Cleaning Unstructured Data

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

Data analysis for the most part has focused on numerical values that represented characteristics that are more measurable and moreover computable. The data analysis of the modern age where source, volume and type of data are abundant and diverse, faces a bigger challenge because data has to be transformed to numeric data..

The purpose of this paper is to discuss unstructured data and the techniques used in cleaning and preparing datasets that contain them.

Cleaning Unstructured Data

Data Cleaning methods

What ends up as zero's and one's that cause the circuitry in the computer chips to toggle one way or the other and achieve a result are arranged by computation. Computation deals with numeric data that we can ultimately perform mathematical operations on. When faced with data that is not numeric, we are left with two options:

1. Remove from computation
2. Convert to numeric value

Option one may make sense, and the decision may very well be made with little or no justification or concern. However, more often than not, option 2 is the only option. It may have made sense to ignore non-numeric data in the past, but with diversity in type of available data, we have no choice but treat this as a challenge and address it. The transformation of non-numeric data or what it has become to be "unstructured data', is a mathematical challenge that has been addressed quite effectively in the recent years. Advances have been made in the following areas (Jeetu Patel):

- **Image and video**: Image recognition including landmark, facial(human/animal), motion detection, etc.
- **Audio:** Natural language processing, sentiment analysis, etc.
- **Text:** content recognition, sentiment analysis (tweeter conversations), language translation, etc.

Most of the applications of unstructured data are in machine learning and artificial intelligence domains. The transformation of the data is part of the data wrangling phase of the project and must go through some sort of cleanup and preparation. Audio and video data must go

through some sort of filtration depending on the application. Text data will have to be cleaned up, categorized, words counted, number of occurrence of words, which are useful words which are redundant are all part of the cleanup.

For example, encoding text as a bag of words would give us the number of times an observed text contains a particular word. It also makes features (or independent variable) out of the words in the text. The text could be anything from comments on items purchased to conversations in social media,

The code snippet below shows how we can transform the text into a sparse matrix with three rows (one for each piece of text/conversation), and as many columns as the algorithm could find features.

```
# Load library
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer

# Create text
text_data = np.array(['I love Brazil. Brazil!',
                      'Sweden is best',
                      'Germany beats both'])



# Create the bag of words feature matrix
count = CountVectorizer()
bag_of_words = count.fit_transform(text_data)

# Show feature matrix
bag_of_words
```

```
<3x8 sparse matrix of type '<class 'numpy.int64'>'
        with 8 stored elements in Compressed Sparse Row format>
```

In this case the number of columns/features is determined by the following list:

```
# Show feature names
count.get_feature_names()
```

```
['beats', 'best', 'both', 'brazil', 'germany', 'is', 'love', 'sweden']
```

The resulting sparse matrix looks like this:

```
bag_of_words.toarray()
```

```
array([[0, 0, 0, 2, 0, 0, 1, 0],
       [0, 1, 0, 0, 0, 1, 0, 1],
       [1, 0, 1, 0, 1, 0, 0, 0]], dtype=int64)
```

Put this matrix with feature names as shown below, we can see that the word Brazil was used twice in the 1st sentence, 0 times in 2nd and 3rd row. Multiply this conversation by 100's, we can decipher some sort of sentimentality toward different countries regarding their nation's superiority in the game of soccer.

Beats	Best	both	brazil	germany	is	love	sweden
0	0	0	2	0	0	1	0
0	1	0	0	0	1	0	1
1	0	1	0	1	0	0	0

Once data is transformed in this way, it can be further transformed and modeled. For example, the 1st step is to create a dense matrix from this sparse matrix. This is especially important when we are dealing with large volume of data. The next steps could be to find some correlation between features. These are all part of data wrangling.

Conclusion

Combination of the advances in mathematics with the availability of high-speed computational power, programming environments such as Python and R with a rich set of open source libraries, leave no choice to the analysis but to go where no one has gone before. The

speed at which data is accumulating plus the diversity of the data almost makes this a necessity as we go forward in this science that is of data and for data.

References

Albon, Chris. Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning (p. 121). O'Reilly Media. Kindle Edition.

Understanding Feature Engineering (Part 2)?—?Categorical Data (2018). Towards Data Science- https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63

Why One-Hot Encode Data in Machine Learning? (2017). Machine Learning Mastery - https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

Machine learning: Unlocking the Power of Unstructured Data (2018). Bloomberg Professional Services -https://www.bloomberg.com/professional/blog/machine-learning-unlocking-power-unstructured-data/

How Machine Learning Will Tame the Explosion of Unstructured Data (2018). CMS Wire - https://www.cmswire.com/information-management/how-machine-learning-will-tame-the-explosion-of-unstructured-data/

http://www.copyright.com/blog/machine-learning-understanding-difference-unstructuredstructured-data/ - Machine Learning: Understanding the Difference Between Unstructured/Structured Data (2018). Copyright Clearance Center -