

**Author: Hendrik A. Dreyer**

**Student ID: 13622464**

**Course: Master of Data Science**

**Faculty: School of Science, Engineering and IT**

# ****Subject: MA5851 – Data Science Master Class 1****

**Assessment: 4 – Strategic Insights Report**

**Due Date: 9 December 2019**

# Understanding the relationship between

# media sentiment and developers

# 

## **Executive summary**

Clearly articulate to the reader an overview of the insights reports.

## Introduction

Discuss the problem and the context of why it is important with respect to the dataset. If you are using the dataset from your Assessment 3 submission, explain why the discussed problem can be addressed through this dataset.

## Data collection and methods

Provide an explanation of the data sources used from your Assessment 3 submission and/or any extra or different data sources that you have found to try to address the problem and improve the overall insights, including links to your data and any other relevant documents. A brief overview of collecting, cleaning, modelling and methods used in the collection should be provided. Where possible, discuss: o Any known interventions on the dataset or text pre-processing applied before analysis o Spark transformation and action operation used and justification of using those o Use of external NLP resources such as WordNet and NLTK.

## Proposed high-level solution design

A simple schematic with clear explanation proposing the final high-level solution that would be deployable for a production data science pipeline. The solution must support data updates taking into consideration: o All scraping requirements o Processing for NLP matching o Data cleansing o Data enhancements and visualisation to deliver an information product o Use of Spark transformation and action (where required).

**The final information product is aimed at providing ongoing regular insights into the problem that is being addressed.**

## Analysis

Your analysis tables and graphs. This should be interspersed with commentary so that it can be read as a document. Add appropriate interpretations and discussions of your results and model selection/performance where appropriate. The codebase and necessary resources should be shared to the reader via a repository link with at least five (5) commits as evidence of code management.

## Discussion

Your conclusions from your project. Restate the original objectives and/or problems and contrast this against the obtained achievements. Discuss the limitations of the analysis, such as what you were able to show or what you couldn't show. Include suggestions for further work, such as including other data sources that might be useful for a future analysis and/or more things you could have done if given more time.

# In this task, you will also be assessed for your writing, in that the report must be:

## References