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

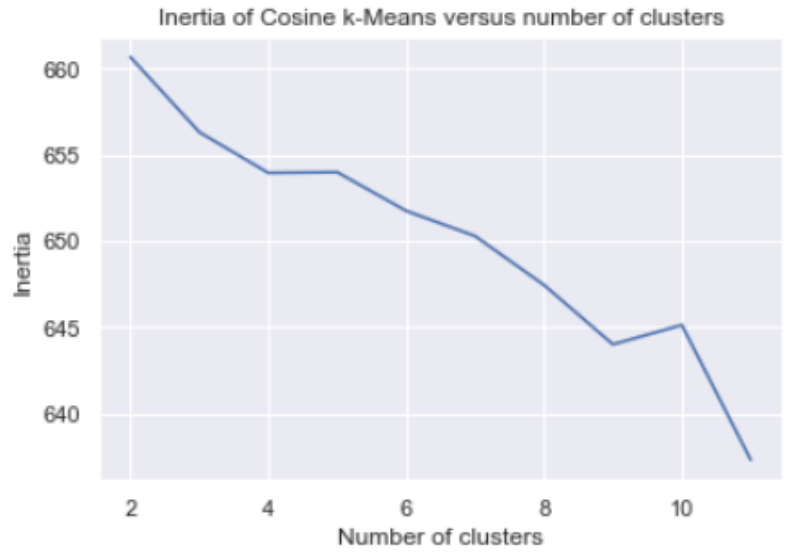
Word Cloud

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

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

Determine number of Tweet Clusters¶

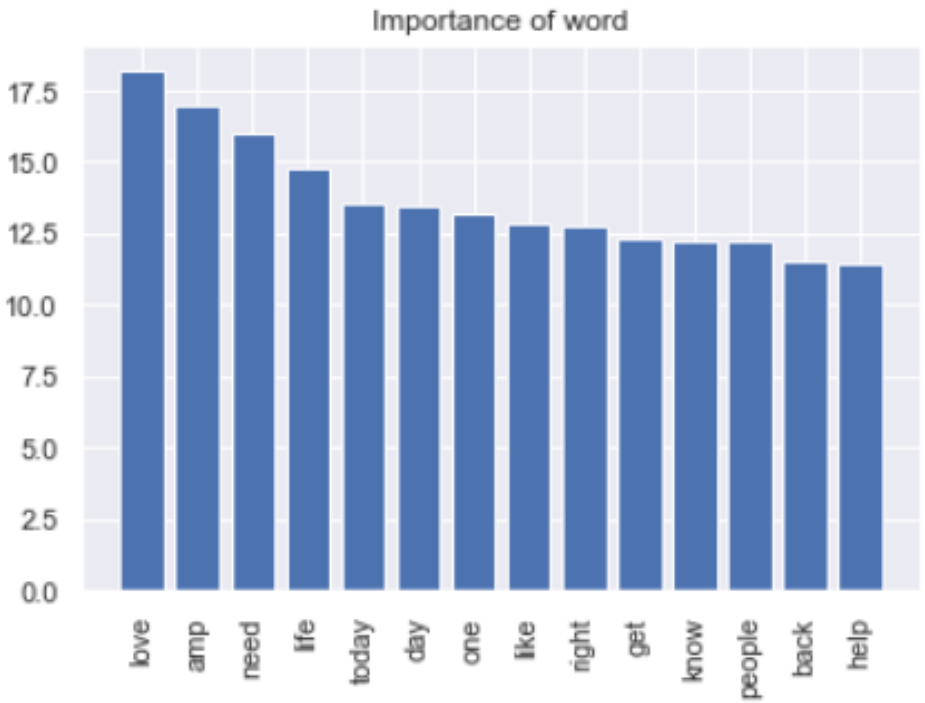
```
:  #high dimension data
   from sklearn import preprocessing
   from sklearn.cluster import KMeans
   normalized_vectors = preprocessing.normalize(tv_transform)
   scores = [KMeans(n_clusters=i+2).fit(normalized_vectors).inertia_
             for i in range(10)]
   sns.lineplot(np.arange(2, 12), scores)
   plt.xlabel('Number of clusters')
   plt.ylabel("Inertia")
   plt.title("Inertia of Cosine k-Means versus number of clusters")
   plt.savefig("intertia_cosine_kmeans.jpg", dpi=300)
```



There should be 4 clusters according to the Elbow method

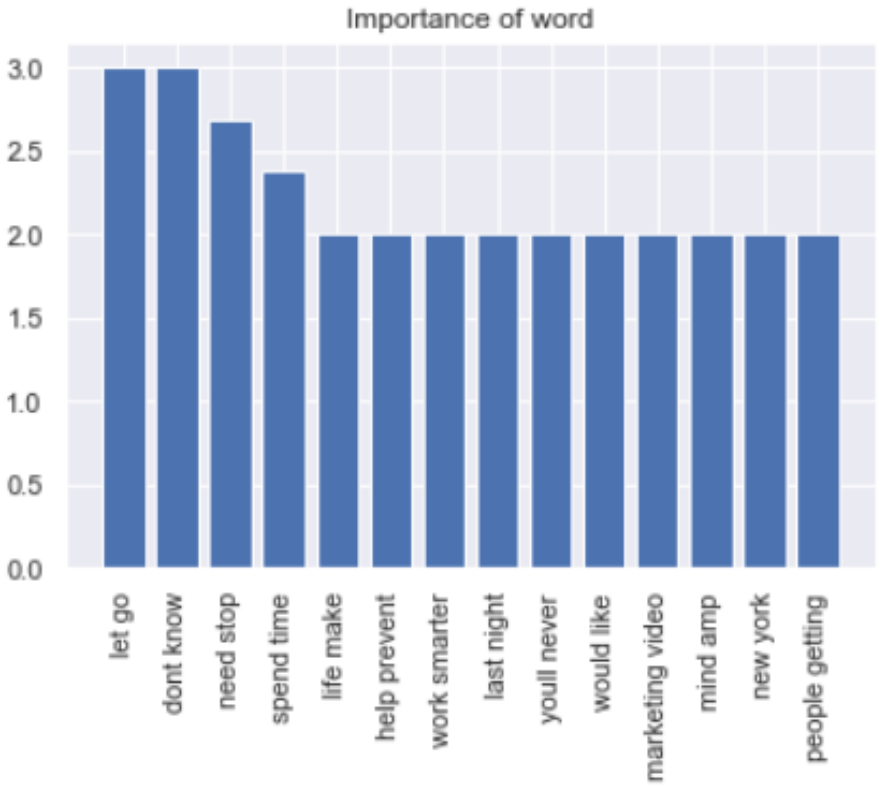
Unigram Cluster

```
['love', 'amp', 'need']
['nan', 'sure', 'always']
['know', 'dr', 'week']
['want', 'dont', 'work']
```



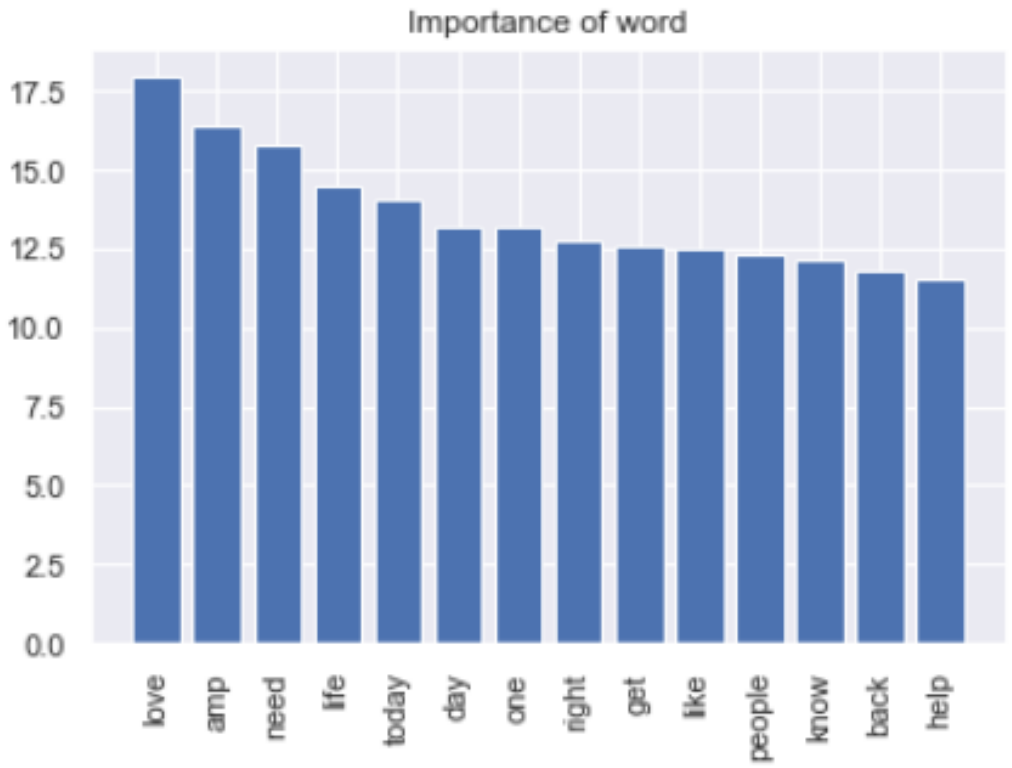
Bigram Cluster

```
['altura de', 'avoid food', 'award winning']
['piece dispelling', 'way always', 'really cute']
['way always', 'youll never', 'altura de']
['rest peace', 'way always', 'wearing mask']
```

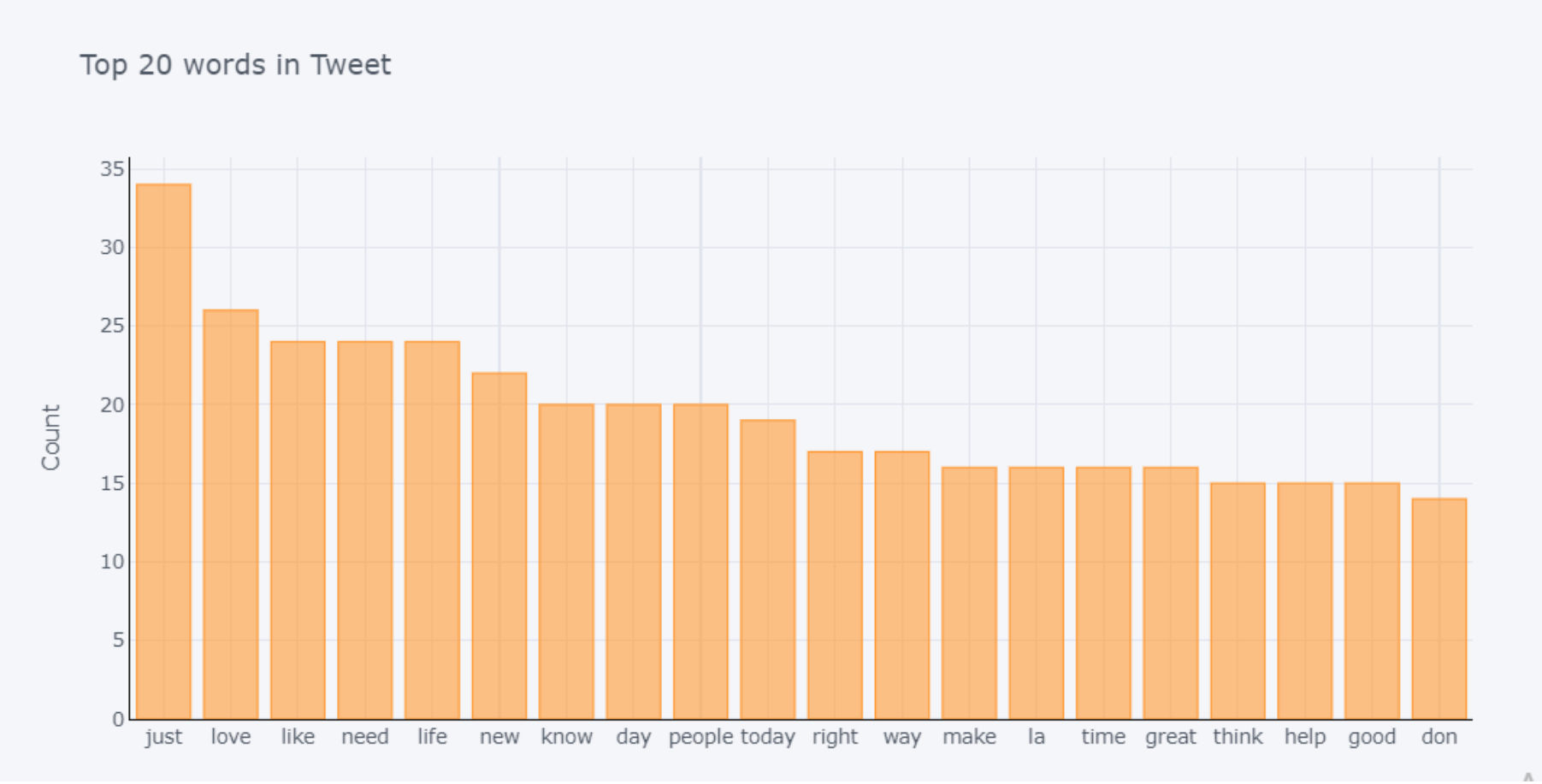


Unigram & Bigram

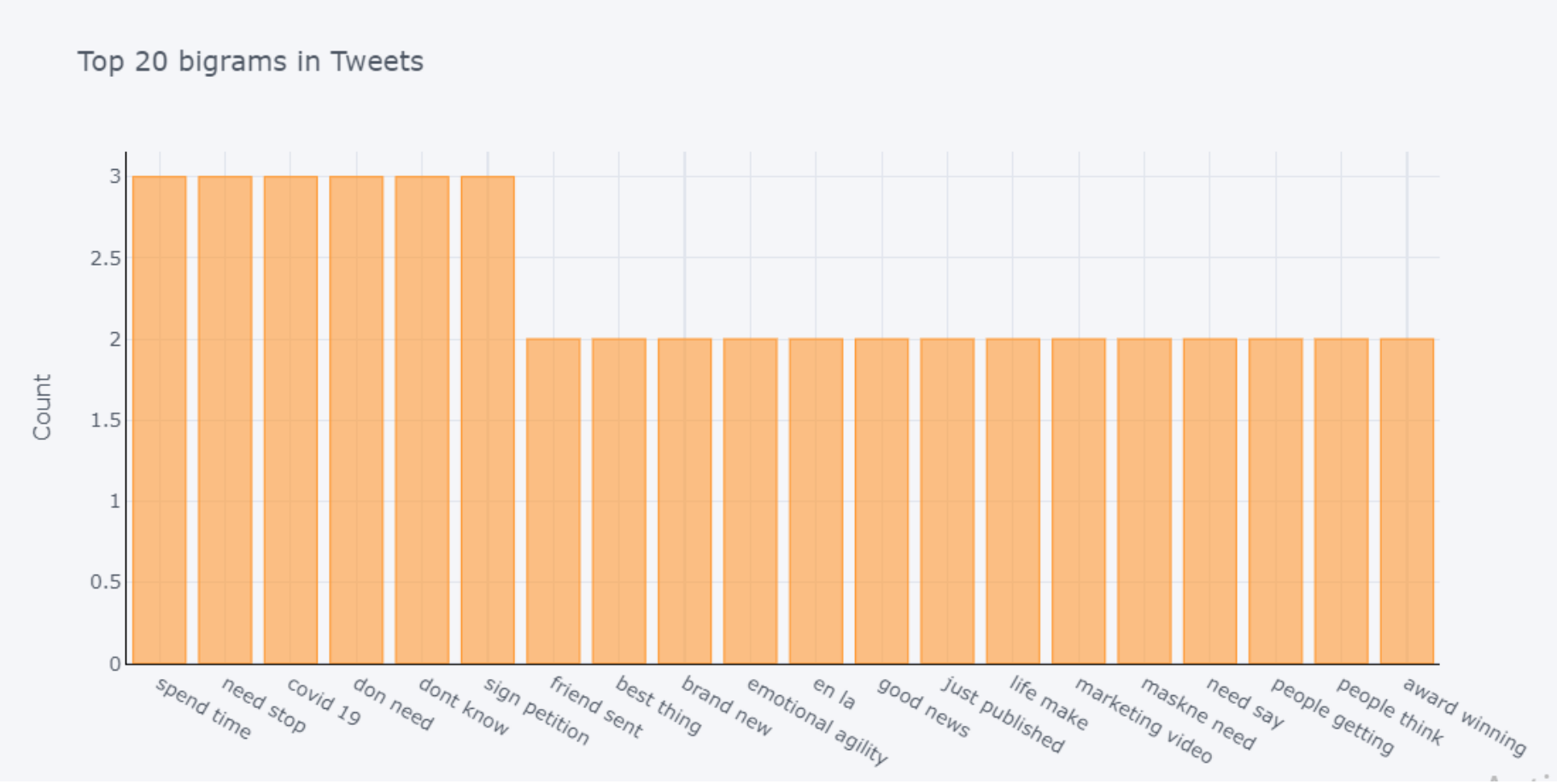
```
['always', 'amp', 'around', 'back', 'beautiful']
['product', 'week', 'thank', 'could', 'well']
['week', 'youre', 'always', 'amp', 'around']
['thing', 'week', 'well', 'via', 'want']
```



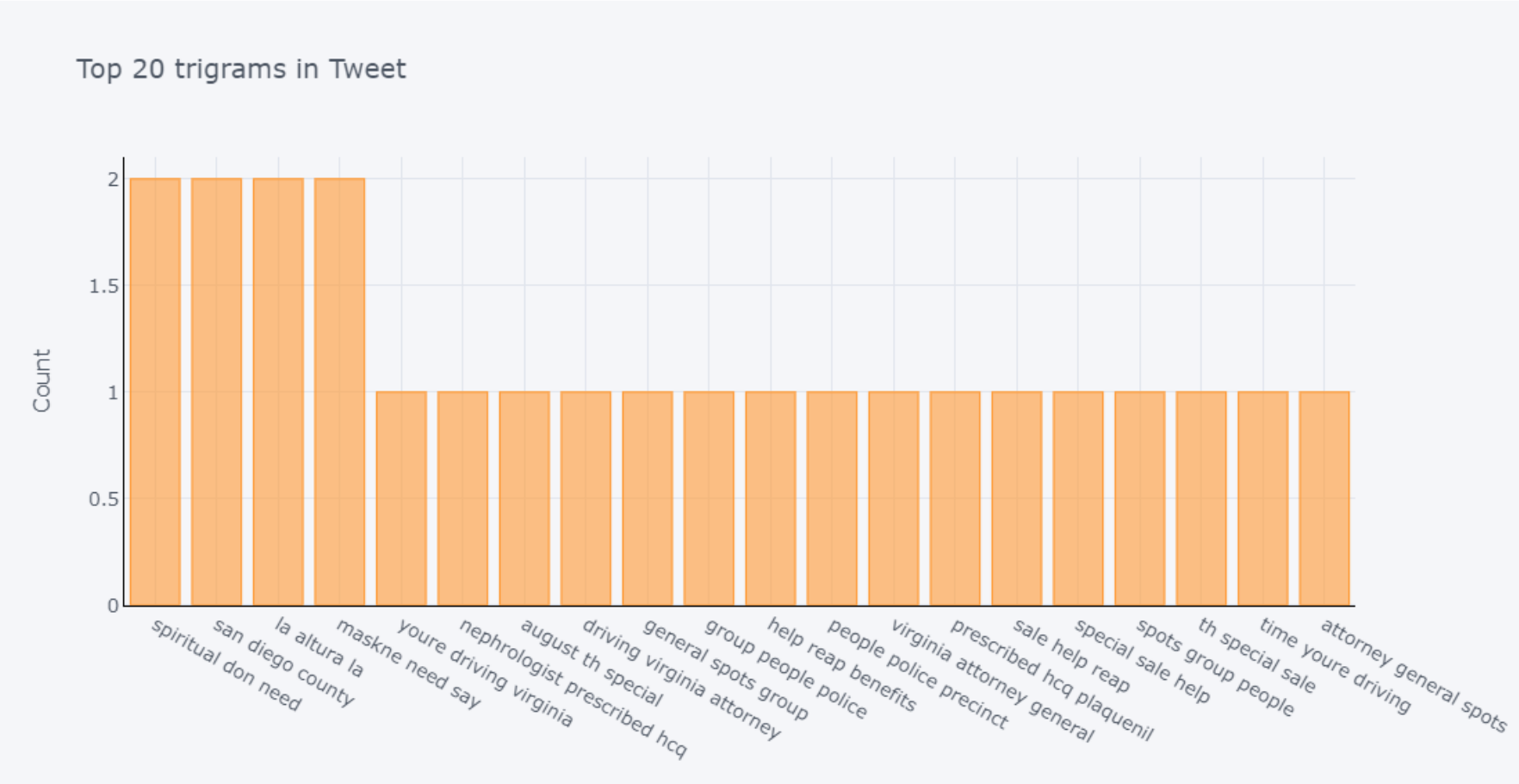
Distribution of Word - Unigram



Distribution of Word - Bigram



Distribution of Word - Trigram



Latent semantic analysis

```
Topic 1:  love society fear
Topic 2:  just like day
Topic 3:  good pretty health
Topic 4:  need people today
Topic 5:  thank don sent
Topic 6:  right life beautiful
```