CCT

MSc In Data Analytics CA2

Contents

[1 Executive Summary 3](#_Toc135678245)

[1.1 Introduction 3](#_Toc135678246)

[1.2 Project Planning 3](#_Toc135678247)

[2 Data Understanding and Cleaning Phase 3](#_Toc135678248)

[2.1 Raw Data Acquisition 3](#_Toc135678249)

[2.2 Positive Aspects of Research and Acquisition 3](#_Toc135678250)

[2.3 Negative Aspects of Research and Acquisition 4](#_Toc135678251)

[2.4 Data Cleaning 4](#_Toc135678252)

[3 Statistical Analysis 6](#_Toc135678253)

[3.1 General analysis 6](#_Toc135678254)

[3.2 Inferential Statistics 9](#_Toc135678255)

[4 Exploratory Data Analysis: EDA 13](#_Toc135678256)

[5 Machine learning 15](#_Toc135678257)

[5.1 KNN Model 15](#_Toc135678258)

[5.2 Linear Regression 16](#_Toc135678259)

[5.3 Decision Tree Regressor 17](#_Toc135678260)

[5.4 Random Forest Regressor 17](#_Toc135678261)

[6 Data Manipulation Comparison 19](#_Toc135678262)

[7 Testing & Optimisation 19](#_Toc135678263)

[8 Reference 20](#_Toc135678264)

[Annexture 21](#_Toc135678265)

# Executive Summary

## Introduction

This project is an integrated assessment used for analysing a large dataset derived from the Twitter API named “ProjectTweets.csv”. The dataset comprises od 1,600,00.0 tweets, and the primary objective was to investigate the sentiment shift over the recorded time period. A time series forecast was mandated fir projecting the sentiment across 1 week, 1 month, and 3 months into the future. The forecast results representation in a dynamic dashboard.

## Dataset description

The Twitter dataset comprises of five key attributes: ”ids” (tweet ID), “date” (tweet date), “flag” (query status), “user” (tweeting user), and “text” (tweet content).

# Data Storage and Processing

The dataset “ProjectTweets.csv” was stored into the Hadoop Distributed File System (HDFS), a reliable and scalable storage system designed to span across multiple clusters of commodity hardware (Databricks, 2023). The storage procedure followed these key steps:

* HDFS Initialization: Using the command “start-dfs.sh”, the HDFS was initialized. This command ensured all background processes have been successfully started and are currently active.
* YARN Initialization: The YARN (Yet Another Resource Negotiator) resource management platform was initiated using the “start-yarn.sh” command. YARN is very important for efficient resource allocation and task scheduling within Hadoop ecosystem.
* Dataset storage: The dataset was saved in HDFS using the command “-put” and using “user1” folder.
* Pyspark Integration: To facilitate data processing and analytics, Pyspark was used. Pyspark is an interface for Apache Spark in Python with powerful big data processing capabilities.



Figure 2.1: Data access denied for free users.

# Reference

# Annexture

A screenshot of a computer

Description automatically generated

Note: the total count includes section title, references etc.

Github project location: https://github.com/cuculicu/CA2-Second-semester