 : INFO : worker thread finished; awaiting finish of 3 more threads
2019-02-21 17:18:04,821 : INFO : worker thread finished; awaiting finish of 2 more threads
2019-02-21 17:18:04,822 : INFO : worker thread finished; awaiting finish of 1 more threads
2019-02-21 17:18:04,846 : INFO : worker thread finished; awaiting finish of 0 more threads
2019-02-21 17:18:04,847 : INFO : EPOCH - 2038 : training on 5386246 raw words (10000 effective
2019-02-21 17:18:04,848 : WARNING : EPOCH - 2038 : supplied example count (1) did not equal ex
2019-02-21 17:18:04,856 : INFO : worker thread finished; awaiting finish of 3 more threads

```

```

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

```

4 Test

```

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

```

```

In [ ]: print(result)

```

5 Save

```

In [ ]: import save_notebook

```

```

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

```

```

In [ ]: 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
            f.close

```

```

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

```

```

In [ ]: name_notebook_exported

```