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 [3]:

```
import gensim
import pandas as pd
import numpy as np
import spacy
import string
```

2) Import Dataset

In [4]:

```
data = pd.read_csv("train.csv")
```

In [6]:
```
data
```

Out[6]:

	tweets	class
0	Be aware dirty step to get money #staylight ...	figurative
1	#sarcasm for #people who don't understand #diy...	figurative
2	@IminworkJeremy @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 [7]:
```
data['class'].value_counts()
```

Out[7]:
```
class
figurative    21238
irony         20894
sarcasm       20681
regular       18595
Name: count, dtype: int64
```

3) Preprocessing (cleaning and tokenizing)

In [11]:
```
clean_tweets = data.tweets.apply(gensim.utils.simple_preprocess)
```

In [12]:
```
clean_tweets
```

Out[12]:
```
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 [13]:
```
from gensim.models import Word2Vec, KeyedVectors
import gensim.downloader as api
```

In [14]:
```
# Initilize Model
model_cbow = gensim.models.Word2Vec(clean_tweets, workers=4, min_count=5, vect

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

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

Out[14]:
```
(4369020, 6291135)
```

In [20]:
```
# load vector
model_w2v = gensim.models.ldamodel.LdaModel.load("word2vec_model.model")
```

```
WARNING:root:random_state not set so using default value
WARNING:root:failed to load state from word2vec_model.model.state: [Errno 2]
No such file or directory: 'word2vec_model.model.state'
```

In [18]:
```
def sent_vec_w2v(sent):
    vector_size = 150
    wv_res = np.zeros(vector_size)
    # print(wv_res)
    ctr = 1
    for w in sent:
        if w in model_w2v.wv:
            ctr += 1
            wv_res += model_w2v.wv[w][:vector_size]
    wv_res = wv_res/ctr
    return wv_res
```

Save tokens

In [15]:
```
data['tokens'] = clean_tweets
```

```
In [16]: data
```

Out[16]:

	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	@IminworkJeremy @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 [21]: data['word2vec'] = data['tokens'].apply(sent_vec_w2v)
```

In [26]:
```
data
```

Out[26]:

	tweets	class	tokens	word2vec
0	Be aware dirty step to get money #staylight ...	figurative	[be, aware, dirty, step, to, get, money, stayl...	[-0.0819224606339748, 0.4479990784938519, 0.19...
1	#sarcasm for #people who don't understand #diy...	figurative	[sarcasm, for, people, who, don, understand, d...	[-0.39415668305009605, 0.14868877977132797, -0...
2	@IminworkJeremy @medsingle #DailyMail readers ...	figurative	[iminworkjeremy, medsingle, dailymail, readers...	[-0.3844123475253582, 0.36639655753970146, -0....
3	@wilw Why do I get the feeling you like games?...	figurative	[wilw, why, do, get, the, feeling, you, like, ...	[-0.46358012628148904, 0.12668053052303466, 0....
4	-@TeacherArthurG @rweingarten You probably jus...	figurative	[teacherarthurg, rweingarten, you, probably, j...	[-0.19652243633754551, 0.19875611818861216, -0...
...	...	...	...	...
81403	Photo: Image via We Heart It http://t.co/ky8Nf...	sarcasm	[photo, image, via, we, heart, it, http, co, k...	[-0.25477669693322647, 0.29173677621616256, 0....
81404	I never knew..I better put this out to the Uni...	sarcasm	[never, knew, better, put, this, out, to, the,...	[-0.32126067590434104, 0.12298504258929328, 0....
81405	hey just wanted to say thanks @ puberty for le...	sarcasm	[hey, just, wanted, to, say, thanks, puberty, ...	[-0.23387800608761608, 0.13634898184609484, -0...
81406	I'm sure coverage like the Fox News Special “T...	sarcasm	[sure, coverage, like, the, fox, news, special...	[-0.12494273199454735, 0.004117939082023345, 0...
81407	@skeyno16 at u13?! I won't believe it until I ...	sarcasm	[skeyno, at, won, believe, it, until, see, it,...	[-0.028264301518599193, -0.10591727081272337, ...

81408 rows × 4 columns

Save vectors

In [25]:
```
data.to_csv("final_dataset.csv", index=False)
```