Mind Body Green's Audience Tweet Analysis

Overview of Data

```
mbg_aud_tweets
```

	Id	Time	Tweets
0	101619625	Thu Aug 06 20:54:57 +0000 2020	
1	1287731622352125953	Thu Aug 06 15:30:03 +0000 2020	August th is and we have a special sale to hel...
2	2182048904	Sat Aug 08 01:31:57 +0000 2020	
3	1291892079086501888	Sat Aug 08 01:56:19 +0000 2020	: That time youre driving in Virginia and the ...
4	1289256723107196929	Mon Aug 03 13:50:24 +0000 2020	: As a nephrologist, I prescribed HCQ (plaquen...
...	...	...	...
752	1673239740	Tue Aug 04 21:20:38 +0000 2020	: Saw Texas Roadhouse trending and thought som...
753	567244654	Wed Aug 05 15:52:41 +0000 2020	: Coronavirus timeline
754	984442240960413697	Fri Aug 07 12:41:52 +0000 2020	I just sent a letter to asking her to please j...
755	1039559305589084162	Wed Aug 05 13:39:53 +0000 2020	Making the Transition A defining factor in the...
756	947177384620515328	Sat Jul 25 21:57:31 +0000 2020	He was the best, a legend indeed. Rest in Peac...

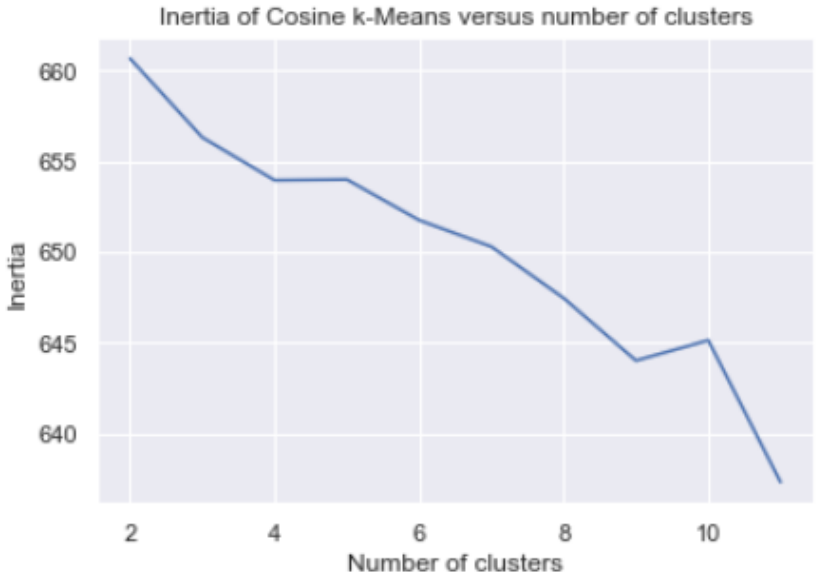
757 rows × 3 columns

Word Cloud

Most of Tweets are about Love, Need, New and Life. Some meaningless words include:

- “Nan” means null value of text
- “Amp” means a tag of HTML

Determine number of Tweet Clusters




```
from sklearn.metrics import silhouette_score
# Prepare models
kmeans = KMeans(n_clusters=4).fit(tv_transform)
normalized_vectors = preprocessing.normalize(tv_transform)
normalized_kmeans = KMeans(n_clusters=4).fit(normalized_vectors)
min_samples = tv_transform.shape[1]+1

# Print results
print('kmeans: {}'.format(silhouette_score(tv_transform, kmeans.labels_,
                                           metric='euclidean')))
print('Cosine kmeans:{}'.format(silhouette_score(normalized_vectors,
                                                 normalized_kmeans.labels_,
                                                 metric='cosine')))
```

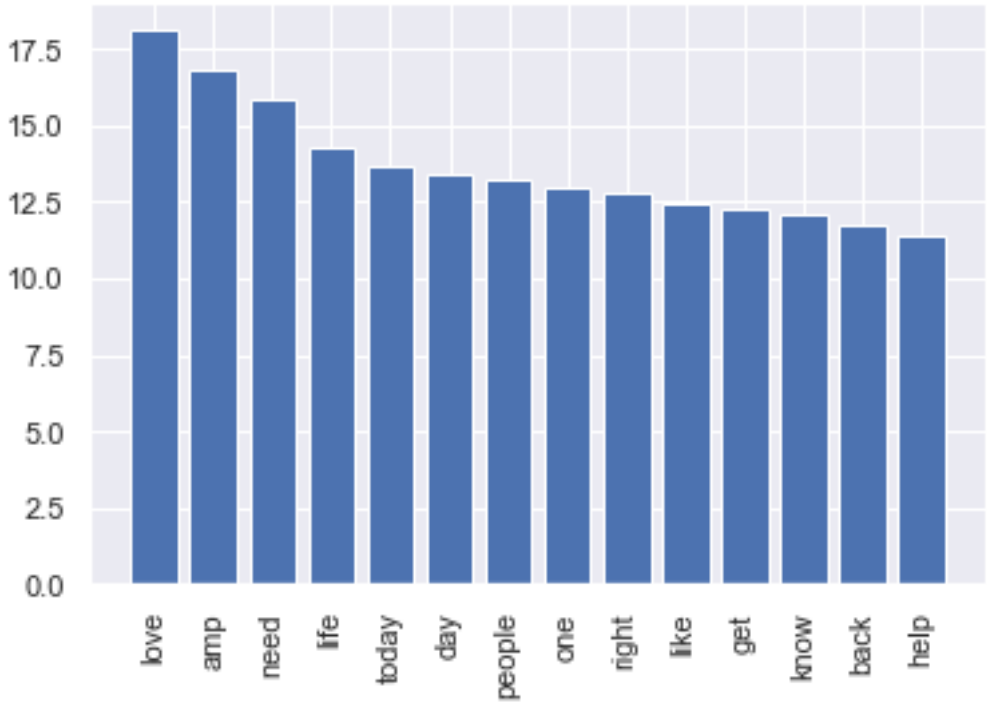
```
kmeans: 0.23320274961013976
Cosine kmeans:0.11398615197129527
```

According to the score above, Kmeans performed better than Cosine. So we choose to go with Kmeans chart. According to the Elbow method, the number of clusters is 6.

Unigram Cluster

```
['always', 'amp', 'around', 'back', 'beautiful', 'begin']
['could', 'product', 'thing', 'always', 'amp', 'around']
['de', 'amp', 'say', 'different', 'im', 'want']
['mean', 'want', 'new', 'make', 'body', 'around']
['week', 'well', 'via', 'thing', 'food', 'going']
['check', 'put', 'everything', 'peace', 'week', 'always']
```

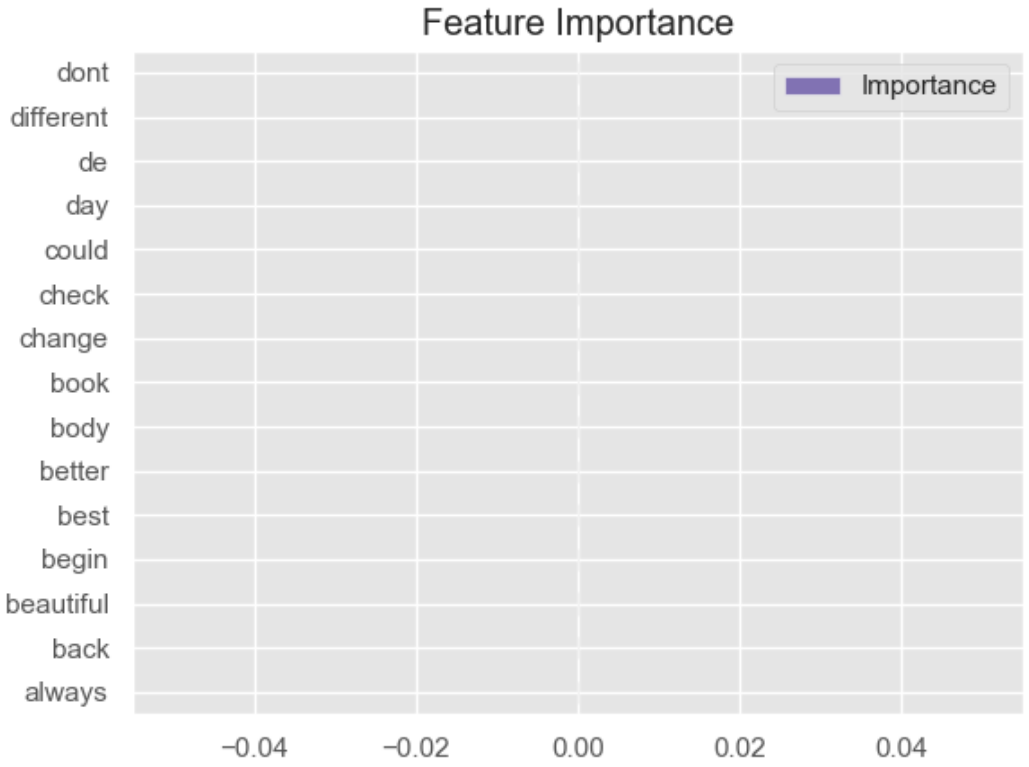
Importance of Word



Cluster 1: Positivity
Cluster 2: Product
Cluster 3: Uniqueness; stand firm
Cluster 4: Standing firm
Cluster 5: Food; how a person is doing
Cluster 6: Peacefullness

Love, Need, Life & Today are the most important words

Unigram – Cluster 1 Feature Importance

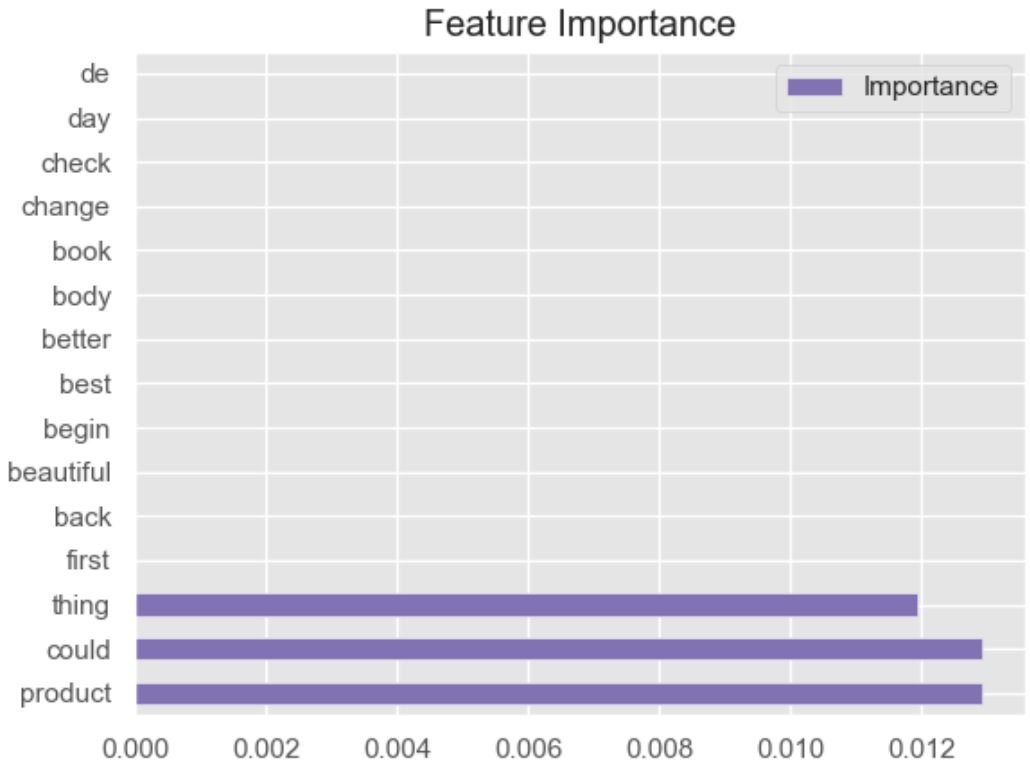


	Words	Importance
0	always	0.0
63	put	0.0
73	something	0.0
72	someone	0.0
71	show	0.0
...	...	...
30	good	0.0
29	going	0.0
28	go	0.0
27	get	0.0
99	youre	0.0

100 rows × 2 columns

Cluster 1 consists of words that are useless and irrelevant.

Unigram – Cluster 2 Feature Importance

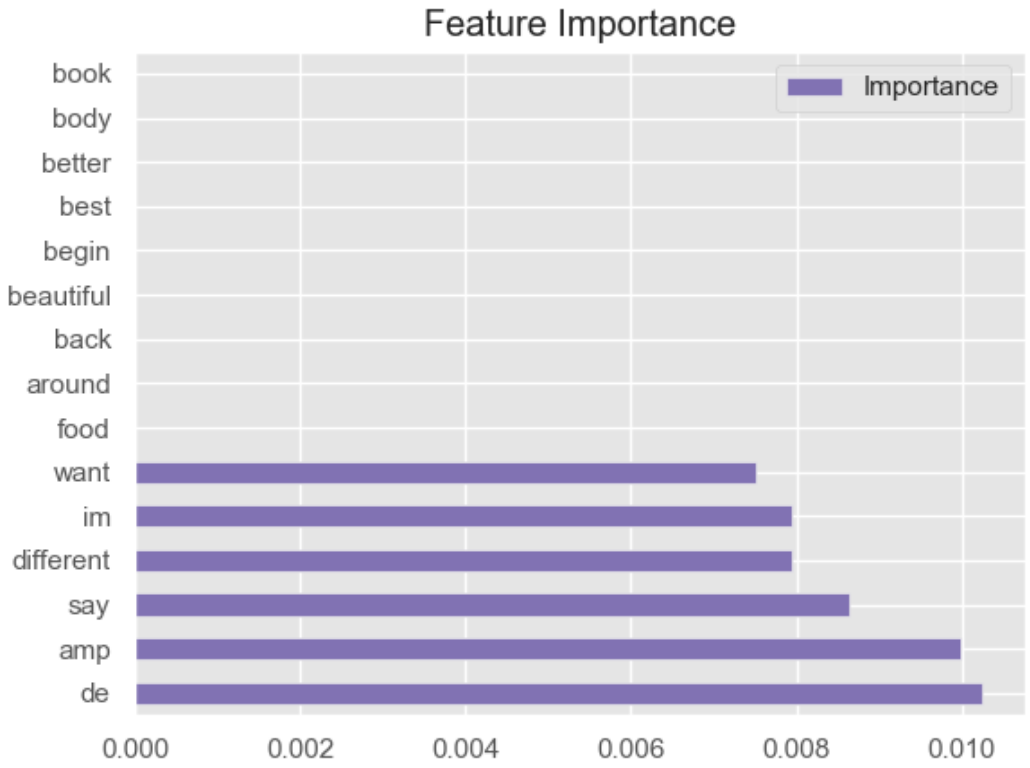


	Words	Importance
62	product	0.012931
12	could	0.012931
81	thing	0.011922
0	always	0.000000
64	que	0.000000
...	...	...
31	great	0.000000
30	good	0.000000
29	going	0.000000
28	go	0.000000
99	youre	0.000000

100 rows × 2 columns

In cluster 2, only the words product, could and thing are important.

Unigram – Cluster 3 Feature Importance

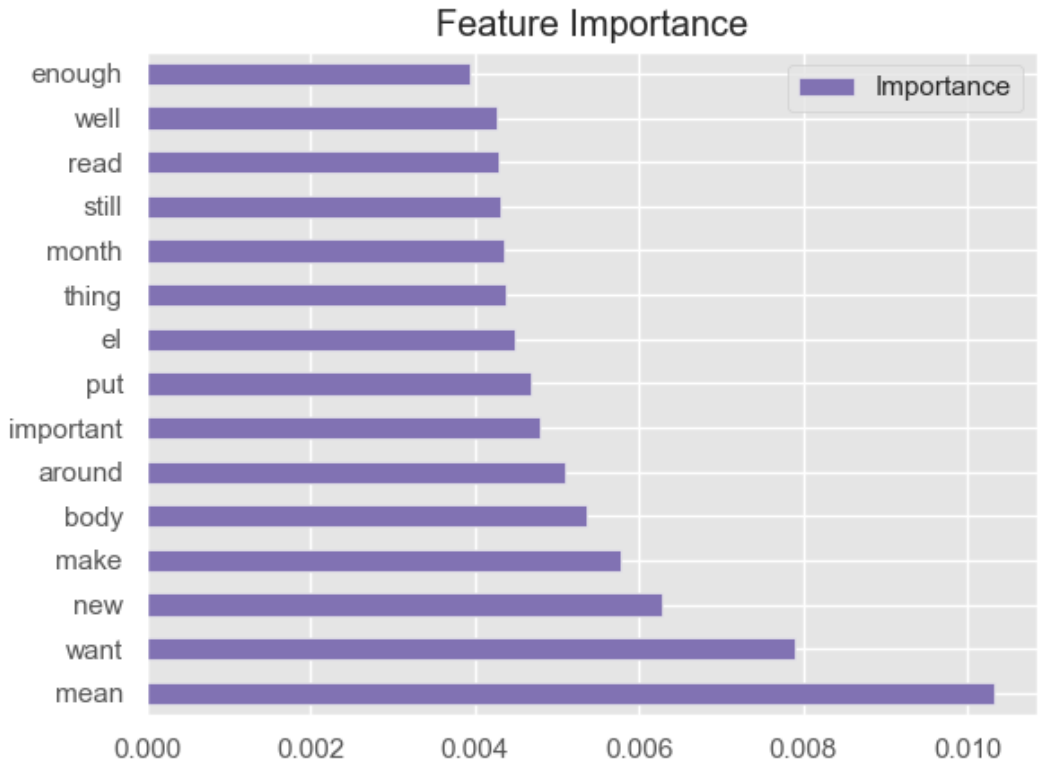


	Words	Importance
14	de	0.010251
1	amp	0.009972
68	say	0.008626
15	different	0.007934
35	im	0.007934
...	...	...
31	great	0.000000
30	good	0.000000
29	going	0.000000
28	go	0.000000
99	youre	0.000000

100 rows × 2 columns

Cluster 3 consists of 5 words (want, I'm, different, say, de) that are deemed to be important.

Unigram – Cluster 4 Feature Importance

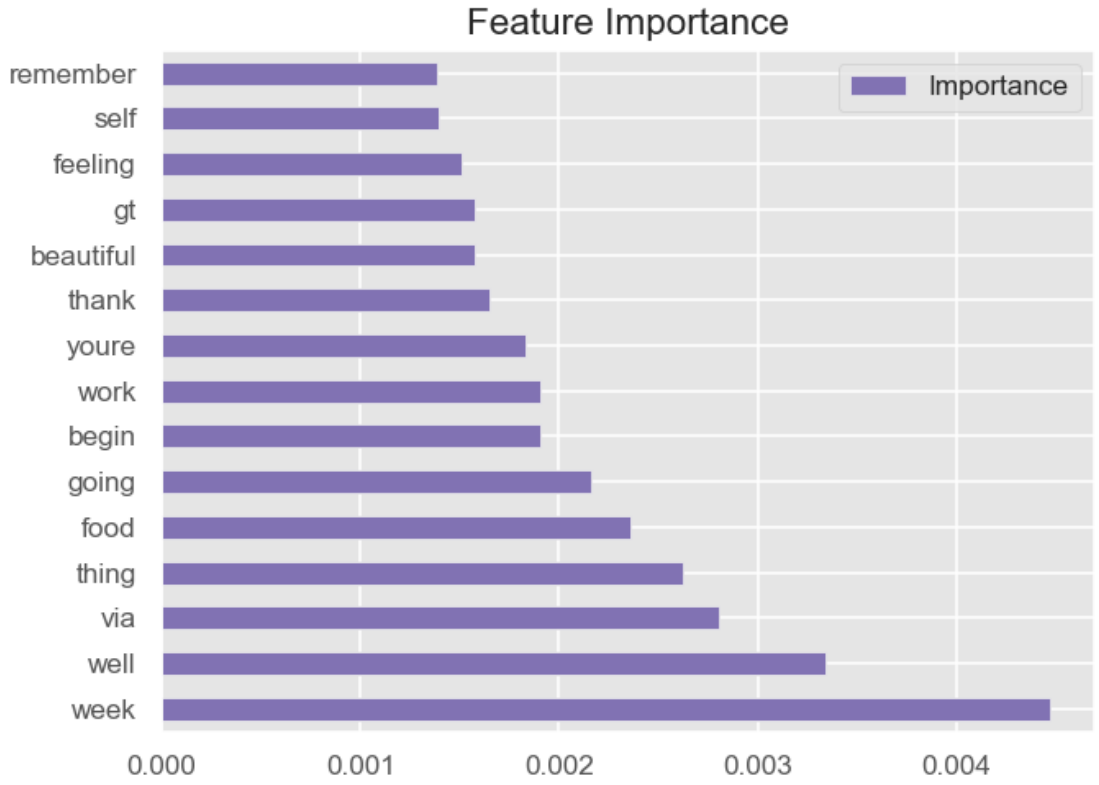


	Words	Importance
51	mean	0.010328
91	want	0.007892
58	new	0.006282
47	make	0.005764
8	body	0.005361
...	...	...
32	gt	0.000000
31	great	0.000000
30	good	0.000000
29	going	0.000000
99	youre	0.000000

100 rows × 2 columns

Cluster 4 contains more important features than other clusters. The top words in this cluster are mean, want, new and make.

Unigram – Cluster 5 Feature Importance

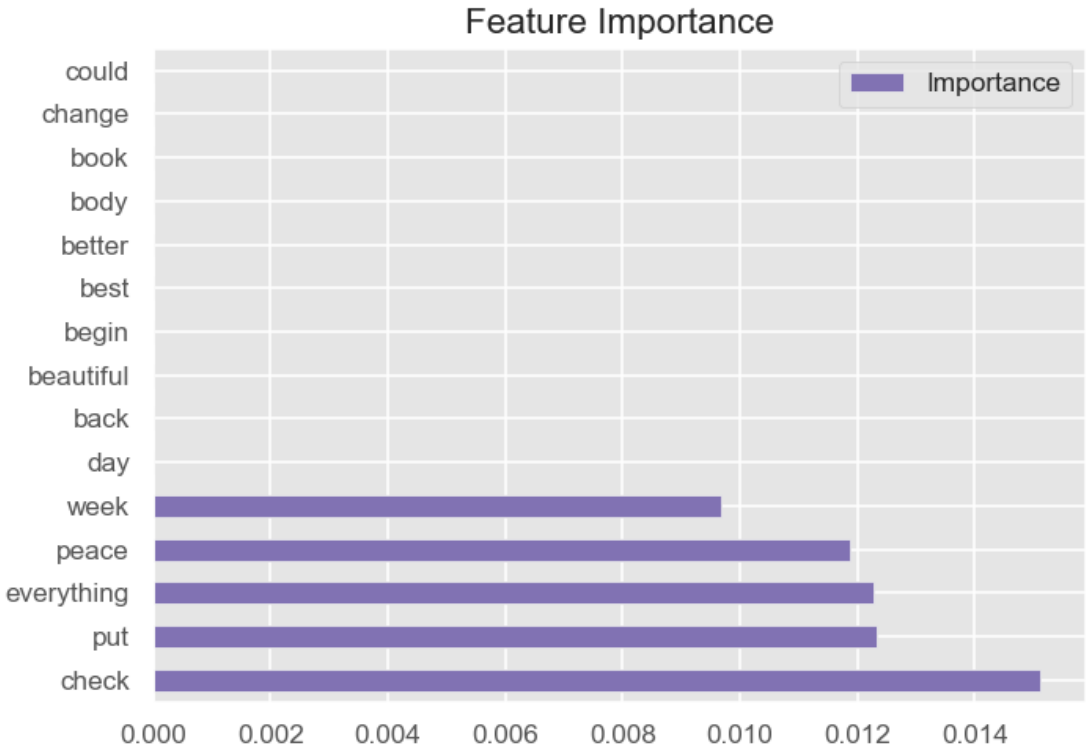


	Words	Importance
94	week	0.004470
95	well	0.003341
89	via	0.002801
81	thing	0.002619
25	food	0.002361
...	...	...
60	peace	0.000000
58	new	0.000000
53	month	0.000000
51	mean	0.000000
21	everything	0.000000

100 rows × 2 columns

Cluster 5 also contains more important features than other clusters. The top words in this cluster are week, well, via and thing.

Unigram – Cluster 6 Feature Importance



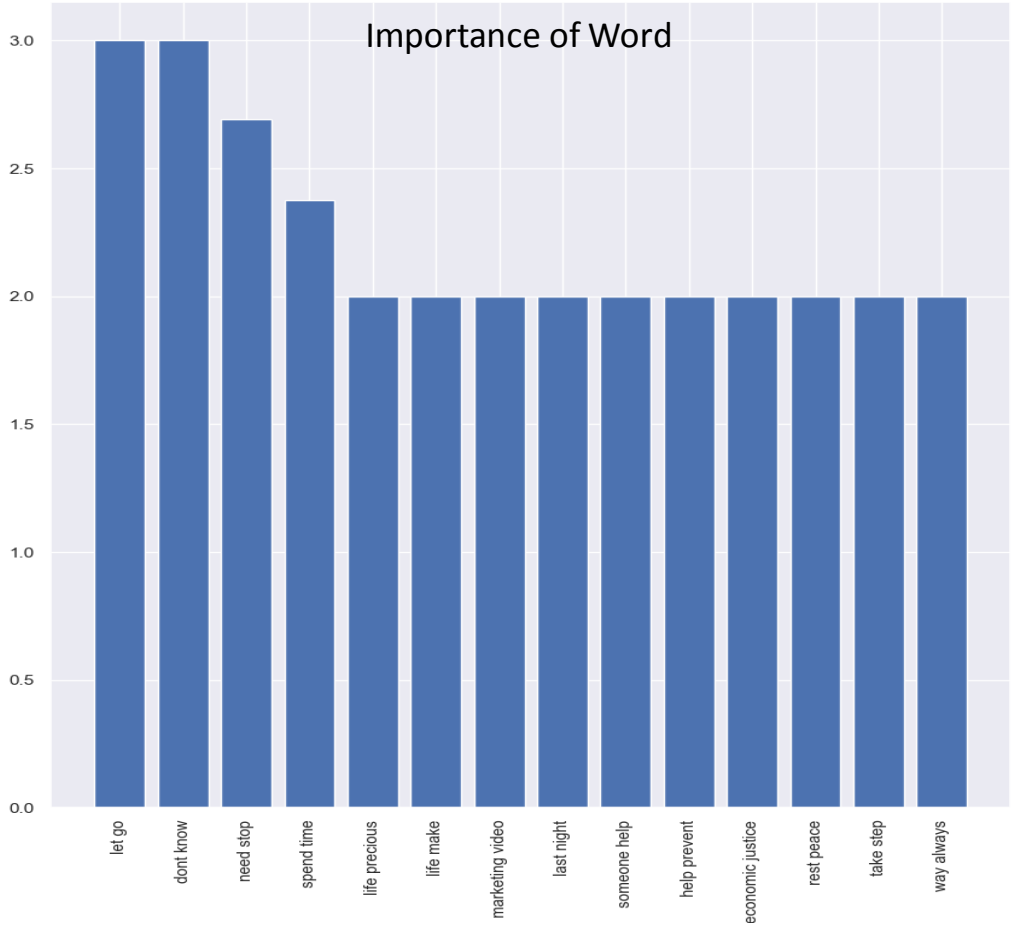
	Words	Importance
11	check	0.015121
63	put	0.012317
21	everything	0.012267
60	peace	0.011868
94	week	0.009680
...	...	...
32	gt	0.000000
31	great	0.000000
30	good	0.000000
29	going	0.000000
99	youre	0.000000

100 rows × 2 columns

This cluster only contains 5 important features which are ‘check’, ‘put’, ‘everything’, ‘peace’ and ‘week’.

Bigram Cluster

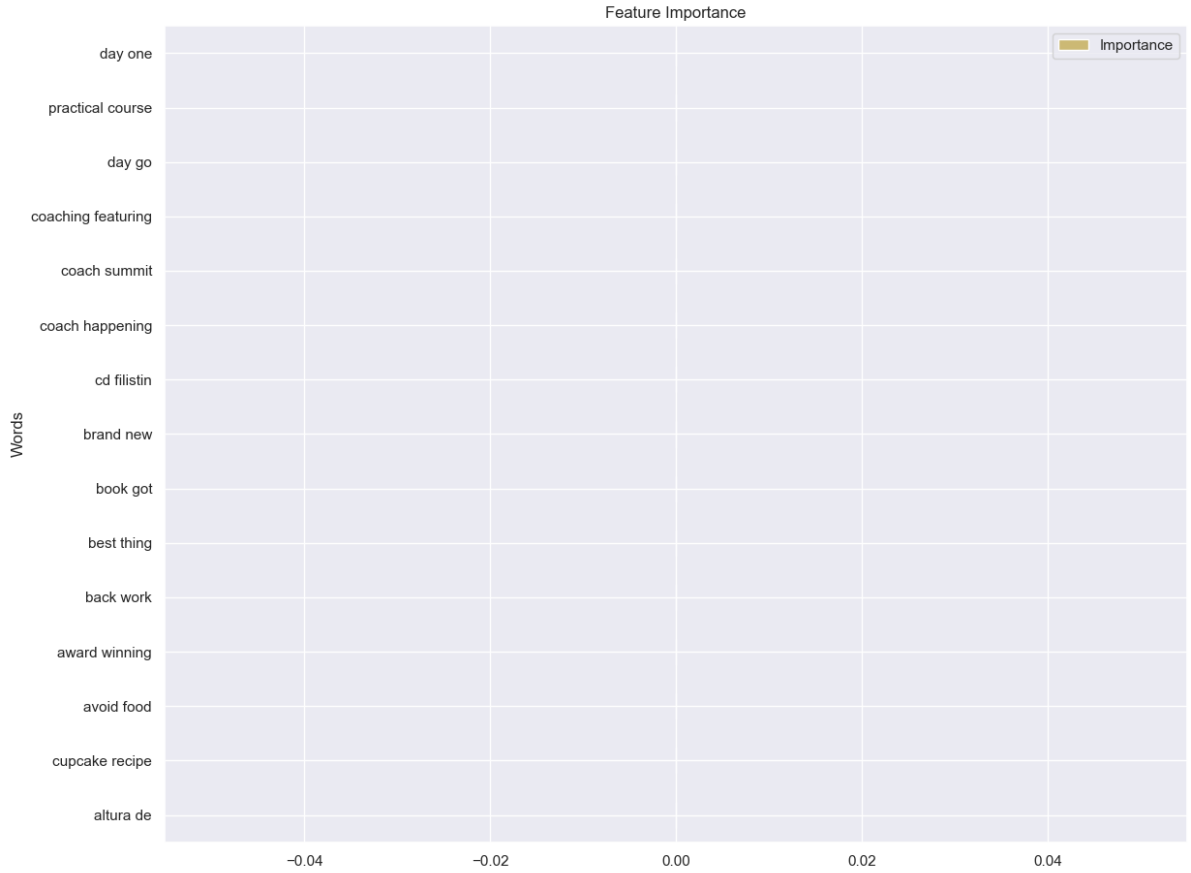
```
['altura de', 'arjantin cd', 'avoid food', 'award winning', 'back work']
['cupcake recipe', 'power change', 'spiritual need', 'altura de', 'arjantin cd']
['day one', 'arjantin cd', 'reminder matter', 'de la', 'great workplace']
['may seem', 'wait coach', 'people getting', 'made feel', 'cd filistin']
['week product', 'work smarter', 'value coaching', 'spiritual need', 'filistin sk']
['coaching featuring', 'practical course', 'excited share', 'petition via', 'week product']
```



- Cluster 1: Arjantin Street in Turkey; avoiding unhealthy food
- Cluster 2: Positive motivation; recipes
- Cluster 3: Healthy working environment; reminding one's self
- Cluster 4: Filistin Street in Ankara, Turkey; coaching
- Cluster 5: Living a healthy life physically, emotionally, spiritually
- Cluster 6: Petitions; coaching

"Let go", "Don't know", "Need stop" and "Spend time" are the most important words.

Bigram – Cluster 1 Feature Importance

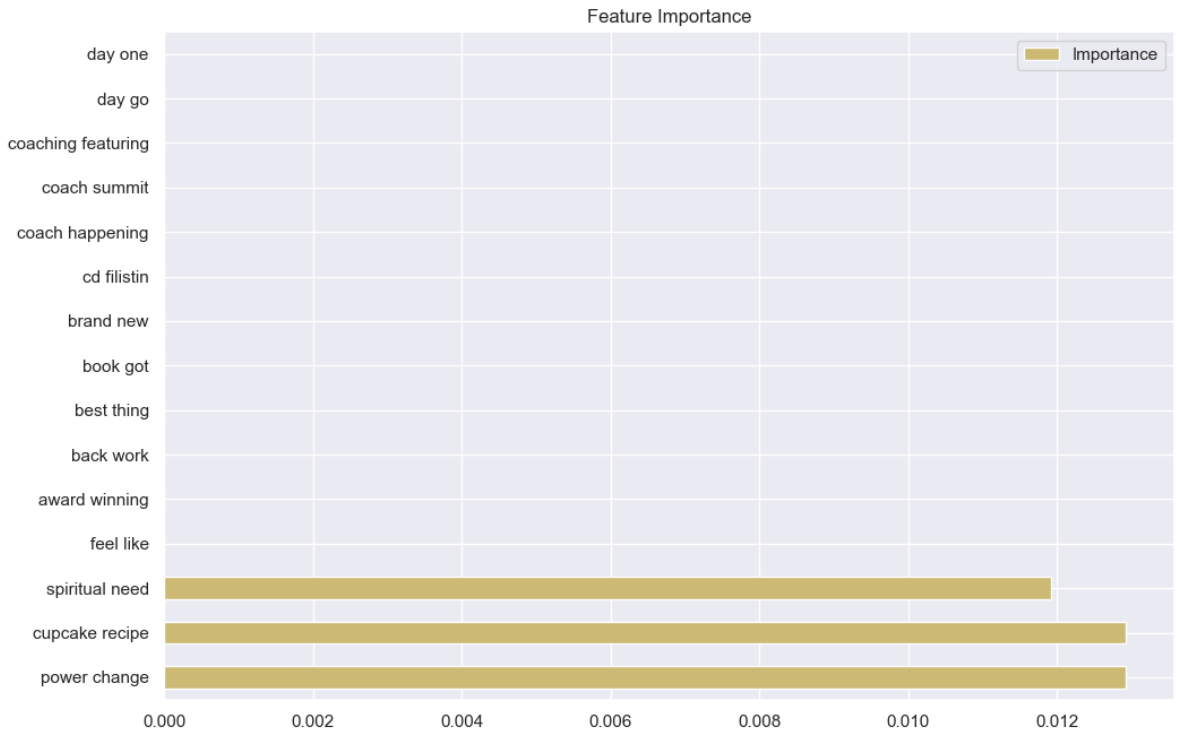


	Words	Importance
0	altura de	0.0
63	practical course	0.0
73	schne tochter	0.0
72	san diego	0.0
71	review week	0.0
...	...	...
30	go back	0.0
29	friend sent	0.0
28	friday night	0.0
27	freelance translation	0.0
99	yummiest snack	0.0

100 rows × 2 columns

Cluster 1 consists of words that are useless and irrelevant.

Bigram – Cluster 2 Feature Importance

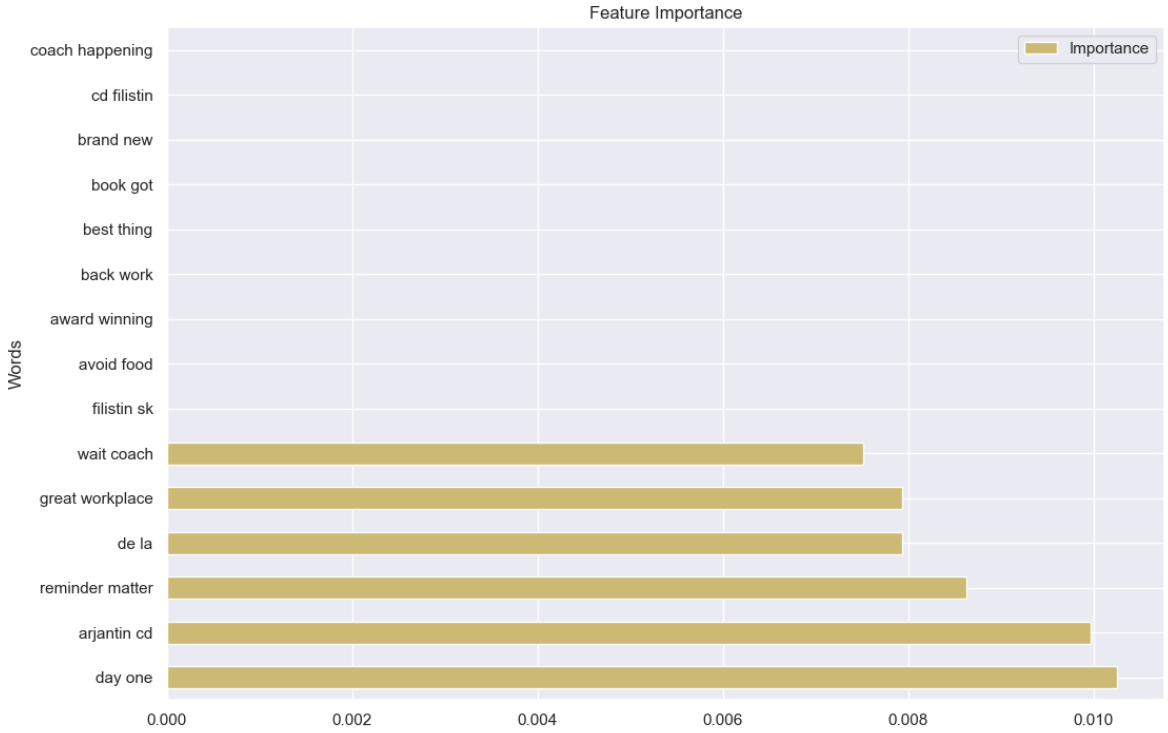


	Words	Importance
62	power change	0.012931
12	cupcake recipe	0.012931
81	spiritual need	0.011922
0	altura de	0.000000
64	product great	0.000000
...	...	...
31	good health	0.000000
30	go back	0.000000
29	friend sent	0.000000
28	friday night	0.000000
99	yummiest snack	0.000000

100 rows × 2 columns

Cluster 2 consists of 3 phrases (spiritual need, power change and cupcake recipe) that are deemed to be important.

Bigram – Cluster 3 Feature Importance

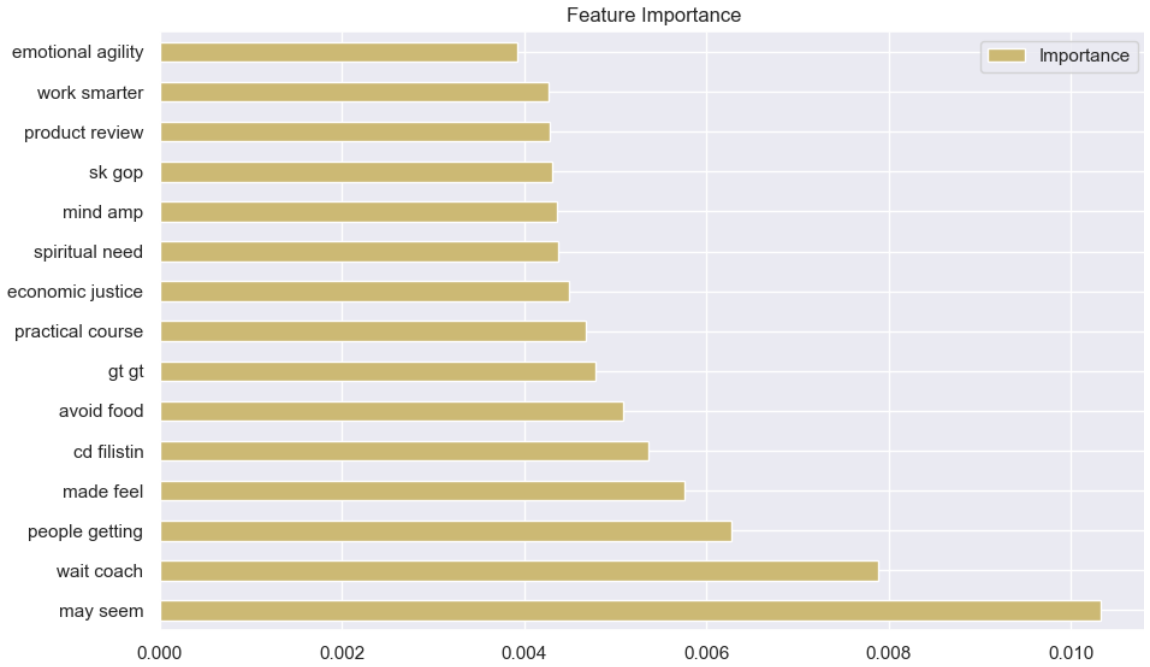


	Words	Importance
14	day one	0.010251
1	arjantin cd	0.009972
68	reminder matter	0.008626
15	de la	0.007934
35	great workplace	0.007934
...	...	...
31	good health	0.000000
30	go back	0.000000
29	friend sent	0.000000
28	friday night	0.000000
99	yummiest snack	0.000000

100 rows × 2 columns

Cluster 3 consists of 5 words that are important (wait coach, great workplace, de la, reminder matter, arjantin cd, day one).

Bigram – Cluster 4 Feature Importance

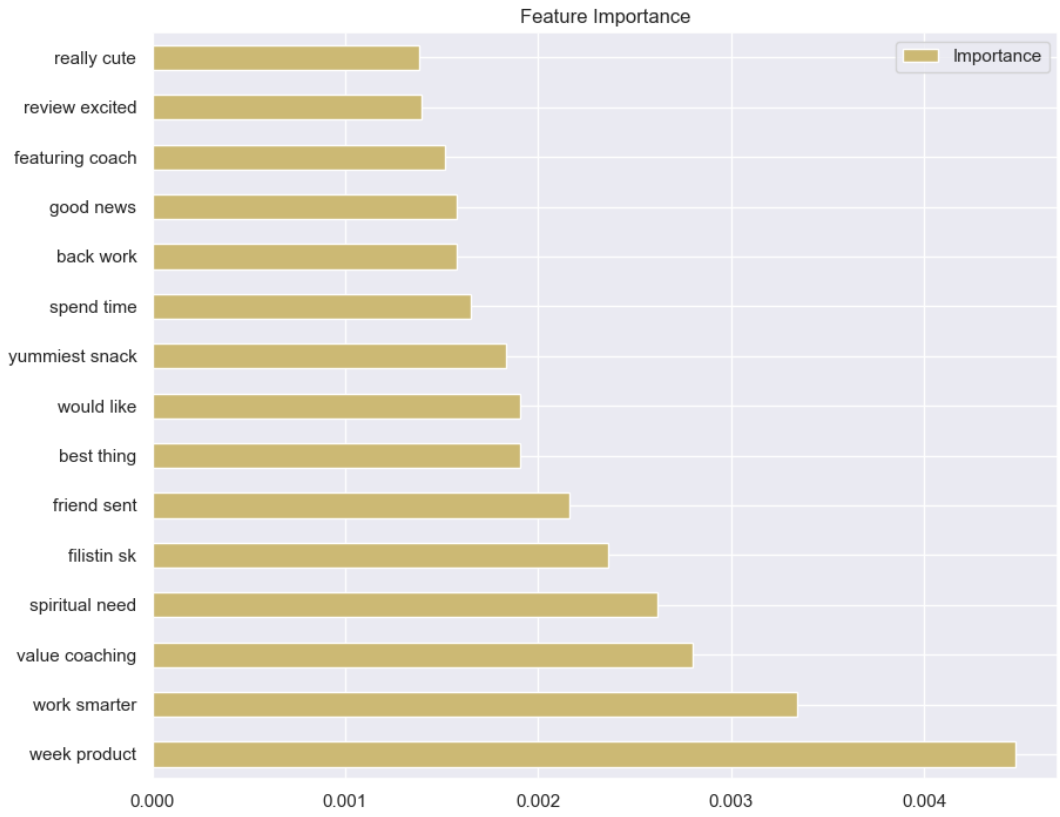


	Words	Importance
51	may seem	0.010328
91	wait coach	0.007892
58	people getting	0.006282
47	made feel	0.005764
8	cd filistin	0.005361
...	...	...
32	good news	0.000000
31	good health	0.000000
30	go back	0.000000
29	friend sent	0.000000
99	yummiest snack	0.000000

100 rows × 2 columns

The top words/phrases for cluster 4 are ‘may seem’, ‘wait coach’, ‘people getting’, ‘made feel’, ‘cd filistin’ and ‘avoid food’.

Bigram – Cluster 5 Feature Importance

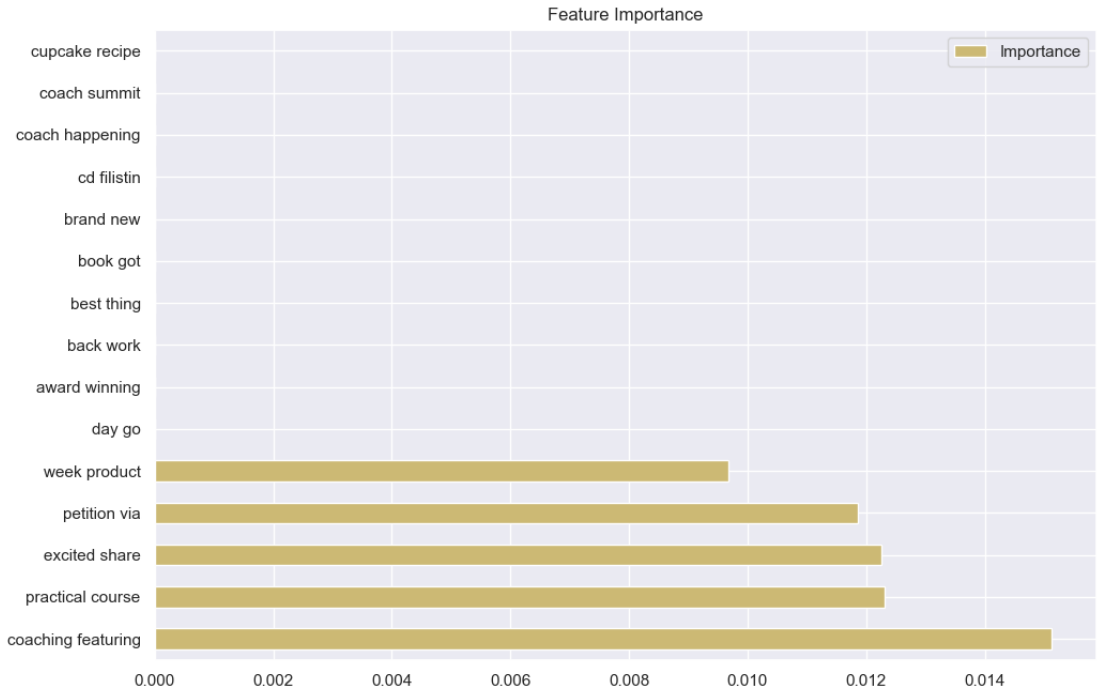


	Words	Importance
94	week product	0.004470
95	work smarter	0.003341
89	value coaching	0.002801
81	spiritual need	0.002619
25	filistin sk	0.002361
...	...	...
60	petition via	0.000000
58	people getting	0.000000
53	mind amp	0.000000
51	may seem	0.000000
21	excited share	0.000000

100 rows × 2 columns

Cluster 5's top words/phrases are 'week product', 'work smarter', 'value coaching' and 'spiritual need'.

Bigram – Cluster 6 Feature Importance



	Words	Importance
11	coaching featuring	0.015121
63	practical course	0.012317
21	excited share	0.012267
60	petition via	0.011868
94	week product	0.009680
...	...	...
32	good news	0.000000
31	good health	0.000000
30	go back	0.000000
29	friend sent	0.000000
99	yummiest snack	0.000000

100 rows × 2 columns

Cluster 6, on the other hand, contains only 5 important features.

Conclusion

Some insights can be inferred:

- Clusters 4 and 5 have the most number of important features.
- The audience frequently talks about work, positivity, peacefulness, food, self-love, healthy living and some streets in Turkey.