Processing Tweets & Analyzing Sentiments for Business Insights

*IS 7034 – Data Warehousing & Business Intelligence*

*By*

*Onkar Suhas Samant*

*Pravallika Mulukuri*

*Stephen O. Yeboah*

*Mohammed Nifaullah Sailappai*

Contents

Introduction ................................................................................................................................................. 2

Use Case…………….………..................................................................................................................... 2

Software Details………….……………...................................................................................................... 2

BI & IT Architecture Integration…………................................................................................................. 2

Software Recommendation……………….................................................................................................. 3

Future Developments………………...……………………........................................................................ 3

Conclusion………….………………...……………………........................................................................ 3

References..................................................................................................................................................... 3

Introduction

This report involves the application of Data Warehousing (DW) and Business Intelligence (BI) ideas to a real world scenario. We present a prototypical demonstration of how opinions from public social media platforms, such as tweets from twitter, can be used to drive business decisions, using various tools and technologies from the DW & BI domain. We’ve restricted our scope to only tweets related to two TV shows namely Money Heist and Tiger King of Netflix. Essentially, we extract key indicators from these tweets to drive business decisions with respect to future shows at Netflix. We first use the GetOldTweets3 library of python to extract tweets, we then manually insert these extracted and tabularized tweets into a RDBMS Table, defined inside the MySql Instance of the Google Cloud. We then use PySpark to load data into a Jupyter notebook and then perform some elementary descriptive analysis on the tweets. Further, we make use of TextBlob & NLTK libraries in python to classify the sentiments for each of the tweets.

Use Case

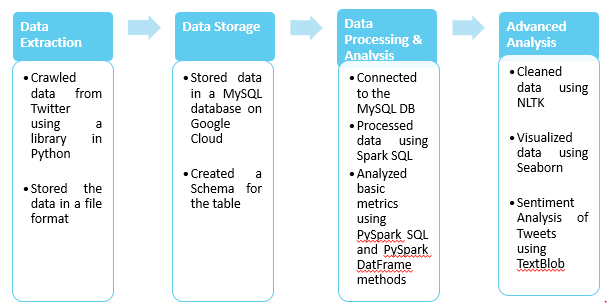
Make use of publically available opinions to drive business insights such as below and many others.

1. How different shows compare with one another?
2. What are the sentiments with respect to a TV show?
3. How people are reacting to announcements & trailers?
4. What is the distribution of tweets pre-release and post-release with respect to a show?
5. How does the distribution of different compare with one another?
6. Is there a high demand for a particular type of Genre during a particular period?
7. What are the most popular Genres?
8. Does releasing 2 shows together benefit Netflix?
9. Location specific insights for a show

Software Details

* Data Extraction: Python library GetOldTweets3
* Data Storage: MySQL database on Google Cloud
* Data Processing and Basic Analysis: PySpark SQL and PySpark DataFrame
* Advanced Analysis: Python libraries like Seaborn, NLTK, TextBlob

BI & IT architecture integration



Software Recommendation

* Data Extraction: Use licensed official Twitter API to extract data.
* Data Storage: NoSQL Databases are recommended because these do not require a fixed schema, avoid joins, and are easy to scale.
* Data Processing and Basic Analysis: PySpark SQL and PySpark DataFrame
* Advanced Analysis: Python libraries like Seaborn, Tensor flow 2.0

Future Developments

* Scale the implementation to accommodate more shows
* Create a windows service or a job to automatically scrape data, classify the data and directly store the data in the physical database.
* Develop a tableau dashboard or a Power BI based reporting system to display live analysis based on the processed data.

Conclusion

Thus with the analysis, supplemented along with the report, we managed to successfully demonstrate how we can make use multiple tools such as python, spark and SQL within the Data Warehousing & Business Intelligence umbrella to create a workflow involving public opinions of customers to provide a Business (i.e. Netflix in this case) with helpful insights in order to drive key business decisions.

References:

1. [GetOldTweets3](https://pypi.org/project/GetOldTweets3/)
2. [Google Cloud SQL](https://cloud.google.com/sql/?utm_source=google&utm_medium=cpc&utm_campaign=na-US-all-en-dr-bkws-all-all-trial-e-dr-1008076&utm_content=text-ad-none-any-DEV_c-CRE_79747410847-ADGP_Hybrid+%7C+AW+SEM+%7C+BKWS+%7C+US+%7C+en+%7C+EXA+~+Google+Cloud+SQL-KWID_43700009739675353-kwd-28489936691&utm_term=KW_google%20cloud%20sql-ST_google+cloud+sql&gclid=Cj0KCQjws_r0BRCwARIsAMxfDRgITu7m4khDemy4Yu7mjrRj2PcZfHGUgU3WhOHIyK5iu9kNnGSARmkaAm6vEALw_wcB)
3. [Install pyspark for windows](https://changhsinlee.com/install-pyspark-windows-jupyter/)