Article-Similarity Search on 20newsgroup dataset

Bharath Shamasundar

Department of Computer Science

bbst59@mst.edu

14 February 2018

Abstract

In this project the 20newsgroup dataset is analyzed with the help of data preprocessing methods in R to give a document term matrix and then some of the best known Machine Learning methods in Python are applied in order to make sense of the documents present in the dataset with respect to the Term frequencies that are present in these documents. Finally three distance metrics namely Cosine, Jaccard and Euclidean are applied on the tern frequency matrix to get the similarity matrix on which the articles are then ranked.

1 Introduction and Problem to be Solved

Making meaning from raw data is one of the toughest challenges facing the computer science community. This not only requires excellent data preprocessing techniques but also one must have meaningful strategies to tackle the cleaned data. Such an interesting problem is present as part of the project here. The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The challenge lies in trying to get a clear picture from this raw data and making sense of it by using various data preprocessing and machine learning techniques. The rest

of the report is organized as follows: Section 2 contains a brief overview of the data, Section 3 contains the data preprocessing procedure and the resulting meaningful features that have been obtained from it, Section 4 contains the similarity matrix analysis, including the heat map analysis. Section 5 contains the linear curve fit analysis of all the three similarity analysis. Section 6 contains the standard deviation analysis of the similarity features when the number of features are varied, as well as the similarity pair ranking analysis when the top three article pairs are selected from each of the articles. The report ends in Section 7 with the conclusion and possible future work.

2 Data Overview

The 20newsgroup is a corpus document which contains a bunch of information but mainly deals with 18828 articles related to over 20 groups. The categories are as follows:

- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- sci.crypt
- sci.electronics

- sci.med
- sci.space
- misc.forsale
- talk.politics.misc
- talk.politics.guns
- talk.politics.mideast
- talk.religion.misc
- alt.atheism
- soc.religion.christian

Some of the categories are closely related, like for example (comp.sys.ibm.pc.hardware / comp.sys.mac.hardware) while some of them are not, like for example (sci.med and sci.space). The raw data has 4 main sub-headings and they are

- The target names meaning the various categories for the entire dataset
- The target values(in the range of the 20 categories)
- Data in terms of characters
- Data description for each article.

3 Data Preprocessing Procedure

In the data preprocessing step the term frequency matrix is generated with the help of R. Initially the corpus document is gotten rid of insignificant terms like prepositions, adverbs, pronouns, whitespace, punctuations, numbers, common English stop words and articles. Once that procedure is completed the term frequency matrix is generated from the filtered corpus document. Once the tf matrix has been generated the sparse terms are then removed to reduce the size

of the tf matrix. One this is completed the top 100 frequent terms are calculated and these become the column names for the final matrix containing all the 18828 matrices. The top 100 frequent terms are written to separate file whereas a matrix upon which analysis will be done is written to another file. This final matrix has rows which correspond to the document/article numbers and the columns are the top 100 features that have been taken from the transformed corpus document. Each entry in this final matrix corresponds to the frequency of a term in that document. The histogram analysis of the term frequencies as well as the term count analysis for each of the top 100 terms is given is Figure 1 and 2 respectively.

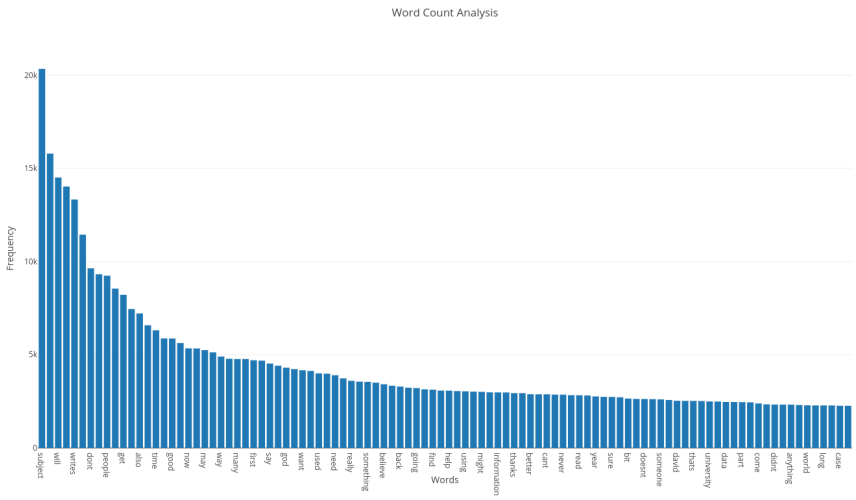

Figure 1: Frequency Analysis of top 100 words

4 Matrix Similarity Analysis

Once the tf matrix has been obtained, the next step is to calculate the similarity distances of this tf matrix on the basis of three metrics namely Euclidean, Cosine and Jaccardian. The definition of the following metrics is as follows

1. Cosine Similarity: Cosine distance is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of

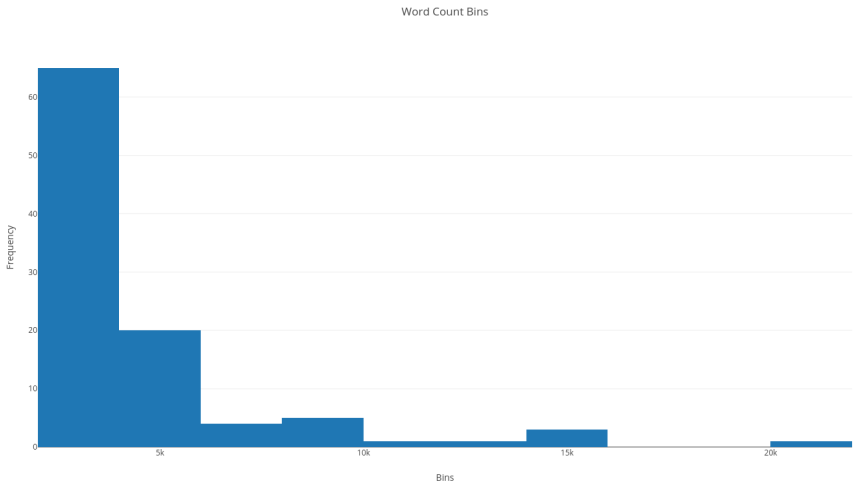

Figure 2: Histogram Analysis of top 100 words

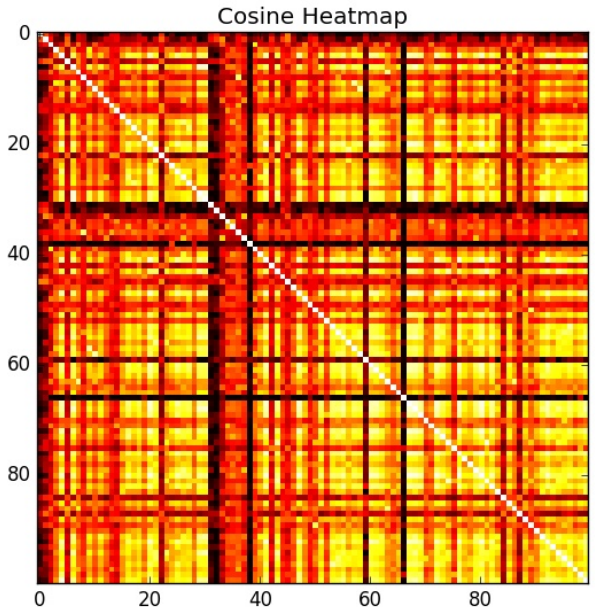

Figure 3: Cosine Similarity Distance matrix heat map

the angle between them. The cosine of 0 is 1, and it is less than 1 for any other angle in the interval [0,2). It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine

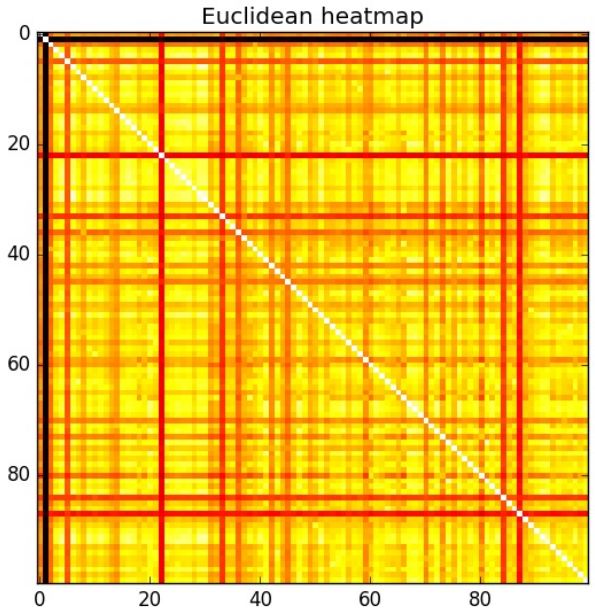

Figure 4: Euclidean Similarity Distance matrix heat map

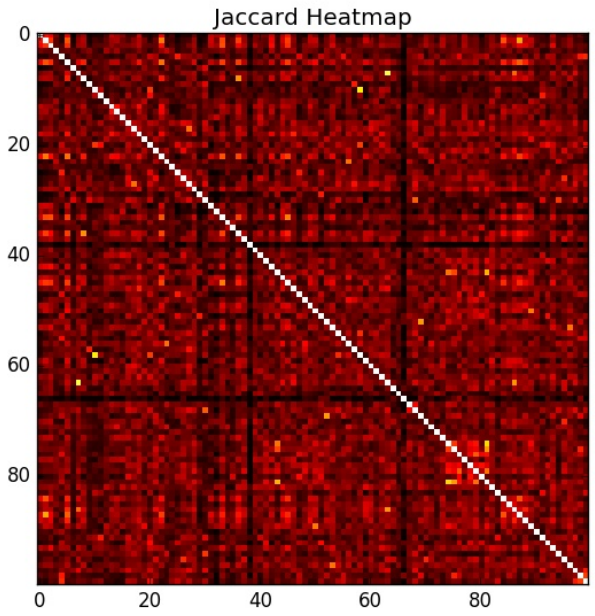

Figure 5: Jaccard Similarity matrix heat map

similarity of 1, two vectors at 90 have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. Cosine similarity is nothing but 1 - Cosine Distance.

2. Euclidean Similarity: In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" straight-line distance between two points in Euclidean space. With this distance, Euclidean space becomes a metric space.Euclidean Similarity is the inverse of Euclidean Distance plus one.

3. Jaccard Similarity: The Jaccard index, also known as Intersection over Union and the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. This gives the Jaccard Distance for each article rows. The Jaccard similarity is calculated by subtracting Jaccard distance from 1.

For the tf matrix obtained during the first stage when applied the distance metrics we generate matrix of size 18828 x 18828. The heat map analysis for all the three metrics is as shown in Figure 3, 4, and 5. Only a sub-sample of the tf matrix is analyzed here since the original matrix is of high dimension, and hence it becomes compute intensive.

5 Linear Regression to fit the three similarity matrices

Once the similarity matrices have been calculated we want to see as to how close these values are matching with each other. In order to do that Linear Regression is used. In statistics, linear regression is a linear approach for model ling the relationship between a scalar dependent variable y and one or more explanatory

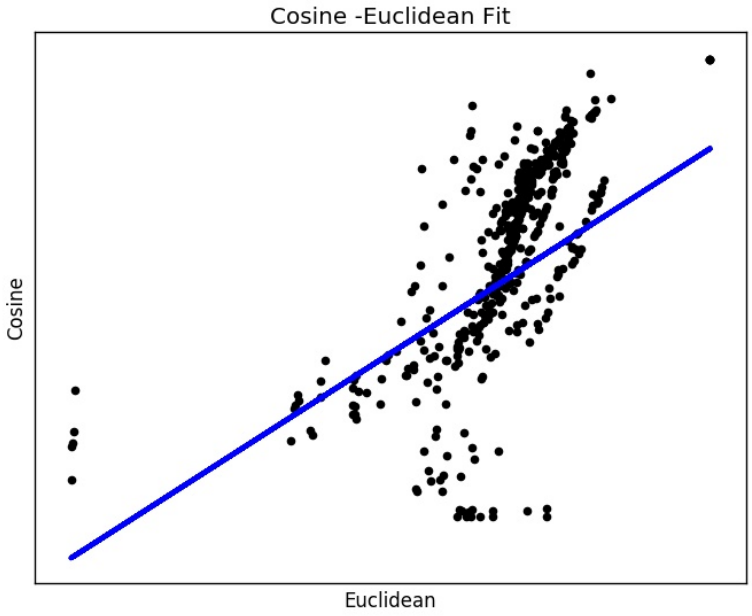

Figure 6: y = Cosine, x = Euclidean

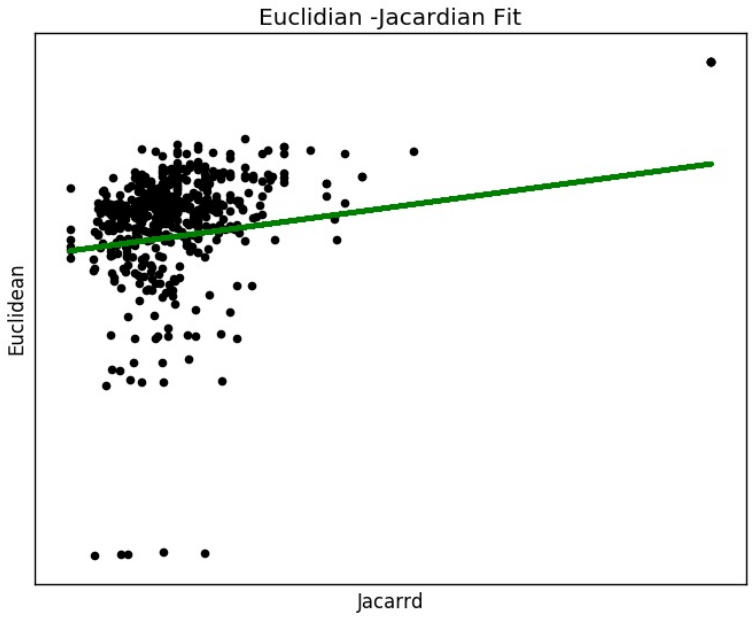

Figure 7: y = Euclidean, x = Jaccard

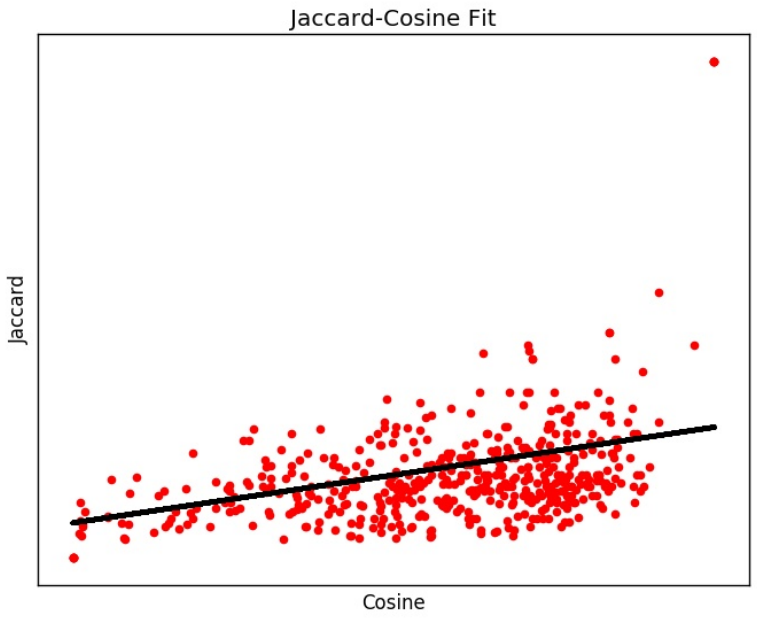

Figure 8: y = Jaccard, x = Cosine

variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.[1] (This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. The three similarity pair matrices are then fitted on the basis of this concept. Using the equation y = ax + b and applying it to the three similarity distance matrices we get the following linear equations:

1. Cosine= a* Euclidean + b
2. Euclidean = a* Jarcard + b
3. Jarcard = a*Cosine + b. The similarity matrices are split into test and train buckets and then linear regression is applied. When tested on the test data corresponding to x we get the straight line, and the scatter plots are points for the test data of x and y. The scatter plots and the fitted

lines for the three similarity pairs are as shown in Figure 6, 7 and 8.

6 Standard Deviation Analysis

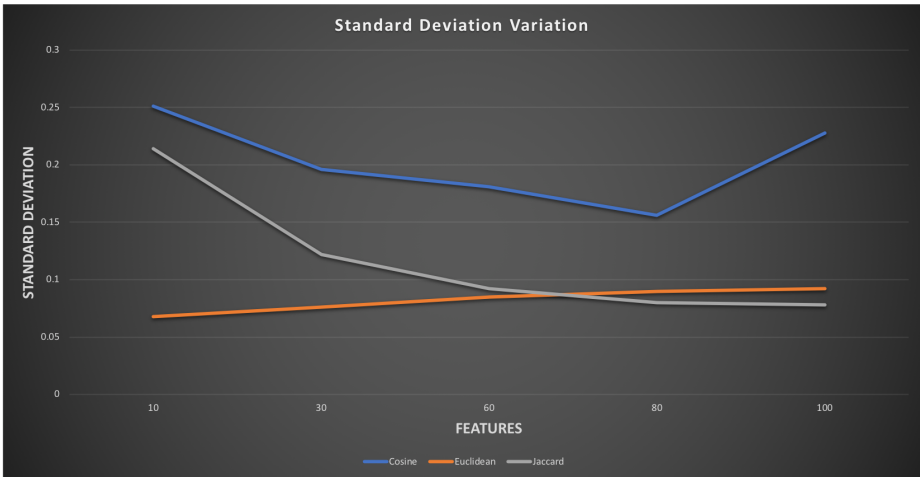

Figure 9: Standard Deviation variation plot

Table 1: Top 3 article pair comparison

Cosine	**Euclidean**	**Jaccard**
(980,330)	(980,330)	(567,11)
(230,456)	(230,456)	(67,88)
(1,11000)	(1,11000)	(1,9)

The standard deviation for each of the similarity matrices with respect to varying features is calculated and then plotted as two dimensional plot with the x axis indicating the features and the y axis indicating the standard deviation. This is shown in Figure 9. The similarity search matrices are then sorted in a decreasing fashion and from each similarity matrix the top three article pairs are selected. All these article pairs will have a score of 1 since they are similar in their distance metric. The top article pair rank is as shown in Table 1. Though Cosine and Euclidean seem to be giving almost the same similarity pairs Jaccard distance seems to be totally giving contradictory similarity pairs. This seems to a flaw in the analysis and needs some introspection.

7 Conclusion

This project mainly dealt with data preprocessing procedures as well as how to calculate similarity distances for various metrics. Finally the top three pairs are ranked for each metric and comparison is done as to which of the three seems to more similar. Although Euclidean and Cosine seem to be giving the same top 3 similarity pairs and are more accurate, the jaccard similarity seems to be giving contradictory results. This seems to suggest there is some flaw in the jaccard analysis and needs correction

8 Acknowledgment

Some of the definitions have been taken from Wikipedia.