Department of Physics and Astronomy

Adrian Vasu

Dr. Campbell

PHY 299: Scientific Methods with Python

April 8th, 2020

TABLE OF CONTENTS

Abstract

Data Science is a multi-disciplinary field and as such makes it very hard to narrow down a single project within the scope. The project chosen to demonstrate the programming skills learned in PHY 299 was an implementation of a Machine Learning algorithm to determine a person's political affiliation based on their twitter data. The results of this project demonstrate a XX% accuracy rate when comparing to the stated political affiliation of the US Senate. Overall the development of this algorithm demonstrates Machine Learning and Data Science fundamentals while remaining relatively accurate.

Introduction

Background

Data Science, “Big Data”, Analytics, Information Analysis, Business Intelligence, these are all fields that have arisen within the past few years. They all share a common basis in informatics and the use of information to develop knowledge and make educated and informed approximations once a baseline has been set. For this project Data Science fundamentals will be utilized to create a machine learning algorithm which is able to determine a user’s political affiliation depending on their twitter history.

Intro to Data Science

Data Science is defined as the field that utilizes different methods, algorithms, systems, and applications to extract knowledge from both structured and unstructured data [1]. This statement demonstrates that Data Science is an interdisciplinary field with many moving components. Data Science is heavily involved in pulling information from data that would otherwise be inaccessible.

As aforementioned Data Science involves the collection, study, analysis, and utilization of large data sets. These data sets are typically in the range of hundreds of terabytes to tens of petabytes for a supremely accurate algorithm. Data science in its fundamental approach makes use of the Law of Large Numbers (LLN). The LLN states that the average of the results obtained from a large number of trials should be close to the expected value and the error will reduce as the number of trials increases [2]. This is fundamental in the field of Data Science because it demonstrates that what was once viewed as a random event can be studied and a pattern can be formed. Once this pattern or algorithm is deduced and understood it can be applied to new data sets and similar deductions can be made.

This algorithm can be created through the application of Machine Learning. Machine Learning is the process of creating algorithms that computers can utilize to complete tasks without an explicit set of instructions. This mainly happens through the use of pattern recognition and inferences made throughout [3]. There are many Machine Learning applications in use today such as computer vision or email filtering. Machine Learning is just one of the many applications of Data Science and will be one area of focus in this project.

Intro to Twitter

Twitter is a social media platform developed by Jack Dorsey which was founded on March 21, 2006. Since then 14 years have passed and Twitter has become, and application used daily by more than 145 million people across the world. The differentiating factor between Twitter and its competitors is that Twitter only allows for the sharing of small 280-character messages (including photos or videos sometimes) to a network of followers which one would accumulate over time. These messages are known as "tweets" and are sent out instantaneously when one user decides to post. While Twitter has a set of guidelines which one is supposed to follow regarding what they are able to post, within the past few years Twitter has become a political hotbed with many people engaging in heated discussions about controversial topics. This recent influx of political material into the platform is also motivated by many politicians taking to the platform to share their messages with the "world".

Intro to Python

Python 3.9 is a programming language which many scientists and programmers take advantage of because of the various useful tools which Python allows one to utilize in various projects. One module which is particularly important; especially in the field of Data Science is SciPy. SciPy stands for Scientific Python and has various Machine Learning methods built in that one can use to analyze and report trends in data sets. Taking advantage of the SciPy toolkit also allows for the creation of graphics which represent the trends in the data.

How Does All of This Fit Together?

Programming, Data Science, Machine Learning, Twitter; how do all of these seemingly disjointed topics come together to produce a cohesive project? They join together in the development of a Machine Learning algorithm implementing Data Science fundamentals and principles utilizing Python 3.9 to import Twitter data, train and develop the Machine Learning algorithm, and finally apply this algorithm to other twitter users' data to determine their political stance.

Motivation

The motivation for this project is largely due to the aforementioned recent influx of political messages that have become prevalent on Twitter. Through this I have noticed some discrepancies between what a person states their political alignment is and the material which they are posting

and supporting on social media. This discrepancy should be quantified, and I would like to know how accurate someone's assessment of their own political beliefs is when compared to a quantified value determined by an objective algorithm. This combined with my general interest in politics and programming/computer science makes a Machine Learning implementation the logical next step in my Data Science education.

Results

The project is currently undergoing development and as such there are no results yet (not even preliminary) to report.

Methods

High-Level Approach

This project will make use of a standard Software Engineering program layout. There will be several classes each with their own definition which will perform the various tasks required. There will be a data requisition class which brings in all of the requested data from Twitter. There will be a data cleaning class which removes all of the erroneous information from the tweet object that is brought in from twitter. There will also be two classes which control the training and application of the Machine Learning algorithm. Finally, there will be a controller class which oversees all of the other classes and ensures that the program is running smoothly.

Code Overview

There are many moving pieces in this project and this section is designed to give an overview of the many pieces and how they interact. As the project is still under development there are going to be major holes throughout this section because I don't know how they're going to fill in yet.

Modules Utilized

The modules I have begun utilizing are python-twitter, SciPy, and all the standard inclusions within Python.

Major Methods

There are no major methods yet.

Data Generation

The data has not begun to be generated yet.

Discussion

Recap

Project Motivation

Results Interpretation

There are no results to interpret at this time.

Results Summary

There are no results to summarize at this time.

Conclusion

In conclusion this project demonstrates that through the use of a user's Twitter data their political affiliation can be approximated to a reasonable conclusion.

References

Works Cited

[1] V. Dhar, Data science and prediction, ACM, 2013.

[2] G. R. Grimmett and Stirzaker, D. R., Probability and Random Processes, Oxford: Clarendon Press, 1992.

[3] A. Samuel, How can computers learn to solve problems without being explicitly programmed?, Dordrecht: Springer, 1959.