# Assignment 3

## Submitted by:

Abirami Dhayalan - axd210002
Harish Srinivasan - hxs200044

## Question 2:

The input data is loaded onto the Databricks cluster and then the output file. gets stored in the Databricks FileStore

**Results analysis:**

The OutDegree nodes are subreddits which usually reference other subreddits, the first value - Subredditdrama is about discussing posts in other subreddits so it
makes sense as this would have the most out going links

The InDegree nodes are subreddits which are the most popular on the site, such as askreddit, ask me anything (ama) and pictures (pics). These subreddits
contain the most number of followers so they have the most incoming links

The PageRank algorithms results also coincide with the InDegree nodes as they are the most popular subreddits on reddit platform. Askreddit, videos and ima
have around 25-30 million subscribers

The connected componenets show that component id 0 have 26492 nodes and remaining components are not connected to the first one and have around 5 nodes

The triangle count algorithm shows the top 5 vertices such as askreddit, subredditdrama which have the higest triangle count