## horizontal line



**Machine Learning Project Proposal - YouTube**

By

Martin Walsh - C00170339

Osama Abou Hajar - C00220135

Rhyder Quinlan - C00223030

# 

*27/09/2019 Fourth Year Data Science Greg Doyle*

[**Introduction**](#_6qaaym4s6uit) **3**

[**Purpose and Goals**](#_46l9bw85bfqn) **3**

[Research questions](#_wj0bogawzeu) 3

[Learning Outcomes](#_3zrgcbbfyzhq) 4

[**Project Strategy**](#_q4znlzww70bf) **4**

[**Scope**](#_8h6496qqivm4) **5**

[**Deliverables**](#_7kuf9328h55w) **5**

[**Functional Interfaces**](#_qj8lthutj4rf) **6**

[**Standards**](#_citrf9yp9tl6) **6**

[**Conclusion**](#_yoi88peq113l) **6**

[**Reference/Research Documents**](#_90x4oq6csc9r) **6**

# 

# **Introduction**

The purpose of this project is to analyse YouTubes public sets. This work will be carried out by first breaking the data sets down by region. Then using the Python programming language paired with machine learning algorithms to retrieve the required results. As the data sets are large, the use of visualisation tools, such as Tableau, will be used to help see and understand the data better. In addition, the data sets may need to be divided among multiple laptops due to processing power needs.

The following report will be broken into multiple sections; Purpose and Goals, Project Strategy, Scope, Deliverables, Functional Interfaces, Standards, Conclusion and References in order to accurately address all substantial areas of this project before progress is started.

# **Purpose and Goals**

## **Research questions**

1. The most common languages in each region excluding its native language?
   1. This will give an estimated percentage of its populations nationality, allowing more specific marketing strategies.
2. Algorithm to automatically generate tags based on analysis of top performing videos.
   1. This will allow a user to generate the most effective tags to increase their videos reach.
3. Most popular categories in each region based on interactions with the video? i.e. likes and comments.
   1. This will allow individuals to choose the most effective category for video growth as interactions directly correlate to a videos popularity.

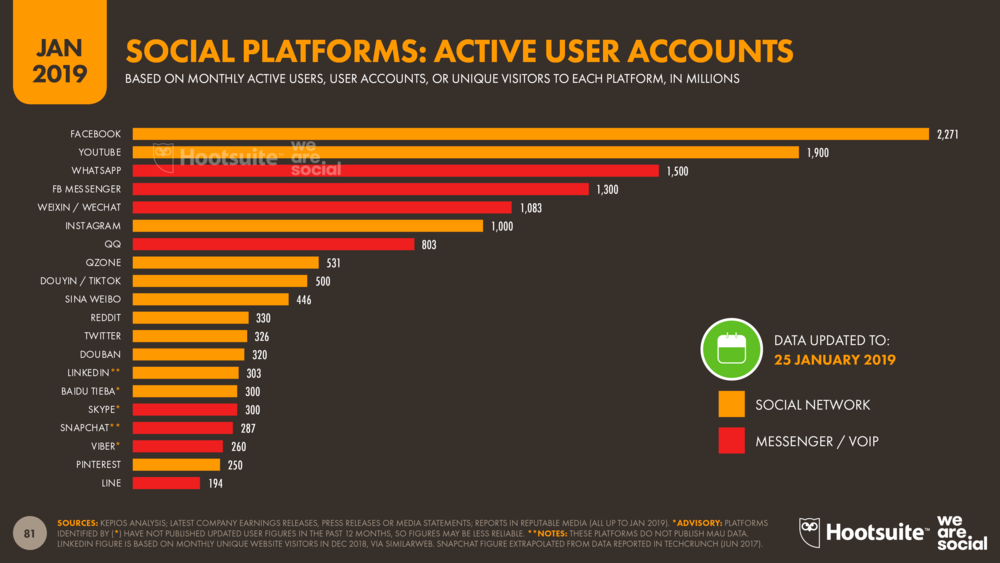
## **Learning Outcomes**

1. How to use Python programming language to handle large data sets.
2. How to code and implement Machine Learning algorithms.
3. Working with data visualisation tools.
4. Working with large data sets over multiple machines.
5. Insights into each regions YouTube metrics.

# **Project Strategy**

The aim is to analyse the YouTube Data Sets of each region. As YouTube is used by over 1.9 billion people worldwide[*Figure 1.0*] through a large variety of devices, ranging from mobile phones to smart TVs, it allows us to gain a real insight into a regions population. With this information we can see a regions most used languages, most popular categories and amount of time each region spends watching YouTube videos, etc.

The approach to analysing this data, as it is large, is to divide the data sets between the three members laptops in case the processors cant handle it. We will also try to blend multiple data sets from different resources and combine them to give a more detailed and informative data set.



**Figure 1.0: Social Platforms: Active User Accounts [Hootsuite]**

# **Scope**

The project will encompass analysing the data sets of multiple regions, by dividing them into smaller sets(ie countries). Once the country specific data has been analysed and the above research questions have been answered, the results of which could then be applied to larger geographical areas, i.e. Europe, Asia, etc. This will be carried out using python(Jupyter Notebook) and machine learning algorithms yet to be decided.

# **Deliverables**

Throughout our project we aim to deliver each of the items listed below;

1. Project Specification
2. Research / Application & Design
3. Mid Point Project Presentation & Feedback
4. Project Implementation
   1. Datasets will be represented via excel or CSV sheets
   2. Jupyter Notebook to represent the algorithms used
   3. Localhost webpage to display the final results
5. Final Report & Presentation

# **Functional Interfaces**

At this point we are using;

1. Python standard libraries and JSON to analyse the data sets.
2. Excel to organise it.
3. Tableau to visualise, helping to see and understand the data.
4. Doxygen for documentation.
5. Are looking into Power BI to create interactive reports.

# **Standards**

There is no known legal or GDPR concerns with the data sets as they are all for public use and do not appear to contain any confidential information.

# **Conclusion**

The purpose and goals of the project have been specified but may shift during the progression of the project. There is now a strategy in place on how to best approach progressing through the tasks. The scope and deliverables of the project have been well determined, clarifying what needs to be done to successfully complete the project. Lastly, there is no known legal or GDPR concerns with the data sets being used. In conclusion, all aspects of the project have been defined.

# **Reference/Research Documents**

[1] J, M. (2017). *Trending YouTube Video Statistics*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasnaek/youtube-new [Accessed 7 Oct. 2019].

[2] Google.com. (2000). *Google Scholar*. [online] Available at: https://scholar.google.com/.

[3] Google.com. (2019). *YouTube-8M: A Large and Diverse Labeled Video Dataset for Video Understanding Research*. [online] Available at: https://research.google.com/youtube8m/index.html [Accessed 7 Oct. 2019].

‌[4] Kemp, S. ed., (2019). *Social Platforms: Active User Accounts*. Available at: https://blog.hootsuite.com/simon-kemp-social-media/ [Accessed 1 Oct. 2019].

‌