

# word2vec\_mowgli

July 10, 2020

## 1 Working with the test data

### 1.0.1 Installing the packages

```
[3]: #!pip install gensim  
#!pip install nltk  
from gensim.models import Word2Vec  
from nltk import sent_tokenize  
from nltk import word_tokenize  
import nltk  
nltk.download('punkt')
```

Collecting gensim

Downloading [https://files.pythonhosted.org/packages/0b/66/04faeedb98bfa5f241d0399d0102456886179cabac0355475f23a2978847/gensim-3.8.3-cp37-cp37m-win\\_amd64.whl](https://files.pythonhosted.org/packages/0b/66/04faeedb98bfa5f241d0399d0102456886179cabac0355475f23a2978847/gensim-3.8.3-cp37-cp37m-win_amd64.whl)  
(24.2MB)

Collecting Cython==0.29.14 (from gensim)

Downloading [https://files.pythonhosted.org/packages/1f/be/b14be5c3ad1ff73096b518be1538282f053ec34faaca60a8753d975d7e93/Cython-0.29.14-cp37-cp37m-win\\_amd64.whl](https://files.pythonhosted.org/packages/1f/be/b14be5c3ad1ff73096b518be1538282f053ec34faaca60a8753d975d7e93/Cython-0.29.14-cp37-cp37m-win_amd64.whl)  
(1.7MB)

Requirement already satisfied: numpy>=1.11.3 in

c:\users\miklp\anaconda3\lib\site-packages (from gensim) (1.16.4)

Requirement already satisfied: scipy>=0.18.1 in

c:\users\miklp\anaconda3\lib\site-packages (from gensim) (1.2.1)

Collecting smart-open>=1.8.1 (from gensim)

Downloading [https://files.pythonhosted.org/packages/0b/8e/464b06f5efd26f2dc16ce7bd1662c2f31cadf9104fdbcbf5994674cc3a51/smart\\_open-2.1.0.tar.gz](https://files.pythonhosted.org/packages/0b/8e/464b06f5efd26f2dc16ce7bd1662c2f31cadf9104fdbcbf5994674cc3a51/smart_open-2.1.0.tar.gz) (116kB)

Requirement already satisfied: six>=1.5.0 in c:\users\miklp\anaconda3\lib\site-packages (from gensim) (1.12.0)

Requirement already satisfied: requests in c:\users\miklp\anaconda3\lib\site-packages (from smart-open>=1.8.1->gensim) (2.22.0)

Requirement already satisfied: boto in c:\users\miklp\anaconda3\lib\site-packages (from smart-open>=1.8.1->gensim) (2.49.0)

Collecting boto3 (from smart-open>=1.8.1->gensim)

Downloading <https://files.pythonhosted.org/packages/3c/f4/41c1d8a69b07b2a087a7e552cbcd2111ff36706fec2f1ba9983fba95771/boto3-1.14.20-py2.py3-none-any.whl>

```

(128kB)
Requirement already satisfied: certifi>=2017.4.17 in
c:\users\miklp\anaconda3\lib\site-packages (from requests->smart-
open>=1.8.1->gensim) (2019.6.16)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in
c:\users\miklp\anaconda3\lib\site-packages (from requests->smart-
open>=1.8.1->gensim) (3.0.4)
Requirement already satisfied: idna<2.9,>=2.5 in
c:\users\miklp\anaconda3\lib\site-packages (from requests->smart-
open>=1.8.1->gensim) (2.8)
Requirement already satisfied: urllib3!=1.25.0,!1.25.1,<1.26,>=1.21.1 in
c:\users\miklp\anaconda3\lib\site-packages (from requests->smart-
open>=1.8.1->gensim) (1.24.2)
Collecting boto3<1.18.0,>=1.17.20 (from boto3->smart-open>=1.8.1->gensim)
  Downloading https://files.pythonhosted.org/packages/87/a6/1710181d97a6763cccd
7f69fff8beea751633af2a101c3d02826cf4acce/boto3-1.17.20-py2.py3-none-any.whl
(6.3MB)
Collecting s3transfer<0.4.0,>=0.3.0 (from boto3->smart-open>=1.8.1->gensim)
  Downloading https://files.pythonhosted.org/packages/69/79/e6afb3d8b0b4e96cefbd
c690f741d7dd24547ff1f94240c997a26fa908d3/s3transfer-0.3.3-py2.py3-none-any.whl
(69kB)
Collecting jmespath<1.0.0,>=0.7.1 (from boto3->smart-open>=1.8.1->gensim)
  Downloading https://files.pythonhosted.org/packages/07/cb/5f001272b6faeb23c1c9
e0acc04d48eaaf5c862c17709d20e3469c6e0139/jmespath-0.10.0-py2.py3-none-any.whl
Requirement already satisfied: docutils<0.16,>=0.10 in
c:\users\miklp\anaconda3\lib\site-packages (from
boto3<1.18.0,>=1.17.20->boto3->smart-open>=1.8.1->gensim) (0.14)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in
c:\users\miklp\anaconda3\lib\site-packages (from
boto3<1.18.0,>=1.17.20->boto3->smart-open>=1.8.1->gensim) (2.8.0)
Building wheels for collected packages: smart-open
  Building wheel for smart-open (setup.py): started
  Building wheel for smart-open (setup.py): finished with status 'done'
  Stored in directory: C:\Users\miklp\AppData\Local\pip\Cache\wheels\25\6c\db\7d
cb26f19fb260c5629af85ed1c8ef9641143444fc7ec1fa08
Successfully built smart-open
Installing collected packages: Cython, jmespath, boto3, s3transfer, boto3,
smart-open, gensim
  Found existing installation: Cython 0.29.12
  Uninstalling Cython-0.29.12:
    Successfully uninstalled Cython-0.29.12
Successfully installed Cython-0.29.14 boto3-1.14.20 boto3-1.17.20
gensim-3.8.3 jmespath-0.10.0 s3transfer-0.3.3 smart-open-2.1.0

[nltk_data] Downloading package punkt to
[nltk_data]      C:\Users\miklp\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt.zip.

```

[3]: True

```
[4]: import numpy as np
import pandas as pd
import emoji
import nltk
from nltk import sent_tokenize
from nltk import word_tokenize
from nltk.corpus import stopwords
```

## 1.0.2 Reading in the data

I had issues with reading in the data in the traditional ways “pd.read...”

I decided to use with with open.... I used this in another project.

```
[14]: path = "/Users/miklp/Documents/GitHub/Student-Projects/datasets_483_982_spam.
→csv"
#path_1 = "https://raw.githubusercontent.com/mowgl-i/Student-Projects/master/
→datasets_483_982_spam.csv"
with open(path) as file:
    data = pd.read_csv(file)
data

#data = pd.read_csv(path_1)
```

```
[14]:      v1      v2 Unnamed: 2  \
0      ham  Go until jurong point, crazy.. Available only ...  NaN
1      ham                Ok lar... Joking wif u oni...      NaN
2      spam  Free entry in 2 a wkly comp to win FA Cup fina...  NaN
3      ham  U dun say so early hor... U c already then say...  NaN
4      ham  Nah I don't think he goes to usf, he lives aro...  NaN
5      spam  FreeMsg Hey there darling it's been 3 week's n...  NaN
6      ham  Even my brother is not like to speak with me. ...  NaN
7      ham  As per your request 'Melle Melle (Oru Minnamin...  NaN
8      spam  WINNER!! As a valued network customer you have...  NaN
9      spam  Had your mobile 11 months or more? U R entitle...  NaN
10     ham  I'm gonna be home soon and i don't want to tal...  NaN
11     spam  SIX chances to win CASH! From 100 to 20,000 po...  NaN
12     spam  URGENT! You have won a 1 week FREE membership ...  NaN
13     ham  I've been searching for the right words to tha...  NaN
14     ham                I HAVE A DATE ON SUNDAY WITH WILL!!  NaN
15     spam  XXXMobileMovieClub: To use your credit, click ...  NaN
16     ham                Oh k...i'm watching here:)      NaN
17     ham  Eh u remember how 2 spell his name... Yes i di...  NaN
18     ham  Fine if thatâs the way u feel. Thatâs the wa...  NaN
19     spam  England v Macedonia - dont miss the goals/team...  NaN
20     ham                Is that seriously how you spell his name?  NaN
21     ham  IÛem going to try for 2 months ha ha only joking  NaN
```

22	ham	So Ì_ pay first lar... Then when is da stock c...	NaN
23	ham	Aft i finish my lunch then i go str down lor. ...	NaN
24	ham	Ffffffffff. Alright no way I can meet up with ...	NaN
25	ham	Just forced myself to eat a slice. I'm really ...	NaN
26	ham	Lol your always so convincing.	NaN
27	ham	Did you catch the bus ? Are you frying an egg ...	NaN
28	ham	I'm back & we're packing the car now, I'll...	NaN
29	ham	Ahhh. Work. I vaguely remember that! What does...	NaN
...	...	...	...
5542	ham	Armand says get your ass over to epsilon	NaN
5543	ham	U still havent got urself a jacket ah?	NaN
5544	ham	I'm taking derek & taylor to walmart, if I...	NaN
5545	ham	Hi its in durban are you still on this number	NaN
5546	ham	Ic. There are a lotta childporn cars then.	NaN
5547	spam	Had your contract mobile 11 Mnths? Latest Moto...	NaN
5548	ham	No, I was trying it all weekend ;V	NaN
5549	ham	You know, wot people wear. T shirts, jumpers, ...	NaN
5550	ham	Cool, what time you think you can get here?	NaN
5551	ham	Wen did you get so spiritual and deep. That's ...	NaN
5552	ham	Have a safe trip to Nigeria. Wish you happines...	NaN
5553	ham	Hahaha..use your brain dear	NaN
5554	ham	Well keep in mind I've only got enough gas for...	NaN
5555	ham	Yeh. Indians was nice. Tho it did kane me off ...	NaN
5556	ham	Yes i have. So that's why u texted. Pshew...mi...	NaN
5557	ham	No. I meant the calculation is the same. That ...	NaN
5558	ham	Sorry, I'll call later	NaN
5559	ham	if you aren't here in the next &#x26; hou...	NaN
5560	ham	Anything lor. Juz both of us lor.	NaN
5561	ham	Get me out of this dump heap. My mom decided t...	NaN
5562	ham	Ok lor... Sony ericsson salesman... I ask shuh...	NaN
5563	ham	Ard 6 like dat lor.	NaN
5564	ham	Why don't you wait 'til at least wednesday to ...	NaN
5565	ham	Huh y lei...	NaN
5566	spam	REMINDER FROM 02: To get 2.50 pounds free call...	NaN
5567	spam	This is the 2nd time we have tried 2 contact u...	NaN
5568	ham	Will Ì_ b going to esplanade fr home?	NaN
5569	ham	Pity, * was in mood for that. So...any other s...	NaN
5570	ham	The guy did some bitching but I acted like i'd...	NaN
5571	ham	Rofl. Its true to its name	NaN

Unnamed: 3 Unnamed: 4

0	NaN	NaN
1	NaN	NaN
2	NaN	NaN
3	NaN	NaN
4	NaN	NaN
5	NaN	NaN

6	NaN	NaN
7	NaN	NaN
8	NaN	NaN
9	NaN	NaN
10	NaN	NaN
11	NaN	NaN
12	NaN	NaN
13	NaN	NaN
14	NaN	NaN
15	NaN	NaN
16	NaN	NaN
17	NaN	NaN
18	NaN	NaN
19	NaN	NaN
20	NaN	NaN
21	NaN	NaN
22	NaN	NaN
23	NaN	NaN
24	NaN	NaN
25	NaN	NaN
26	NaN	NaN
27	NaN	NaN
28	NaN	NaN
29	NaN	NaN
...	...	...
5542	NaN	NaN
5543	NaN	NaN
5544	NaN	NaN
5545	NaN	NaN
5546	NaN	NaN
5547	NaN	NaN
5548	NaN	NaN
5549	NaN	NaN
5550	NaN	NaN
5551	NaN	NaN
5552	NaN	NaN
5553	NaN	NaN
5554	NaN	NaN
5555	NaN	NaN
5556	NaN	NaN
5557	NaN	NaN
5558	NaN	NaN
5559	NaN	NaN
5560	NaN	NaN
5561	NaN	NaN
5562	NaN	NaN
5563	NaN	NaN

5564	NaN	NaN
5565	NaN	NaN
5566	NaN	NaN
5567	NaN	NaN
5568	NaN	NaN
5569	NaN	NaN
5570	NaN	NaN
5571	NaN	NaN

[5572 rows x 5 columns]

Looks like only two of the columns are useful to us.

That would be the “class” of message that we get. Either “ham” or “spam” and the actual “message” we get.

Let’s rename the columns and select only the two columns we need.

```
[15]: data.columns = ["class", "text", "none", "none", "none"]

data = data[["class", "text"]]

data
```

```
[15]:      class      text
0      ham  Go until jurong point, crazy.. Available only ...
1      ham                Ok lar... Joking wif u oni...
2      spam  Free entry in 2 a wkly comp to win FA Cup fina...
3      ham  U dun say so early hor... U c already then say...
4      ham  Nah I don't think he goes to usf, he lives aro...
5      spam  FreeMsg Hey there darling it's been 3 week's n...
6      ham  Even my brother is not like to speak with me. ...
7      ham  As per your request 'Melle Melle (Oru Minnamin...
8      spam  WINNER!! As a valued network customer you have...
9      spam  Had your mobile 11 months or more? U R entitle...
10     ham  I'm gonna be home soon and i don't want to tal...
11     spam  SIX chances to win CASH! From 100 to 20,000 po...
12     spam  URGENT! You have won a 1 week FREE membership ...
13     ham  I've been searching for the right words to tha...
14     ham                I HAVE A DATE ON SUNDAY WITH WILL!!
15     spam  XXXMobileMovieClub: To use your credit, click ...
16     ham                Oh k...i'm watching here:)
17     ham  Eh u remember how 2 spell his name... Yes i di...
18     ham  Fine if that's the way u feel. That's the wa...
19     spam  England v Macedonia - dont miss the goals/team...
20     ham                Is that seriously how you spell his name?
21     ham  IÜøem going to try for 2 months ha ha only joking
22     ham  So Ì_ pay first lar... Then when is da stock c...
23     ham  Aft i finish my lunch then i go str down lor. ...
24     ham  Fffffff. Alright no way I can meet up with ...
25     ham  Just forced myself to eat a slice. I'm really ...
```

```

26      ham                Lol your always so convincing.
27      ham  Did you catch the bus ? Are you frying an egg ...
28      ham  I'm back & we're packing the car now, I'll...
29      ham  Ahhh. Work. I vaguely remember that! What does...
...      ...
5542     ham                Armand says get your ass over to epsilon
5543     ham                U still havent got urself a jacket ah?
5544     ham  I'm taking derek & taylor to walmart, if I...
5545     ham                Hi its in durban are you still on this number
5546     ham                Ic. There are a lotta childporn cars then.
5547     spam  Had your contract mobile 11 Mnths? Latest Moto...
5548     ham                No, I was trying it all weekend ;V
5549     ham  You know, wot people wear. T shirts, jumpers, ...
5550     ham                Cool, what time you think you can get here?
5551     ham  Wen did you get so spiritual and deep. That's ...
5552     ham  Have a safe trip to Nigeria. Wish you happines...
5553     ham                Hahaha..use your brain dear
5554     ham  Well keep in mind I've only got enough gas for...
5555     ham  Yeh. Indians was nice. Tho it did kane me off ...
5556     ham  Yes i have. So that's why u texted. Pshew...mi...
5557     ham  No. I meant the calculation is the same. That ...
5558     ham                Sorry, I'll call later
5559     ham  if you aren't here in the next &#x26; hou...
5560     ham                Anything lor. Juz both of us lor.
5561     ham  Get me out of this dump heap. My mom decided t...
5562     ham  Ok lor... Sony ericsson salesman... I ask shuh...
5563     ham                Ard 6 like dat lor.
5564     ham  Why don't you wait 'til at least wednesday to ...
5565     ham                Huh y lei...
5566     spam  REMINDER FROM 02: To get 2.50 pounds free call...
5567     spam  This is the 2nd time we have tried 2 contact u...
5568     ham                Will Ì_ b going to esplanade fr home?
5569     ham  Pity, * was in mood for that. So...any other s...
5570     ham  The guy did some bitching but I acted like i'd...
5571     ham                Rofl. Its true to its name

```

[5572 rows x 2 columns]

```
[16]: data.shape
```

```
[16]: (5572, 2)
```

```
[17]: data.dtypes
```

```
[17]: class    object
text      object
dtype: object
```

```
[18]: data.isnull().sum()
```