DSCI 510: Principles of Programming for Data Science
Final Report
Armand Patel: 9357-3932-46

1. The title of my project is "Exploring Sentiments: An In-depth Analysis of Discussions about Artificial Intelligence on Reddit."

I chose to do this project because I was curious about what the population of people who are Reddit users think about the recent rise of artificial intelligence in media, technology, and academia. The research question I am trying to answer with this project is, how do Reddit users express sentiments in discussions related to Artificial Intelligence, and what insights can be gained from analyzing the sentiment patterns? The data was collected dynamically from Reddit itself and an analysis was done to see what kinds of words were used in posts and comments about artificial intelligence on reddit. My original project idea was to scrape data from Zillow and analyze housing prices across all fifty neighborhoods in Manhattan, New York, but it proved to be more difficult to get access to a Zillow API. That project proposal can be found in the project folder; however, it is the not the project proposal for this project. I did not submit one for this project as I independently decided to start working on it.

2. The data I collected was collected using a Safari Web Driver from selenium, a library that allows you to dynamically scrape data from a static website like Reddit. The web driver allowed me to collect a large amount of data because it loaded the web page from the URL and then scrolled through the page ten times. The Reddit feed for posts only allowed you to scroll ten times, so this was appropriate for my project. After collected all the HTML script if the status code from the requests library was 200, I quit the driver and parsed the HTML content using the Beautiful Soup library and an HTML parser. In the following parts, I'll explain what each link represented from the Reddit website.
    a. When searching for posts on reddit, you have the option to filter your search by five parameters, "relevance", "Hot", "Top", "New", and "Most Comments." Given this, I had a set of five links that corresponded to each of these parameters after typing in "artificial intelligence" into the search bar on Reddit. I then passed all five of those links into my Safari Web Driver and collected all the HTML content from those links. To parse the content, collected the tag that corresponded to the title of the post and then joined all the words into a string before cleaning the text and splitting it into hundreds of tokens. I then created five dictionaries that contained two columns, one for the word, and one for the count. This resulted in five dictionaries for each of the five links corresponding to the post search. These five dictionaries were then exported to CSV files titled, dict1_newest_post_titles.csv, dict3_most_comment_post_titles_word_counts_dict_final.csv, dict6_top_post_titles_word_counts_dict_final.csv, dict7_hot_post_titles_word_counts_dict_final.csv, and dict8_hot_post_titles_past_year_word_counts_dict_final.csv. I then collected two more results from sorting from posts from the past year and from all time, sorted by "relevance" and "most comments". Those were titled

dict9_relevance_post_titles_past_year_word_counts_dict_final.csv and dict10_most_comments_post_titles_past_year_word_counts_dict_final.csv.

b. The next part was to sort the search by comments. When doing this, you only have three parameters to choose from, which were “newest”, “Top”, and “relevance.” The process was different because I was searching for the comments on the posts rather than the titles of the posts themselves. So, after implementing the driver, I had to create a for loop to iterate through all the comment content and collect all the words from all of the comments. The same process was done to tokenize the text and three dictionaries were created then exported to CSV files. They are titled, dict2_newest_comments.csv, dict4_relevance_comments_word_counts_dict_final.csv, and dict5_top_comment_word_counts_dict_final.csv.

c. Finally, I wanted to pull a data set that I could create more visualizations on. So, I searched for posts on “artificial intelligence” and sorted them from all time by “relevance”, “Top”, and “most comments.” Those were the only three parameters that I could search by when sorting posts from all time, however, it left me with a lot of data to work with. I implemented the driver and searched for the time the post was made (this included the day of the week, the hour and minute, and the time zone I was in), the amount of votes it received, the title of the subreddit it was posted to, and the title of the post. I organized that data into a dictionary and had three dictionaries as a result titled, final_top_date_time_vote_comment_dict, final_most_coments_date_time_vote_comment_dict, and final_relevance_date_time_vote_comment_dict. I then combined all those dictionaries into one using the extend method and exported that dictionary to a CSV using the Dict Writer method and titled it, post_data_from_all_time_top_relevance_most-comments.csv.

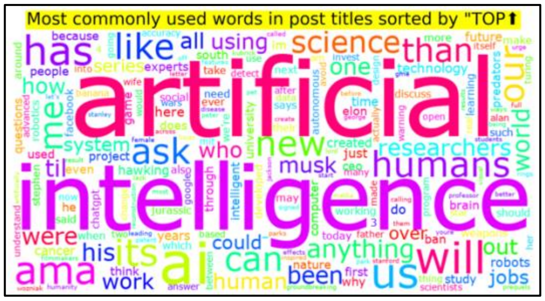
d. The data was cleaned by eliminating some stop words I had arbitrarily chosen.

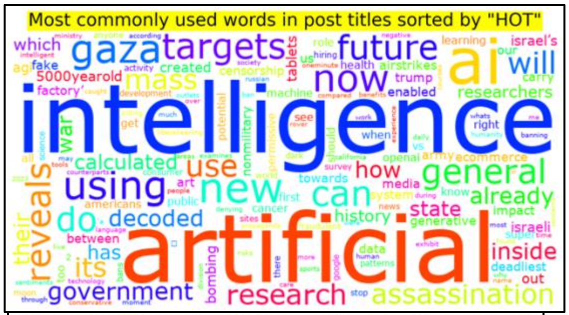
Problems encountered:

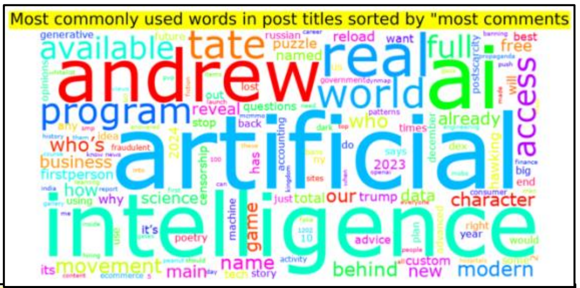
I encountered some problems when I initially tried to flatten all the text and separate each word so it could be counted. I decided to then use the Counter method from collections after tokenizing the text, which proved to be easier because it was more accurate at splitting all the text. I documented this change in the Jupyter Notebook file I placed in the results folder. Overall, this was a very sophisticated setup for web scraping. Selenium is rather difficult to use, and I had to scrape using multiple tags and iterate over multiple classes for each HTML file. Also, before collecting a count of all the words, all the words had to be lowercased. For example, “The” and “the” should be the same word and not counted separately.

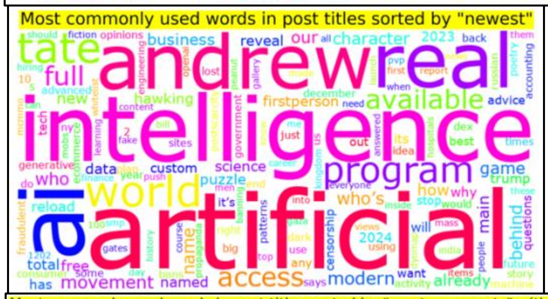
3. For my analysis, I used Pandas and NumPy to first find the total unique word count, the total word count, the most common words, and the average word length for each of my ten CSV files that I mentioned earlier. Those results were stored in a dictionary and then I converted the dictionary to a JSON file that is housed in the results folder. They’re titled, aggregate_results_for_comments.json and aggregate_results_for_posts.json. Besides the JSON files, I created a word cloud for each of my ten CSV files, and then the three separate ones I created for the posts sorted by all time. I also combined all my CSV files for posts and comments, and I created two bar graphs that show the top 10 most used words other than “artificial” and “intelligence.” Lastly, I created a heat map of the CSV

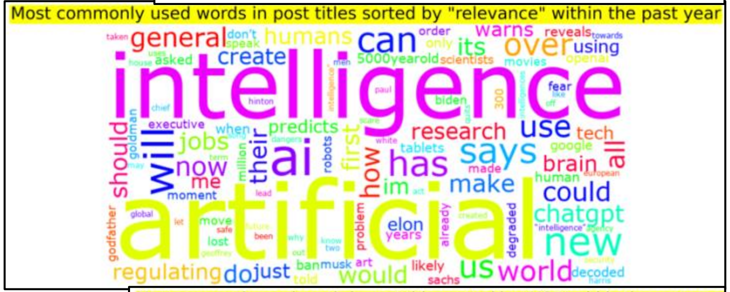
file that contain all the post data like time of day, and created a visual that showed the heaviest posting activity about artificial intelligence on Reddit by time of the day and day of the week. I will paste my word clouds following this paragraph. The following images depict the results of my Word Cloud analysis on the CSV files containing word counts for post titles.

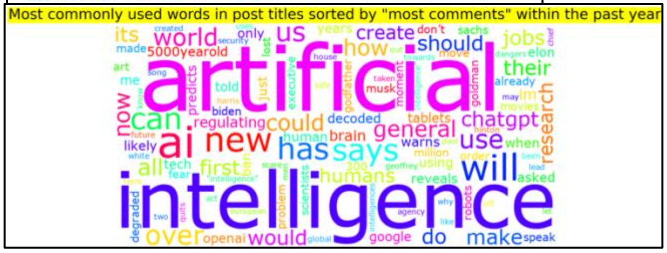

Most commonly used words in post titles sorted by "TOP"

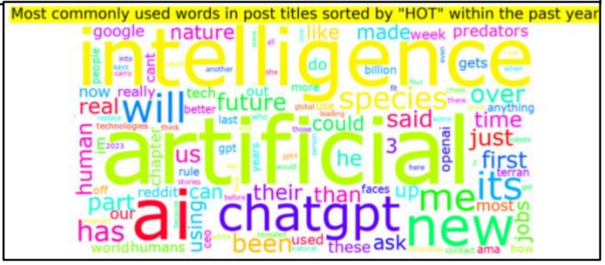

Most commonly used words in post titles sorted by "HOT"


Most commonly used words in post titles sorted by "most comments"

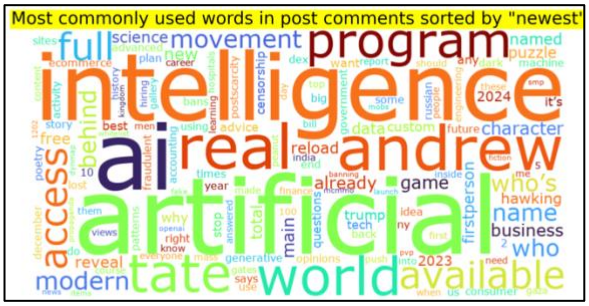

Most commonly used words in post titles sorted by "newest"

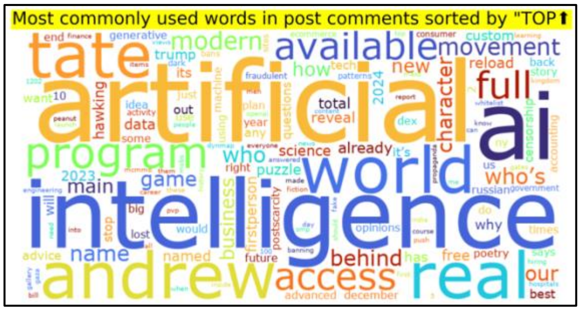

Most commonly used words in post titles sorted by "relevance" within the past year

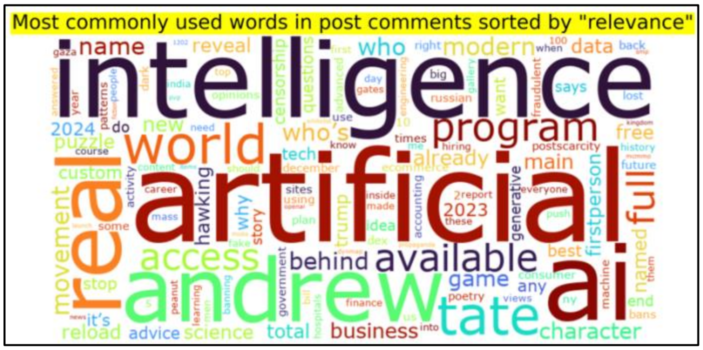

Most commonly used words in post titles sorted by "most comments" within the past year


Most commonly used words in post titles sorted by "HOT" within the past year

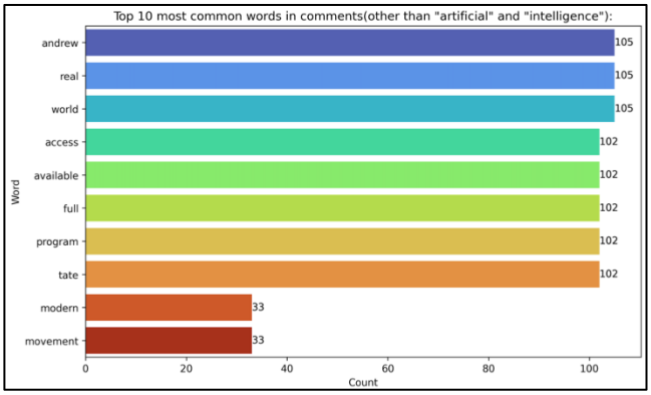
After visualizing all these word clouds, it appears that the most commonly used words in post titles about artificial intelligence are words like “Andrew” and “tate” along with a lot of words that center around “threaten” and “warn.” I also saw a lot of words that would suggest that there were a lot of job postings relating to AI like “research”, “san”, “Francisco”, “110k-150K”, “jobs”, “tensorflow”, and others. Andrew Tate is a former professional kickboxer and a businessman. He gained public attention outside of the sports world when he appeared on the UK reality show "Big Brother" in 2016. Since then, he has been involved in various ventures, including fitness training programs, motivational speaking, and online business. He also is an extremely controversial figure in the media and holds and speaks on his opinions openly. He has misogynistic, sexist, homophobic, and transphobic views and used harmful rhetoric in the media. My hypothesis as to why so many posts reference Andrew Tate is simply because he is such a hot topic right now, and artificial intelligence is also controversial. Given the nature of individuals who post on Reddit, it is likely that they'd be talking about both topics in their posts. There is also a lot of language surrounding the Israeli–Palestinian conflict with words like “gaza”, “targets”, “government”, “Israel”, and strangely enough, the word “assassination” appeared very frequently. The sentiment around the Israeli–Palestinian conflict could be because it is such a hot topic right now, and posts will mention artificial intelligence to get bumped up higher. Also, users may use discussions about artificial intelligence as a platform to express their political or social views. In such cases, they might bring up unrelated topics, like geopolitical conflicts, to convey broader perspectives or opinions. Depending on the context of the discussions, users from different cultural or regional backgrounds might interject discussions about artificial intelligence with references to geopolitical issues. This could be driven by personal experiences, concerns, or perspectives. Some online communities develop their own internal dynamics, jokes, or references that might not be immediately clear to outsiders. If there's a community-specific reason for the mentions, it could be related to the history or culture of that online space. This is something that is also very specific to the Reddit platform. The following images are the word cloud analysis for the comments.

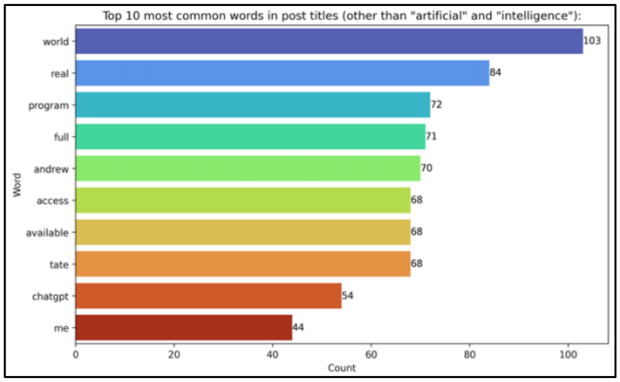




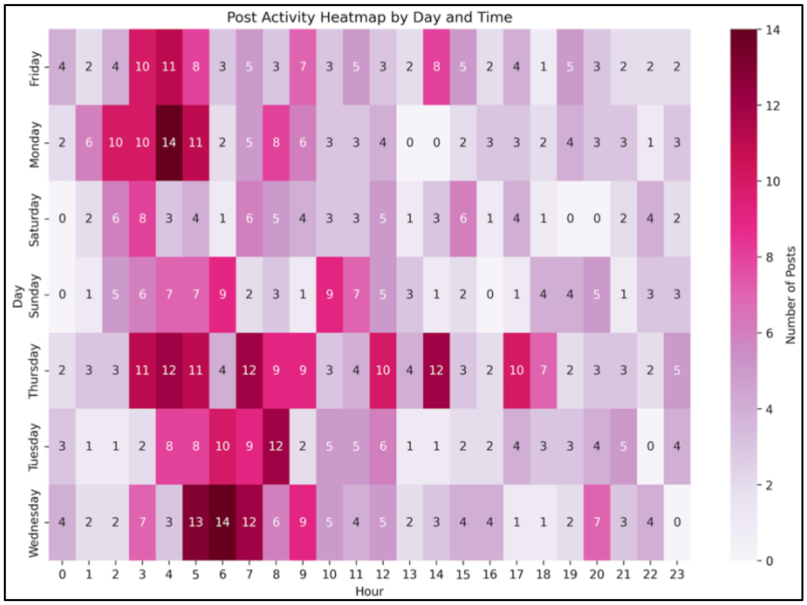


It appears that in discussion threads and comments on posts about AI, people are still mostly talking about Andrew Tate. There are also mentions of words like “sentient”, “future”, “puzzle”, “censorship”, “fraudulent”, and “fake.” This would suggest that in discussions about AI on Reddit, people are fearful of the ability of AI to become sentient or become controlling in the future. This also suggests a general fear around the unknown that surrounds AI. Since there is still so much to be discovered about its capabilities, people often use language that suggests that they're afraid. The following images show the top ten most used words in posts and comments. Before this analysis was done, stop words were removed. These are words like "a", "and", and "the." These words are not useful for my analysis and should be removed.

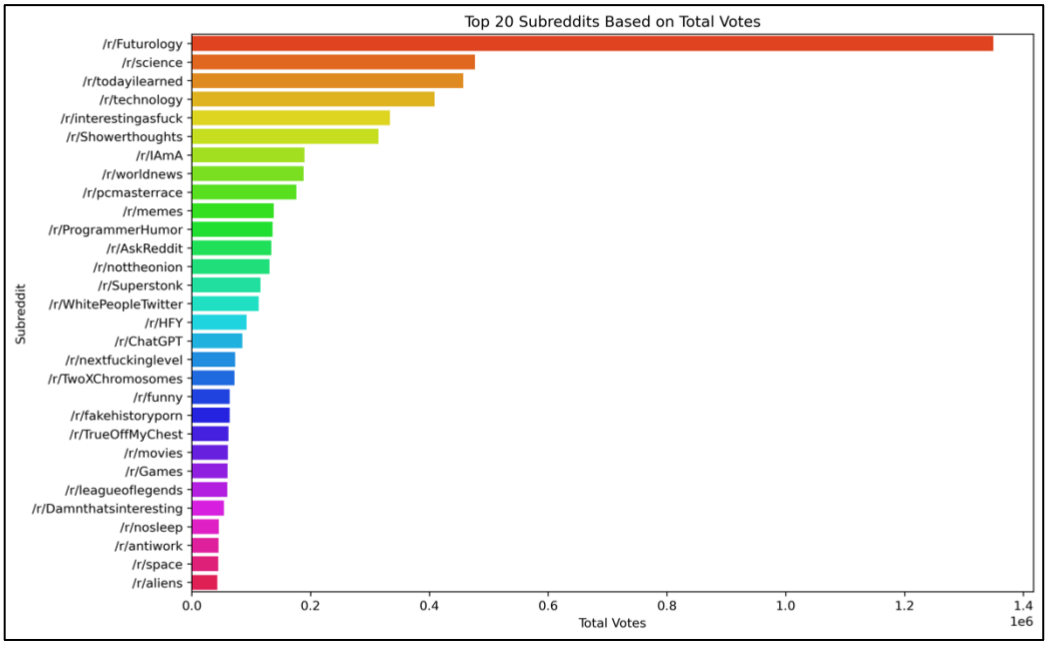




These figures also suggest that people think of what the government could do with AI, and they are fearful of its capabilities. There are also many references to Andrew Tate and ChatGPT. The next figure is a heat map analysis that shows the time of day and the day of the week that posts about artificial intelligence are made, and this was done using my data set that contained information about the subreddit, time, and title of the post.



This figure suggests that people make the most posts about AI on Mondays, Tuesdays, and Wednesdays, and between the third and seventh hours of the days. There is some variance on other days but the posts are generally centered around that area. These posts are also sorted by "all time." This may be due to the nature of the average Reddit user, who may be someone who stays up late at night and browses the internet. The following figure shows the names of the Subreddits that received the most traffic in reference to AI. This analysis was done by looking at the most amount of votes each post received by the subreddit it was posted in.



The names of these subreddits are things like “Futurology”, “Science”, “Technology”, “World News”, and “Aliens.” These make sense because AI falls under most of these categories. My general conclusion after my analysis is that there is a fear around the rise of artificial intelligence, and people are not sure of what it could bring. This analysis does center around Reddit users, however, for the sake of my analysis, we are using it as a representation of the general population. This means that more work could be done to shed light on AI and its capabilities. As well as the scope of its ability to replace people and their jobs, which is something that is a source of worry for many.

4. If given more time, I would’ve liked to conduct a sentiment analysis on these posts using a machine learning model. Sentiment analysis in machine learning involves training a model with a labeled dataset, where text samples are paired with sentiment labels like positive, negative, or neutral. After preprocessing the text by cleaning and converting it into numerical features, a suitable model is chosen—commonly Naive Bayes, SVM, or deep learning architectures. The model is trained to recognize patterns associating text with sentiments. Evaluation on a separate test dataset determines its performance. Fine-tuning and adjustments may be made, and once satisfied, the model is deployed to predict sentiments in new text inputs in a real-world setting. This could be done to further deduce sentiment about artificial intelligence.