VISUALIZING AND DISPLAYING DATA

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

Machine Learning involves complex mathematical computations which are possible today through the advents in computer technology and related fields. Finding and detecting patterns in data has always been a challenge to data scientists and alike, and again with the advent in computer technology, we can now visualize data in unprecedent ways.

In this paper, we will review some of the most effective ways and associated tools to visualize data when dealing with textual data. We will outline the importance of data visualization in the analysis and the challenges in the exercise.

Keywords:

Machine Learning

Visualizing and Displaying data

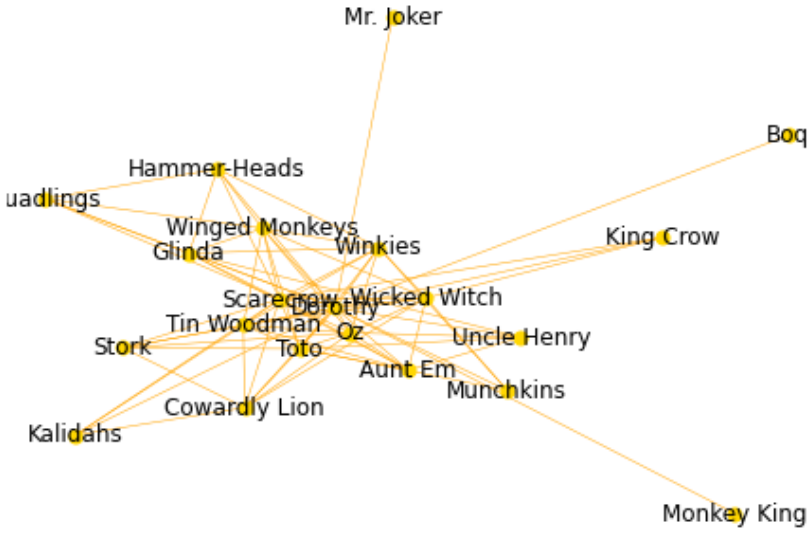
Visualizing Text Data

When dealing with numeric data with a finite set of features and outcome, we can use mathematical methods to find correlation between features and outcomes and build models and fine tune them based on tests. Scatter plots, line and other 3 or 3 dimensional graphs usually suffices the requirements laid out for the project. When dealing with unstructured data such as text, each letter, word, or sentence becomes a feature and the juxtaposition of those features can yield different outcome. In natural language processing, we need the algorithm to decipher context of the text. If misinterpreted, the consequences could be undesirable. Imagine speaking to an automated customer service.

With text visualization, we can see the features, and once we can see patterns, we can create models based on a feature set that can best predict the outcome. Traditional methods that we use in visualizing numerical data would fall short when dealing with high number of features. It therefore becomes a mandatory task to reduce the features which we can accommodate by analyzing the feature visually. For example, we can dissect text into n-grams and see what the juxtaposition of the n-grams would reveal the most accurate information about the document. In this case, we could visualize the number of times an n-gram is repeated over the course of a document. An example given in Bengfort and Bilbro shows the number of times the names of ‘trump’,’clinton’ and ‘bernie’ were found in documents gathered from March 10, 2016 until May 19th 2016. This sort of graph doesn’t give us any context of the documents, but informative, nonetheless. Other techniques are:

Network visualization: This type of visualization shows the relationship between the features in the document under analysis. For example, the story of Wizard of Oz given the name

of the characters in the story and the network nodes and the text as defining the relationship between the character will reveal the relationship of Dorothy with other characters. Shown below, we can see how close in proximity are the main characters such as Toto and Tin Man are to Dorothy. The algorithm for this graph uses a co-occurrence technique that yields the distance between nodes (or number of times each node/character have appeared together) based on the number for times they appear in the same sentence.



Co-Occurrence plots: We used co-occurrence algorithm to create network plot, co-occurrence plots show a heatmap of features where the color intensified for higher number of co-occurrences and less intense for lower number of co-occurrences.

Text x-rays and dispersion plots: This plot shows the overall narrative of the document under investigation. In the case of Wizard of Oz story, the characters that are mentioned in the document during certain events are visualized. For example, Dorothy and Toto are mentioned more during the tornado where Tim Man is not.

Luckily, in Python, there are tools that we can use to create algorithms and graph the results. Matplotlib, Scikit-Learn,NLTK,Gensim,SpaCy, NetworkX, and, YellowBrick, provide functions and utilities to analyze text.

Conclusions

From the perspective of data science, the world has more unstructured data than structured. This is especially true when dealing with text which involves spoken as well as written text. We must be able to make an accurate meaning of the data; otherwise natural language processing would not be possible. This is the biggest challenge in the analysis of text. The other challenge is the applications that we can create when we can manage textual data. With the tools, algorithms, and computing power available to us today, and the ones upcoming or improving, we can address the challenges before us.

References

1. Bengfort, Benjamin; Bilbro, Rebecca; Ojeda, Tony. Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning. O'Reilly Media. Kindle Edition.
2. PosTag Visualization - https://www.scikit-yb.org/en/latest/api/text/postag.html
3. Albon, Chris. Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning (p. 161). O'Reilly Media. Kindle Edition.