Reddit Data Science Sentiment Analysis

OMAR ABDEL-KEREM

M.S DATA SCIENCE, UNIVERSITY OF ST. THOMAS

Purpose

- This project is to analyze the content of the Data Science subreddit in order to gain insights into the skills, tools, and trends in the data science field that are relevant for job seekers.

- By examining the most discussed topics, the types of questions. and discussions that are prevalent in the subreddit, this project aims to help data science job seekers identify the areas of expertise that are in high demand, the types of roles and companies that are hiring, and the best practices for networking and learning from peers in the field.

Project Plan

- Write Python script to fetch data from Reddit API
- Clean and transform data using Python
- Store the data in Snowflake data warehouse
- Preform Sentiment analysis using Python libraries:
  - textBlob
  - Matplotlib
  - sklearn.feature_extraction.text → TfidfVectorizer
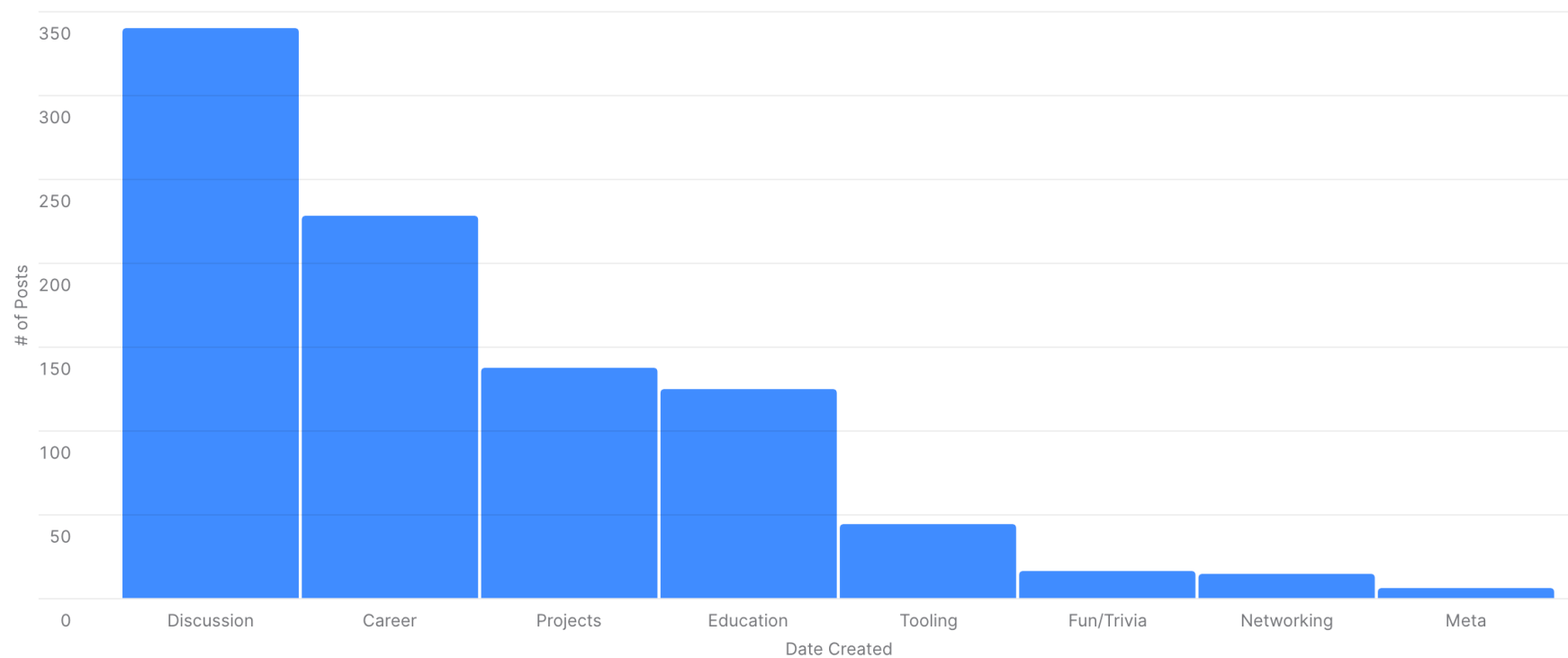 - sklearn.decomposition → NMF

Data Statistics

- 930 posts were extracted from the Reddit API.
  - More posts were not able to be extracted due to API restrictions
- Columns:
  - Post Creation Date
  - Post Flair
  - Post Title
  - Post Body Text
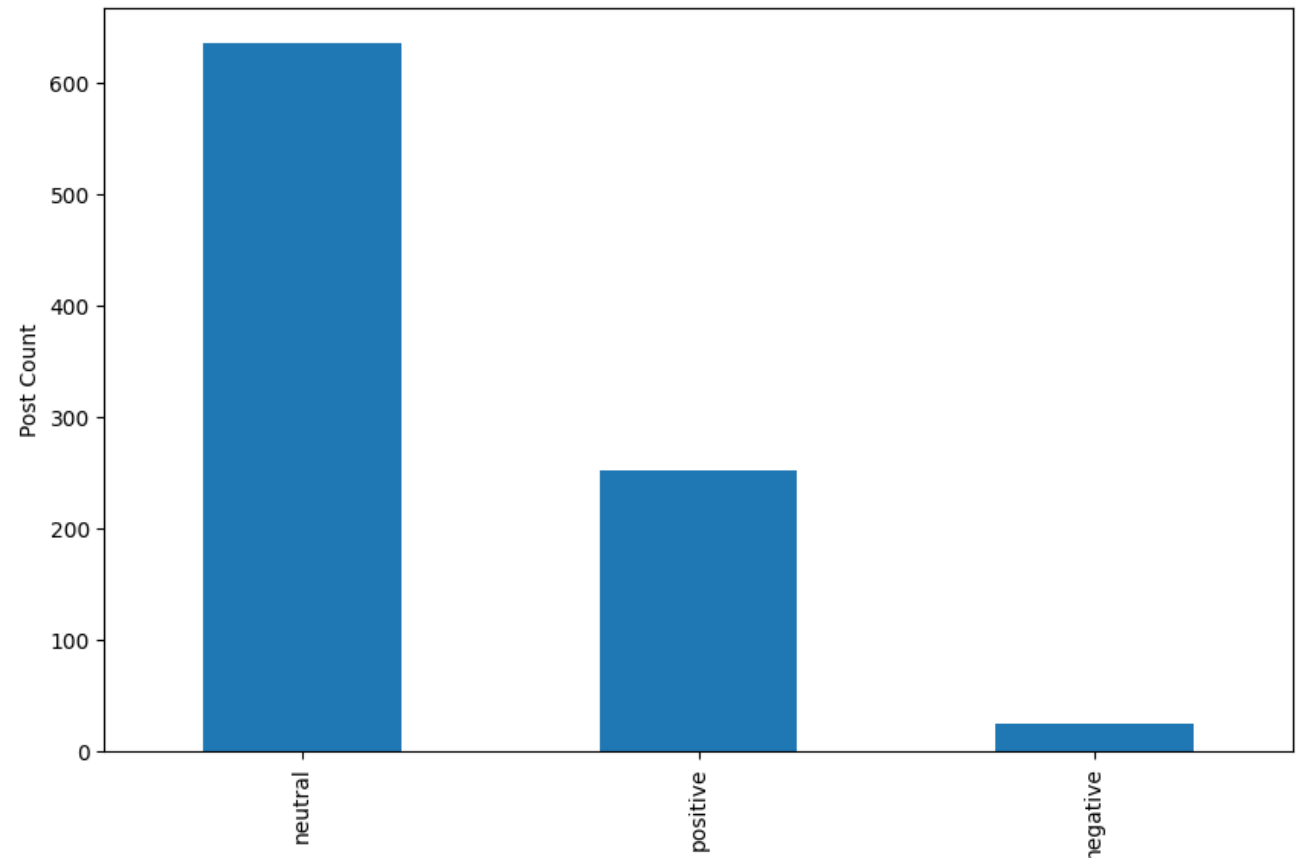 - Upvotes
  - Downvotes
  - Subreddit Name

Data Statistics

	Post Flair	Avg Sentiment	Min Sentiment	⋯ Max Sentiment	# of Posts	Sum Upvotes
1	Networking	0.1934200983	-0.0875	0.4444805195	14	34
2	Education	0.1552272866	-0.55	1	125	1,454
3	Discussion	0.1342230208	-0.6	1	340	7,922
4	Fun/Trivia	0.1313374256	-0.3	0.5	16	4,377
5	Meta	0.1282671958	-0.06111111111	0.2857142857	6	15
6	Projects	0.1276514097	-0.253125	0.6641414141	137	1,462
7	Career	0.1099259472	-0.5	0.6666666667	228	7,362
8	Tooling	0.1009945946	-0.35	0.4375	44	1,152

of Posts by Flair
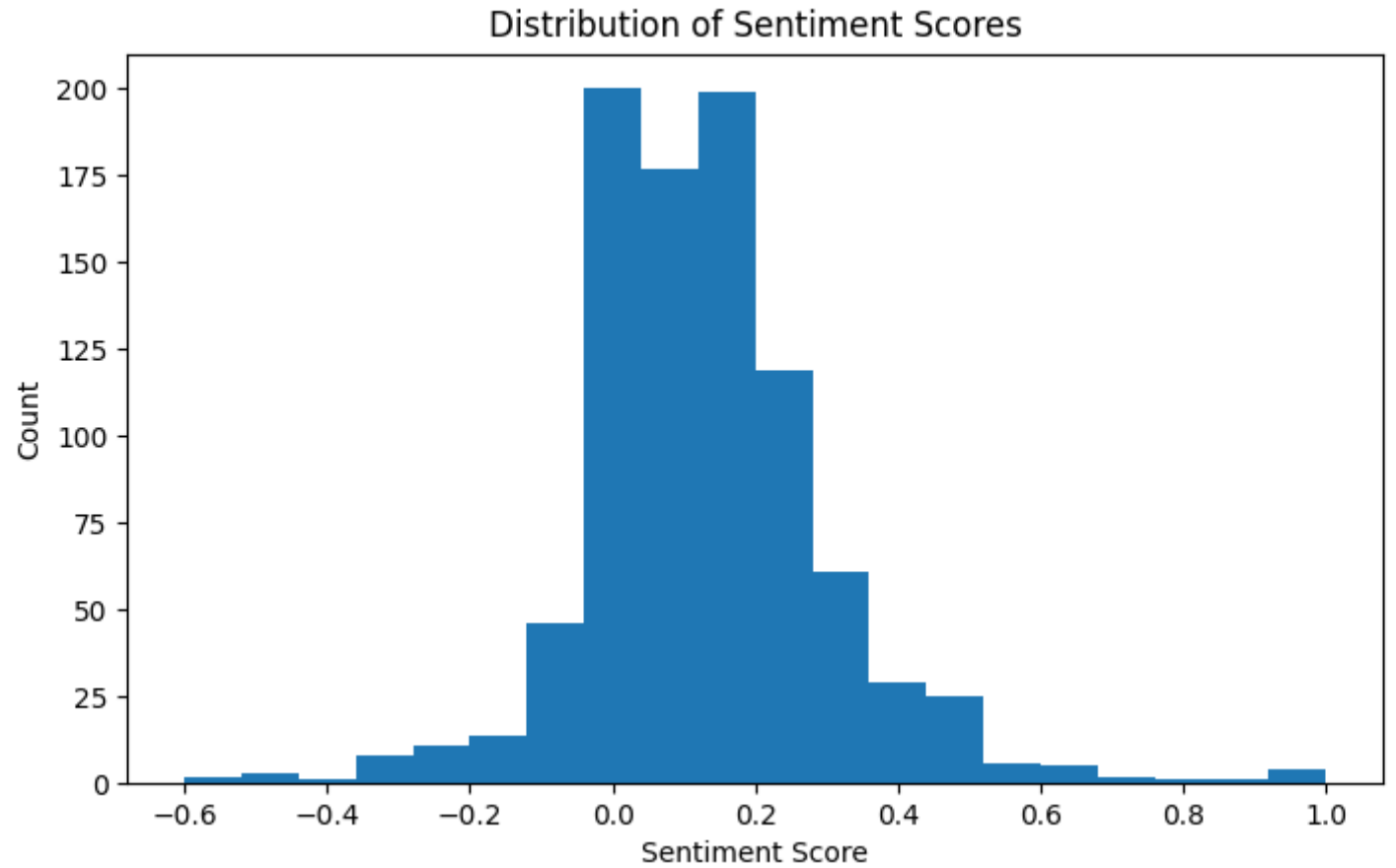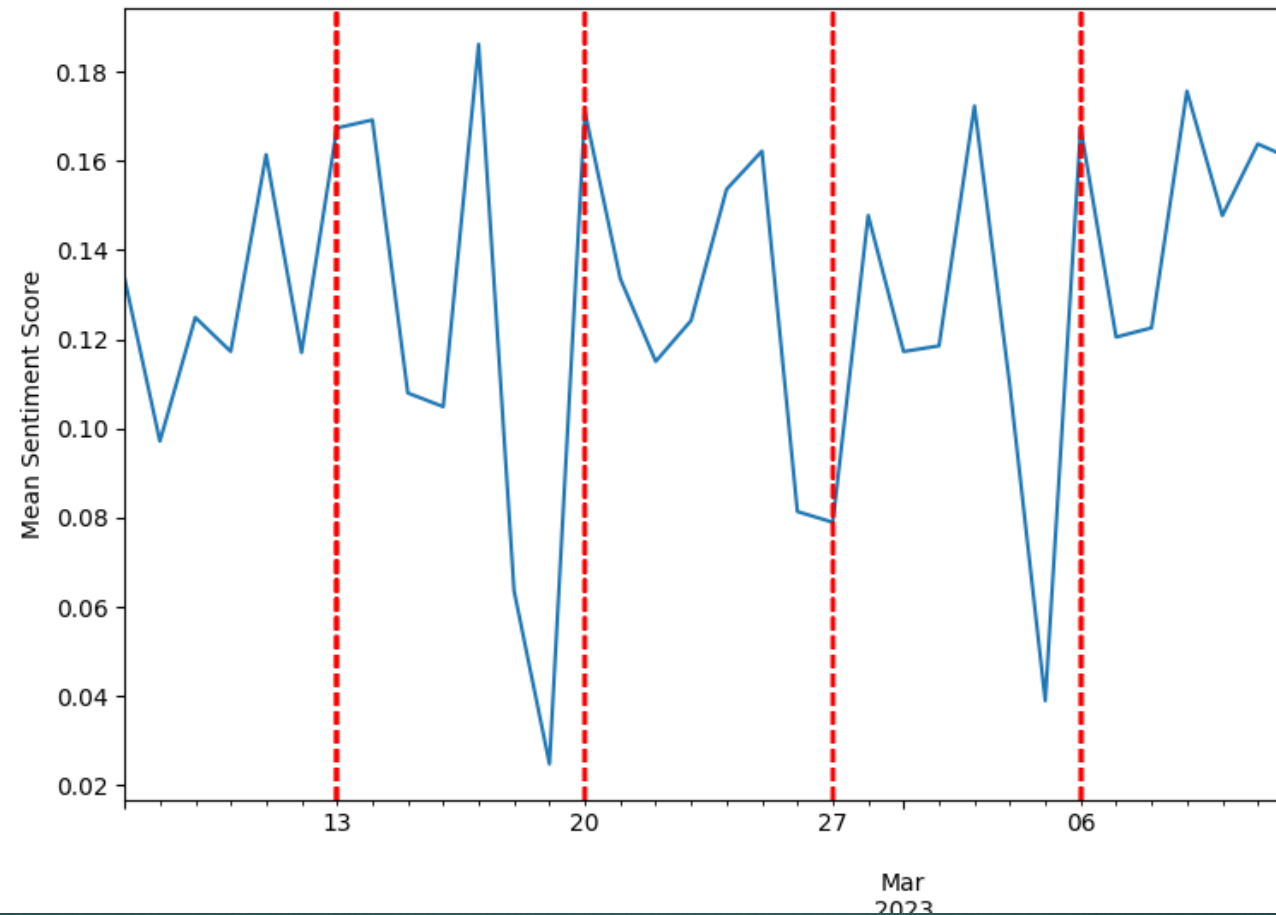
Post Count by Category

Distribution of Sentiment Scores

Sentiment by Week (Red shows Mondays)

Findings

- Keywords Per Flair:
- Career:
  - data science, job, career, field, program, confused, graduate, learning, python, learn, machine, course, good, math, courses, programming
- Education:
  - python, data, languages, scientist, learning, SQL, Kaggle, tasks, program, data, masters, science, related, years, research, analysis
- Discussion:
  - model, new, holdout, score, time, train, credit, science, data, degree, business, field, read, career, wondering, related, experience