Real Time Analysis for Reddit

Final Project Report

BADM 558 Big Data Infrastructure

Professor Zilong Liu

Group 12

Jahyeon Jee, Pi-Te Chen, Audrey Jung,

Seunggyun Shin, Hyun Ji Lee

Index

Abstract

This project endeavors to construct an extensive Big Data Pipeline geared towards scrutinizing social media interactions. Utilizing AWS infrastructure, our pipeline is designed to interface with the Job Search Hacks subreddit on Reddit. Our primary objective is to extract invaluable insights into innovative methodologies within the realm of job hunting. Through a systematic approach encompassing data collection, preprocessing, analysis, and visualization, we aim to uncover trends, patterns, and best practices prevalent in the job search process. Our results illuminate various strategies, techniques, and success stories shared by the community, providing actionable intelligence for both job seekers and recruiters. By leveraging advanced analytics on social media data, our project contributes to a deeper understanding of contemporary job search dynamics, ultimately empowering individuals and organizations in navigating the ever-evolving employment landscape.

Introduction

In the contemporary landscape of employment, the advent of social media platforms has transformed the dynamics of job searching. With millions of users engaging in online communities, platforms like Reddit serve as fertile ground for sharing insights, experiences, and strategies related to various aspects of professional life. Recognizing the significance of these virtual forums as repositories of valuable information, our project embarks on a journey to harness the power of Big Data analytics to dissect social media interactions, specifically focusing on the Job Search Hacks subreddit.

The problem at hand lies in the vast expanse of unstructured data scattered across online platforms, holding within it a wealth of untapped knowledge. Traditional methods of manual analysis are inadequate in efficiently sifting through this vast trove of information, necessitating the development of automated systems capable of extracting actionable insights. Thus, the overarching objective of our project is twofold: to develop a robust Big Data Pipeline and to leverage it to uncover innovative methodologies in the domain of job searching as shared within the Job Search Hacks subreddit community.

By constructing a comprehensive AWS-based workflow, our project seeks to streamline the process of data collection, preprocessing, analysis, and visualization. Through this systematic approach, we aim to address key research questions surrounding the efficacy of various job search strategies, the prevalence of emerging trends, and the success stories shared by individuals within the community. Ultimately, our endeavor is driven by the pursuit of empowering both job seekers and recruiters with actionable intelligence derived from the rich tapestry of social media interactions.

In essence, this project represents a convergence of technology and social science, aiming to unlock valuable insights hidden within the digital discourse of job seekers. By bridging the gap between Big Data analytics and online community engagement, we aspire to contribute meaningfully to the advancement of knowledge in the field of employment dynamics while offering practical solutions to real-world challenges faced by individuals navigating the job market.

Methodology

Pipeline

Figure 1 illustrates the analysis pipeline of the project. Data was collected from subreddit "*jobsearchhacks*" utilizing Reddit API on AWS EC2 instance. Collected data was stored in the AWS S3 bucket. Connecting the S3 bucket to Amazon SageMaker, the data was loaded on a Jupyter notebook for data manipulation and analysis.

Figure 1. Analysis Pipeline

Data Collection and Explanation

In order to collect data from the subreddit, getting access to the Reddit API was the initial step. By creating an account and registering for the API usage, access to the data was authorized. Launching Python on Linux using EC2 instance, dataframe was created by specifying information of each post - title, score, upvotes, downvotes, number of comments, created date, and content. Final dataframe was transferred into a csv file for storage.

Figure 2 shows the sample of the data. The data contains 9 columns which explains information about the data. For rows of Reddit posts with only images, values in the *Content* column were converted into NA values and rows without content text were eliminated as the

objective of this project is to get insights from Reddit posts through text analysis. The final data had 615 rows.

	title	score	upvotes	downvotes	num_comments	submission_id	url	created_utc	Content
9	I got put on PIP…is it too late to tell my bos...	88	88	0	90	1c6zq86	https://www.reddit.com/r/jobs/comments/1c6zq86...	1.713435e+09	I started working for a nonprofit a year ago. ...
568	Companies need to learn to build from the bott...	0	0	0	0	1c5b2q7	https://www.reddit.com/r/jobs/comments/1c5b2q7...	1.713256e+09	Look I get it, everybody is lazy and wants to ...
216	Scams used to believable	2	2	0	2	1c6vls2	https://i.redd.it/wqec23xec6vc1.jpeg	1.713418e+09	Copywriter/research assistant/ typer/ and edit...
320	Haven't heard back from a job after calling th...	1	1	0	1	1c6gg6j	https://www.reddit.com/r/jobs/comments/1c6gg6j...	1.713376e+09	I had an interview on Tuesday and a working in...
265	Is it time to move on?	3	3	0	2	1c6fs7i	https://www.reddit.com/r/jobs/comments/1c6fs7i...	1.713375e+09	I recently had a job interview on the 8th, see...

Figure 2. Data

Data Pre-processing and Analysis

Regarding the objective of the project, text mining was prioritized during the data manipulation process. After converting texts into lowercase, Regular Expression (re package) was initially used to remove punctuations and white spaces from texts. In addition, looking through each text, the stopword dictionary (nltk package) was optimized and used to remove superfluous stopwords from texts. Binary dataframe indicating whether each post contains certain words was created through tokenization, lemmatization, and vectorization, and added to the original dataframe. Lastly, the *created_utc* column (post created date) was converted into date format from timestamp for further analysis.

With the cleaned data, data visualizations, sentiment analysis, bivariate analysis, and trend analysis were conducted to extract meaningful insights from the data.

Results

Data analysis using Reddit API mainly focused on word frequency to identify common words and trends. We employed techniques such as word clouds, sentiment analysis, bivariate analysis based on sentiment, and trend analysis to explore the frequency and trends of words over time.

Word Cloud Analysis

We used the wordcloud package to generate a word cloud from our basic dataset, filtering general stopwords as seen in figure 3. Then, to refine the results and focus more on the topic of interest, we further filtered the words to obtain a refined version of the word cloud, extracting words that were more closely related to the specific topic. The resulting word cloud can be seen in figure 4.

Figure 3. Word cloud from basic dataset

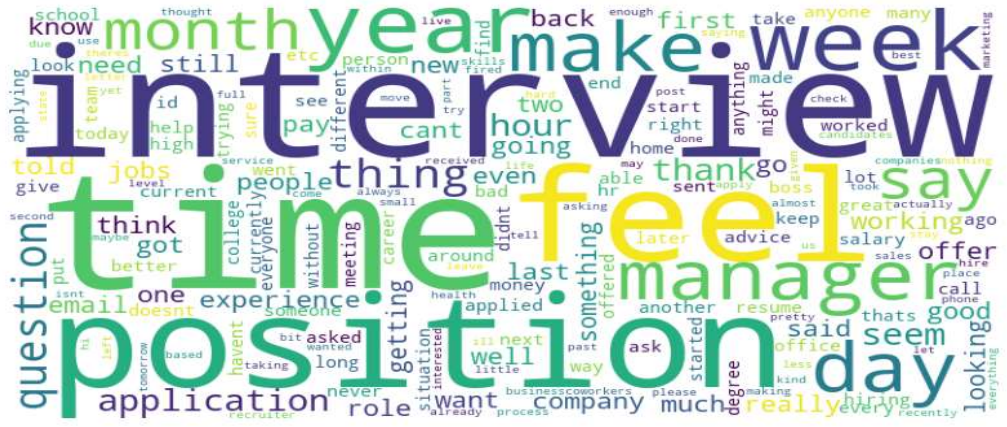


Figure 4. Refined word cloud

Sentiment Analysis

Firstly, we saw the distribution of the sentiment score or its labels and their frequencies. Then, we applied bivariate analysis to see the potential relationship between the sentiment score and post score on the Reddit platform.

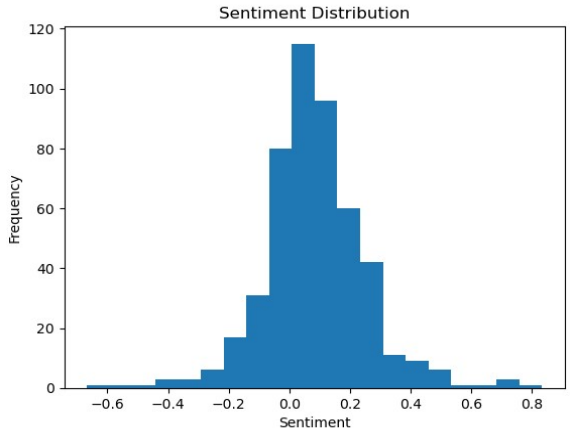


Figure 5. Sentiment score distribution

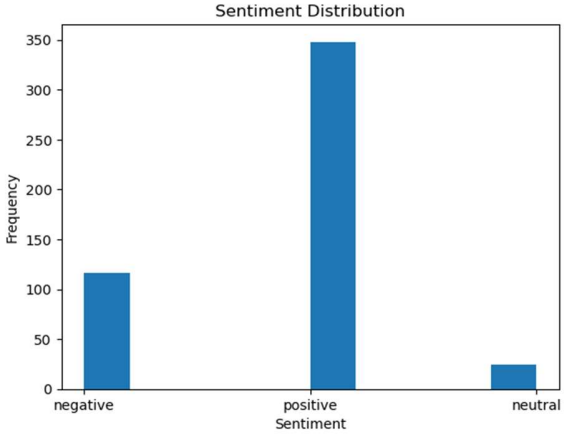


Figure 6. Sentiment distribution

Figure 5 shows the distribution of the score based on each score's frequency. The distribution closely resembles a normal distribution but is slightly skewed towards the positive side. Figure 6 shows the frequency of each sentiment label. We observed more positive posts with scores above 0 compared to negative posts, which was surprising given the common perception of online opinions being predominantly negative, especially considering the perceived challenges in the job market.

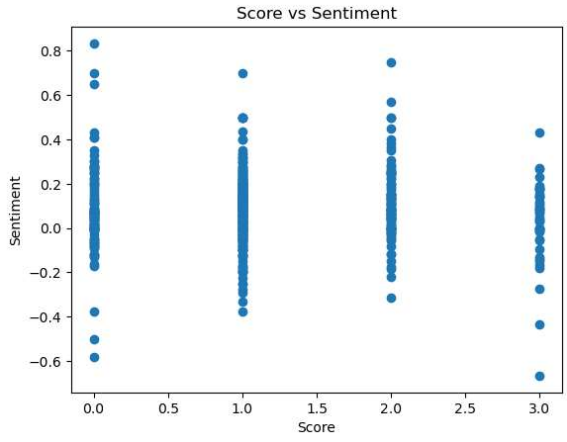


Figure 7. Score vs Sentiment

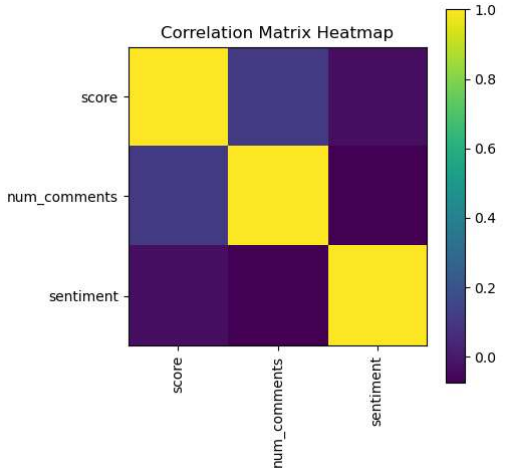


Figure 8. Correlation Matrix Heatmap

We further explored the sentiment analysis by conducting bivariate analysis using the scores. According to figure 7, we observed a weak pattern indicating a decrease in sentiment score as post score increases. In figure 8, we noted that there is a very weak relationship between the post's score and the number of comments.

Trend Analysis

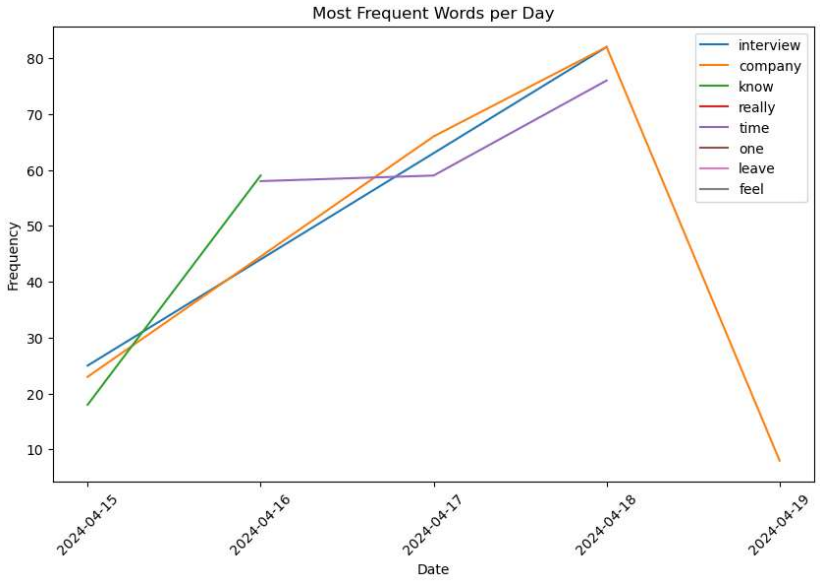


Figure 9. Word frequency by date

Figure 9 above illustrates the fluctuation in the number of mentions for specific words over several days. Some words show overlapping trends across multiple days, reflecting a similar pattern to the total number of posts.

Performance Metrics

We utilized the t2.micro instance for EC2, leveraging its free tier and operating on the Linux platform. Regarding S3, we managed a total storage size of 582.2 KB with 1 object, resulting in an average object size of 582.2 KB over a 2-week period. For SageMaker, we employed a free ml.t2.medium instance, running on Linux with Jupyter Lab 3, and utilized a 5 GB EBS volume. All services used were cost-effective and incurred no charges.

Discussion

The analysis of the Job Search Hacks subreddit via our Big Data pipeline has provided a multi-faceted view into the dynamics of job searching as shared by users on this platform. The positive sentiment predominance in the posts is particularly intriguing, suggesting that despite the challenges often associated with job searching, the overall tone of discussions remains optimistic. This could potentially influence the way job seekers approach their search, instilling a sense of hope and motivation.

The sentiment analysis indicated a surprising skew towards positive sentiments. Typically, online forums, especially those revolving around challenging topics like job searching, might be expected to showcase a mix of emotions, with significant expressions of frustration or anxiety. However, our findings suggest that the Job Search Hacks subreddit serves as a supportive community where users predominantly share successful strategies and encouraging stories. This could be indicative of a self-selecting nature of the community, where individuals who have had positive experiences are more likely to share, or it could reflect a culture that actively promotes positive reinforcement among peers.

Furthermore, the bivariate analysis of sentiment scores and post scores reveals a weak relationship, indicating that the engagement of a post (as reflected by its score) does not necessarily correlate with the sentiment it expresses. This could suggest that users value content for reasons beyond just emotional alignment, such as the usefulness of the advice or the novelty of the strategy discussed.

The trend analysis highlights the ebb and flow of certain keywords over time, such as "interview" and "company," which likely correlate with peak job searching periods or specific events in the broader economic landscape. Monitoring these fluctuations can provide recruiters and job seekers with insights into when their strategies might be most effective or when engagement might be highest.

From a technical perspective, the use of AWS tools like EC2, S3, and SageMaker has demonstrated a scalable approach to handling and analyzing large datasets in real-time. The performance metrics suggest that our infrastructure is not only capable but also cost-effective,

enabling continuous data analysis without significant financial investment. This aspect of our study underscores the accessibility of advanced analytics tools to broader audiences, potentially democratizing data-driven insights.

To summarize, our study not only sheds light on the nature of job searching discourse on a popular online platform but also illustrates the capability of modern technology to capture and interpret complex human interactions at scale. For future work, exploring deeper integrations of machine learning models to predict job market trends or the impact of economic shifts on job searching behaviors could offer further valuable contributions to both academic research and practical applications in human resources and recruitment strategies.

Conclusion

In conclusion, our project represents an effort to harness the power of Big Data analytics and social media interactions to gain valuable insights into the domain of job searching, particularly within the Job Search Hacks subreddit community. Through the construction of an extensive AWS-based Big Data Pipeline, we successfully collected, preprocessed, analyzed, and visualized data from the subreddit, unveiling trends, patterns, and best practices prevalent in the contemporary job search landscape.

Our methodology encompassed a systematic approach, starting from data collection utilizing the Reddit API on AWS EC2 instances, to preprocessing and analysis employing techniques such as text mining, sentiment analysis, bivariate analysis, and trend analysis. By prioritizing text analysis and employing advanced techniques, we were able to extract actionable intelligence from the vast trove of unstructured social media data.

The results of our analysis provided valuable insights into the sentiment distribution, word frequency, and trends within the Job Search Hacks subreddit. Through word cloud analysis, sentiment analysis, and trend analysis, we gained a deeper understanding of the prevailing sentiments, common words, and evolving trends within the community. Surprisingly, we observed a predominance of positive sentiments and identified weak relationships between sentiment scores, post scores, and the number of comments, challenging common perceptions regarding online discourse.

Furthermore, our project demonstrated cost-effective utilization of AWS services, ensuring efficient processing and analysis without incurring substantial charges. Leveraging the t2.micro instance for EC2, S3 for storage, and SageMaker for analysis, we managed to operate within the free tier limits while maintaining a high level of performance.

In essence, our project serves as a bridge between technology and social science, offering both theoretical insights and practical solutions to the challenges faced by individuals and organizations navigating the job market. By uncovering hidden patterns and trends within social media interactions, we contribute to a deeper understanding of contemporary employment dynamics, ultimately empowering stakeholders with actionable intelligence to make informed decisions in the ever-evolving landscape of job searching.

APPENDICES

Figure 1. Analysis Pipeline

	title	score	upvotes	downvotes	num_comments	submission_id	url	created_utc	Content
9	I got put on PIP...is it too late to tell my bos...	88	88	0	90	1c6zq86	https://www.reddit.com/r/jobs/comments/1c6zq86...	1.713435e+09	I started working for a nonprofit a year ago. ...
568	Companies need to learn to build from the bott...	0	0	0	0	1c5b2q7	https://www.reddit.com/r/jobs/comments/1c5b2q7...	1.713256e+09	Look I get it, everybody is lazy and wants to ...
216	Scams used to believable	2	2	0	2	1c6vls2	https://i.redd.it/wqec23xec6vc1.jpeg	1.713418e+09	Copywriter/research assistant/ typer/ and edit...
320	Haven't heard back from a job after calling th...	1	1	0	1	1c6gg6j	https://www.reddit.com/r/jobs/comments/1c6gg6j...	1.713376e+09	I had an interview on Tuesday and a working in...
265	Is it time to move on?	3	3	0	2	1c6fs7i	https://www.reddit.com/r/jobs/comments/1c6fs7i...	1.713375e+09	I recently had a job interview on the 8th, see...

Figure 2. Data

Figure 3. Word cloud from basic dataset

Figure 4. Refined word cloud

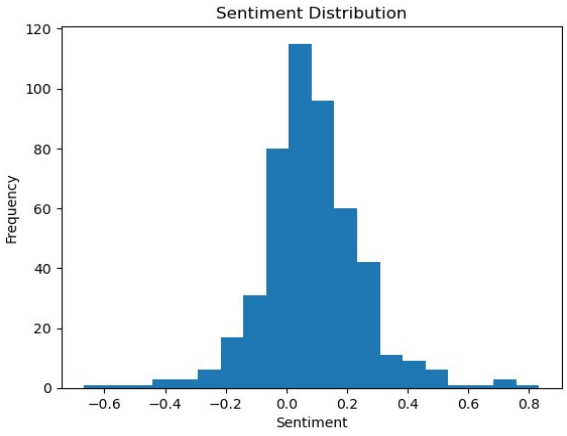


Figure 5. Sentiment score distribution

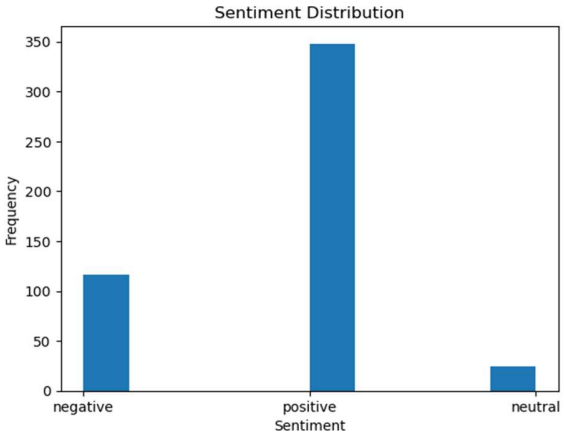
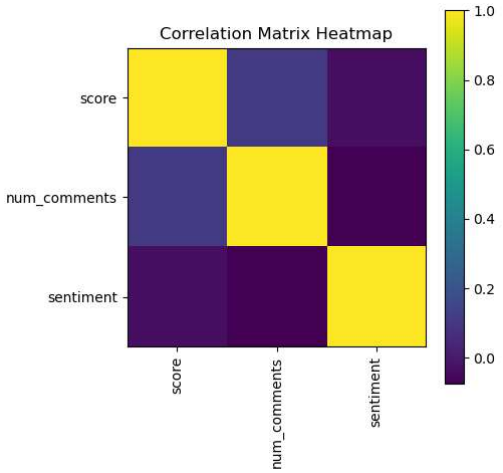


Figure 6. Sentiment distribution



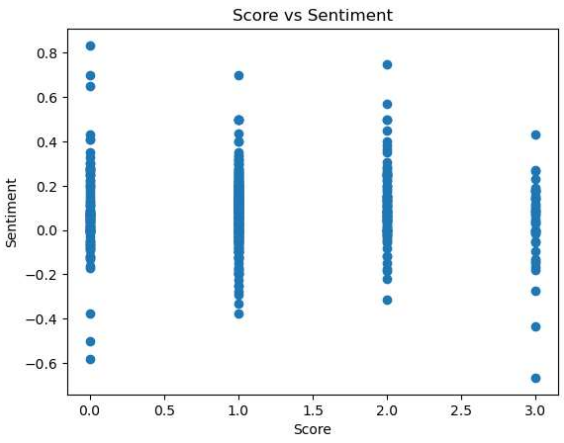


Figure 7. Score vs Sentiment

Figure 8. Correlation Matrix Heatmap

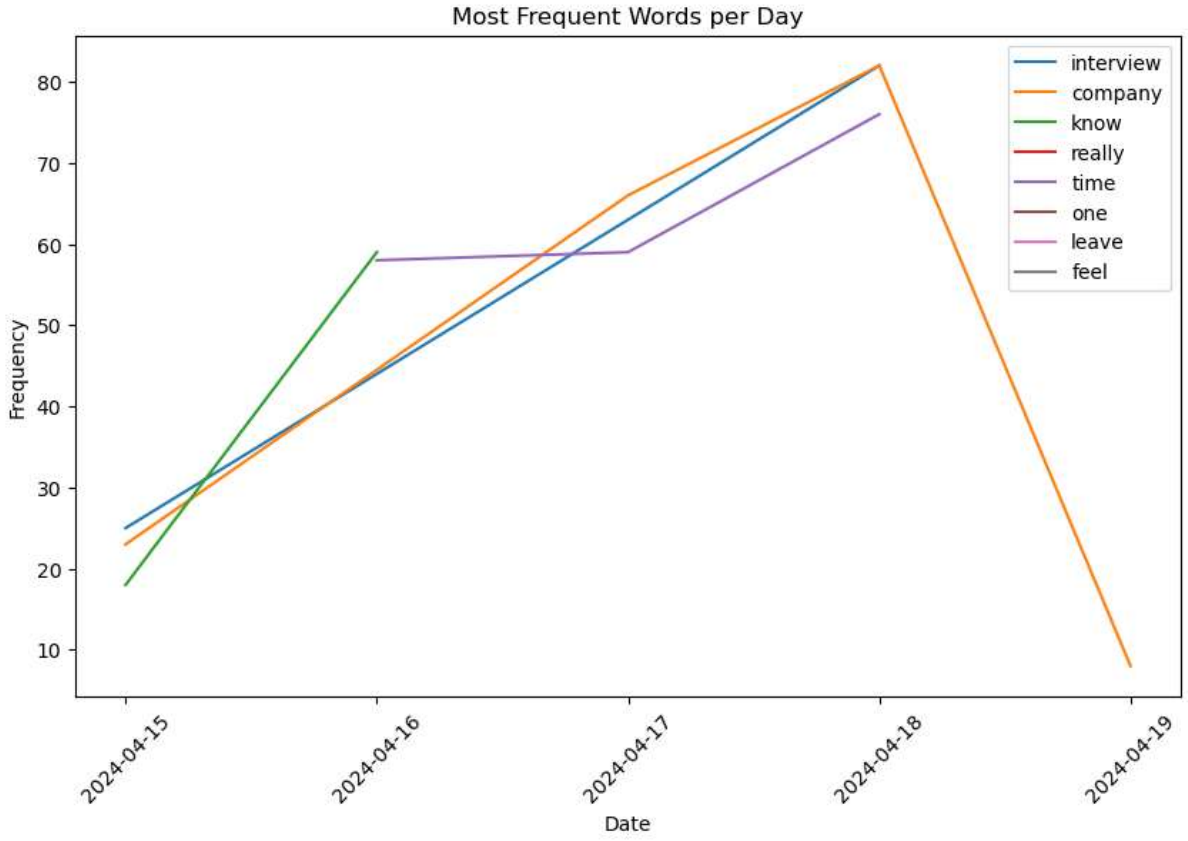


Figure 9. Word frequency by date