Adidas Project Weekly Report

April 4, 2021

Jinhang Jiang

1. Recap last week meeting
   a. Korean case study (8 korean idol groups/singers)
2. Word2vec for clustering and similarity
3. Case study on adidas' list, korean list, and combined

Word2vec - Determine the number of dimensions

Google Developer Blog: Introducing TensorFlow Feature Columns :

the following "formula" provides a general rule of thumb about the number of embedding dimensions:

embedding_dimensions = number_of_categories**0.25

E.g. in the blog, they only used 3 vectors for 81 words “pow(81,0.25) = 3”

Word2vec - Other parameters

The interpretation of dream meaning: Resolving ambiguity using Latent Semantic Analysis in a small corpus of text :

A parameter selection was made, obtaining the best performance for LSA in 200 dimensions and for **Skip-gram in win = 15 and neg = 10** (Table 3, Table 4).

win \ neg	5	10	15
5	−0.63	−0.96	−0.99
10	−0.95	−1.06	−1.01
15	−0.95	**−1.12**	−1.02

The scores are the slopes in the log-linear regression of the escape/chase (to word “run”) rank distance vs escape/chase fraction.

Word2vec - Other parameters


```
model1.wv.most_similar('adidas',topn=15)
```

```
[('yzy', 0.7066393494606018),
 ('harden', 0.684260368347168),
 ('arrive', 0.6721178293228149),
 ('pharrellwilliams', 0.6636843681335449),
 ('hoops', 0.6501753330230713),
 ('cf', 0.6455749273300171),
 ('wondering', 0.6418324708938599),
 ('uj', 0.6399343013763428),
 ('dayum', 0.6361621618270874),
 ('speed', 0.6353594660758972),
 ('ni', 0.6286124587059021),
 ('washing', 0.6265524625778198),
 ('nmd', 0.6249428987503052),
 ('snowing', 0.6168386340141296),
 ('narrow', 0.614669680595398)]
```

Case study on adidas' list

```
file.Celebrity.unique()
```

```
array(['BlackPink', 'Naeun Son', 'Kerwin Frost', 'Beyonce', 'Zoe Saldana',
       'Karlie Kloss', 'Yara Shahidi', 'Pharrell Williams',
       'Adriene Mishler', 'Ninjas Hyper', 'Bad Bunny', 'Jerry Lorenzo',
       'Chinae Alexander', 'Ally Love'], dtype=object)
```

```
num_words = len(str(df_merge['clean txt'].tolist()))
num_words
```

```
1969589
```

```
dim_size = num_words**0.25
"{:.8f}".format(float(dim_size))
```

```
'37.46225386'
```

Case study on adidas' list


```
# Create skip-gram model
model1 = gensim.models.Word2Vec(data,
                                min_count = 5,
                                size = 38,
                                window = 15,
                                negative=10,
                                sg=1)
model1.train(data, total_examples=model1.corpus_count,epochs=10)
```

Case study on adidas' list

```
print(model1.wv.similarity('kpop','idol'))
print(model1.wv.similarity('kpop','movie'))
print(model1.wv.similarity('kpop','fortnite'))
print(model1.wv.similarity('ninjahyper','fortnite'))
```

```
0.4852233
0.22420447
0.19395979
0.3617612
```

```
model1.wv.most_similar('ninjashyper',topn=15)
```

```
[('your', 0.8933946490287781),
 ('ninjahyper', 0.8607471585273743),
 ('imagined', 0.8289879560470581),
 ('vids', 0.8085686564445496),
 ('noscopes', 0.785987377166748),
 ('subscribers', 0.7619348168373108),
```

```
model1.wv.most_similar('yzy',topn=15)
```

```
[('harden', 0.9197931289672852),
 ('dayum', 0.9056742191314697),
 ('lillard', 0.871384859085083),
 ('hoops', 0.8373474478721619),
 ('moving', 0.8362677097320557),
 ('besides', 0.8068371415138245),
 ('basketball', 0.7969784736633301),
 ('nmd', 0.7660571336746216),
 ('according', 0.7623190879821777),
 ('speed', 0.7588093280792236),
 ('brought', 0.7358499765396118),
 ('struggling', 0.7249552607536316),
 ('signature', 0.7191159725189209),
 ('especially', 0.7170764803886414),
 ('adidas', 0.7066393494606018)]
```

Case study on adidas' list

```
embedding.columns=list(range(skip_gram.shape[1]+1))
embedding.rename(columns={ embedding.columns[0]: "Celebrity" }, inplace = True)
embedding
```

	Celebrity	1	2	3	4	5	6	7	8	
0	Adriene Mishler	0.883938	0.077463	-0.242327	-0.476454	1.035660	0.383212	0.343211	0.335006	0
0	Ally Love	0.463414	-0.036454	-0.206449	-0.450776	0.474303	0.557577	0.584207	0.170215	0

Case study on adidas' list

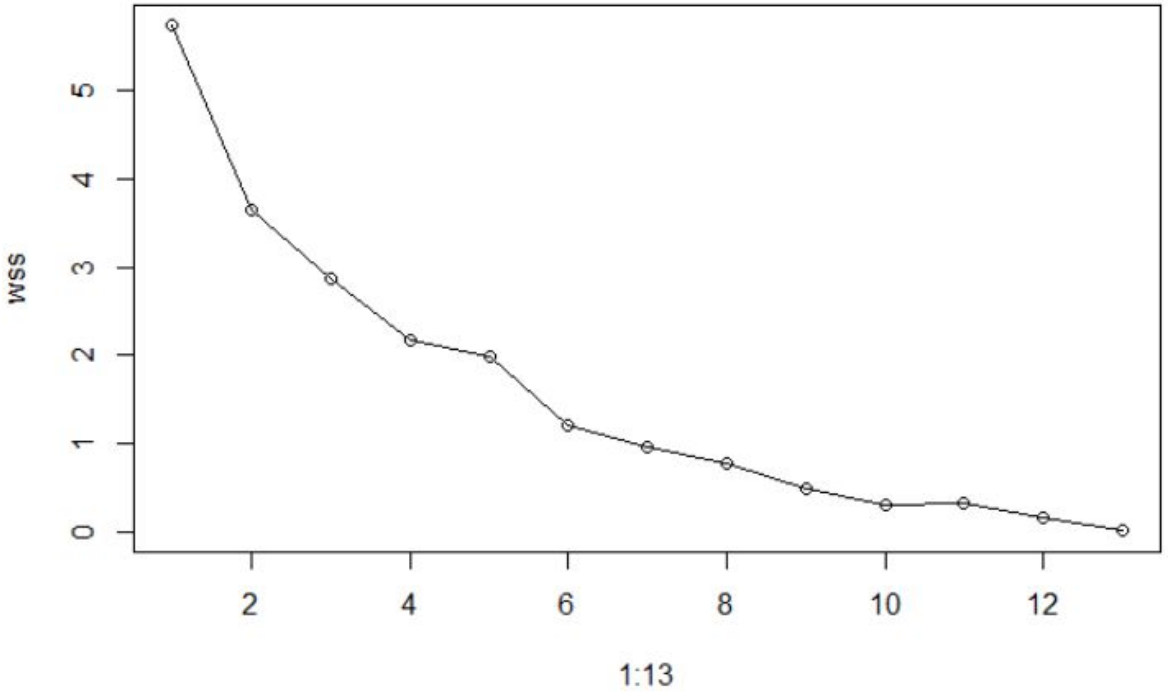
```
similarity(embedding,'Celebrity','Naeun Son','TopN')
```

```
[('BlackPink', array([[0.9771761]], dtype=float32)),
 ('Ally Love', array([[0.974951]], dtype=float32)),
 ('Bad Bunny', array([[0.9689709]], dtype=float32)),
```

```
similarity(embedding,'Celebrity','Ninjas Hyper','TopN')
```

```
[('Ally Love', array([[0.9561324]], dtype=float32)),
 ('Bad Bunny', array([[0.9524962]], dtype=float32)),
 ('Jerry Lorenzo', array([[0.9410826]], dtype=float32)),
 ('BlackPink', array([[0.93595386]], dtype=float32)),
```

Case study on adidas' list



K = 10

```
> fit$betweenss/fit$totss
[1] 0.9430811
```

Case study on adidas' list

Grou	Name	Category	Gende	Country	Profession
1	Kerwin Frost	FASHION	M	U.S.	Talkshow Host
2	NinjasHyper	ESPORTS & TECH	M	U.S.	Gamer
3	Yara Sayeh Shahidi	VIP	F	U.S.	Actress
4	Adriene Mishler	WOMENS	F	U.S.	Actress
5	Pharrell Williams	Top Creators	M	U.S.	Singer
6	Karlie Kloss	WOMENS	F	U.S.	Fashion model
7	JERRY LORENZO	Top Creators	M	U.S.	Sneaker Designer
8	Beyonce	Top Creators	F	U.S.	Singer
	Zoe Saldana	VIP	F	U.S.	Actress
9	CHINAE ALEXANDER	WOMENS	F	U.S.	Instgram Star
10	ALLY LOVE	WOMENS	F	U.S.	Fitness instructor
	BadBunny	MUSIC	M	Puerto Rico	Rapper
	BlackPink	MUSIC	F	Korea	Girl Group
	naeun	MUSIC	F	Korea	Singer

	Category	Gender	Country	Profession
Group 1				
BadBunny	MUSIC	M	Puerto R	Rapper
JERRY LORENZO	Top Creators	M	U.S.	Sneaker Designer
BlackPink	MUSIC	F	Korea	Girl Group
naeun	MUSIC	F	Korea	Singer
Group 2				
Karlie Kloss	WOMENS	F	U.S.	Fashion model
Beyonce	Top Creators	F	U.S.	Singer
Pharrell Williams	Top Creators	M	U.S.	Singer
Yara Sayeh Shahidi	VIP	F	U.S.	Actress
Group 3				
NinjasHyper	ESPORTS & T	M	U.S.	Gamer
Group 4				
Kerwin Frost	FASHION	M	U.S.	Talkshow Host
Group 5				
Zoe Saldana	VIP	F	U.S.	Actress
ALLY LOVE	WOMENS	F	U.S.	Fitness instructor
Adriene Mishler	WOMENS	F	U.S.	Actress
CHINAE ALEXANDER	WOMENS	F	U.S.	Instgram Star

Extra finding

```
model1.wv.most_similar('adidas',topn=15)
```

```
[('yzy', 0.715119719505310 1),
 ('speed', 0.708686351776123),
 ('hu', 0.690321683883667),
 ('basketball', 0.685364365577697 8),
```

```
model1.wv.most_similar('nike',topn=15)
```

```
[('yeezys', 0.832863628864288 3),
 ('alcoholism', 0.751563787460327 1),
 ('python', 0.744528889656066 9),
 ('brown', 0.734571576118469 2),
```

Case study on korean list

```
num_words = len(str(df_merge['clean txt'].tolist()))
num_words
```

```
3013433
```

```
dim_size = num_words**0.25
"{:.8f}".format(float(dim_size))
```

```
'41.66442427'
```

```
# Create skip-gram model
model1 = gensim.models.Word2Vec(data,
                                min_count = 5,
                                size = 42,
                                window = 15,
                                negative=10,
                                sg=1)
model1.train(data,  total_examples=model1.corpus_count,epochs=10)
```

Case study on korean list

```
similarity(embedding,'Celebrity','Naeun Son','TopN')[0:2]
```

```
[('Solar', array([[0.9933375]], dtype=float32)),
 ('iZone', array([[0.9925412]], dtype=float32))]
```

```
similarity(embedding,'Celebrity','NCT','TopN')[0:2]
```

```
[('iZone', array([[0.9952595]], dtype=float32)),
 ('Solar', array([[0.9952272]], dtype=float32))]
```

```
similarity(embedding,'Celebrity','BlackPink','TopN')[0:2]
```

```
[('NCT', array([[0.9935911]], dtype=float32)),
 ('Solar', array([[0.9884211]], dtype=float32))]
```

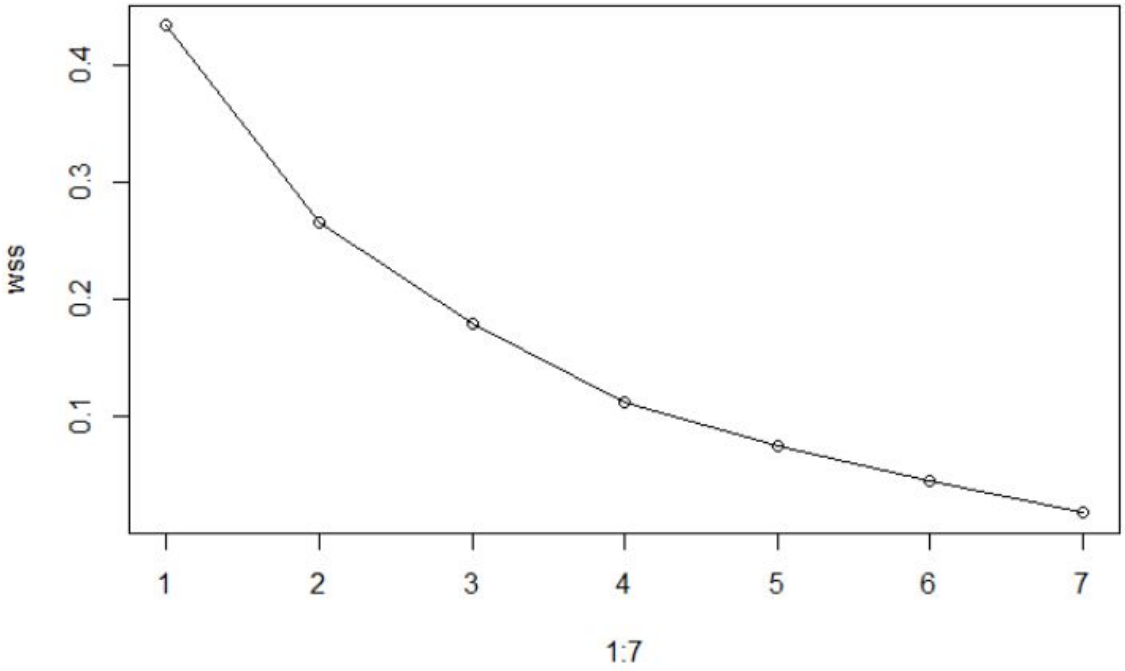
```
model.best_sub("Naeun Son",n=1)
```

```
['Seolhyun']
```

```
model.best_sub("NCT",n=1)
```

```
['Solar']
```

Case study on korean list



K=4 (Same with network analysis)

```
> fit$betweenss/fit$totss
[1] 0.7406216
```

Case study on korean list

Not very correlated...

Word2vec

Group 1: BTS, NCT, Naeun Son, Solar, iZone

Group 2: Seolhyun

Group 3: GFriend

Group 4: BlackPink

Node2vec

Group 1: Solar, BTS

Group 2: Seolhyun, Naeun Son

Group 3: GFriend, iZone

Group 4: BlackPink, NCT

Case study on full list (14+6)

```
similarity(embedding,'Celebrity','Naeun Son','TopN')[0:2]
```

```
[('iZone', array([[0.9938302]], dtype=float32)),
 ('Solar', array([[0.99303406]], dtype=float32))]
```

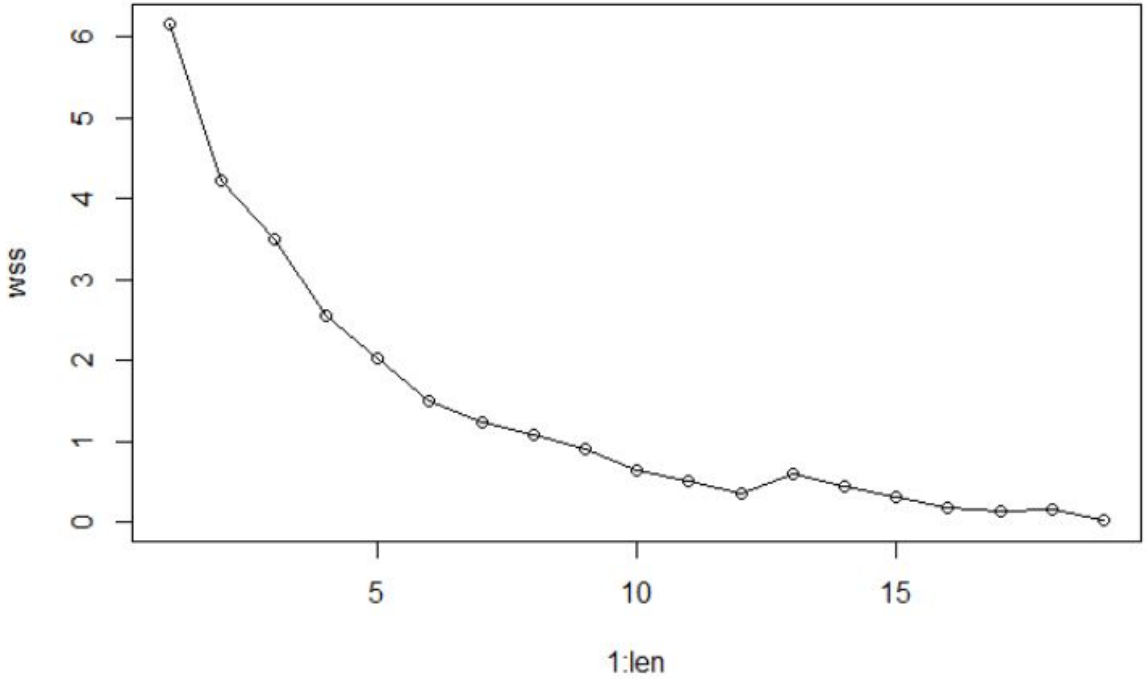
```
similarity(embedding,'Celebrity','Beyonce','TopN')[0:2]
```

```
[('Zoe Saldana', array([[0.9993838]], dtype=float32)),
 ('Bad Bunny', array([[0.97338504]], dtype=float32))]
```

```
similarity(embedding,'Celebrity','Ninjas Hyper','TopN')[0:2]
```

```
[('Ally Love', array([[0.96392196]], dtype=float32)),
 ('Jerry Lorenzo', array([[0.95860016]], dtype=float32))]
```

Case study on full list (14+6)



K = 12

BSS/TSS = 0.94

Case study on full list (14+6)

Group	Name	Category	Gende	Country	Profession
1	BTS	MUSIC	M	Korea	Boy Group
	NCT	MUSIC	M	Korea	Boy Group
	Naeun Son	MUSIC	F	Korea	Singer
	Solar	MUSIC	F	Korea	Girl Group
	iZone	MUSIC	F	Korea	Girl Group
2	BlackPink	MUSIC	F	Korea	Girl Group
3	GFriend	MUSIC	F	Korea	Girl Group
4	Seolhyun	MUSIC	F	Korea	Singer
5	NinjasHyper	ESPORTS & TECH	M	U.S.	Gamer
6	Adriene Mishler	WOMENS	F	U.S.	Actress

Group	Name	Category	Gender	Country	Profession
7	Chinae Alexander	WOMENS	F	U.S.	Instgram Star
8	Kerwin Frost	FASHION	M	U.S.	Talkshow Host
9	BadBunny	MUSIC	M	Puerto Rico	Rapper
10	ALLY LOVE	WOMENS	F	U.S.	Fitness instructor
	JERRY LORENZO	Top Creatc	M	U.S.	Sneaker Designer
11	Karlie Kloss	WOMENS	F	U.S.	Fashion model
	Yara Sayeh Shahidi	VIP	F	U.S.	Actress
12	Zoe Saldana	VIP	F	U.S.	Actress
	Beyonce	Top Creatc	F	U.S.	Singer
	Pharrell Williams	Top Creatc	M	U.S.	Singer

Group 1				
BadBunny	MUSIC	M	Puerto R	Rapper
JERRY LORENZO	Top Creators	M	U.S.	Sneaker Designer
BlackPink	MUSIC	F	Korea	Girl Group
naeun	MUSIC	F	Korea	Singer
Group 2				
Karlie Kloss	WOMENS	F	U.S.	Fashion model
Beyonce	Top Creators	F	U.S.	Singer
Pharrell Williams	Top Creators	M	U.S.	Singer
Yara Sayeh Shahidi	VIP	F	U.S.	Actress

Group 3				
NinjasHyper	ESPORTS &	M	U.S.	Gamer
Group 4				
Kerwin Frost	FASHION	M	U.S.	Talkshow Host
Group 5				
Zoe Saldana	VIP	F	U.S.	Actress
ALLY LOVE	WOMENS	F	U.S.	Fitness instructor
Adriene Mishler	WOMENS	F	U.S.	Actress
CHINAE ALEXANDER	WOMENS	F	U.S.	Instgram Star

Case study on full list (14+6)


```
dictionary = ["adidas","yeezys","yzy","nmd","hu","harden","lillard"]
```

```
df[df["flag"]>0].groupby(["Celebrity"]).size()
```

```
Celebrity
BTS                1
Bad Bunny          6
BlackPink          6
Jerry Lorenzo     12
Kerwin Frost       4
NCT                5
Pharrell Williams  4
dtype: int64
```

Word2vec - Pre-trained

Pre-trained model vs. train model on our data

“even if the vocabulary is just 300 words, using pre-trained embeddings will probably yield better results than training the embeddings directly on the dataset.” --- stackoverflow

Google's trained Word2Vec model in Python

3.5 G file, 3M words vectors with 300 dimensions

Test of pre-trained word2vec model

for	-0.011780	-0.047363	0.044678	0.063477
that	-0.015747	-0.028320	0.083496	0.050293
is	0.007050	-0.073242	0.171875	0.022583
...	...	...	...	...
RAFFAELE	0.009277	-0.050537	-0.018799	0.029785
Bim_Skala_Bim	0.012573	0.045410	-0.043213	-0.001495
Mezze_Cafe	-0.019653	-0.090820	-0.019409	0.019653
pulverizes_boulders	0.032715	-0.032227	0.036133	0.001175
snowcapped_Caucasus	0.045166	-0.045166	-0.003937	0.048828

3000000 rows × 300 columns

Test of pre-trained word2vec model

Does it include stop words?

Answer: Some stop words like “a”, “and”, “of” are excluded, but others like “the”, “also”, “should” are included.

Does it include misspellings of words?

Answer: Yes. For instance, it includes both “mispelled” and “misspelled”–the latter is the correct one.

Does it include commonly paired words?

Answer: Yes. For instance, it includes “Soviet_Union” and “New_York”.

Does it include numbers?

Answer: Not directly; e.g., you won’t find “100”. But it does include entries like “###MHz_DDR2_SDRAM” where I’m assuming the ‘#’ are intended to match any digit.

https://mccormickml.com/2016/04/12/googles-pretrained-word2vec-model-in-python/

Test of pre-trained word2vec model

```
model.wv.most_similar('Adidas',topn=15)
```

```
[('adidas', 0.8445298671722412),
 ('Nike', 0.7950947284698486),
 ('Adidas_ADDDY.PK_news', 0.6852840185165405),
 ('Reebok', 0.6841334104537964),
 ('Puma', 0.6774643659591675),
 ('Adidas_Salomon', 0.659113883972168),
 ('spokesman_Jan_Runau', 0.6242237091064453),
 ('adidas_Salomon_AG', 0.6223492622375488),
 ('sportswear', 0.614206075668335),
 ('Adidas_Salomon_AG', 0.6134569048881531),
```

```
Round # 20 for: iZone completed

'3:47:53.287497'
```

Test of pre-trained word2vec model

Pre-trained

```
similarity(embedding1,'Celebrity','Naeun Son','TopN')[0:2]
```

```
[('iZone', array([[0.9967277]], dtype=float32)),
 ('NCT', array([[0.9953402]], dtype=float32))]
```

```
similarity(embedding1,'Celebrity','Beyonce','TopN')[0:2]
```

```
[('Zoe Saldana', array([[0.9997813]], dtype=float32)),
 ('BlackPink', array([[0.98949885]], dtype=float32))]
```

```
similarity(embedding1,'Celebrity','Ninjas Hyper','TopN')[0:2]
```

```
[('Ally Love', array([[0.98195165]], dtype=float32)),
 ('NCT', array([[0.9788385]], dtype=float32))]
```

Trained on collected data

```
similarity(embedding,'Celebrity','Naeun Son','TopN')[0:2]
```

```
[('iZone', array([[0.9938302]], dtype=float32)),
 ('Solar', array([[0.99303406]], dtype=float32))]
```

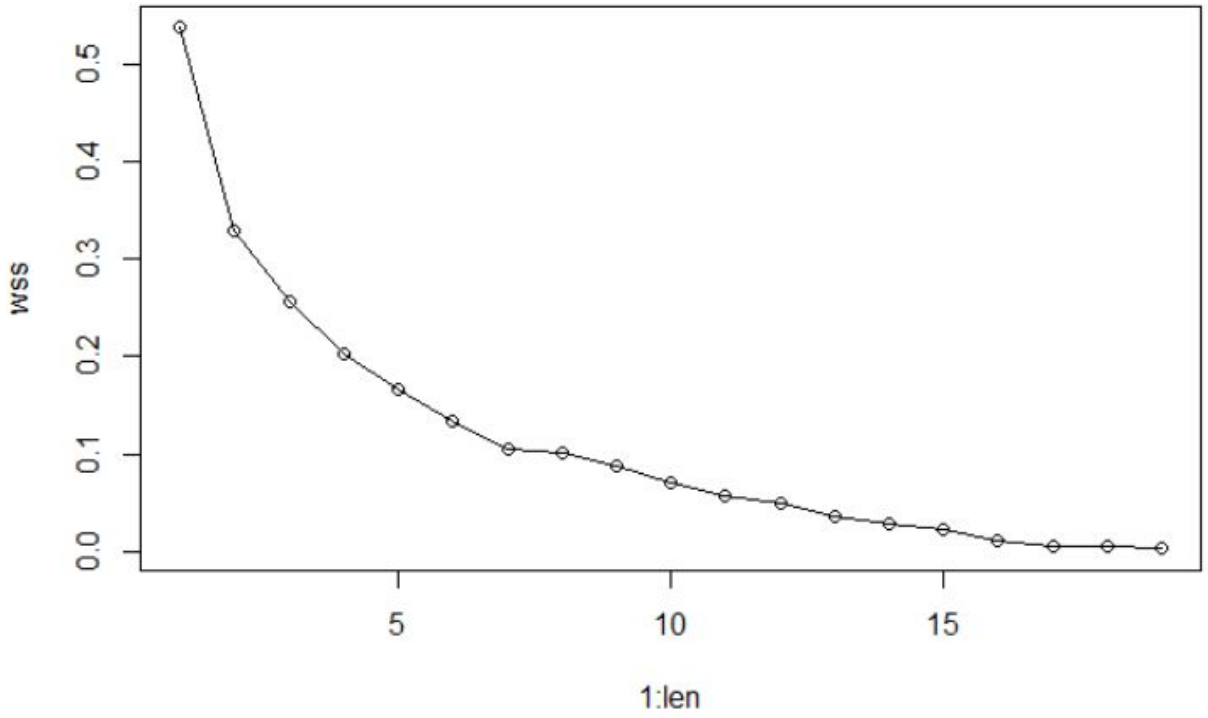
```
similarity(embedding,'Celebrity','Beyonce','TopN')[0:2]
```

```
[('Zoe Saldana', array([[0.9993838]], dtype=float32)),
 ('Bad Bunny', array([[0.97338504]], dtype=float32))]
```

```
similarity(embedding,'Celebrity','Ninjas Hyper','TopN')[0:2]
```

```
[('Ally Love', array([[0.96392196]], dtype=float32)),
 ('Jerry Lorenzo', array([[0.95860016]], dtype=float32))]
```

Test of pre-trained word2vec model



K=12

BSS/TSS = 0.916

Test of pre-trained word2vec model

ASU W. P. CAREY SCHOOL of BUSINESS
ARIZONA STATE UNIVERSITY

Group	Name	Category	Gende	Country	Profession
1	BTS	MUSIC	M	Korea	Boy Group
	NCT	MUSIC	M	Korea	Boy Group
	Naeun Son	MUSIC	F	Korea	Singer
	Solar	MUSIC	F	Korea	Girl Group
	iZone	MUSIC	F	Korea	Girl Group
2	BlackPink	MUSIC	F	Korea	Girl Group
3	GFriend	MUSIC	F	Korea	Girl Group
4	Seolhyun	MUSIC	F	Korea	Singer
5	NinjasHyper	ESPORTS & TECH	M	U.S.	Gamer
6	Adriene Mishler	WOMENS	F	U.S.	Actress

Group	Name	Category	Gender	Country	Profession
7	Chinae Alexander	WOMENS	F	U.S.	Instgram Star
8	Kerwin Frost	FASHION	M	U.S.	Talkshow Host
9	BadBunny	MUSIC	M	Puerto Rico	Rapper
10	ALLY LOVE	WOMENS	F	U.S.	Fitness instructor
	JERRY LORENZO	Top Creatc	M	U.S.	Sneaker Designer
11	Karlie Kloss	WOMENS	F	U.S.	Fashion model
	Yara Sayeh Shahidi	VIP	F	U.S.	Actress
12	Zoe Saldana	VIP	F	U.S.	Actress
	Beyonce	Top Creatc	F	U.S.	Singer
	Pharrell Williams	Top Creatc	M	U.S.	Singer

Group	Name	Category	Gende	Country	Profession
1	Kerwin Frost	FASHION	M	U.S.	Talkshow Host
	BadBunny	MUSIC	M	Puerto Rico	Rapper
2	Naeun Son	MUSIC	F	Korea	Singer
	Solar	MUSIC	F	Korea	Girl Group
	iZone	MUSIC	F	Korea	Girl Group
	NCT	MUSIC	M	Korea	Boy Group
3	Pharrell Williams	Top Creators	M	U.S.	Singer
4	Seolhyun	MUSIC	F	Korea	Singer
5	NinjasHyper	ESPORTS & TECH	M	U.S.	Gamer
6	Adriene Mishler	WOMENS	F	U.S.	Actress

Group	Name	Category	Gender	Country	Profession
7	Chinae Alexander	WOMENS	F	U.S.	Instgram Star
8	GFriend	MUSIC	F	Korea	Girl Group
	BTS	MUSIC	M	Korea	Boy Group
9	ALLY LOVE	WOMENS	F	U.S.	Fitness instructor
10	JERRY LORENZO	Top Creatc	M	U.S.	Sneaker Designer
11	Karlie Kloss	WOMENS	F	U.S.	Fashion model
	Yara Sayeh Shahidi	VIP	F	U.S.	Actress
	BlackPink	MUSIC	F	Korea	Girl Group
12	Beyonce	Top Creatc	F	U.S.	Singer
	Zoe Saldana	VIP	F	U.S.	Actress

Textblob Sentiment Analysis


```
df.groupby(["Celebrity"])['Sentiment_Polarity'].mean().sort_values(ascending=False)
```

```
Celebrity
Seolhyun              0.197389
BTS                   0.190309
BlackPink             0.168621
Karlie Kloss          0.162711
Ally Love             0.162413
NCT                   0.161409
Yara Shahidi          0.150044
Pharrell Williams     0.138143
GFriend               0.125707
Solar                 0.124558
Jerry Lorenzo         0.118302
Naeun Son             0.116083
```

End