This project was about implementing natural language processing using the bag-of-words model on data for over 500,000 user reviews on Amazon.

First step was to process the reviews and transform the text into a format that can be used to implement the bag-of-words model.

Using the most frequently used 500 words, a bag-of-word vector is created for each review.

To measure the distance between reviews, I calculated the Euclidean distance between different reviews using the bag of words vector. Since the bag of words vector has 1s and 0s, I just measured the distance on the basis of existence of the word in the review. I applied this method for the first 100 reviews and then displayed them sorted from smallest to largest distances.

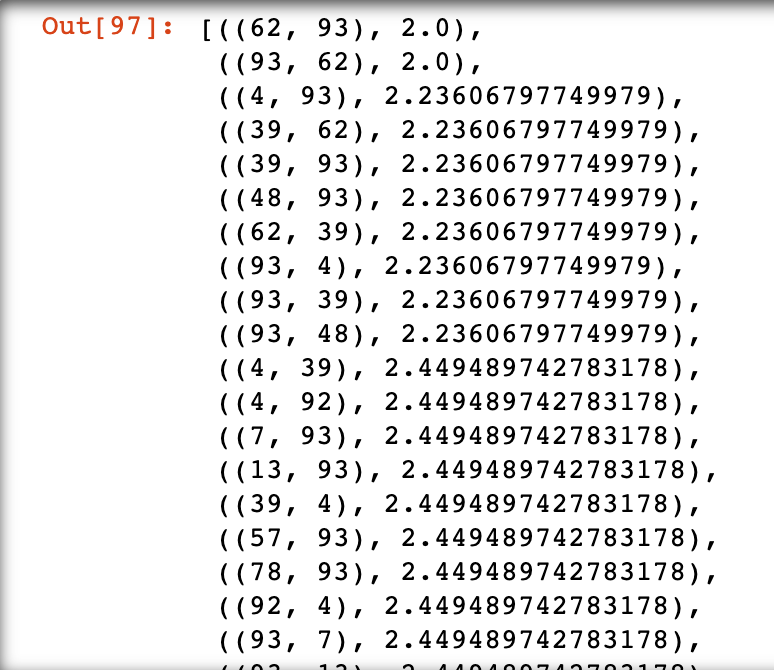


Fig.1 Euclidean distance for first 100 reviews

I ran a PCA for the first 100 reviews to validate my previous findings, using their respective bag of words vectors as the attributes and then chose the first two PCs to plot a graph for the reviews. I observed that the graph largely shows similar findings as the previous exercise in terms of determining which two reviews are similar, but the distances are not as accurate here, because 500 attributes have been reduced to two PCs, so a lot of the variation will be unexplained.

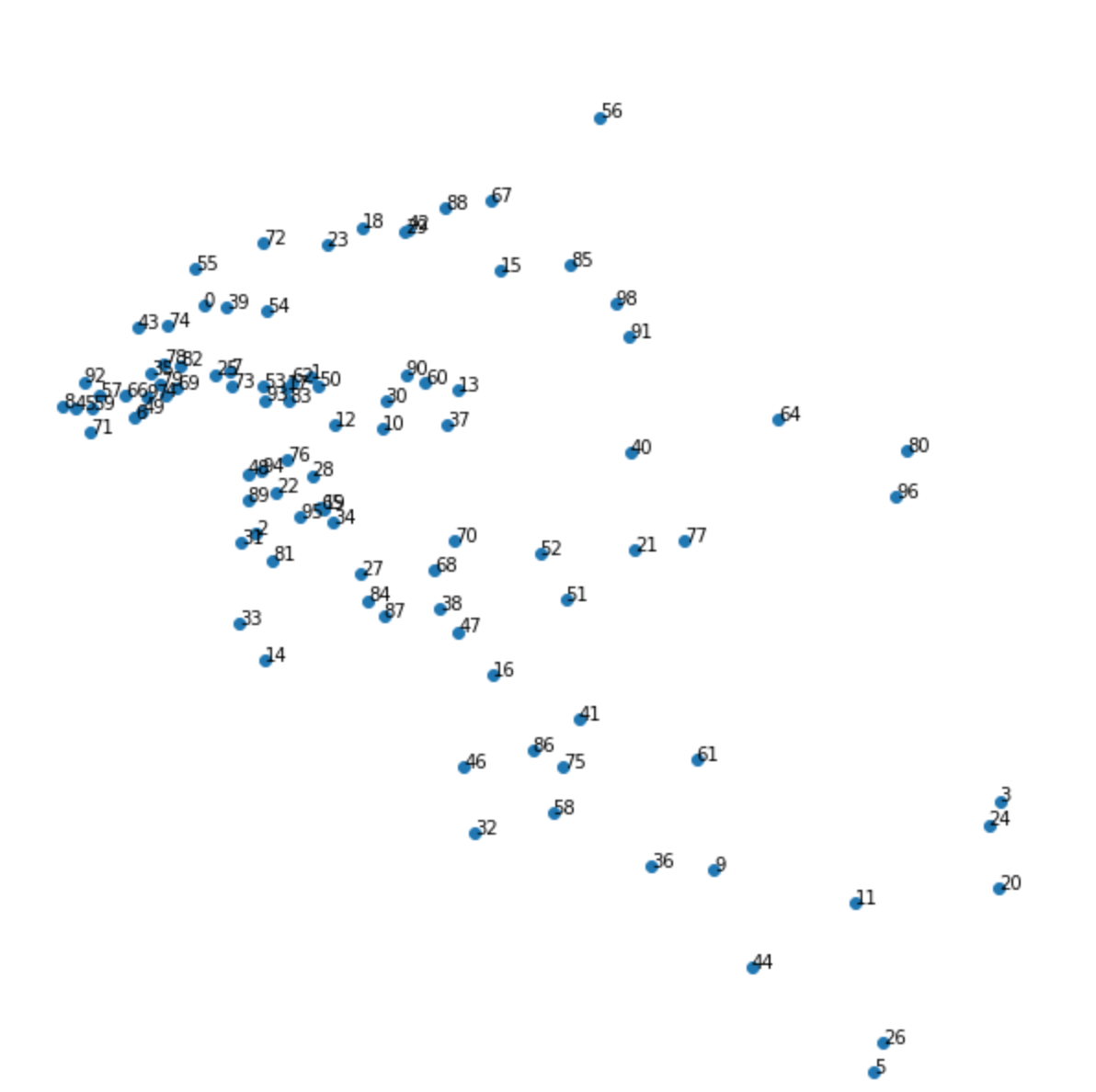


Fig.2: PCA 2D graph

In order to represent each product as a single vector for accurate comparison, I had to aggregate the reviews for each product.

For aggregation I summed the values for the corresponding words in the bag of words vector for each review to get the total frequency of the word in all reviews and divided that by the number of reviews for the particular product in order to normalize the aggregated vector.

After calculating the aggregated vectors, each product has the same aggregate vector. Therefore, each aggregate vector represents a single product. Using this logic and the similar method of utilizing Euclidean distances, I calculated the distances between the 10 products from the dataset. The products are represented by the ‘asin’ IDs.

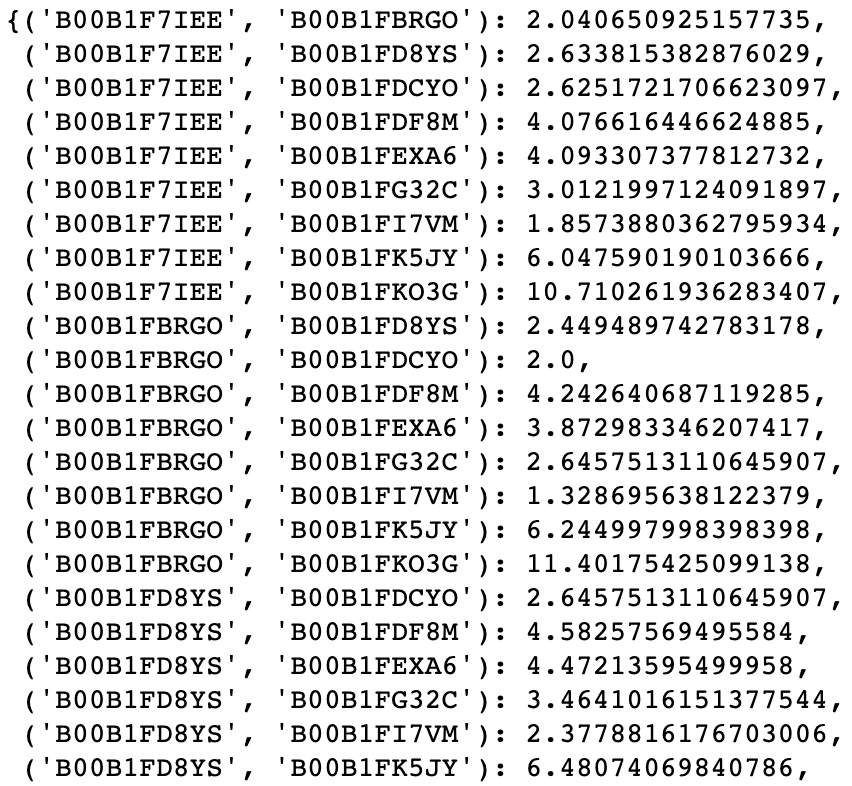


Fig.3: The Euclidean distance between products