DSCI 552 – Assignment 3: PCA and FastMap

Team Members	Sanjana Gopnal Swamy (gopnalsw@usc.edu)
	Amit Sankhla (asankhla@usc.edu)

Implementation:

For implementation of both the algorithms, we have used standard python libraries such as numpy and python list/dict. To plot the words in a 2D plane for FastMap, we have used matplotlib.pyplot library.

PCA

The data structure we use to store the data is a 2D nested array which is easy for matrices calculation. We follow the steps of PCA, normalize every dimension of data, calculate covariance matrix, get the eigenvalues and eigenvectors, and choose the first k's as the principal components. Finally, we dot product the normalized n-by3 data matrix with 3-by-2 vectors and get the n-by-2 result. We have directly used functions of numpy to reduce some transformations and simplify the code.

Output for PCA:

Direction for Principal Component 0:

[0.8666713670843745, -0.23276481647597735, 0.441249681798192]

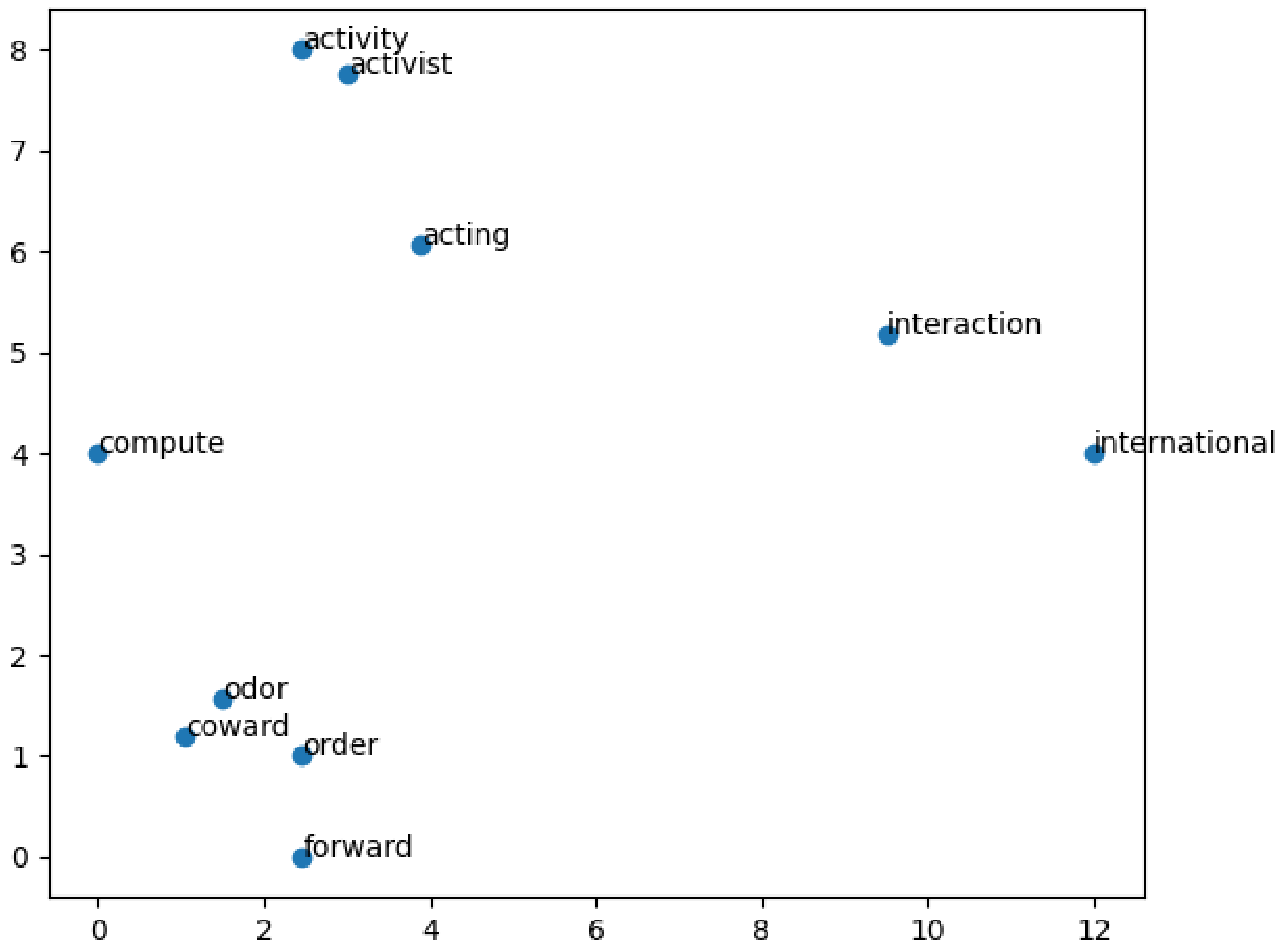
Direction for Principal Component 1:

[-0.4962773044983004, -0.4924792045083853, 0.714963684509003]

FastMap

We constructed a distance dictionary to store the distances of all points with key as the point pair and value as the distance. We initially thought of using numpy or pandas library to read the data file but soon realized that it would be much easier and simpler to use a dictionary and store the pair of points as key. Rest we followed the algorithm by finding the two farthest points and then calculating the perpendicular distance of all the points the line joining the farthest points .The biggest challenge was to understand the result that we got after plotting the points on a 2D plane. Since we were not very familiar with FastMap, we were not sure how to review the result. After carefully analyzing the result, we understood that the similar words should be closer than other words. Now we understand how this plot/data can be used to create meaningful clusters.

Output of words plotted in a 2D plane:



Contributions:

1: Sanjana’s contribution

- Researched about numpy functions to calculate eigen values and vectors.
- Implemented those functions to write the PCA algorithm code.
- Wrote the Report

2: Amit’s contribution

- Implemented FastMap algorithm
- Learned how to plot points using matplotlib.pyplot
- Worked on understanding the results of FastMap.