Assignment 1

09/24/2020

Name: Yusuf Elnady PID: 906280028

Q1.a If we have the vector X that has a mean $\hat{X} = 0$ and the vector Y that has a mean $\hat{Y} = 0$.

$$\cos(\theta) = \frac{\sum_{i=1}^{n} x_i y_y}{\|X\|\|Y\|}$$

The correlation has the formula

$$corr(X, Y) = \frac{S_{XY}}{S_X S_Y}$$

As the mean = 0, we can rewrite the formula of the covariance of the vector X as

$$S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 0)} = \sqrt{\frac{1}{n-1}}\|X\|$$

, the same will be S_Y.
The covariance between X and Y will be given by the formula

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i y_i)$$

We can now rewrite the correlation formula to be

$$corr(X, Y) = \frac{\sum_{i=1}^{n}(x_i y_i)}{\|X\|\|Y\|}$$

Then it's clear that if two vector has a mean = 0, then the cosine similarity between them is equal to the correlation between them.

Q1.b

If the two vectors has magnitute of 1, that means $\|X\| = \|Y\| = 1$.
The cosine similarity formula is

$$\cos(X, Y) = \frac{\sum_{i=1}^{n} x_i y_y}{\|X\|\|Y\|} = \sum_{i=1}^{n} x_i y_y$$

The eculidean distance is

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = \sqrt{\sum_{i=1}^{n} (x_i^2 - 2x_i y_i + y_i^2)} = \sqrt{1 - 2\cos(X, Y) + 1} = \sqrt{2(1 - \cos(X, Y))}$$

Q2

The solution is quite similar to the TF-IDF Weighting. So, basically in TF-IDF, we have the formula $w_{ij} = tf_{ij} * \log(N/df_i)$, where df_i is the number of documents containig term i, and f_{ij} is the frequency of term i in document j, then we normalize by dividing by the maximum frequency because the length may vary.
For this problem, and using the same approach: The set of all articles is $\mathbf{A}$, and one article is A. So, first, I will define the document frequency of w_i in A to be $df_i = ||A \in \mathbf{A}\text{: } w_i \in A|| + 1$ Then, we can define

$$C_{w_i,D} = \frac{f(w_i, A_D)}{\max(f(w_i, A_D) : w_i \in A_D)} * \log(\frac{||A||}{df_i})$$

The first part of the equation is the normalized term frequency of word w_i in the domain set A_D, and the second part is the inverse document frequency of w_i in the set $\mathbf{A}$ of all the articles.

For the second question and using the same idea, it results in $df_i = ||A \in A_D : w_i \in A|| + 1$. Therefore

$$E_{W_i,e} = \frac{f(w_i, A_e)}{\max(f(w_i, A_e) : w_i \in A_e)} * \log(\frac{||A_D||}{df_i})$$

Again, the first part of the equation is the normalized term frequency of word w_i in the domain set A_e, and the second part is the inverse document frequency of w_i in A_D.

Q3.1

As John is waiting at the center of the track, so we can consider that he is in the position $(0, 0)$. Given that the radius of the circle is 1, so the equation for this circle will be $x^2 + y^2 = 1$ and we can consider the position of Mike is $(\cos(\theta), \sin(\theta))$ and this is using the parametric equation while θ can be ranging from 0 to 2π.
Manhattan Distance: By applying its formula, the distance will be $d = |\cos(\theta) - 0| + |\sin(\theta) - 0| =$

$|\cos(\theta)| + |\sin(\theta)|$. To find the maximum and the minimum distance, and for simplicity, I will consider the positive values for now (the first quadrant). By taking the derivative of the distance, it yields $\cos(\theta) - \sin(\theta) = 0$. The possible values for θ is $0°$ or $45°$, then the maximum distance will be $\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$ and the minimum distance will be $1 - 0 = 0$. **As the circle is symmetric across all the four quadrants, so the maximum Manhattan distance is $\sqrt{2}$ miles and the minimum distance is 1 miles**

Eculidean Distance: $d = \sqrt{(\cos(\theta) - 0)^2 + (\sin(\theta) - 0)^2} = \sqrt{\cos(\theta)^2 + \sin(\theta)^2} = 1$. **So the minimum and maximum euclidean distance is 1 mile (constant).**

Chebyshev Distance: $d = \max(|\cos(\theta) - 0|, |\sin(\theta) - 0|) = \max(|\cos(\theta)|, |\sin(\theta)|)$. Again, using the symmetry of the circle, I will work on the first quadrant and generalize the solution for the whole circle. First, we want to find some θ that gives the minimum distance, and it's clear that when decreasing $\cos(\theta)$, the $\sin(\theta)$ will be increasing, and vice versa. **So, the minimum distance will correspond to $\theta = 45°$ as $(\cos(\theta) = \sin(\theta))$, with a distance of $\frac{1}{\sqrt{2}}$ miles. The maximum distance can be found at $\theta = 0°, 90°$ with a distance of 1 miles.**

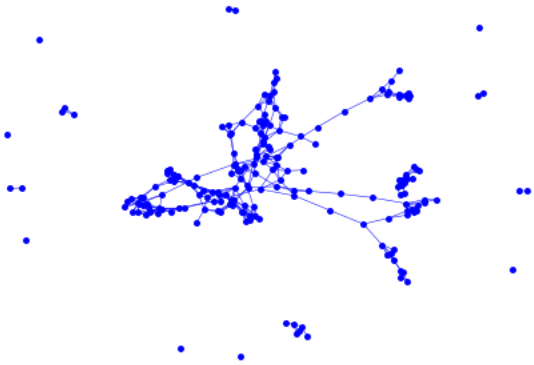
Q4.1

The euclidean distance matrix among the first 8 data points

	0	1	2	3	4	5	6	7
0	0.000000	19.785404	18.106390	15.303795	16.152285	9.020344	20.270749	23.205982
1	19.785404	0.000000	16.302516	20.059622	6.049194	18.741645	3.729406	27.115826
2	18.106390	16.302516	0.000000	15.545197	16.331315	23.908451	19.658936	13.989015
3	15.303795	20.059622	15.545197	0.000000	20.975731	20.904075	21.736345	13.203144
4	16.152285	6.049194	16.331315	20.975731	0.000000	14.554975	6.479995	28.031548
5	9.020344	18.741645	23.908451	20.904075	14.554975	0.000000	17.542457	30.944231
6	20.270749	3.729406	19.658936	21.736345	6.479995	17.542457	0.000000	30.048003
7	23.205982	27.115826	13.989015	13.203144	28.031548	30.944231	30.048003	0.000000

(a) Graph using knn = 5

(b) Graph using distance ¡ 6



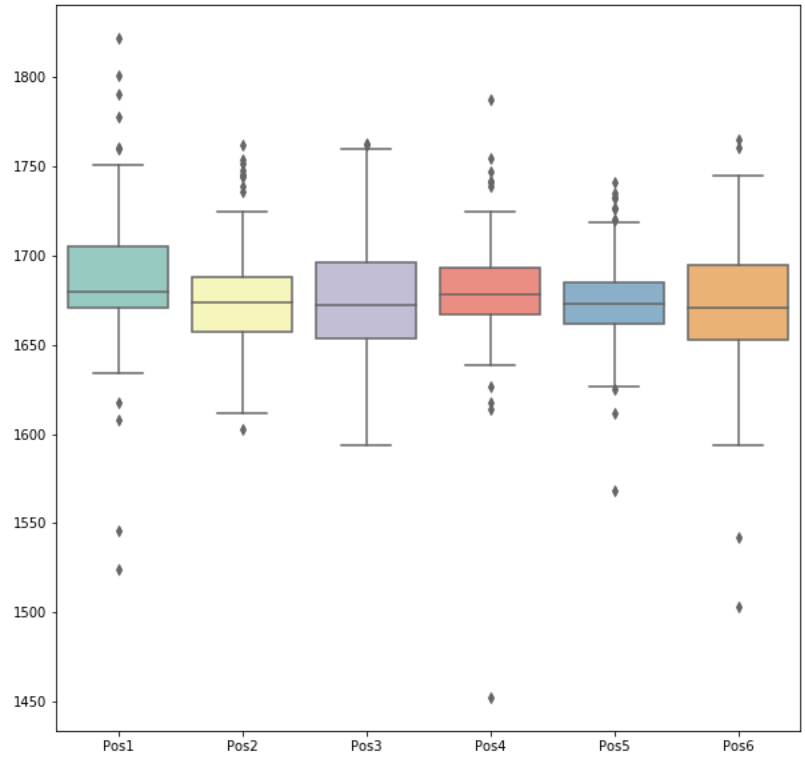
Q4.2

Geodesic distance between each pair of these points using Dijkstra's shortest path algorithm

	0	1	2	3	4	5	6	7
0	0.000000	29.624514	30.848992	15.303795	23.575319	9.020344	30.055315	28.506939
1	29.624514	0.000000	60.473505	44.928308	6.049194	20.604169	3.729406	58.131453
2	30.848992	60.473505	0.000000	15.545197	54.424311	39.869336	60.904306	13.989015
3	15.303795	44.928308	15.545197	0.000000	38.879114	24.324139	45.359109	13.203144
4	23.575319	6.049194	54.424311	38.879114	0.000000	14.554975	6.479995	52.082258
5	9.020344	20.604169	39.869336	24.324139	14.554975	0.000000	21.034970	37.527283
6	30.055315	3.729406	60.904306	45.359109	6.479995	21.034970	0.000000	58.562254
7	28.506939	58.131453	13.989015	13.203144	52.082258	37.527283	58.562254	0.000000

See the code.

Q5.1



Q5.2

The thick center (median) line in some of the box plots is not symmetrical with the outer edges of the box because the data is skewed. The outer edges refers to the Q1 and Q3.

Q6.1

See the code.

Q6.2

If we have vector X, then centering it will follow the formula $X := X - \hat{X}$, $\hat{X}$ is the mean of the vector X. Scaling is dividing vector by its standard deviation, which will be $X := \frac{X}{\sigma(X)}$. The whole process will be given by the formula $X := \frac{X-\hat{X}}{\sigma(X)}$.

See the code.

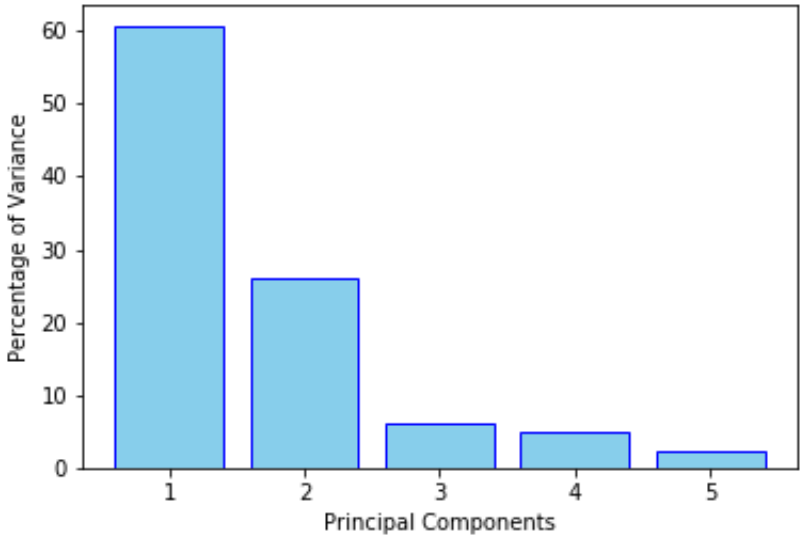
Q6.3

What we did in step 2 is called the standardization (centering and scaling), and we do that so all features can be on the same scale to be able to calculate the eigen vectors correctly without having a feature dominating the other one.
For the interpretation of the data, it won't have an effect on the meaning the data. We don't lose any information, so if two variables were positively correlated they will have the same correlation after the standardization (normalization).

Q6.4
There are many formulas to calculate the correlation squared matrix, each cell will follow this correlation formula $corr(x, y) = \frac{S_{xy}}{S_x S_y}$. But I have used another formula in my code (quicker way) since the data are already centered and has zero mean, so I calculated it using $X^T X$, where X is a matrix with objects as rows and features as columns.
See the code.

Q6.5
See the code.

Q6.6

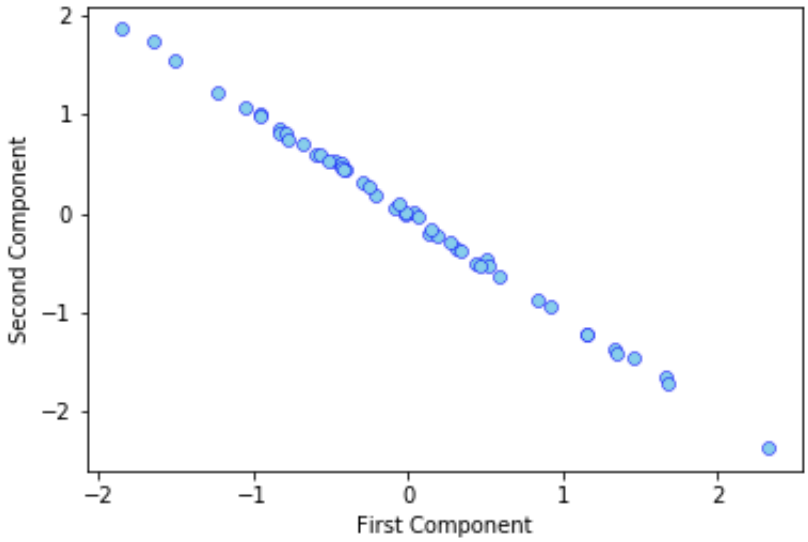


See the code.

Q6.7 I would choose the first three Principal Components because they have a total variance of 92.73 percent which captures most of the data.
I carried out the steps of PCA algorithm using eigenvalue decomposition.
See the code.

Q6.8

Scatter plot of the first two components obtained using PCA.



Q6.9

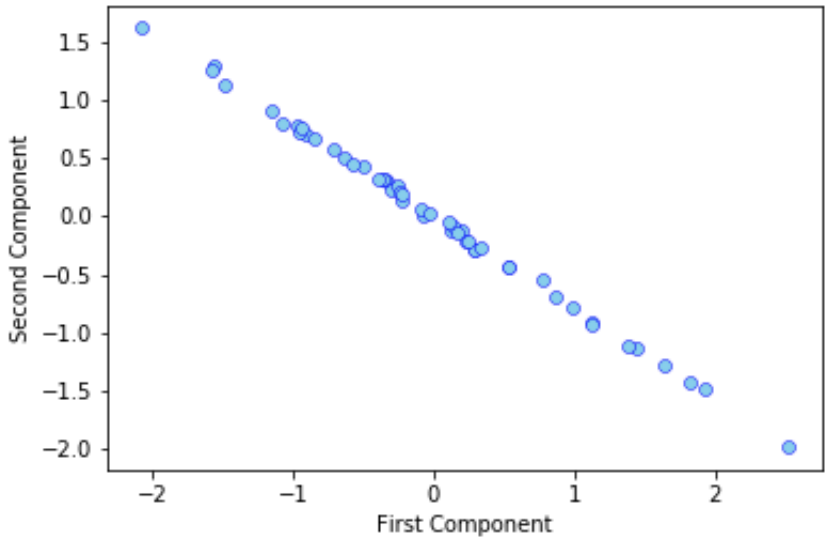
It is not recommended to use the PCA Algorithm since it calculates all the eigenvectors for the given correlation/covariance matrix, although we are going to use only some of these eigenvectors that corresponds to the highest eigenvalues.

Q6.10

The method is SVD.

See the code.

Q6.11



Q6.12

By comparing the two scatter plots we got above, It is clear that the decomposition of both of the algorithms gives (almost) the same new dimensions (components). From mathematical perspective: In PCA, we do the eigen decomposition using $X = TP^T$ and in SVD we use $X = USV^T$, so $T = US$ and $P = V$, so the two ways are similar to each other