



# REDDIT TOPIC TRENDS IN r/datascience

By Burak Basogul

# CONTENTS

- 1.INTRODUCTION
- 2.METHODOLOGY
- 3.ANALYSIS
- 4.CONCLUSIONS
- 5.FUTURE WORK
- 6.APPENDIX



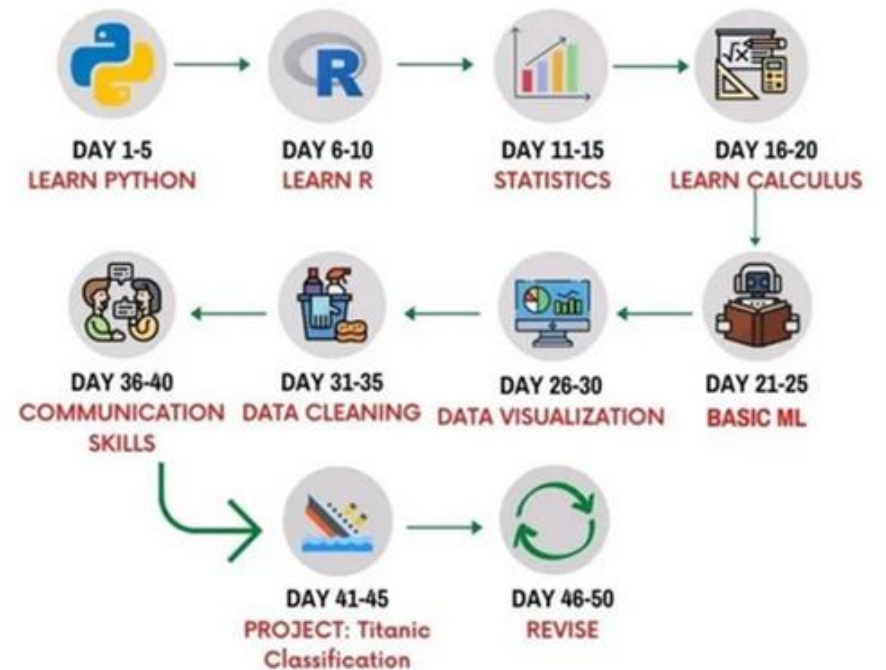
3063

Posted by u/sdfipvcmsvkoj 4 months ago

Meta Guys, we've been doing it wrong this whole time

[i.imgur.com/TAex5z...](https://i.imgur.com/TAex5z...)

## BECOME A DATA SCIENTIST IN 50 DAYS



380 Comments



Award



Share



Save



# INTRODUCTION

---

- Reddit shows top posts but no topic trends
- Top posts are usually memes or cross-posted fun/trivia content
- Reddit can benefit from deploying a Trending Topics under subreddits
- Propose to identify top 5 topics under subreddits

# METHODOLOGY -1

- DATA: Used PushshiftAPI to scrape post and comment data for six months
- 8,715 Submissions & 86,212 Comments
- After EDA, iterative clean-up and tokenization ended up with +10K feature terms.



# METHODOLOGY -2

Posted by u/Kent-Clark- 1 year ago 3 3 2 3 3  
3577 Fun/Trivia The pain and excitement

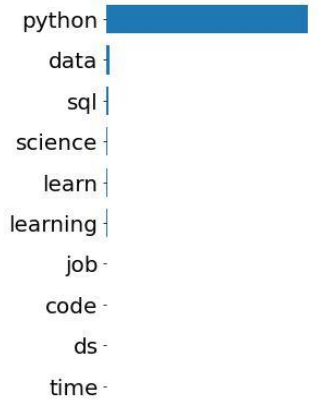


173 Comments Award Share Save ...

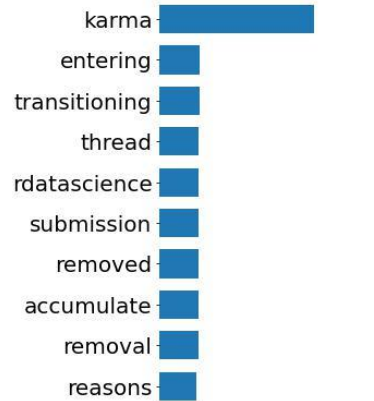
- Used LSA, NMF & LDA to generate 10 topic components each with 10 terms
- Visualize components based on weight / probability
- Interpret results and assign topic names

# ANALYSIS – LSA

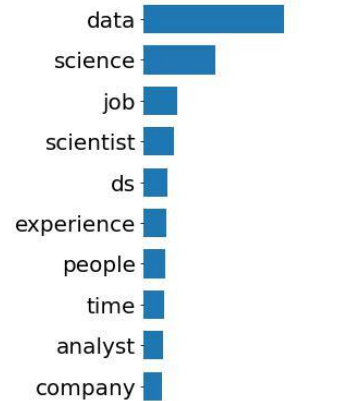
## DS Fundamentals



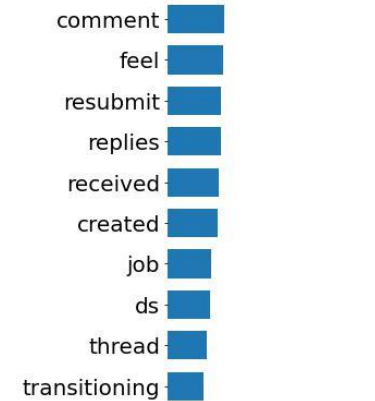
## Post Removal



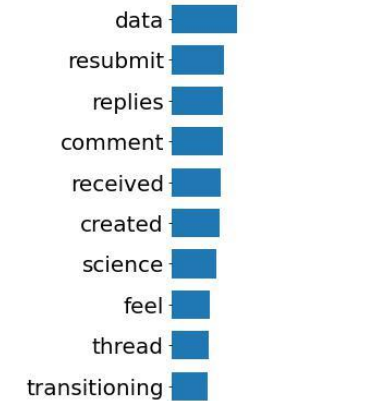
## Entry / Transitioning



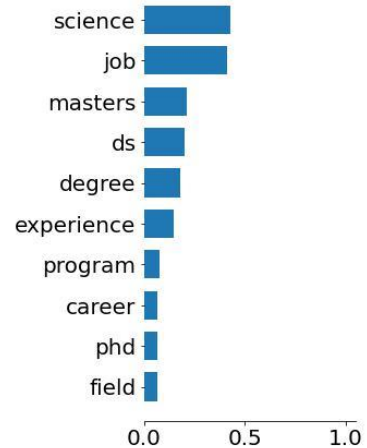
## Weekly Thread



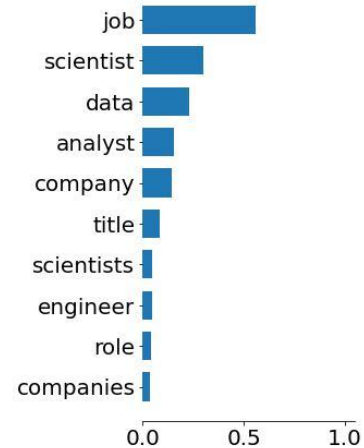
## Weekly Thread



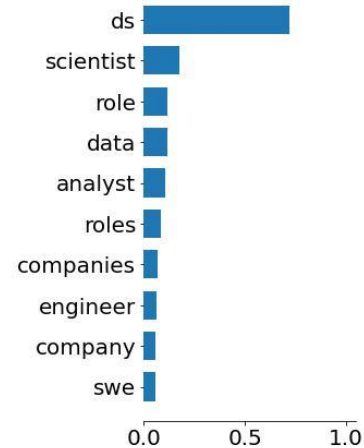
## Higher Education



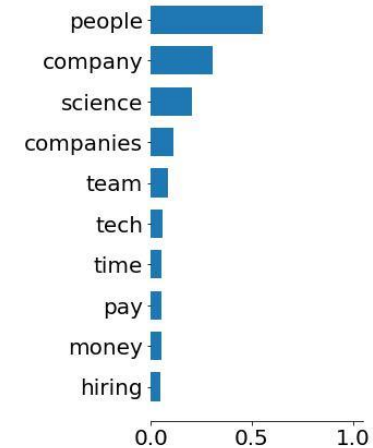
## Entry / Transitioning



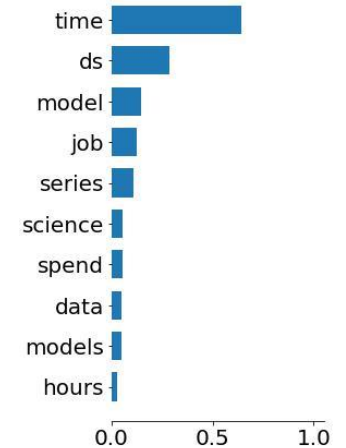
## Entry / Transitioning



## Career

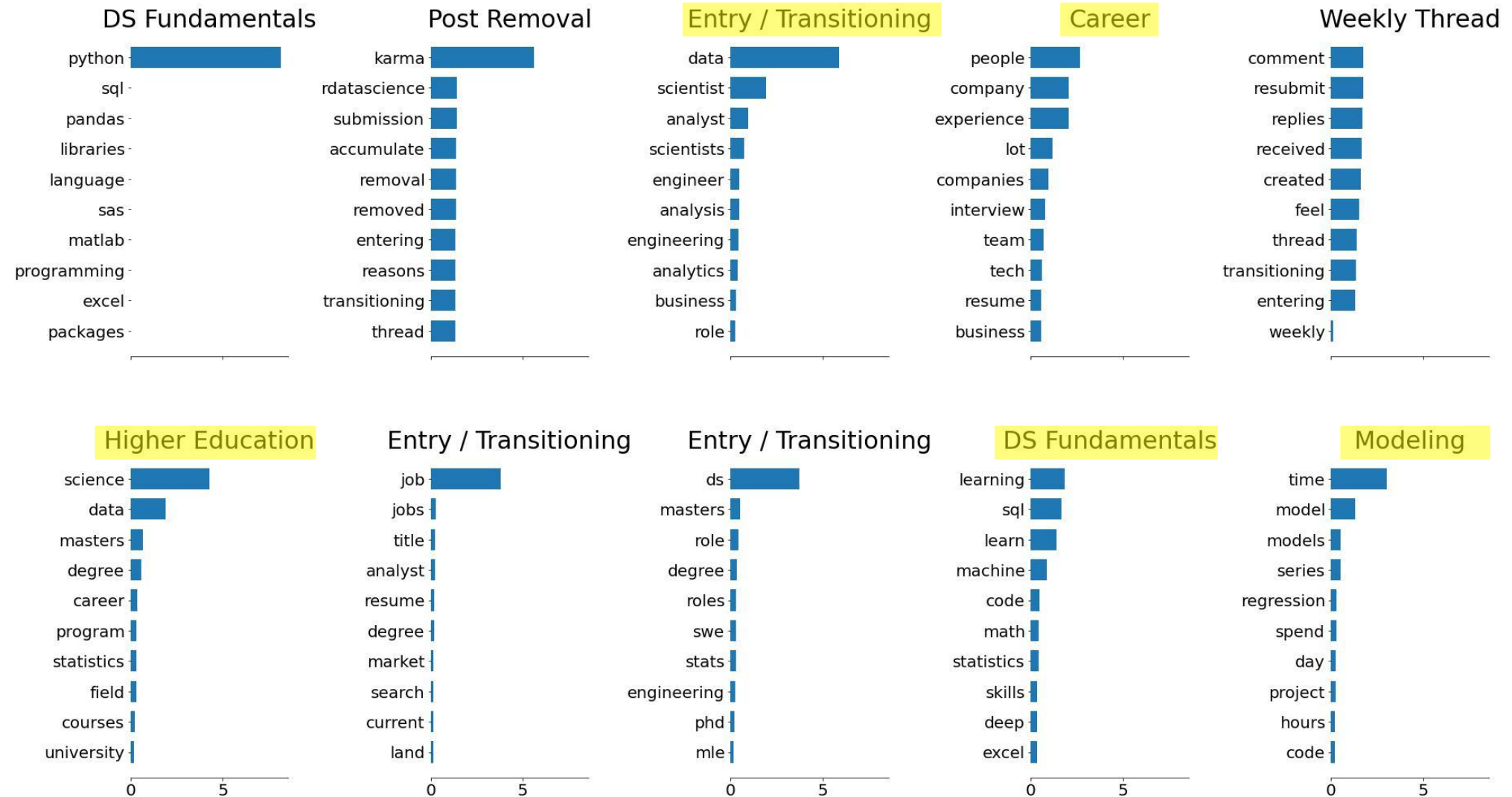


## Modeling

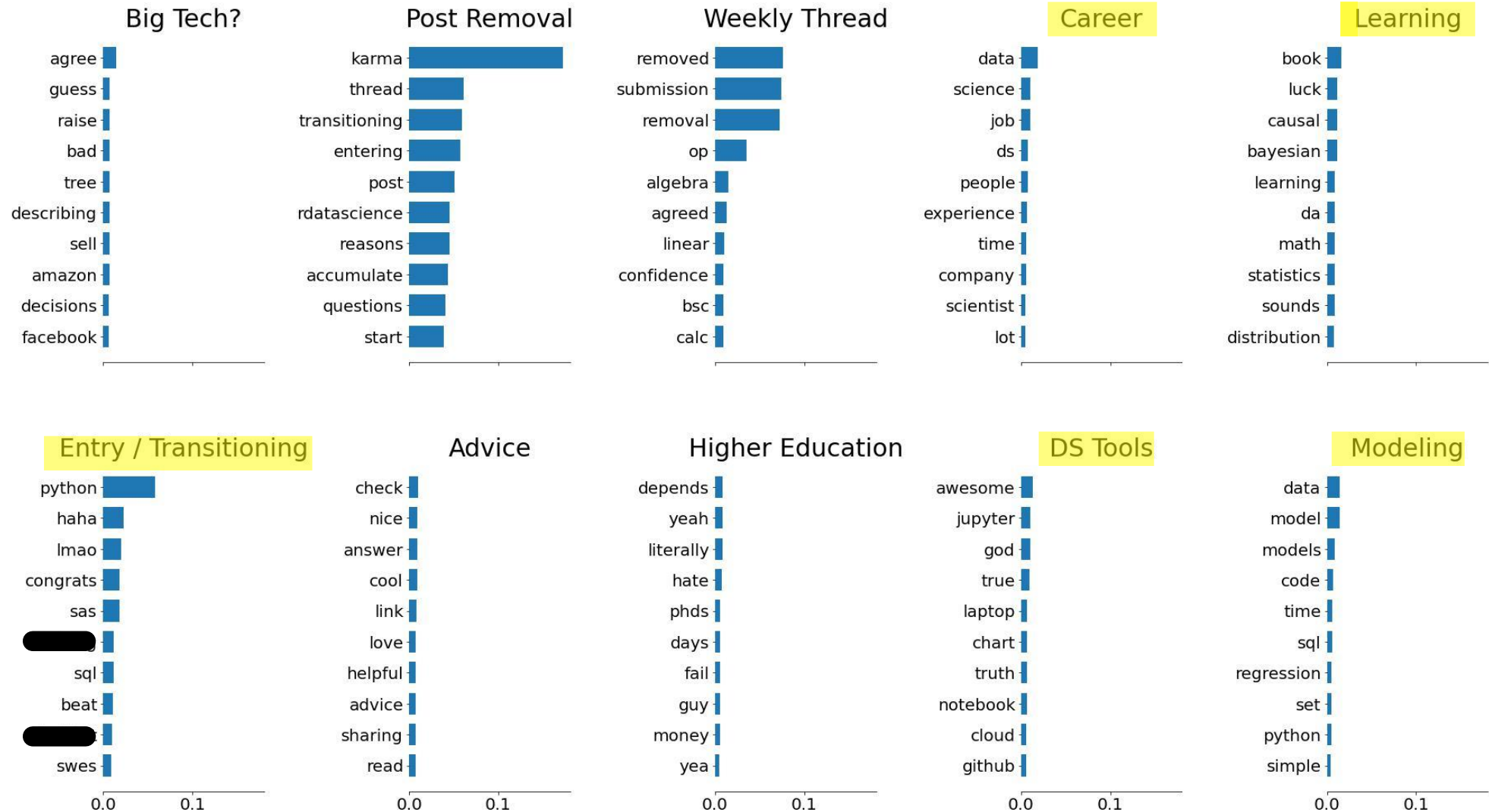




# ANALYSIS – NMF



# ANALYSIS – LDA

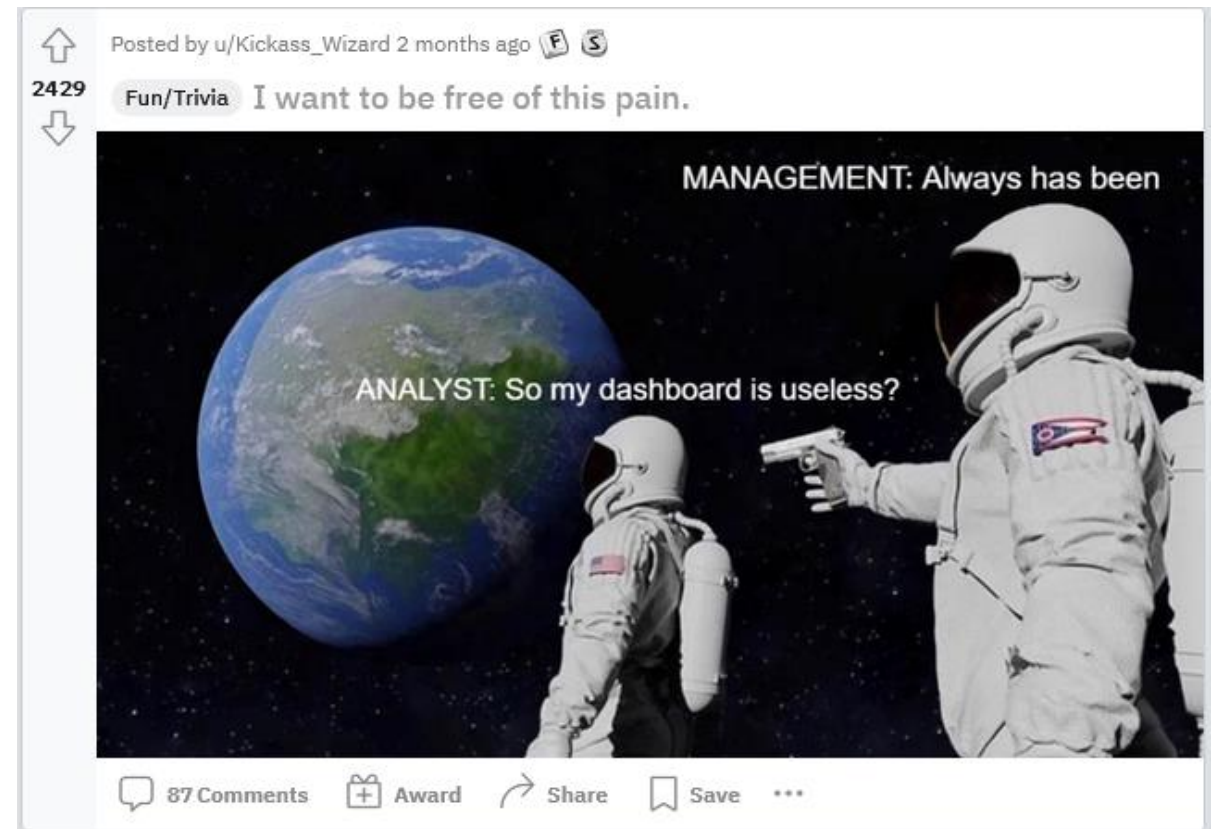




# CONCLUSIONS

---

- LSA is the model of choice due to its speed and accuracy for this use case
- Trending Topics will potentially draw new users
- Reddit can make use of Trending Topics along with Filter by Flair



# FUTURE IMPROVEMENTS

---

- Exclude LDA in the next iteration of this project
- Explore correlation between a post's / comment's score in reddit vs its term weight / probability in document term matrix.
- Conduct a topic trending analysis in r/MachineLearning for more insightful topics

# APPENDIX

---

- <https://www.reddit.com/r/datascience/top/?t=all>
- [https://scikit-learn.org/stable/auto\\_examples/applications/plot\\_topics\\_extraction\\_with\\_nmf\\_lda.html#sphx-glr-auto-examples-applications-plot-topics-extraction-with-nmf-lda-py](https://scikit-learn.org/stable/auto_examples/applications/plot_topics_extraction_with_nmf_lda.html#sphx-glr-auto-examples-applications-plot-topics-extraction-with-nmf-lda-py)
- <https://towardsdatascience.com/scraping-reddit-data-1c0af3040768>
- <https://medium.com/swlh/how-to-scrape-large-amounts-of-reddit-data-using-pushshift-1d33bde9286>
- <https://medium.com/mlearning-ai/extracting-philosophical-topics-from-reddit-posts-via-topic-modeling-affbaaa8a0b9>