

# Average Word Vector

Created by: Aniket Pasi

From: SVNIT, Gujarat

Mtech Data Science - p23ds020 (2023-25)

Subject: NLP Project

Last Updated: 29/03/2024

## 1) Import Libraries

```
In [1]: import gensim
import pandas as pd
import numpy as np
import spacy
import string
```

## 2) Import Dataset

```
In [2]: data = pd.read_csv("train.csv")
```

```
In [3]: data
```

Out[3]:

	tweets	class
0	Be aware dirty step to get money #staylight ...	figurative
1	#sarcasm for #people who don't understand #diy...	figurative
2	@lminworkJeremy @medsingle #DailyMail readers ...	figurative
3	@wilw Why do I get the feeling you like games?...	figurative
4	-@TeacherArthurG @rweingarten You probably jus...	figurative
...	...	...
81403	Photo: Image via We Heart It http://t.co/ky8Nf...	sarcasm
81404	I never knew..I better put this out to the Uni...	sarcasm
81405	hey just wanted to say thanks @ puberty for le...	sarcasm
81406	I'm sure coverage like the Fox News Special "T...	sarcasm
81407	@skeyno16 at u13?! I won't believe it until I ...	sarcasm

81408 rows × 2 columns

```
In [4]: data['class'].value_counts()
```

```
Out[4]: class
figurative    21238
irony         20894
sarcasm       20681
regular       18595
Name: count, dtype: int64
```

### 3) Preprocessing (cleaning and tokenizing)

```
In [5]: clean_tweets = data.tweets.apply(gensim.utils.simple_preprocess)
```

```
In [6]: clean_tweets
```

```
Out[6]: 0      [be, aware, dirty, step, to, get, money, stayl...
1      [sarcasm, for, people, who, don, understand, d...
2      [iminworkjeremy, medsingle, dailymail, readers...
3      [wilw, why, do, get, the, feeling, you, like, ...
4      [teacherarthurg, rweingarten, you, probably, j...
...
81403  [photo, image, via, we, heart, it, http, co, k...
81404  [never, knew, better, put, this, out, to, the,...
81405  [hey, just, wanted, to, say, thanks, puberty, ...
81406  [sure, coverage, like, the, fox, news, special...
81407  [skeyno, at, won, believe, it, until, see, it,...
Name: tweets, Length: 81408, dtype: object
```

### 4) Create Word2vec Model

```
In [7]: from gensim.models import Word2Vec, KeyedVectors
import gensim.downloader as api
```

```
In [9]: ## Initilize Model
## model_cbow = gensim.models.Word2Vec(clean_tweets, workers=4, min_count=5,

## workers = number of threads
## windows = number of words in considerations to predict the word similar
## vector_size = number of features in input layer
## min_count = number of minimum words to consider the windows, if less th

## Build Vocabulary
## model_cbow.build_vocab(clean_tweets, progress_per=1000)

## Train Model
## model_cbow.train(clean_tweets, total_examples=model_cbow.corpus_count, ep

## to save the model and reuse it without training
## model_cbow.save("./word2vec_model.model")
```

```
In [8]: try:
        with open('wiki_model_1.bin', 'r') as f:
            w_model_1 = gensim.models.KeyedVectors.load('wiki_model_1.bin')
        except FileNotFoundError:
            wiki_model_1 = gensim.downloader.load('glove-wiki-gigaword-100')
            wiki_model_1.save('wiki_model_1.bin')
            w_model_1 = gensim.models.KeyedVectors.load('wiki_model_1.bin')
```

```
In [9]: try:
        with open('wiki_model_2.bin', 'r') as f:
            w_model_2 = gensim.models.KeyedVectors.load('wiki_model_2.bin')
        except FileNotFoundError:
            wiki_model_2 = gensim.downloader.load('glove-wiki-gigaword-50')
            wiki_model_2.save('wiki_model_2.bin')
            w_model_2 = gensim.models.KeyedVectors.load('wiki_model_2.bin')
```

```
In [17]: # # Load vector
        # model_w2v = gensim.models.Ldamodel.LdaModel.Load("word2vec_model.model")
```

```
In [15]: def sent_vec_w2v(sent):
        vector_size_1 = 100
        vector_size_2 = 50

        vec_1 = np.zeros(vector_size_1)
        vec_2 = np.zeros(vector_size_2)

        ctr = 1

        for w in sent:
            if w in w_model_1 and w in w_model_2:
                ctr += 1
                vec_1 += w_model_1[w][:vector_size_1]
                vec_2 += w_model_1[w][:vector_size_2]

        wv_res = np.concatenate((vec_1, vec_2))
        wv_res = wv_res/ctr

        return wv_res
```

### Save tokens

```
In [11]: data['tokens'] = clean_tweets
```

```
In [12]: data
```

Out[12]:

	tweets	class	tokens
0	Be aware dirty step to get money #staylight ...	figurative	[be, aware, dirty, step, to, get, money, stayl...
1	#sarcasm for #people who don't understand #diy...	figurative	[sarcasm, for, people, who, don, understand, d...
2	@lminworkJeremy @medsingle #DailyMail readers ...	figurative	[iminworkjeremy, medsingle, dailymail, readers...
3	@wilw Why do I get the feeling you like games?...	figurative	[wilw, why, do, get, the, feeling, you, like, ...
4	-@TeacherArthurG @rweingarten You probably jus...	figurative	[teacherarthurg, rweingarten, you, probably, j...
...	...	...	...
81403	Photo: Image via We Heart It http://t.co/ky8Nf...	sarcasm	[photo, image, via, we, heart, it, http, co, k...
81404	I never knew..I better put this out to the Uni...	sarcasm	[never, knew, better, put, this, out, to, the,...
81405	hey just wanted to say thanks @ puberty for le...	sarcasm	[hey, just, wanted, to, say, thanks, puberty, ...
81406	I'm sure coverage like the Fox News Special "T...	sarcasm	[sure, coverage, like, the, fox, news, special...
81407	@skeyno16 at u13?! I won't believe it until I ...	sarcasm	[skeyno, at, won, believe, it, until, see, it,...

81408 rows × 3 columns

5) Create vectors

```
In [16]: data['word2vec'] = data['tokens'].apply(sent_vec_w2v)
```

In [17]: data

Out[17]:

	tweets	class	tokens	word2vec
0	Be aware dirty step to get money #staylight ...	figurative	[be, aware, dirty, step, to, get, money, stayl...	[-0.011905075838932624, 0.11662022568858586, 0...
1	#sarcasm for #people who don't understand #diy...	figurative	[sarcasm, for, people, who, don, understand, d...	[0.029422137746587397, 0.20814320296049119, 0....
2	@lminworkJeremy @medsingle #DailyMail readers ...	figurative	[iminworkjeremy, medsingle, dailymail, readers...	[-0.1235885014757514, 0.05067412555217743, 0.3...
3	@wilw Why do I get the feeling you like games?...	figurative	[wilw, why, do, get, the, feeling, you, like, ...	[-0.021854890137910844, 0.34805419892072675, 0...
4	-@TeacherArthurG @rweingarten You probably jus...	figurative	[teacherarthurg, rweingarten, you, probably, j...	[-0.0528821237385273, 0.16415249928832054, 0.5...
...	...	...	...	...
81403	Photo: Image via We Heart It http://t.co/ky8Nf...	sarcasm	[photo, image, via, we, heart, it, http, co, k...	[-0.1844358862274223, 0.18341966884003746, 0.3...
81404	I never knew..I better put this out to the Uni...	sarcasm	[never, knew, better, put, this, out, to, the,...	[0.01726132275705988, 0.1990272288464687, 0.38...
81405	hey just wanted to say thanks @ puberty for le...	sarcasm	[hey, just, wanted, to, say, thanks, puberty, ...	[-7.330059357311414e-05, 0.011773303960976393,...
81406	I'm sure coverage like the Fox News Special "T...	sarcasm	[sure, coverage, like, the, fox, news, special...	[-0.08389464654028415, 0.10088500231504441, 0....
81407	@skeyno16 at u13?! I won't believe it until I ...	sarcasm	[skeyno, at, won, believe, it, until, see, it,...	[-0.02437943137354321, 0.19550255437692007, 0....

81408 rows × 4 columns

In [19]: `print(len(data['word2vec'][0]))`

150

### Save vectors

In [18]: `data.to_csv("final_dataset.csv", index=False)`