Donal Maher 22178, word count 2757

CA2 Final project

Contents

[Aim 1](#_Toc136324895)

[Introduction 1](#_Toc136324896)

[Objectives 1](#_Toc136324897)

[Discussion 2](#_Toc136324898)

[Conclusions 9](#_Toc136324899)

[References 10](#_Toc136324900)

[Appendix 10](#_Toc136324901)

Aim

To collect at least 1 year's tweets on a topic, this data should be stored as requested below, and you are then required to analyse any change sentiment that occurs over the time period that you have selected.

Following your analysis, you are then required to make a time series forecast of the sentiment for 1 week, 1 month and 3 months going forward. This forecast must be displayed as a dynamic dashboard.

# Introduction

There has been in the last number of the past years data explosion and this vast amount of data has been collected and stored. The analysis is the key to delivering more efficient processes and forecasting demand. This project has been developed to investigate the background elements that are used to improve the analysis of the collected data. Hadoop, is an open-source framework a most powerful distributed computing platform capable of processing massive datasets across clusters of computers.[7]

This project will evaluate these technologies using Twitter tweets which will be implemented on Ubuntu [8]through a virtual machine, with a configured VM image and installed with the latest Hadoop which includes HDFS and YARN.

And display on an interactive dashboard that is built in Python code.

# Objectives

1. Details of the data storage and processing activities carried out, including preparation of the data and processing the data in a MapReduce/ Spark environment.
2. A discussion of the rationale and justification for the choices you have made in terms of data processing and storage, programming language choice, machine learning models and algorithms that you have implemented**.**
3. Comparative analysis for at least two databases using any benchmarking tool. (For example, ycsb)

4. Your analysis of  any change sentiment that occurs over the time period that you have selected.

5.          Your forecast of the sentiment at 1 week, 1 month and 3 months going forward**.**

6. Presentation of results by making appropriate use of figures along with caption, tables, etc and your dashboard for your forecast**.**

# Discussion

1. **Details of the data storage and processing activities carried out, including preparation of the data and processing the data in a MapReduce/ Spark environment.**

Data storage carried out Pyspark running hadoop dfs and yarn. Then using the select query to extract the data and into the pyspark dataframe. From here the data was transformed to a pandas dataset. And then written out the a comma separated file.

The data was not scrapped from the twitter as the standard developers account was insufficient to allow the data to be scrapped. The dataset is a historical dataset [1] with over 2097150 rows which contained the topics:

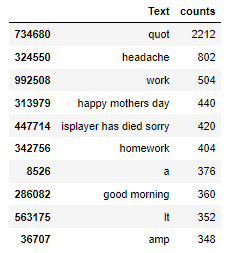


Figure 1

Figure 1 is the results of preforming topic modelling on the twitter data. Used to figure out the topics of each tweet. This is done by preforming a simple Exploratory data Analysis (EDA) on the data. Appling LDA, pandas, numpy seaborn, mathplotlib, sklearn and nltk.

LDA is latent Dirichlet Allocation, is a probabilistic generative model used for topic modelling. This technique of unsupervised learning used to figure out the hidden topics within a collection. LDA does assume the each of the documents is mixed up set of topics and the words contain in is a distribution. LDA allows the discovery of topics without the need of labelling of the data. This can be seen in GetTopic.ipynb file. Or in appendix 1.

Its commonly known that twitter uses hashtags(#). Used before a relevant keyword or phrase in the tweet to categorize those Tweets and help them show more easily in a twitter search.

The process of implementing LDA

**Preparation of the data**:

Cleaning the data was done in the after this part above as the cleaning of the data would have removed the hashtag that is required for this implementation of LDA.

The steps involved with cleaning the data:

1. Handling Missing Data.

Determining if the data contains null values with the function “isnull()”. If the data contain null values.

Fill then with “fillna()” or drop the rows or columns with dropna(). Not normally a good idea to drop that that contains null values. You could also use K-nearnest neighnourn to determine which value the to fill or in the case of numerical filling the values with mean or median values can also be done.

1. Handling Outliers

An outliers is a data point that stew that data and will affect the results and need to be handled. The outlier is detected by using, IQR (interquartile range) or the Z-score (standard score) which measures how many standard deviations the data point is way from the mean will allow.

1. Data Transformation

When the data is read in python assumes the type and sometimes get this wrong, which can affect how code is processes, for example the twitter data has special characters and hashtags which python cannot be certain about and classifies as a object. This has to be handled by using the astype function and declaring the types “tweets['Text'] = tweets['Text'].astype(str)”.

1. Handling Duplicates

Duplication of the data can affect the results, handling the duplicates using the duplicated () function and using drop\_duplicates() if required.

1. **Comparative analysis between MySQL and pyspark using any benchmarking tool.**

Introduction:

When comparing two databases, implementing a benchmarking tool YCSB (Yahoo! Cloud Serving Benchmark) used to evaluate their performance in a NoSQL framework for example mongodb.

YCSB is a popular open-source benchmarking tool designed for evaluating the performance of key-value and cloud data serving stores. [ <https://docs.yugabyte.com/preview/benchmark/ycsb-jdbc/> ]

In this section discussion will focus on the benchmarking using comparison between MySql and Pyspark using Sysbench.

Sysbench is a popular open-source benchmarking used to test open-sources database management [2]

The objective here is to insert the captured data into both mySql and Pyspark and to compare using benchmarking like YCSB. Is just one benchmarking tool, and there are other options available depending on the type of database you want to evaluate. It's important to consider the specific characteristics of your databases and workload requirements to choose the most appropriate tool.

|  |  |  |
| --- | --- | --- |
| **Database** | **MySql** | **Pyspark** |
| **Dataprocesssing and Storage** | Optimized for handling ACID-compliant transactions and has efficent indexing and querying. | Used for big data processing and analytics. Handles large-scale data processing including batch, real-time streaming and Machine learning tasks. |
| **Scalabliity and Performance** | Moderate to large-scale datasets.  Single-node deployments | Massive to complex datasets.  Uses cluster node processing |
| **Programming Language and Ecosystem** | SQL | SQL |
| **Machine Learning** | Needs to have integration between external libraries or framework | Integrates easily with Python library, supports Machine Learning |
| **Community and Adoption** | Has been around for a long time. | Increasing popularity |

Pyspark in the obvious choice just for the documentation. Process would seem to be faster and will allow the interaction with python and pythons machine learning libraries.

By performing a comparative analysis using benchmarking tools like YCSB, you can gain insights into the performance, scalability, and efficiency of different databases, helping you make informed decisions based on your specific use case and requirements.

This became apparent the this was a task that was not possible to the errors that occurred while attempting to set up and inserted the twitter data in the mySql database on the virtual machine not to mention the storage limitations when dealing with large datasets. See the error in Appendix 3

1. **Your analysis of any change sentiment that occurs over the time period that you have selected.**

Introduction

Sentiment analysis uses the natural language processing (NLP) library and techniques to figure out the if the text contains in the dataset is positive, negative or neutral. And determines if the tone or the underlying emotions or attitudes in the text is positive, negative or neutral. Known as opinion mining and is a natural language processing technique.

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Figure 2

Figure 2 shows the sentiment of the entire twitter dataset. Which contains over 400,000 Netural elements, over 350,000 Positive sentiments and over 270,000 negative sentiments.

**A picture containing text, screenshot, line, plot

Description automatically generated**

**Figure 3**

The change in sentiment over time refers to how the emotional tone is expressed in the text data over time.

For example, in figure 1 the shows what is captured by LDA in the form of topics in the twitter data over the period of time how this topic has progresses. Has the topic decrease/ increased or stayed neutral.

This section will detail that the topics over a period of 1 day,1 week and 1 month. By comparing the sentiment distribution over time to show if there is a trend and to observer how the sentiment fluctuates over the time period. This could be used to identify shifts in public opinion, track sentiment during significant events, or monitor changes in customer sentiment. Which will provide valuable insights into how sentiment and attitudes change over time, this is noted to be a key area in which the final portion of this course will help identify changes in customer sentiment of the industry.

|  |  |
| --- | --- |
| Techniques of Sentiment analysis[3] | |
| Text processing | Tokenization , removing stopwords, handling negations, lemmatization and removing special characters , like hashtags. |
| Evaluation Metrics | Accuracy, Precision, recall, F1-score and ROC |
| Aspect-based Sentiment Analysis | The understanding of associated sentiment analysis with different aspects of the product, service or topic. |
| Domain-specific Sentiment Analysis | The training the models on domain data or domain specific sentiment lexicons |
| Machine Learning Approaches | Training models on labelled datasets. Using Naïve Bayes, SVM, decision tree Neural Networds and Random Forest |
| Lexicon-based Approaches | Each word is match against the lexicon. The sentiment is calculated on the polarity of words.  VADER, Senti,WordNet and AFINN |

Choosing the correct technique is the process of evaluation of the data, resources and characteristics of the text data that is being analysed.

Sentiment analysis has limitations, it may find difficult to capture sarcasm etc.

|  |  |  |
| --- | --- | --- |
| A picture containing text, line, diagram, plot  Description automatically generated | A picture containing text, diagram, screenshot, line  Description automatically generated | A picture containing text, line, screenshot, diagram  Description automatically generated |
| Figure 4 | Figure 5 | Figure 6 |

Looking the topic “Work” you would expect that the Work could be a positive negative and neutral sentiment the is 100% neutral.

A picture containing text, screenshot, rectangle, square

Description automatically generated

Figure 7

The figure above shows that the data sentiment above is on the topic of “work” show all the netural values.

**The forecast of the sentiment at 1 week, 1 month and 3 months going forward.**

Forecasting the process of predicting future values, in is bases on data and patterns from the past, forecasting is the process of predicting future values or trends. It is a useful tool for planning, making well-informed decisions, and comprehending potential outcomes. With regard to the models ggiven we investigated estimating opinion values in light of the given dataset.

There are several steps involved in the forecasting:

1. The Preparation of the Data.

Seen in Sentiment analysis “Sentiment Analysis (Opinion Mining) with Python — NLP Tutorial amended.ipynb” Jupiter notebook shows that reading in the data and writing out to a cleaned CSV file is the best method of keeping the working of data preparation to the minimum. The only data preparation done here is the inclusion of the sentiment value along with the dummies that rate the sentiment.

1. Model Selection [4]

|  |  |
| --- | --- |
| **Model Name** | **Brief description** |
| ARIMA | A classic. Uses autoregressive (AR) , integrated and moving average(MV) |
| Prophet | Facebook uses generalised additive model (GAM) incorporates seasonal trend |
| Seasonal Decomposition of Time Series (STL) | Break down into trend and seasonality and recombines to produce the overall forecast |
| Exponential Smoothing (ES) |  |
| Vector Autoregression (VAR) | Used on multiple time series. Captures relationships between variables and can be used for multivariate time series forecasting |
| Generalized Autoregressive Conditional Heteroskedasticity | Forecasting volatility or the conditional variance of a time series. |
| Neural Networks | LSTM or RNN captures complex patterns and dependences in the data |
| Gradient Boosting Models | XGBoost or LightGBM iteratively learn makes predctions for non linear patterns |
| State Space Models | Kalman Filters and structural time series |

How to choose a model

Depends on the requirements of your data , the presence of seasonality or trends, the availability of external factors and the level of complexity wanted.

It is important to note that the code snippets provided were simplified examples, and forecasting accuracy and performance can vary based on a variety of factors, including the dataset, model selection, parameter tuning, and so on. To ensure that forecasts are accurate and meaningful, it is essential to thoroughly comprehend the data's characteristics, investigate a variety of modelling methods, and validate the results.

You can further analyse and interpret the forecasting results by comparing the actual sentiment values to the ones that were forecasted. You can also look for patterns or trends in the forecasts and use the right evaluation metrics to determine how accurate the predictions are. Also, you might need to consider consolidating different elements or variables that could affect feeling to further develop the anticipating exactness and give more exhaustive bits of knowledge.

In general, forecasting is an effective instrument for comprehending and predicting outcomes in the future based on data from the past. Forecasting can assist in making informed decisions, comprehending trends, and anticipating changes in sentiment or other target variables by utilizing the appropriate modelling techniques, evaluating the results, and iteratively refining the models.

Ultimately, the choice will depend on the combination of careful analysis, experimentation, and consideration of the specific problem and requirements.

The results of the forecasting are shown in the Forecasting for 1\_week\_4\_weeks\_3\_months for the topic Work.ipynb Jupiter notebook file. The model uses here is called Arima uses (p,d,q) parameters. Is popular model is widely used for time series forecasting. Enables the learner to have many examples at had to figure it out. The model combines autoregressive. Which captures the linear relationship between the current values of the time series and its past value [5] Will predict based on the past data (AR). Integrated(I) used to address statistical properties such as mean, variance. Which allow the ARIMA to remain constant over time. The moving average (MA) models the dependence between the current values (time series , linear) combinations of past residuals. Which help reduce noise.

The “p” parameter. Found in the (AR) part . The order or the past observations used for prediction

The “d” parameter . Found in the (I) part. The order of differencation required to make series static computing the differences and to remove variances ie trends or seasonality.

The “p” parameter. Found in the (MA) part. The order of the moving average which indicated the number of lagged forecast errors used in the model.

Kind of like a PID controller. Has the (AR)(I)(MA)

As expected, the forecast was neutral and has returned:

|  |  |
| --- | --- |
| Forecasted Sentiments for 1 week | 0.0 |
| Forecasted Sentiments for 1 month | 0.0 |
| Forecasted Sentiments for 3 months | 0.0 |

Assuming that the return from the analysis of the sentiments shows that they were all neutral is the most likely explation for the neutral forecasted of the time periods of the forecasts.

**Presentation of results by making appropriate use of figures along with caption, tables, etc and your dashboard for your forecast.**

This section shows the basic dashboard creates to allow the forecasting to be delivered, implemented with python dashboard plotly to create a simple interactive dashboard that is represented the actual sentiments’ and the forecasted sentiments .

# Conclusions

The project focuses on Hadoop and the power that follows the processing of data on a distributed file system. The comparison between the processes of a sql databases mySql and pyspark in always to prove the superiority of pyspark and even though there were many problem during this module and during this project there always show the power of the distributed file systems over the mover conventional system. But there had to be a system to process the developed to allow the swift process rather than the conventional systems.

The project has been very difficult and on more that one occasion the VM had to be reinstall due to a hard drive space which lead to slowing of my process. I recommend that at the allocation of 100GB to the virtual would help to mitigate the issues caused.

Data is difficult to get as a developer’s account is no longer available without subscription. The alternative is to pay for the data or to find a dataset that had scraped tweets, this took extra time.

Most difficult in delivering the comparison MySQL in the installation of storage when installing MySql on the VM urbanity but research allow me to do not directly a comparison to which the results expected were confirmed.

For forecasting in this project resulting in the interactive dashboard not being able to populate another example using the sentiments was developed and displayed in the panel.ipynt and in appendix 3.

Git hub could not be pushed to the repo create with the collage id which ,means that the repo had been create on personal name.

# References

[1] Usen, O. (2019). Topic Extraction from Tweets using LDA. Medium. Retrieved from <https://medium.com/@osas.usen/topic-extraction-from-tweets-using-lda-a997e4eb0985>

[2] Amazon Web Services. (Year). Optimizing MySQL on EC2 using Amazon EBS: MySQL benchmark observations and considerations. Retrieved from <https://docs.aws.amazon.com/whitepapers/latest/optimizing-mysql-on-ec2-using-amazon-ebs/mysql-benchmark-observations-and-considerations.html>

[3] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press. Chapter 21 covers sentiment analysis and provides a comprehensive introduction to the topic.

[4] Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd Edition). OTexts. Available online: [https://otexts.com/fpp2/](<https://otexts.com/fpp2/>)

[4] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135. This survey paper offers an extensive overview of sentiment analysis techniques and approaches.

[5]. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167. This book by Bing Liu covers the fundamental concepts, techniques, and challenges in sentiment analysis.

[6]. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media. Chapter 6 focuses on sentiment analysis using Python and introduces various techniques and libraries.

[7] Apache Software Foundation. (Year). Apache Hadoop Documentation. Retrieved from <https://hadoop.apache.org/docs/>

[8] Ubuntu. (Year). Ubuntu Documentation. Retrieved from <https://help.ubuntu.com/>

# Appendix

1. GetTopic

A screenshot of a computer program

Description automatically generated with low confidence

1. Clean data



1. Dynamic Dashboard

A screenshot of a computer screen

Description automatically generated with medium confidence