Creating Tools for a Better Understanding of Datasets

Corbin Grosso
Faculty Mentor: Dr. Michael Bloodgood
Computer Science Department at The College of New Jersey

Background/Introduction

In order to use machine learning, data is required to train a model; however, some data may produce a better model than other data. This change in quality is a result of the two datasets containing different information, but it is usually unknown what the differences between the two datasets are. The goal of this research is to create tools that can be used to gain an understanding of what comprises a given dataset.

Method

The tools were written in Python 3 and work best with a dataset that has been normalized; in these experiments, datasets were normalized by having excess whitespace and non-alphanumeric characters removed, setting every character to lowercase, and replacing literal numbers with 'NUMBER' in all capital letters. Each article in the dataset is read in by the program and tokenized, creating a collection of unigrams, bigrams, trigrams, and four-grams. The number of times each unique token occurred throughout the dataset is tracked and used in many of the graphs, which are created using the matplotlib library.

The beginning experiments were dominated by common words that made trends harder to notice, so most experiments ran with these words being removed from each article before it is tokenized. Some examples include, but are not limited to, the following list of words: 'the', 'of', 'a', 'and', and 'if'.

Results

Each tool has the possibility of being far more or far less useful than the rest; the usefulness is often based on the data being processed, its size, and its components, as some common terms that carry little significance may appear often.

The Cumulative n-Gram Frequency Histogram often has the same general shape as it does in Figure 1. Since the majority of tokens have a low number of occurrences, this graph leads a square shape almost instantly. This graph shows how many distinct tokens there are, *y*, that occur *x* times.

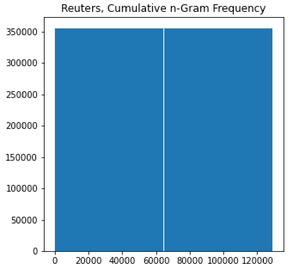

Figure 1:Cumulative n-Gram Frequency Histogram made using all of the data in the Reuters dataset.

Results (Cont.)

The Frequency of Frequency graph looks at how many tokens there are (*y*) that occur *x* times; it is the same data as the Cumulative n-Gram Frequency Histogram, but presented in a new way. Without magnifying the graph at all, as shown in Figure 2, the data seems very linear, but upon magnifying the data near the origin as Figure 3 does, a clear curve makes itself apparent, showing that there is a relationship between how many tokens occur a similar number of times and how many times they each occurred.

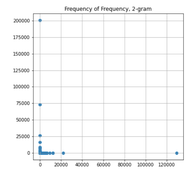

Figure 2:Frequency of Frequency graph made using all of the data in the Reuters dataset.

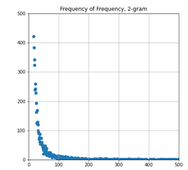

Figure 3:Magnified Frequency of Frequency graph made using all of the data in the Reuters dataset.

There are two different types of Relative Frequency of Frequency graphs: by Type and by Token. Both graphs aim to make the data relative to make it easier to compare datasets of differing sizes. Relative Frequency of Frequency by Type is relative to the number of distinct tokens, whereas Relative Frequency of Frequency by Token is relative to the total number of tokens in the data (for example, 'hello world' and 'hello world', being the same, would count as two total tokens, but only one distinct token). As seen in Figure 4 and Figure 6 respectively, the data is too close together to be understood, but their curve is clearly visible when magnified, as shown in Figure 5 and Figure 7. This curve shows the same correlation witnessed with the Frequency of Frequency graphs

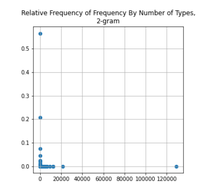

Figure 4:Relative Frequency of Frequency by Types graph made using all of the data in the Reuters dataset.

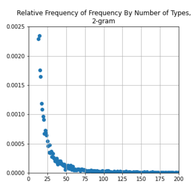

Figure 5:Magnified Relative Frequency of Frequency by Types graph made using all of the data in the Reuters dataset.

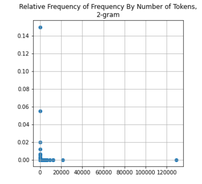

Figure 6:Relative Frequency of Frequency by Tokens graph made using all of the data in the Reuters dataset.

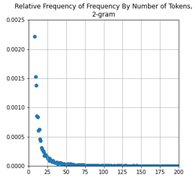

Figure 7:Magnified Relative Frequency of Frequency by Tokens graph made using all of the data in the Reuters dataset.

Results (Cont.)

The Cardinality of F_x Vs x uses a concept we have abbreviated to F_x, which describes which frequencies have *x* types that occur at that frequency. As an example, if there are 5 frequencies that all occur 3 times each, then $|F_3| = 5$. These values are inherently discrete, being the number of times something occurs, causing a series of short lines on many of the lower F_x values. These short lines are not visible in Figure 8, in which the data appears to be linear. When the graph is magnified, as in Figure 9, these short lines become clearly visible.

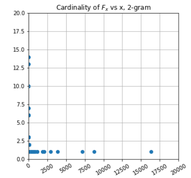

Figure 8:Cardinality of F_x Vs x graph made using all of the data in the Reuters dataset.

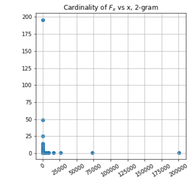

Figure 9:Magnified Cardinality of F_x Vs x graph made using all of the data in the Reuters dataset.

The *x* Times Sum of F_x graph also utilizes F_x, this time used to have all values of F_x summed and multiplied by *x*. This provides the total number of tokens that occur *x* times; for example, if 2 frequencies, *a* and *b*, both occurred 3 times, then there would be $3(a + b)$ total tokens that are part of F_3. This graph starts with a few abnormally high values, due to summing the value of every frequency that only occurred once. It rapidly goes down before rising again as *x* rises in value, as shown in Figure 10. When looking at a magnified version of the graph, as in Figure 11, there are some patterns such as lines that can be drawn between multiple points, where the sum of F_x is similar, so only the value of *x* is changing, causing a few linear patterns to appear.

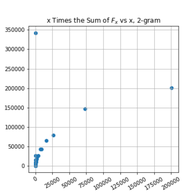

Figure 10:*x* Times Sum of F_x graph made using all of the data in the Reuters dataset.

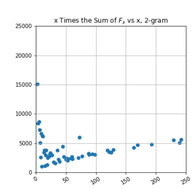

Figure 11:Magnified *x* Times Sum of F_x graph made using all of the data in the Reuters dataset.

Future Work

- Begin taking a more syntax-based approach to dataset analysis, such as performing part of speech tagging on a dataset before using the current tools on the data
- Look into fitting lines to the graphs with a strong recognizable pattern
- Integrate dataset analysis into our current infrastructure

Acknowledgements

The authors acknowledge use of the ELSA high performance computing cluster at The College of New Jersey for conducting the research reported in this poster. This cluster is funded by the National Science Foundation under grant number OAC-1828163.