**1. Development Timeline:**

**Task 1:**

**August 14 to 15: Formulation of regex pattern for the possible tags**

**Challenges:** We faced challenges in writing efficient regex code and failed multiple time to extract all the patterns but with study materials and references we were able to fetch all the patterns.

**August 17 to 22: Parsing data into CSV and JSON file**

**Challenges:** For these steps to be performed the previous regex extraction had to be accurate and with which we had failed but after multiple iteration and cross verification on row counts we were able to parse all the required data into JSON file.

**August 23 to 29: Validation**

**Challenges:** We had few issues with respect to JSON file format and had to perform validation on the go while performing other tasks.

**Task 2:**

**August 20th to 24th: Gmap_id mapping and text preprocessing**

**Challenges: None**

**August 23rd to August 27th: Vocab count and Count Vec**

**Challenges:** We found it a bit hard to understand the steps taken to perform countvec and vocab. We had to perform multiple iteration as we had made mistake by performing stemming for bigram and later we rectified it.

**August 28th to August 29th: Validation and testing**

**Task 3:**

**August 26th to August 27th: Initial data exploration and preprocessing.**

**Challenges: With huge data set we were unsure about on how to go on insights later when considered an exploration question we were able to consider the valid set of columns for the EDA**

**August 27th to August 29th : Data merging and preprocessing**

**August 29th to 30th: Data Insights and video recording .**

**Screnshots:**

https://colab.research.google.com/drive/10zmLc_hFcWCW7I3-7icCa1jz7Dc6vH3W#scrollTo=jgVbnAdbjCBt

https://colab.research.google.com/drive/1nkjgsiHDdITXCRkCYdYC6xUIVU3VIahm#scrollTo=F7J4VuI3dn5J

https://colab.research.google.com/drive/1bghOQ5NYKWxSBQdVG22qmqm98hHHSmS3#scrollTo=1mil3vmbcssa