Data Analysis and Machine learning on data from the EA Forum

This post describes the results from the following project:

1. Scrape all posts from the EA Forum
2. Explore overall trends in the data: posts with the greatest number of comments, authors with the greatest number of posts etc.
3. Build a wordcloud to visualize the most used words
4. Fine-tune GPT2 on the text corpus and generate text from the model
5. Scraping

The robots.txt file of the EA Forum, indicates that the forum disallows crawling/scraping data from /allPosts. To get around this, I used a link extractor add-on from Chrome () and extracted yearly links from the /allPosts page using the extractor. This required some manually clicking of “Load more” but it wasn’t too cumbersome. I extracted data from 01-01-2013 to 05-04-2020.

Next, I used Scrapy to extract the following fields from each link output by the link extractor: Date, author, title, content, and number of comments.

I cleaned the data and restricted further analysis on posts from the period 01-01-2013 to 04-15-2020, since more recent posts may not have many comments.

1. Exploratory Data Analysis