Background

The data wizards at NWO.ai works on predicting trends spanning an array of categories and issues. The predictions feed off of massive data streams from the internet. At NWO.ai, we are always looking to add proprietary data streams that enrich our dataset and give us access to insights that were previously not accessible. Since the Cambridge Analytica debacle, it has become exponentially more difficult to access certain data APIs that were previously accessible. At NWO.ai, we do not wish to use the data with malicious intent. Moreover, we do not wish to store the raw data but extract its metadata to derive value for our customers - small to medium enterprises - in order to give them a fighting chance to identify emerging trends before the Fortune 500.

Challenge: Semantic Search Algorithm

Design and implement a semantic search algorithm that is able to score and rank a set of keywords (trends) by how strongly associated they are to a given query term. The algorithmic approach could borrow techniques from association rule mining to analyze the co-occurrence of terms within a corpora of tweets and reddit posts, and should take into consideration the uniqueness of the trend and the recency of the association. For example, the algorithm should be able to determine that the query ‘iPhone’ is more strongly associated to trends like ‘MagSafe’, ‘5G’, and ‘pacific blue' then it is to “Biden” or “perfume”.

Details:

- The expected input to the method should be a query term, and the output should be an ordered set of trends. The method should be implemented in Python (v3.7).
- You can explore the dataset via the GCP BigQuery WebIDE and you can connect to the database from python using the provided JSON key.
- The sample twitter and reddit datasets can found in the tables `nwo-sample.graph.tweets` and `nwo-sample.graph.reddit` respectively
- The final script should be made available to us on GitHub.