Finding Content Titles within A Sample of Streaming Responses

# Foreword

As per the recommendations of the document outlining this task, I have thoroughly explained my thinking at each step of this exercise. This will hopefully demonstrate my complete understanding of the task and subject matter. Typically, I would develop a solution for a problem such as this, in Visual Studio or Spyder, rather than Jupyter Notebook. Jupyter is great for analytical problems. But for a problem based around automating a task then in my opinion a typical IDE is better. However, Jupyter allows me to clearly annotate my thinking throughout the task and therefore I have opted to use it.

# Introduction

I have created three short Jupyter Notebooks that step through three solutions to this problem, each growing with complexity. I start with simple text matching, then progress to fuzzy text matching and finally, use a python NLP package for the last iteration. The purpose of this structure is to show you exactly how I tackle a data science problem.

The packages I have used for this project are:

* Pandas – For holding the data, loading from the csv files, and performing transformations on the data.
* Re – Pythons built-in regex package.
* Spacy – A commonly used python NLP package.
* Spaczz – A modified version of Spaccy that incorporates fuzzy matching.
* Fuzzywuzzy – A python package for performing simple fuzzy matching operations on text.
* NLTK – Another common python NLP package, my only use for it in this project is for stop words.

# Source Code

The Jupyter Notebooks for this project can be found at the git repository:

You will need to setup a python virtual environment or a conda environment and install the packages listed in the requirements.txt file. Please navigate through the notebooks according to the order of the leading number, they are named as 1\_InitialAttempt, 2\_FuzzyAttempt and 3\_SpacyNLPAttempt. All python code has been written with the Pep8 standards and best practices in mind.