Fall 2023 DATA 220 Mathematical Methods for Data Analytics

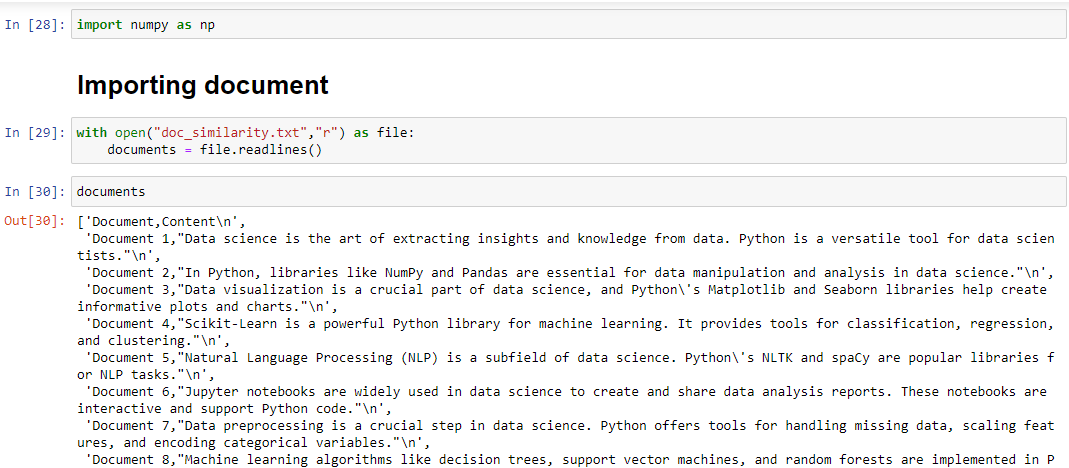
Homework – 3

Name :- Prayag Nikul Purani

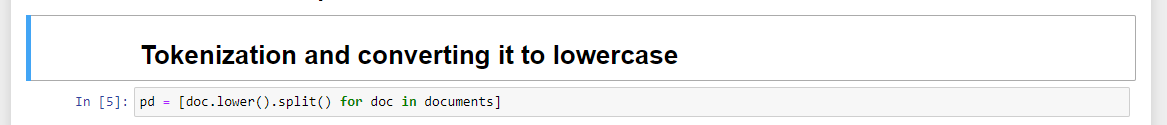
SJSU Id :- 017416737

Problem 2 :–

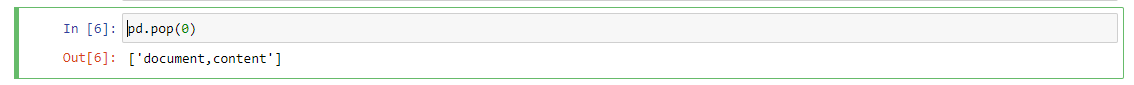
First step is to import the txt file in python using the file object and reading the line using readlines()



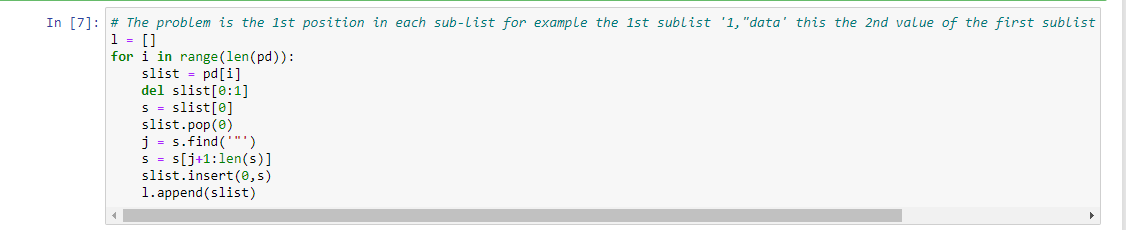
Now we will be tokenization and converting the word to lowercase so the constancy is maintains

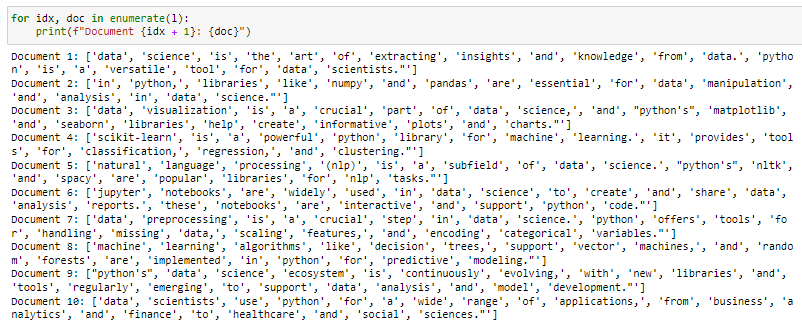


The next is the cleaning if the data the first problem is the first entry in the pd which is 'document,content' but we need to remove which can be done using pop function.



The second problem is the first to element in each sublist like the pd[0][0] is document and the main challenge is pd[0][1] = ’1,”data’ so we need to pop[i][0] for each sub-list and we need to perform the string operation in the 2nd element of each sub-list, so for loop is runs to for each sub-list





Now the problem is that for example in document one as a value called ‘data’ and ‘data.’ Which are treated differently so we need to remove this type of inconsistency.



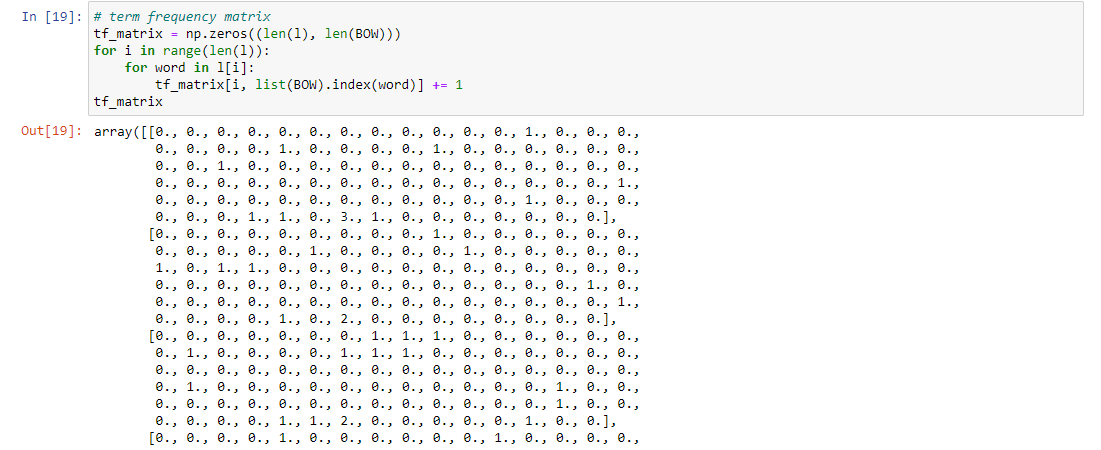
Now we need to remove stop words which are meaningless to us.



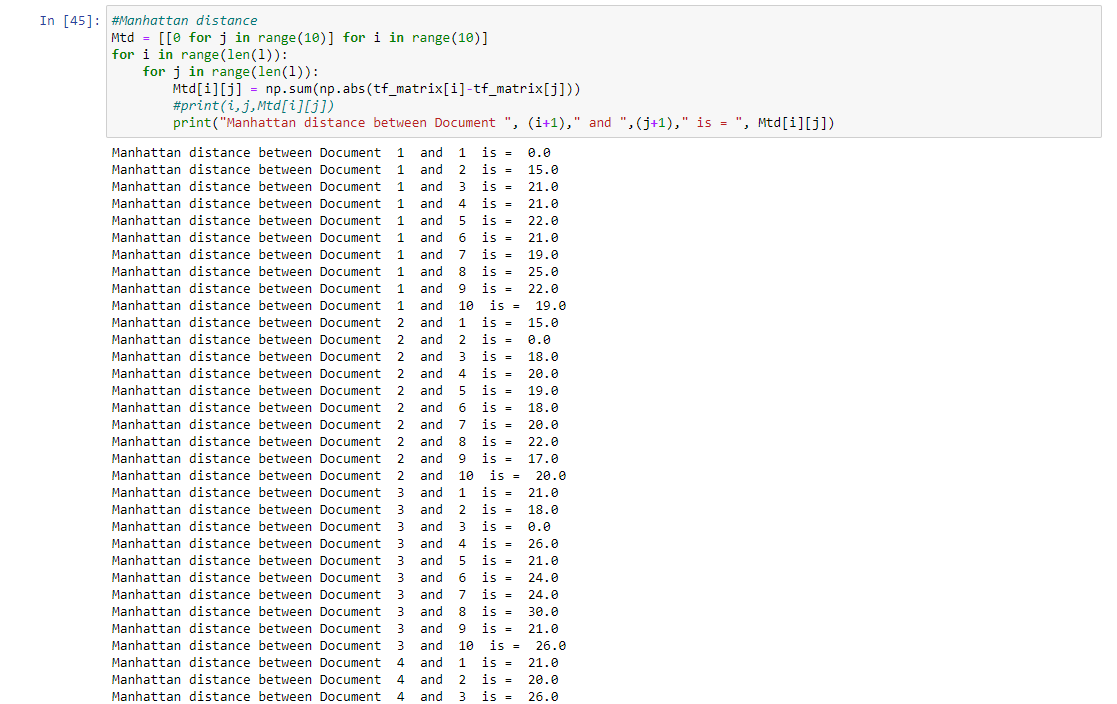
The next step is to create bag of words by using the set data structure.

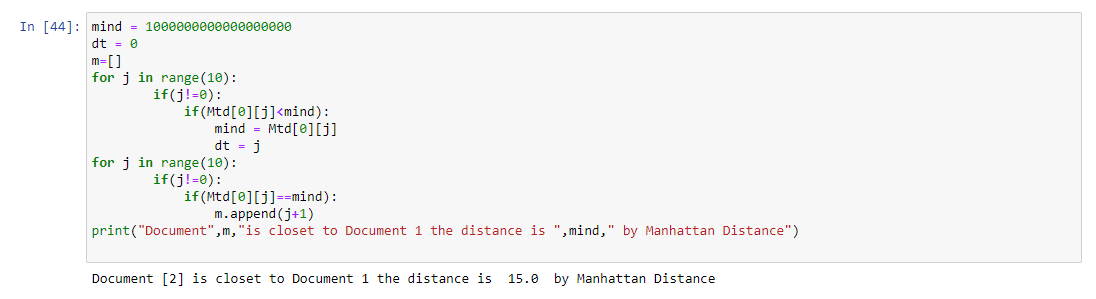


Now the term-frequency table will be created



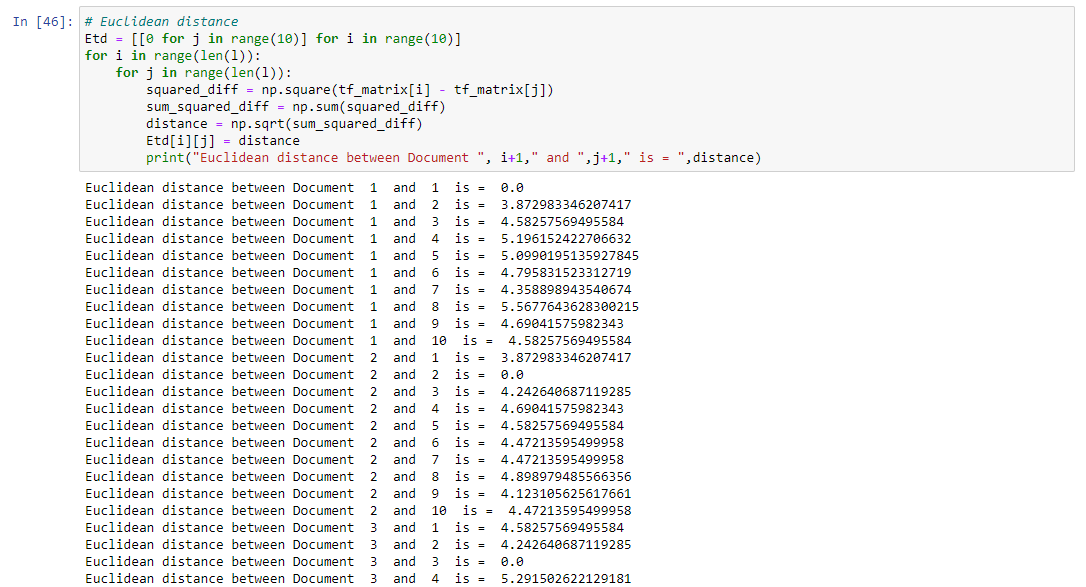
1. Manhattan distance

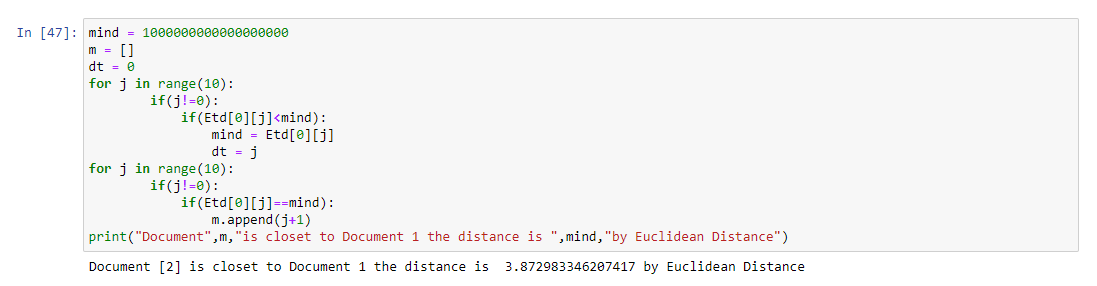




According to Manhattan distance documents 2 has the least distance from document 1 which is 15.

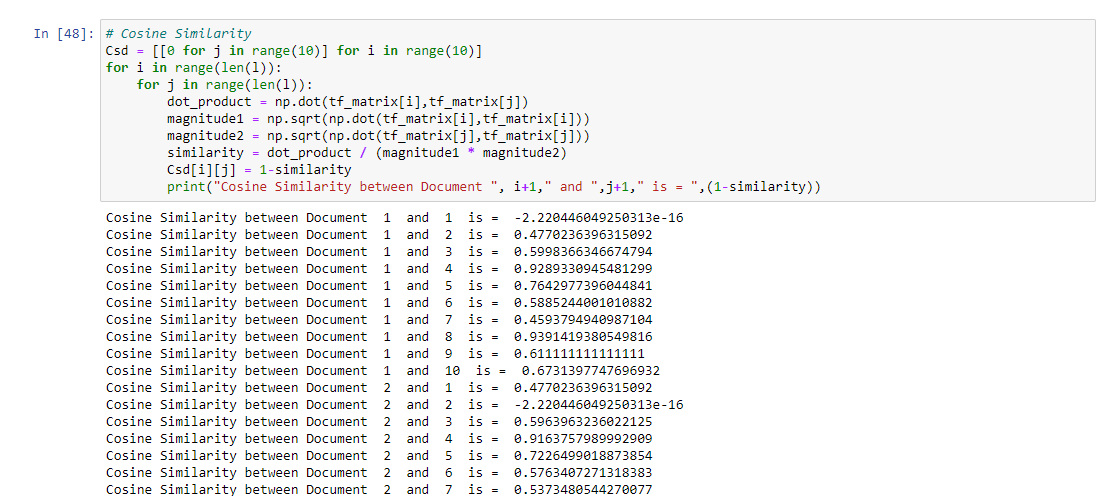
1. Euclidean distance

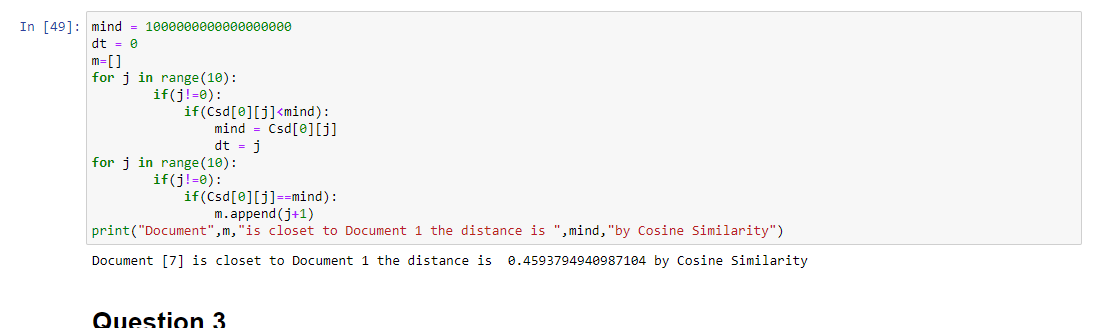




According to Euclidean distance document 2 has the least distance from document 1 which is ~ 3.8729

1. Cosine Similarity



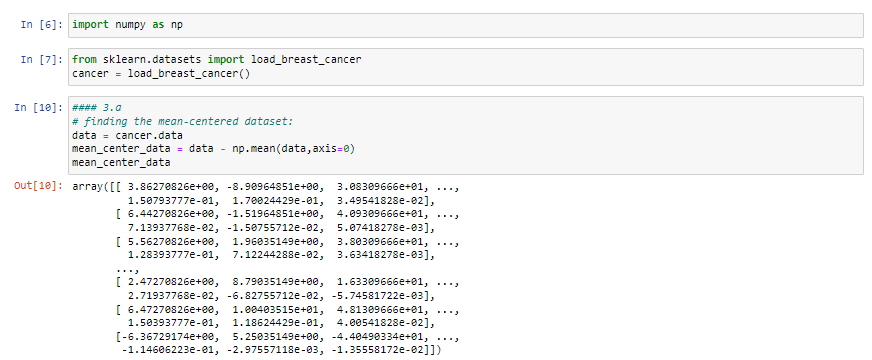


According to Cosine similarity document 7 has the least distance from document 1 which is ~ 0.45937

Problem 3 :–

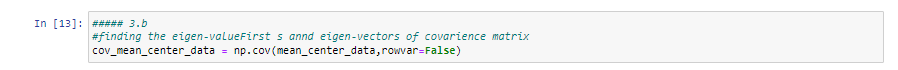
1. Find the mean-centered dataset:

To build a mean-centered dataset, you must first compute the mean of each characteristic and then remove it from each data point.



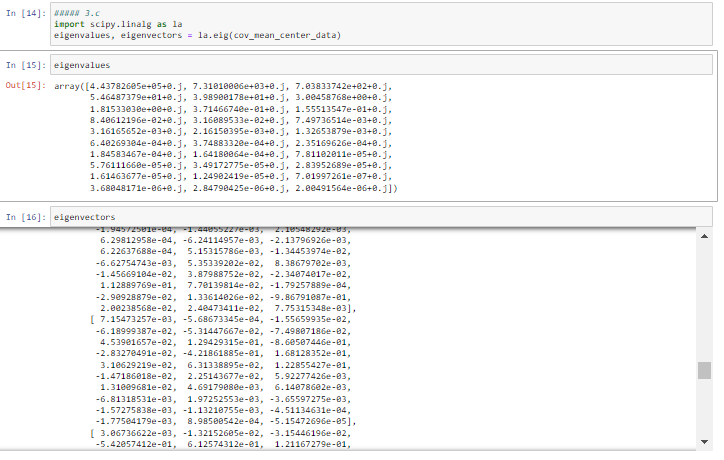
1. Find eigenvalues and associated eigenvectors of the covariance matrix of the mean-centered dataset:

You may determine the eigenvalues and eigenvectors by first computing the covariance matrix.



1. Plot percent of variance explained by each of the principal components (Scree Plot) and explain.

Plotting the percentage of variation explained by each primary component is the first step in making a scree plot.





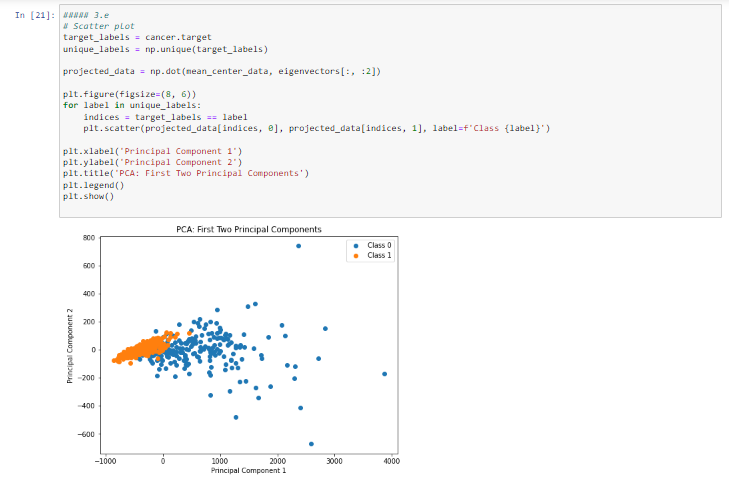
The "elbow" position in the plot may be found with the aid of the Scree Plot. The point at which the explained variance falls off the most is frequently the ideal number of primary components to keep. In this instance, the plot makes it clear.

1. What is the optimal number of principal components to retain for reducing the dimension of the original data?

Through Scree Plot analysis, the ideal number of main components may be ascertained. That's the point at which the explained variance begins to plateau. To find this spot, visually check the plot. Depending on the desired amount of explained variance and dimensionality reduction, one must decide how many main components to keep.

So, from the graph in the previous graph the 1st first feature covers 99% of the data, so we can use the first features is more than sufficient to cover all the data and we can drop the rest of the feature so the dimension will be reduced to only one dimension.

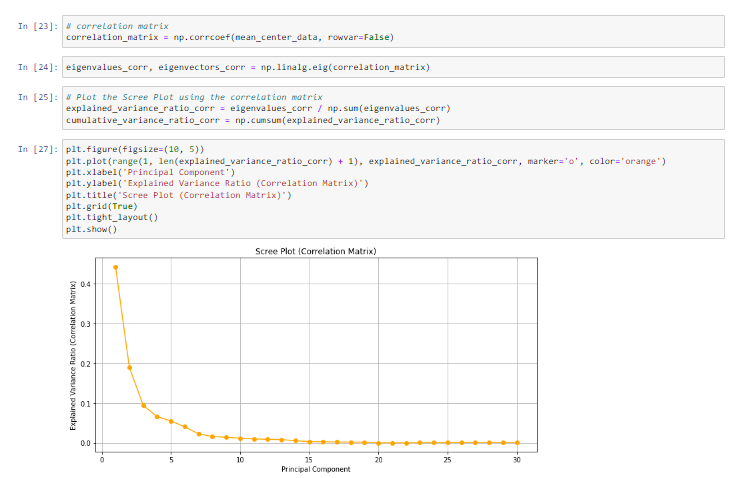
1. Create a scatter plot with each observation projected onto the first two principal components and use different colours for each group, including a legend:



To do this, make a scatter plot after projecting the data onto the first two principal components. The observations projected onto the first two principal components will be displayed in this scatter plot, with distinct colours denoting each class and a legend identifying them. To do this, make a scatter plot after projecting the data onto the first two principal components.

Problem 4 :–

We must first compute the correlation matrix and then determine the eigenvalues and eigenvectors in order to compare the Scree Plots using the correlation matrix (a normalized version of the covariance matrix). Two Scree Plots will be produced by this code; one will use the covariance matrix and the other the correlation matrix. The differences between the two plots in terms of variance explained by each principal component can be observed visually. After adjusting for the variable scales, the correlation matrix Scree Plot's values, which range from 0 to 1, represent the percentage of variance explained by each principal component.



Plotting Covariance Matrix Scree:

Eigenvalues show how much of the variation in the original data each primary component accounts for. Using the covariance matrix usually results in greater eigenvalues and a higher explained variance ratio. If the explained variance ratio drops more sharply, it means that a small number of principle components account for the majority of the variation, with the other components making up the remainder.

Scree plot with a correlation matrix:

Eigenvalues show how much of the variation in the data is explained by each primary component while taking the variable scales into consideration. Since it indicates the percentage of total variation that can be explained, the explained variance ratio will fall between 0 and 1. Because they are normalized, the eigenvalues will usually be less than in the covariance matrix situation.

You should search for variations in the two Scree Plots' shapes and the amount of explained variance when comparing them. In particular: The eigenvalues will be less and the curve may seem smoother in the correlation matrix Scree Plot. There will be a more progressive decrease in the explained variance for each main component. You may see a sharper initial decline in explained variance in the covariance matrix Scree Plot, which suggests that fewer main components account for the majority of the variation in the original data.

References: -

<https://stackoverflow.com/questions/627435/how-to-remove-an-element-from-a-list-by-index>

<https://realpython.com/python-strings/#:~:text=String%20indexing%20in%20Python%20is,of%20the%20string%20minus%20one>.

<https://stackoverflow.com/questions/3221891/how-can-i-find-the-first-occurrence-of-a-sub-string-in-a-python-string>

<https://stackoverflow.com/questions/21939652/insert-at-first-position-of-a-list-in-python> <https://www.geeksforgeeks.org/enumerate-in-python> <https://stackoverflow.com/questions/74206689/syntaxerror-invalid-syntax-list-comprehension-on-enumerated-on-list>

[https://stackoverflow.com/questions/19184335/is-there-a-need-for-rangelena](https://www.google.com/search?q=length+of+list+python&oq=length+of+lis&gs_lcrp=EgZjaHJvbWUqBwgAEAAYgAQyBwgAEAAYgAQyBwgBEAAYgAQyBwgCEAAYgAQyBwgDEAAYgAQyBggEEEUYOTIHCAUQABiABDIHCAYQABiABDIHCAcQABiABDIHCAgQABiABDIHCAkQABiABKgCALACAA&sourceid=chrome&ie=UTF-https://stackoverflow.com/questions/19184335/is-there-a-need-for-rangelena)

<https://www.google.com/search?q=first+index+of+string+in+python&oq=first+index+of+string+in&gs_lcrp=EgZjaHJvbWUqBwgDEAAYgAQyBggAEEUYOTIHCAEQABiABDIHCAIQABiABDIHCAMQABiABDIICAQQABgWGB4yCAgFEAAYFhgeMggIBhAAGBYYHjIICAcQABgWGB4yCAgIEAAYFhgeMggICRAAGBYYHtIBCDg1NzNqMGo3qAIAsAIA&sourceid=chrome&ie=UTF-8>

<https://www.programiz.com/python-programming/methods/list/remove>

<https://www.w3schools.com/python/python_lists_loop.asp>

<https://www.educative.io/answers/how-to-find-the-length-of-a-string-in-python>

<https://vitalflux.com/eigenvalues-eigenvectors-python-examples/>

<https://www.google.com/search?q=how+to+calculate+eigenvectors+from+covariance+matrix+in+python&sca_esv=577233357&ei=hP07ZZ6kKqKuqtsP6dii6Ao&ved=0ahUKEwie9oeJ6paCAxUil2oFHWmsCK0Q4dUDCBA&uact=5&oq=how+to+calculate+eigenvectors+from+covariance+matrix+in+python&gs_lp=Egxnd3Mtd2l6LXNlcnAiPmhvdyB0byBjYWxjdWxhdGUgZWlnZW52ZWN0b3JzIGZyb20gY292YXJpYW5jZSBtYXRyaXggaW4gcHl0aG9uSABQAFgAcAB4AZABAJgBAKABAKoBALgBA8gBAPgBAeIDBBgAIEE&sclient=gws-wiz-serp>

<https://www.youtube.com/watch?v=uJ-9qak7LPY>

<https://machinelearningmastery.com/principal-component-analysis-for-visualization/>

<https://stackoverflow.com/questions/40296058/plotting-two-principal-component-score-vectors-using-a-different-color-to-indic>

<https://www.statology.org/scree-plot-python/>