Principal Component Analysis

Features extraction and representation

Principal Component Analysis

Greater the feature - > need of more training sample

Training set is small - > increase in feature -> Degrades the classifier - >

Peaking Phenomenon

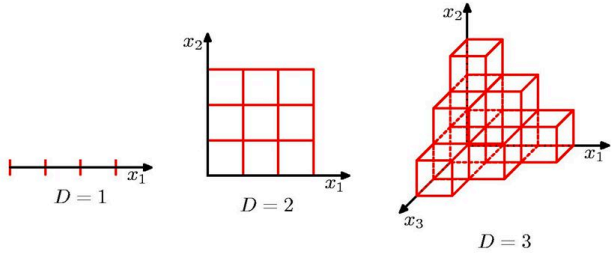
Desirable to have small set of Feature without compromising the classifier quality

Curse of
DIMENSIONALITY

As the dimensionality of the features space increases, the number Configurations can grow exponentially, and thus the number of configurations covered by an observation decreases.

Chris Albon

Curse of Dimensionality

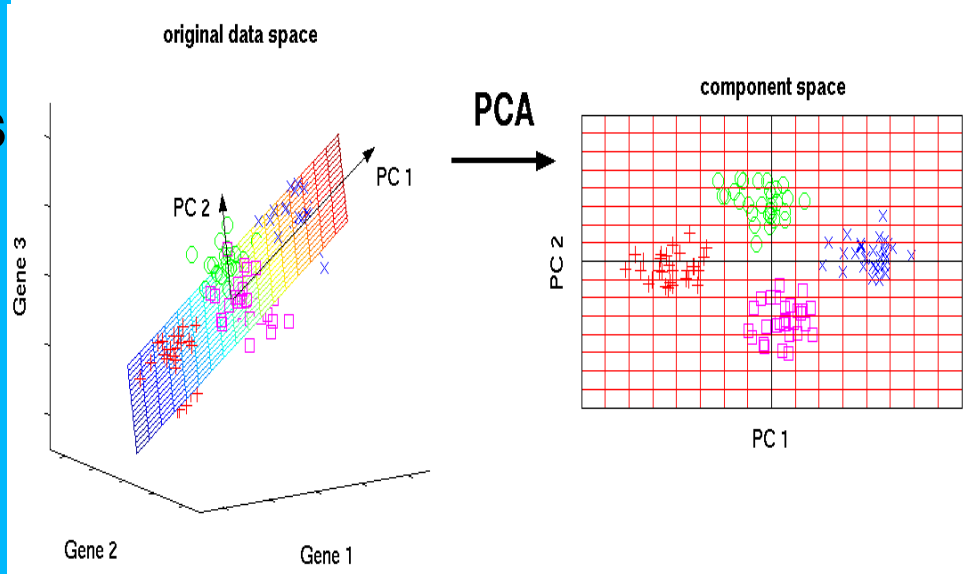


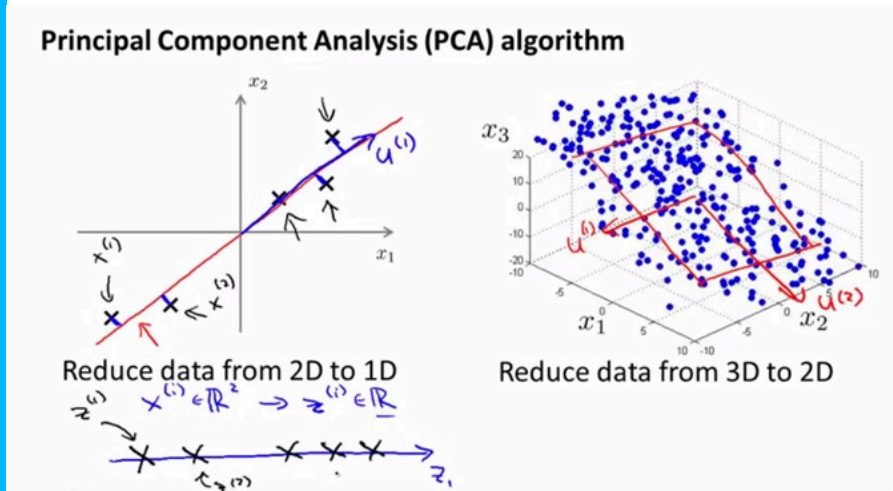
- No. of cells grow exponentially with D
- Need exponentially large no. of training data points
- Not a good approach for more than a few dimensions!

Reference: Christopher M Bishop: Pattern Recognition & Machine Leaning, 2006 Springer

Principal Components Analys

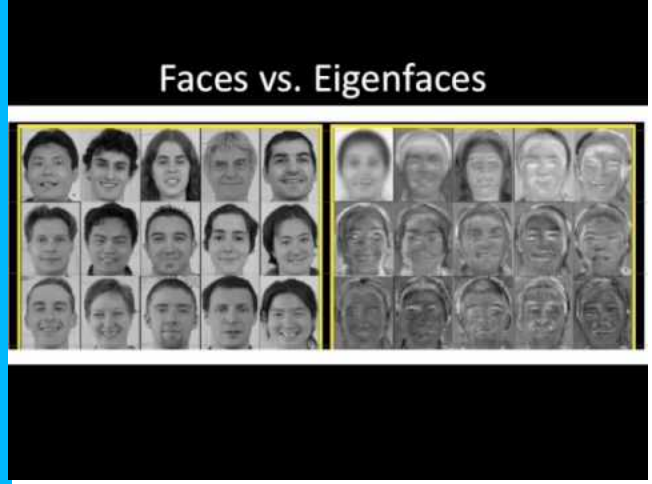
1. Does the data set ‘span’ the whole of d dimensional space?
2. For a matrix of *m* samples x *n* genes, create a new covariance matrix of size *n* x *n*.
3. Transform some large number of variables into a smaller number of uncorrelated variables called principal components (PCs).
4. Developed to capture as much of the variation in data as possible
5. Line drawn through largest variance line
6. PCA2 is orthogonal to first line PCA1





Example Applications

- Face Recognition
- Image Compression
- Gene Expression Analysis
- Data Reduction
- Data Classification
- Trend Analysis
- Factor Analysis
- Noise Reduction



PCA Principle

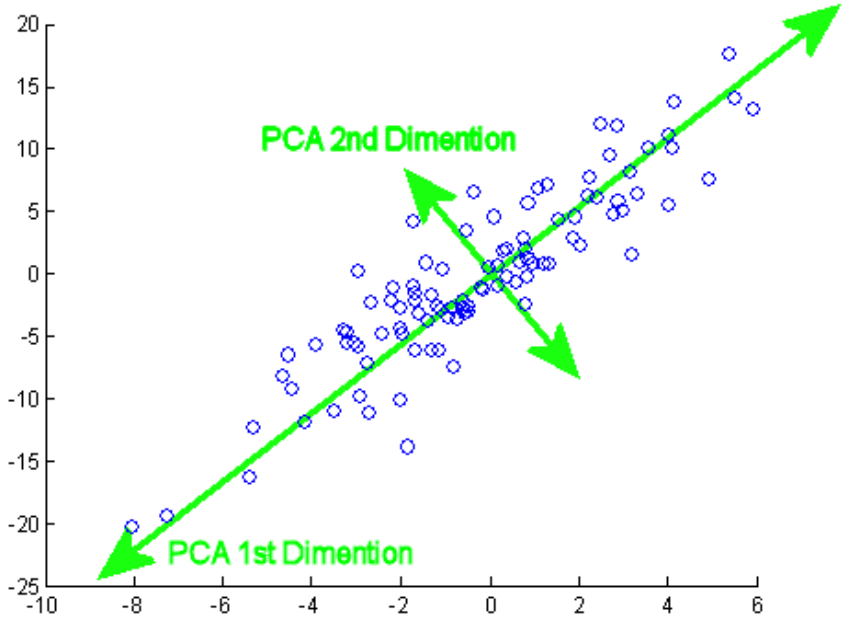
The basic principle or assumption in PCA is:

The eigenvector of a covariance matrix equal to a principal component, because the **eigenvector** with the **largest** eigenvalue is the direction along which the data set has the **maximum variance**.

Each **eigenvector** is associated with a **eigenvalue**;

Eigenvalue è tells how much the variance is;

Eigenvector è tells the direction of the variation;



Step by steps ➔Assume a data with 3 dimensions

Step1: Center the data by subtracting the mean of each column

Original Data

X	Y	Z
2.5	9.4	6.5
0.2	5.6	4.2
6	3.2	0.3
4.2	3.9	6.1
2.3	5	5.2
11	7	0.56
2.6	0.3	0.9
3.4	0.02	1.81
3.3	6.5	2.13
8	2	4.2

Mean(X) = 4.35
Mean (Y) = 4.292
Mean (Z) = 3.19

For example,
X11 = 2.5
Update:
X11 = 2.5-4.35 = -1.85

Transformed Data

X	Y	Z
-1.85	5.108	3.31
-4.15	1.308	1.01
1.65	-1.092	-2.89
-0.15	-0.392	2.91
-2.05	0.708	2.01
6.65	2.708	-2.63
-1.75	-3.992	-2.29
-0.95	-4.272	-1.38
-1.05	2.208	-1.06
3.65	-2.292	1.01

Step2: Compute covariance matrix based on the transformed data

X	Y	Z
-1.85	5.108	3.31
-4.15	1.308	1.01
1.65	-1.092	-2.89
-0.15	-0.392	2.91
-2.05	0.708	2.01
6.65	2.708	-2.63
-1.75	-3.992	-2.29
-0.95	-4.272	-1.38
-1.05	2.208	-1.06
3.65	-2.292	1.01

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

```
>> C=cov(M)

C =

   10.0228    0.0329   -3.0013
    0.0329    9.0331    2.7696
   -3.0013    2.7696    5.4495
```

Step3: Calculate eigenvectors & eigenvalues from covariance matrix

```
>> C=cov(M)

C =

   10.0228    0.0329   -3.0013
    0.0329    9.0331    2.7696
   -3.0013    2.7696    5.4495
```



```
>> [V,D]=eig(C)

V =

    0.3603   -0.5838    0.7275
   -0.3863   -0.8033   -0.4533
    0.8491   -0.1177   -0.5150


D =

    2.9155         0         0
         0    9.4629         0
         0         0   12.1270
```

[V,D] = eig (Covariance Matrix)

V = eigenvectors (each column)

D = eigenvalues (in the diagonal line)

For example, 2.9155 correspond to vector <0.3603, -0.3863, 0.8491>

Each eigenvector is considered as a principle component.

Step4: Order the eigenvalues from largest to smallest, where the eigenvectors will also be re-ordered; and then we can select the top-K one; for example, we set K=2

```
>> [V,D]=eig(C)

V =

    0.3603   -0.5838    0.7275
   -0.3863   -0.8033   -0.4533
    0.8491   -0.1177   -0.5150


D =

    2.9155         0         0
         0    9.4629         0
         0         0   12.1270
```

K=2, it indicates that we will reduce the dimensions to be 2. The last second columns are extracted to have an EigenMatrix

```
VM =

   -0.5838    0.7275
   -0.8033   -0.4533
   -0.1177   -0.5150
```

Step5: Project the original data to those eigenvectors to formulate the new data matrix

Original data, D, 10x3

X	Y	Z
2.5	9.4	6.5
0.2	5.6	4.2
6	3.2	0.3
4.2	3.9	6.1
2.3	5	5.2
11	7	0.56
2.6	0.3	0.9
3.4	0.02	1.81
3.3	6.5	2.13
8	2	4.2

Transformed data, TD, 10x3

X	Y	Z
-1.85	5.108	3.31
-4.15	1.308	1.01
1.65	-1.092	-2.89
-0.15	-0.392	2.91
-2.05	0.708	2.01
6.65	2.708	-2.63
-1.75	-3.992	-2.29
-0.95	-4.272	-1.38
-1.05	2.208	-1.06
3.65	-2.292	1.01

EigenMatrix, EM, 3X2

```
VM =

   -0.5838    0.7275
   -0.8033   -0.4533
   -0.1177   -0.5150
```

FinalData (10xk) = TD (10x3) x EM (3xk), here k = 2

Step5: Project the original data to those eigenvectors to formulate the new data matrix

Original data, D, 10x3

X	Y	Z
2.5	9.4	6.5
0.2	5.6	4.2
6	3.2	0.3
4.2	3.9	6.1
2.3	5	5.2
11	7	0.56
2.6	0.3	0.9
3.4	0.02	1.81
3.3	6.5	2.13
8	2	4.2

Final Data, FD, 10x2

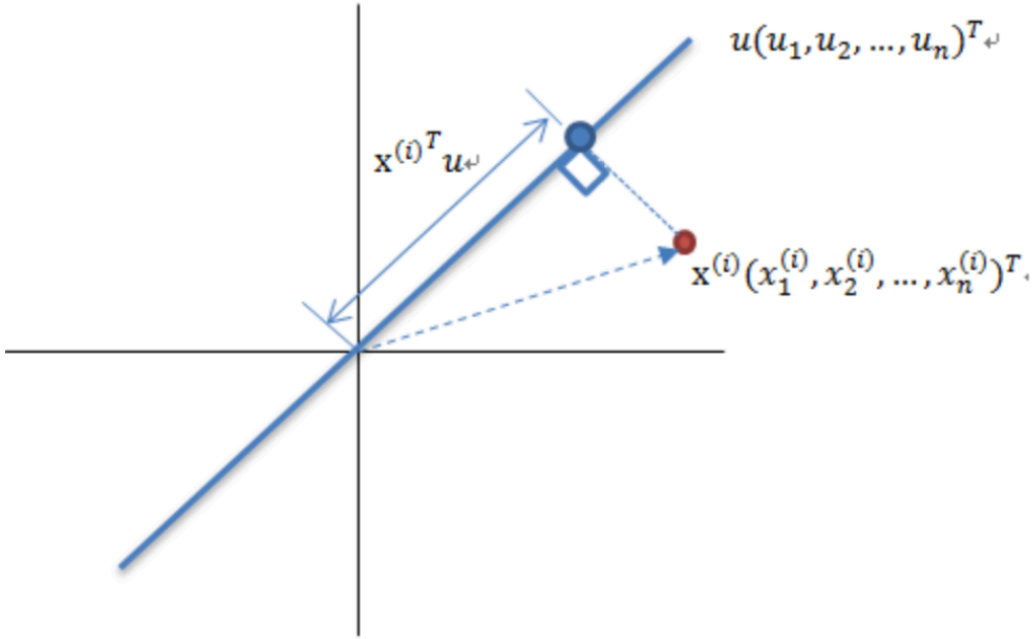
```
FD =

   -3.4128   -5.3660
    1.2532   -4.1322
    0.2541    3.1837
    0.0600   -1.4301
    0.3915   -2.8475
   -5.7481    4.9648
    4.4980    1.7158
    4.1487    1.9561
   -1.0359   -1.2189
   -0.4086    3.1742
```

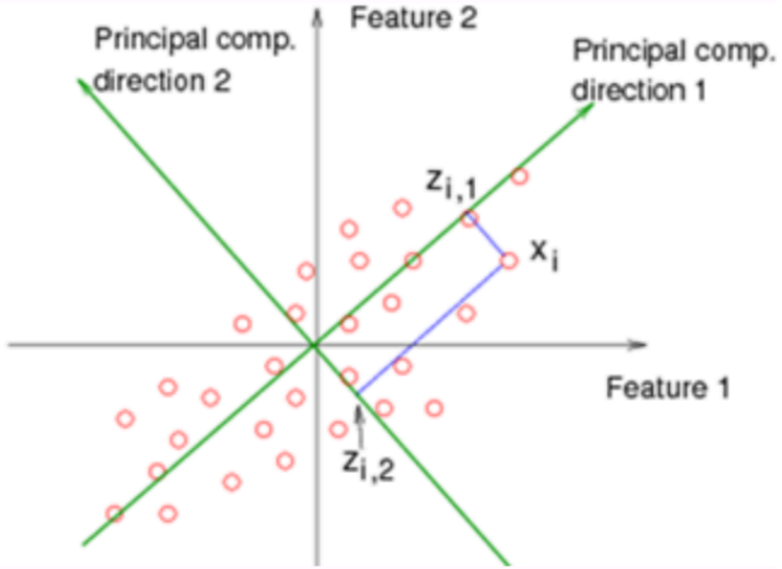
FinalData (10xk) = TD (10x3) x EM (3xk), here k = 2

Step5: Project the original data to those eigenvectors to formulate the new data matrix

The idea of Projection



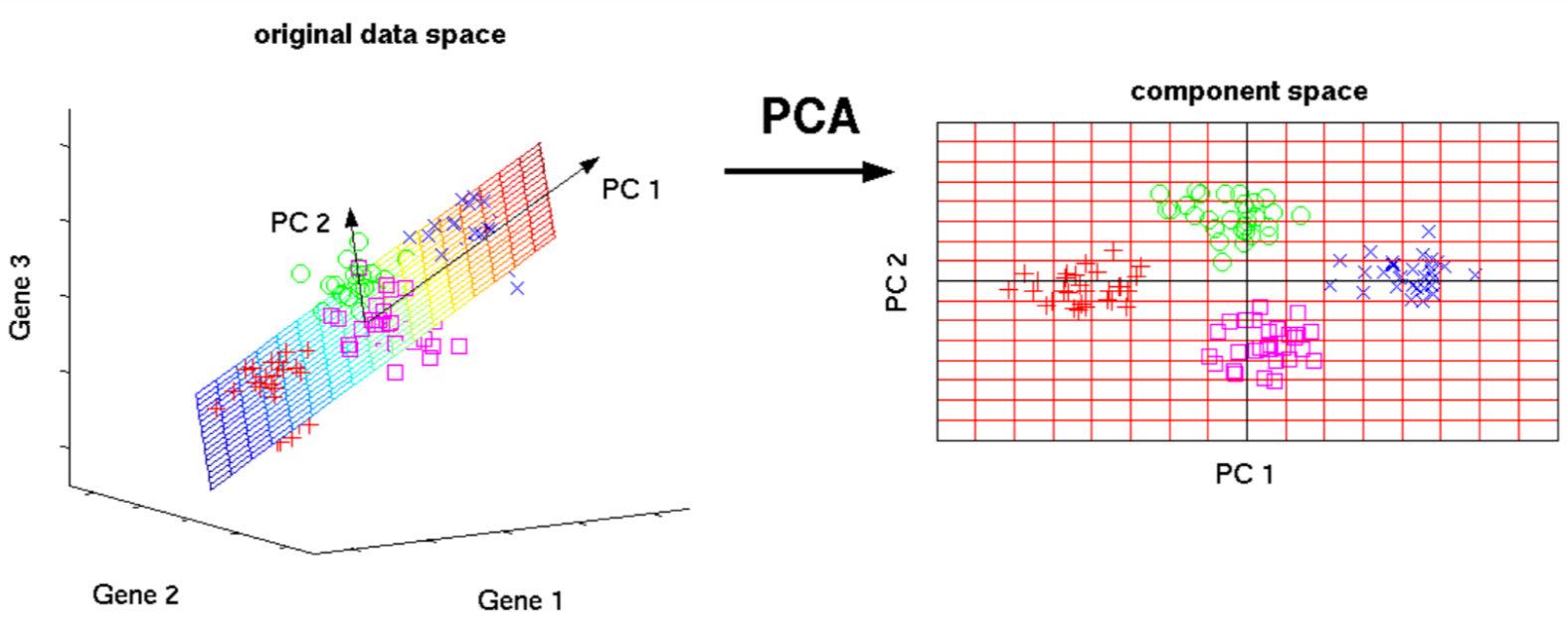
Visualization of PCA



PCA finds a linear projection of high dimensional data into a lower dimensional subspace.

PCA reduces the dimensionality (the number of features) of a data set by maintaining as much variance as possible.

Another example: Gene Expression

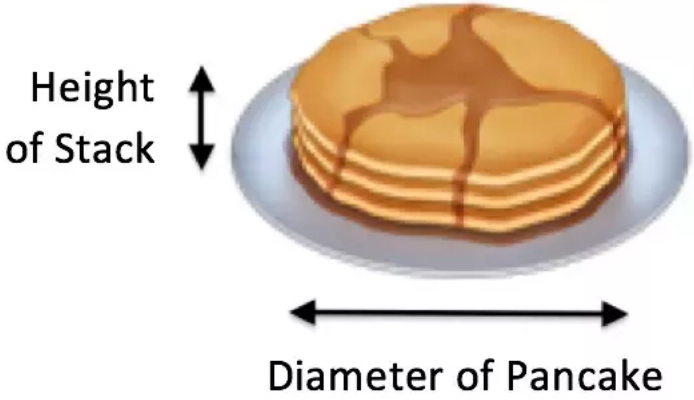


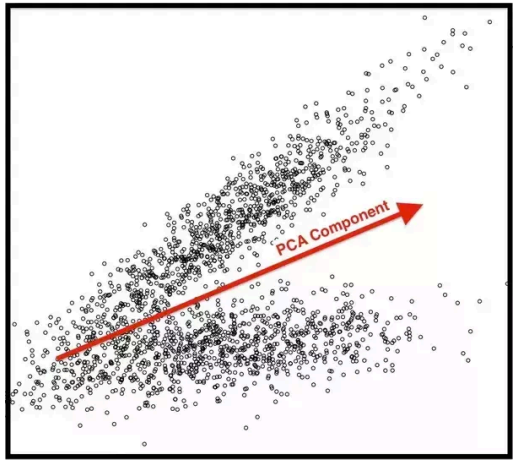
The original expression by 3 genres is projected to two new dimensions, Such two-dimensional visualization of the samples allow us to draw qualitative conclusions about the separability of experimental conditions (marked by different colors).

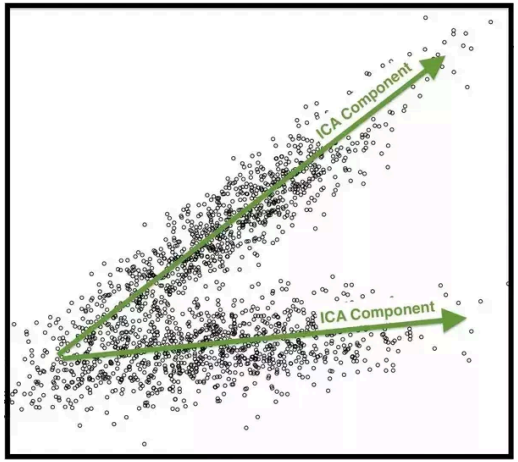
Limitations

Maximizing Spread

Orthogonal Components.







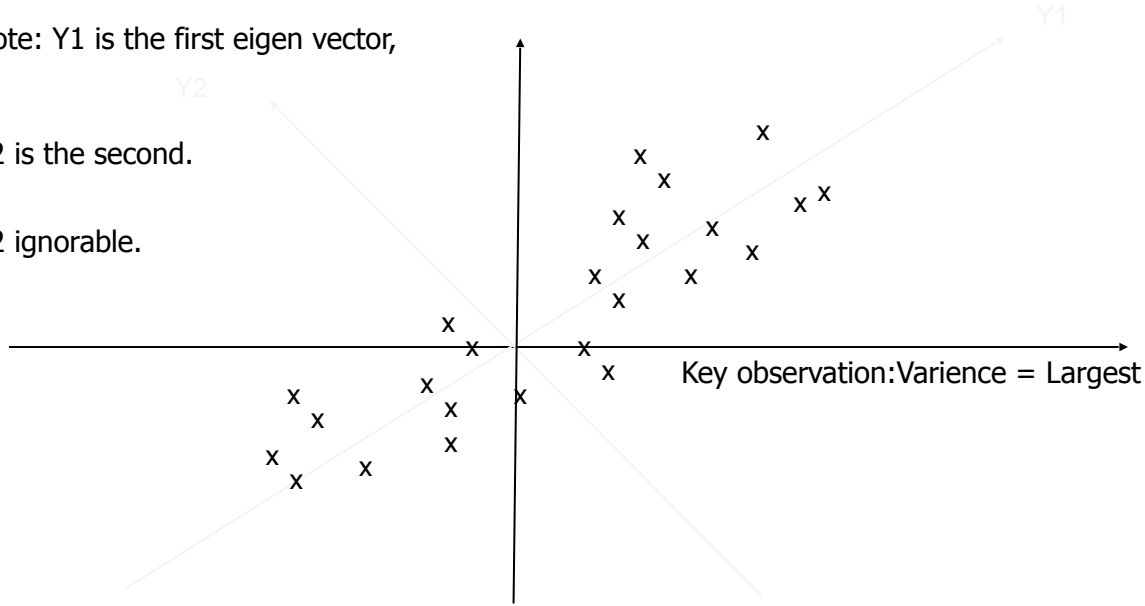
Conclusion

Principal Component Analysis

Note: Y1 is the first eigen vector,

Y2 is the second.

Y2 ignorable.



Principal Component Analysis: one attribute first

Question: how much spread is in the data along the axis? (distance to the mean)

Variance=Standard deviation^2

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{(n-1)}$$

Temperature
42
40
24
30
15
18
15
30
15
30
35
30
40
30

Now consider two dimensions

Covariance: measures the correlation between X and Y

- cov(X,Y)=0: independent
- Cov(X,Y)>0: move same direction
- Cov(X,Y)<0: move opposite direction

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)}$$

X=Temperature	Y=Humidity
40	90
40	90
40	90
30	90
15	70
15	70
15	70
30	90
15	70
30	70
30	70
30	90
40	70
30	90

More than two attributes: covariance matrix

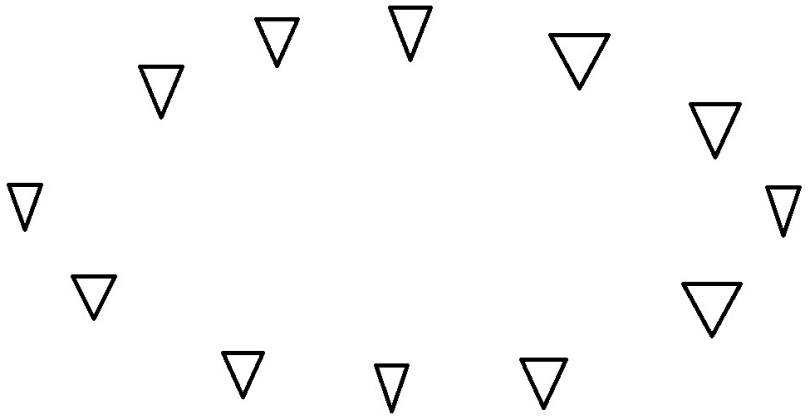
- Contains covariance values between all possible dimensions (=attributes):

$$C^{nxn} = (c_{ij} \mid c_{ij} = \mathrm{cov}(Dim_i, Dim_j))$$

- Example for three attributes (x,y,z):

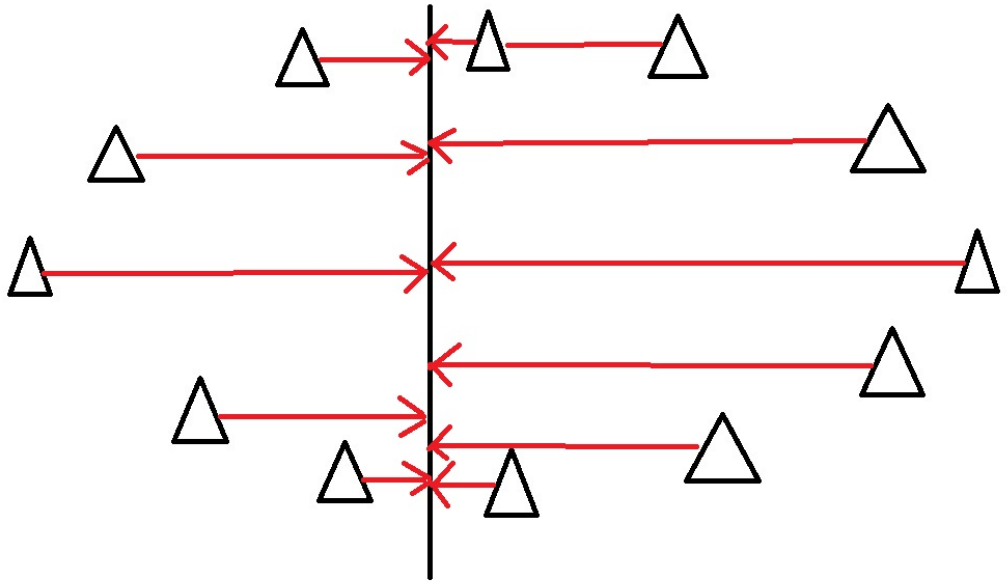
$$C = \begin{pmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) & \mathrm{cov}(x,z) \\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) & \mathrm{cov}(y,z) \\ \mathrm{cov}(z,x) & \mathrm{cov}(z,y) & \mathrm{cov}(z,z) \end{pmatrix}$$

What is Principal Component Analysis?

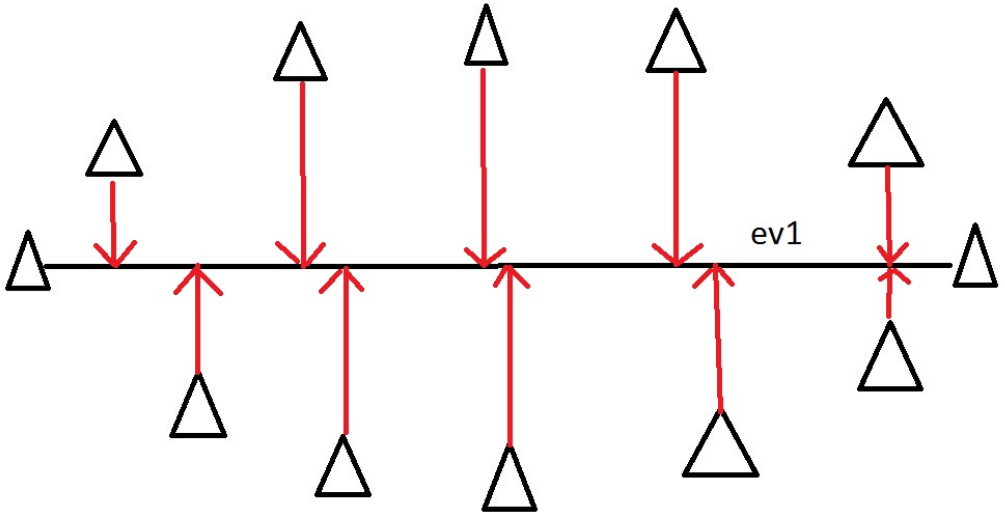


They are the directions where there is the most variance, the directions where the data is most spread out.

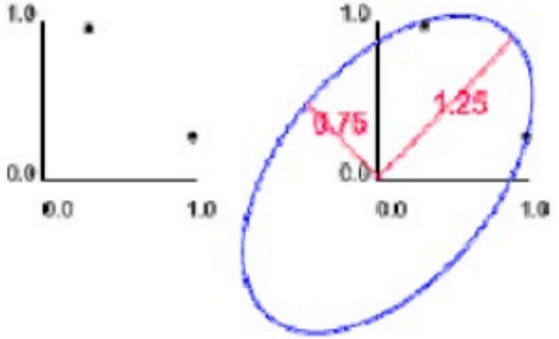
To find the direction where there is most variance, find the straight line where the data is most spread out when projected onto it. A vertical straight line with the points projected on to it will look like this:



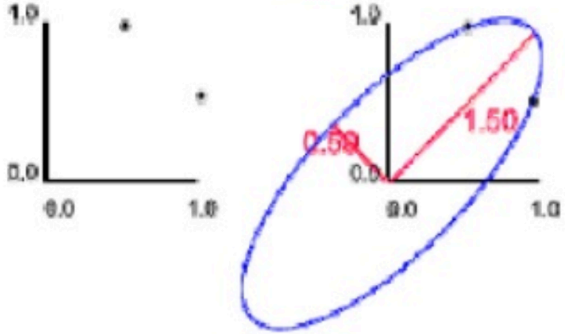
On this line the data is way more spread out, it has a large variance.
In fact there isn't a straight line you can draw that has a larger variance than a horizontal one. A horizontal line is therefore the principal component in this example.



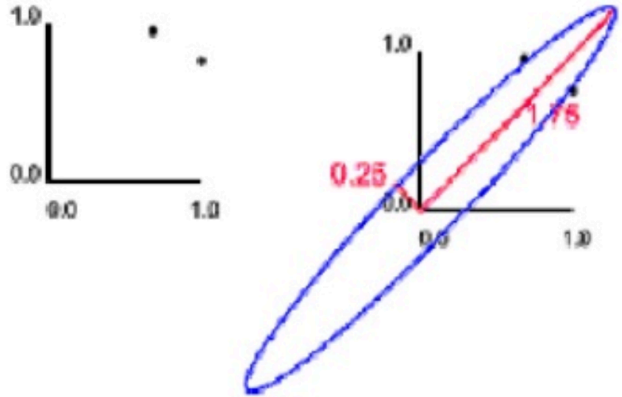
1.00 0.25
0.25 1.00



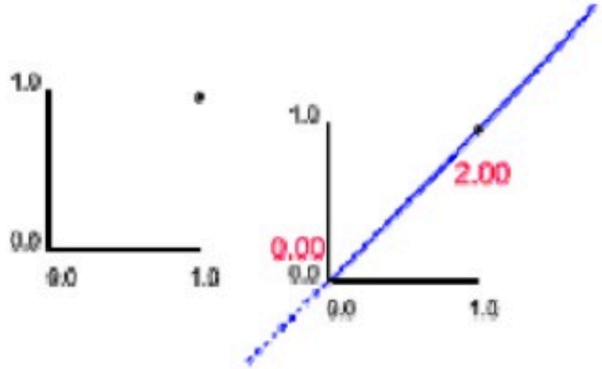
1.00 0.50
0.50 1.00



1.00 0.75
0.75 1.00



1.00 1.00
1.00 1.00



Eigenvalues & Eigenvectors

- Vectors **x** having same direction as *A***x** are called *eigenvectors* of *A* (*A* is an n by n matrix).
- In the equation *A***x**=λ**x**, λ is called an *eigenvalue* of *A*.

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} x \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4x \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Principal components:-

- 1. principal component (PC1)
  - The eigenvalue with the largest absolute value will indicate that the data have the largest variance along its eigenvector, the direction along which there is greatest variation
- 2. principal component (PC2)
  - the direction with maximum variation left in data, orthogonal to the 1. PC
- In general, only few directions manage to capture most of the variability in the data.

Steps of PCA :-

- Let $\overline{X}$ be the mean vector (taking the mean of all rows)
- Adjust the original data by the mean

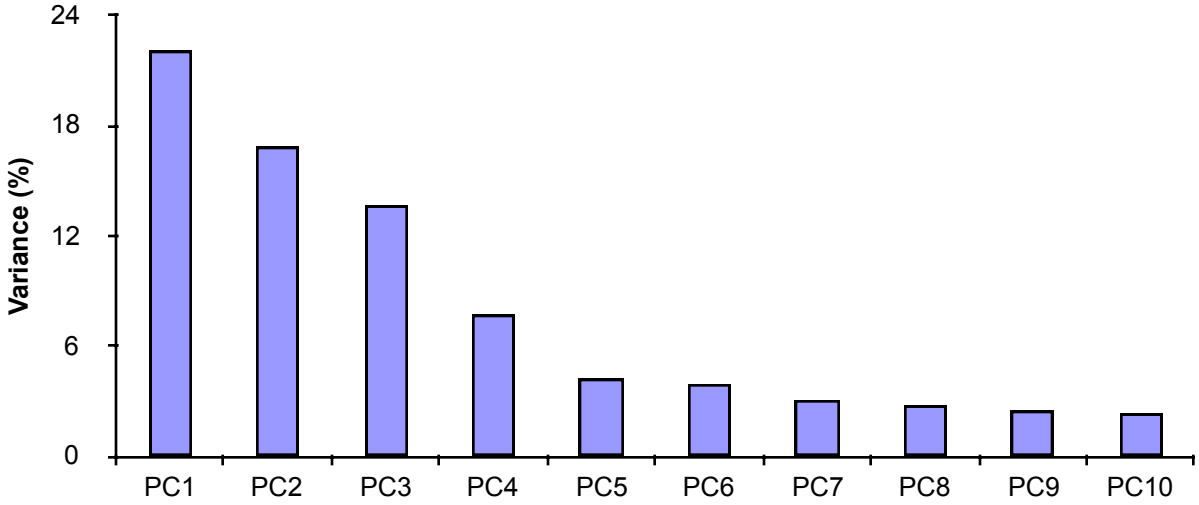
  $X' = X - \overline{X}$
- Compute the covariance matrix C of adjusted X
- Find the eigenvectors and eigenvalues of C.

Eigenvalues

- Calculate eigenvalues λ and eigenvectors **x** for covariance matrix:
  - Eigenvalues λ_j are used for calculation of [% of total variance] (V_j) for each component *j*:

$$V_j = 100 \cdot \frac{\lambda_j}{\sum_{x=1}^{n} \lambda_x} \qquad \sum_{x=1}^{n} \lambda_x = n$$

Principal components - Variance



Transformed Data

- Eigenvalues λ_j corresponds to variance on each component *j*
- *Thus, sort by* λ_j
- Take the first *p* eigenvectors $\mathbf{e_i}$; where p is the number of top eigenvalues
- These are the directions with the largest variances

$$\begin{pmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{ip} \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_p \end{pmatrix} \begin{pmatrix} x_{i1} - \overline{x_1} \\ x_{i2} - x_2 \\ \dots \\ x_{in} - \overline{x_n} \end{pmatrix}$$

PCA –> Original Data

- Retrieving old data (e.g. in data compression)
  - $RetrievedRowData = (RowFeatureVector^T \times FinalData) + OriginalMean$
  - Yields original data using the chosen components

Refrences :-

- https://pdfs.semanticscholar.org/6e72/2fafa3b1191f7c779ddf09b64eda01c94a28.pdf
- https://en.wikipedia.org/wiki/Principal_component_analysis
- principalcomponentanalysis-150314161616-conversion-gate01.pdf
- https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/
- https://www.slideshare.net/CvilleDataScience/az-tecpca-datasciencemeetupkscott20140218?next slideshow=2

THANKYOU

Understanding the right time
to port over to GHub from LGS
using Sentiment Analysis
on Text Data

Data

4 FILES

LGS/G HUB mentions file.xlsx

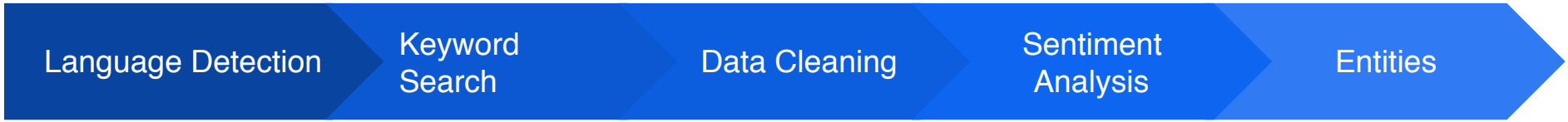
&

LGS/G HUB raw file.xlsx

Data

1. GUID
2. DATE
3. URL
4. CONTENT
5. AUTHOR
6. NAME
7. COUNTRY
8. STATE/REGION
9. CITY
10. CATEGORY
11. BRAND
12. FOLLOWING
13. EMOTION
14. SOURCE
15. GENDER
16. POST
17. FOLLOWERS
18. POST TITLE
19. POST TYPE
20. IMAGE URL

Methodology


Language Detection
Keyword Search
Data Cleaning
Sentiment Analysis
Entities

Language Detection

en	2409
ja	264
vi	168
de	157
es	149
sv	97
fr	75
pt	51
ko	43
zh-cn	39
no	20
pl	18
id	17
ru	16
it	13
tr	6
fi	4
sk	4
cs	3
bg	3
nl	3
et	2
el	2
ar	2
th	2
da	2
tl	2
ro	2
ca	1

Keyword Search

“g hub/ghub”

na	2271
GHUB	1303

Data Cleaning

Removal of hyperlinks & special characters

+

Sentence Splitting

Before -> After

Awhile back I used a hub for my X55 but had ghosting problems. If I recall correctly it was the fact that I was using a 3.0 port instead of a 2.0 (You know, just weird Saitek problems). In the process of changing my gaming space around and was thinking about using a 3.0 hub to easily and neatly be able to swap in sticks, wheels, etc. Will this cause any problems? Virlpil Delta, TM FCS Throttle/ p e d a l s , G 2 9 w h e e l . ==========::========== == Been using a hub with my Thrustmaster pperipherals for years with no iissues, unlike spectrum and android.

->

Awhile back I used a hub for my X55 but had ghosting problems.
If I recall correctly it was the fact that I was using a 3.0 port instead of a 2.0 You know, just weird Saitek problems.
In the process of changing my gaming space around and was thinking about using a 3.0 hub to easily and neatly be able to swap in sticks, wheels, etc.
Will this cause any problems?
Virlpil Delta, TM FCS Throttlepedals, G29 wheel.
:: Been using a hub with my Thrustmaster pperipherals for years with no iissues, unlike spectrum and android.

Sentiment Analysis

VADER

(Valence Aware Dictionary and sEntiment Reasoner)

VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

VADER uses a combination of sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative.

	sentence	review_no	text	Compound	Positive	Neutral	Negative	sentiment	comp_sentiment
0	0	0	Awhile back I used a hub for my X55 but had gh...	-0.5499	0.000	0.738	0.262	neu	neg
1	1	0	If I recall correctly it was the fact that I w...	-0.5267	0.000	0.812	0.188	neu	neg
2	2	0	In the process of changing my gaming space aro...	0.5859	0.156	0.844	0.000	neu	pos
3	3	0	Will this cause any problems?	-0.4019	0.000	0.597	0.403	neu	neg
4	4	0	Virlpil Delta, TM FCS Throttlepedals, G29 wheel.	0.0000	0.000	1.000	0.000	neu	neu
5	5	0	:: Been using a hub with my Thrustmaster pperi...	-0.2960	0.000	0.879	0.121	neu	neg
6	0	2	Random Duos In Blackout!	0.0000	0.000	1.000	0.000	neu	neu
7	1	2	Games Playing Currently: Black Ops 4: PUBG: ...	-0.3182	0.018	0.956	0.026	neu	neg
8	2	2	This helps support the channel and allows me t...	0.7783	0.394	0.606	0.000	neu	pos
9	3	2	Thank you for the support!	0.6696	0.647	0.353	0.000	pos	pos
10	4	2	Subscribe for more vids: Thanks for watching,...	0.8739	0.392	0.608	0.000	neu	pos
11	5	2	Hope you enjoy : All Love	0.8834	0.837	0.163	0.000	pos	pos

The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be:

- **between -1 (most extreme negative) and +1 (most extreme positive).**

This is the most useful metric if you want a single unidimensional measure of sentiment for a given sentence. Calling it a 'normalized, weighted composite score' is accurate.

It is also useful for researchers who would like to set standardized thresholds for classifying sentences as either positive, neutral, or negative. Typical threshold values (used in the literature cited on this page) are:

- **positive sentiment: compound score >= 0.05**
- **neutral sentiment: (compound score > -0.05) and (compound score < 0.05)**
- **negative sentiment: compound score <= -0.05**

The pos, neu, and neg scores are ratios for proportions of text that fall in each category (so these should all add up to be 1... or close to it with float operation).

These are the most useful metrics if you want multidimensional measures of sentiment for a given sentence.

Entities

Nouns

	sentence	review_no	text	Compound	Positive	Neutral	Negative	sentiment	comp_sentiment	entities	nouns
1	1	0	If I recall correctly it was the fact that I w...	-0.5267	0.000	0.812	0.188	Neutral	Negative	[(3.0), (2.0), (Saitek)]	[(I), (it), (the, fact), (I), (a, 3.0, port), ...
2	2	0	In the process of changing my gaming space aro...	0.5859	0.156	0.844	0.000	Neutral	Positive	[(3.0)]	[(the, process), (my, gaming, space), (a, 3.0,...
3	3	0	Will this cause any problems?	-0.4019	0.000	0.597	0.403	Neutral	Negative	[]	[(any, problems)]
4	4	0	Virlpil Delta, TM FCS Throttlepedals, G29 wheel.	0.0000	0.000	1.000	0.000	Neutral	Neutral	[(Virlpil, Delta), (TM, FCS, Throttlepedals), ...	[(Virlpil, Delta), (TM, FCS, Throttlepedals), ...
5	5	0	:: Been using a hub with my Thrustmaster pperi...	-0.2960	0.000	0.879	0.121	Neutral	Negative	[(years)]	[(a, hub), (my, Thrustmaster, pperipherals), (...
6	0	2	Random Duos In Blackout!	0.0000	0.000	1.000	0.000	Neutral	Neutral	[(Random, Duos)]	[(Random, Duos), (Blackout)]
7	1	2	Games Playing Currently: Black Ops 4: PUBG: ...	-0.3182	0.018	0.956	0.026	Neutral	Negative	[(Games, Playing, Currently), (4), (PUBG), (Ha...	[(Games, Playing), (Black, Ops), (PUBG), (Halo...
8	2	2	This helps support the channel and allows me t...	0.7783	0.394	0.606	0.000	Neutral	Positive	[]	[(the, channel), (me), (videos)]
9	3	2	Thank you for the support!	0.6696	0.647	0.353	0.000	Positive	Positive	[]	[(you), (the, support)]
10	4	2	Subscribe for more vids: Thanks for watching,...	0.8739	0.392	0.608	0.000	Neutral	Positive	[( )]	[(more, vids), (Thanks), (guysgals), (it), (me...
11	5	2	Hope you enjoy : All Love	0.8834	0.837	0.163	0.000	Positive	Positive	[]	[(you), (All, Love)]
12	0	3	Welcome to RVR Closet.	0.4588	0.500	0.500	0.000	Positive	Positive	[(RVR, Closet)]	[(RVR, Closet)]
13	1	3	This is my take on Mestizo69s crazy discord ch...	-0.6239	0.102	0.473	0.425	Neutral	Negative	[]	[(my, take), (Mestizo69s, crazy, discord, chal...
14	2	3	Please join me, Ozz and Mestizo69 on our 3D Di...	0.8777	0.253	0.704	0.043	Neutral	Positive	[(Discord, Channel), ( ), (, Sumerian, Daemon...	[(me), (our, 3D, Discord, Channel), (you), (ot...

Demo