Categorizing Self-Improvement Posts on Reddit

Metis Natural Language Processing and Unsupervised Learning Project

By: Maxwell Wood

Background Information

Business Opportunity

The self-improvement market is estimated to grow from $11.3 billion in 2021 to **$14 billion** by 2025*

Challenge

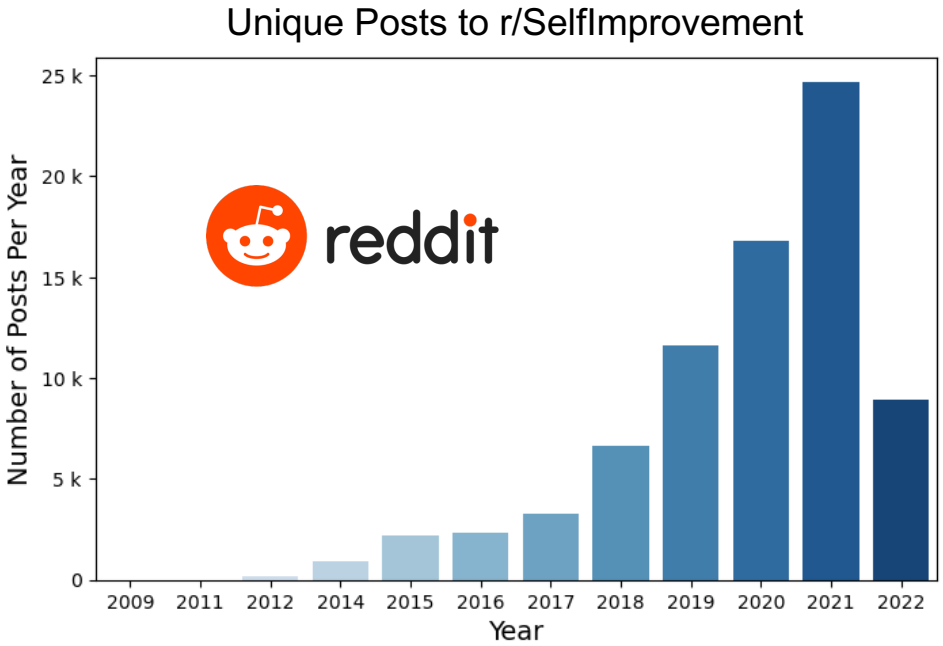
The dominant consumer group in this market is shifting from Baby Boomers to Millennials, who have different needs from their parents

* Numbers according to Marketresearch.com/

Data Science Solution

Data Source: r/selfimprovement

- Subreddit Community
- Founded in 2008
- Currently has 1.1 million active members
- Dedicated to “questions about how to improve any aspect of one’s life, from motivation and procrastination, to social skills and fitness and everything in between”.



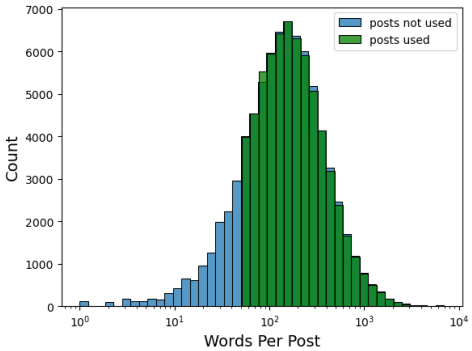
Data Pipeline

1 Collect Posts

- snscrape
- Over 750k posts and comments gathered

2 Data Cleaning

- Posts separated from comments
- Deleted, removed, and empty posts removed
- Posts with fewer than 50 words discarded



3 Text Preprocessing

- Tokenized
- Emojis removed
- Stop words removed
- Parts of speech tagged
- Lemmatized

4 Topic Modeling

Models Tried

- LSA - :(
- NMF - :(
- LDA - :D w/ 14 topics

LDA Topics

Some Make More Sense than Others

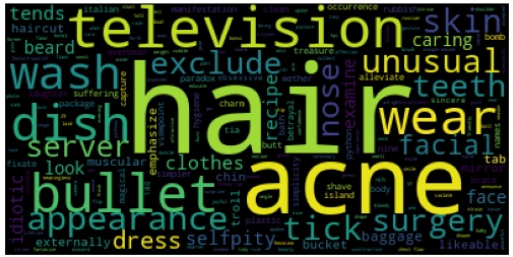
Topic #0

Topic #1

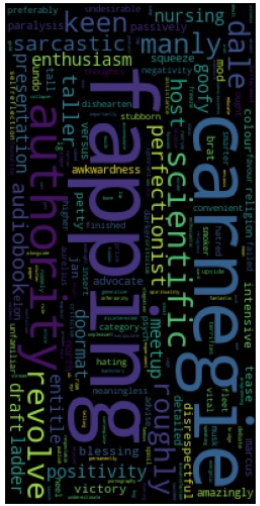
Topic #6

Topic #2

Topic #10



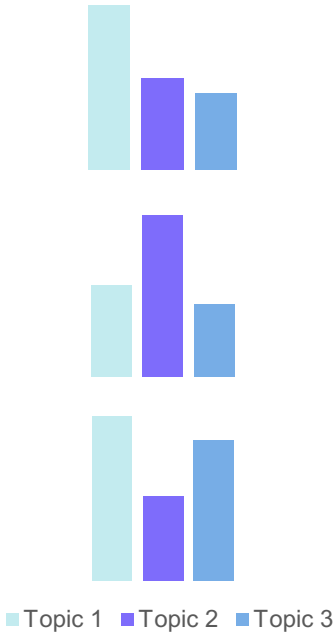
Topic #4



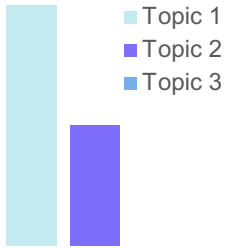
LDA Model Interpretation

Documents in a month:

Topic Distributions:

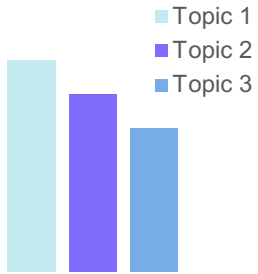


Winner Takes All



A topic for a document is only counted if it's the dominant topic for that post

Sum The Corpus

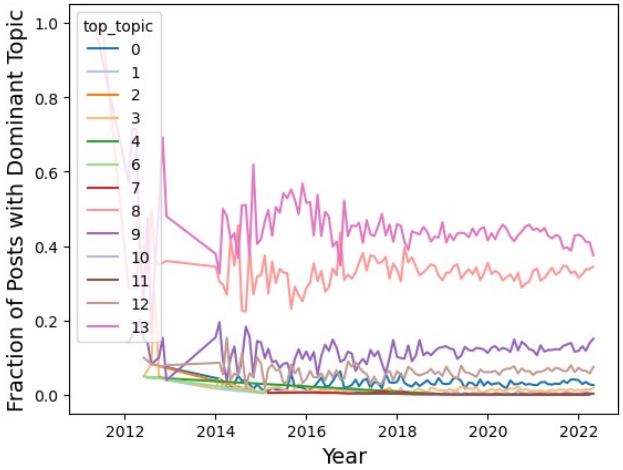


Sum the topic distributions for every document in a month to get a month's topic distribution

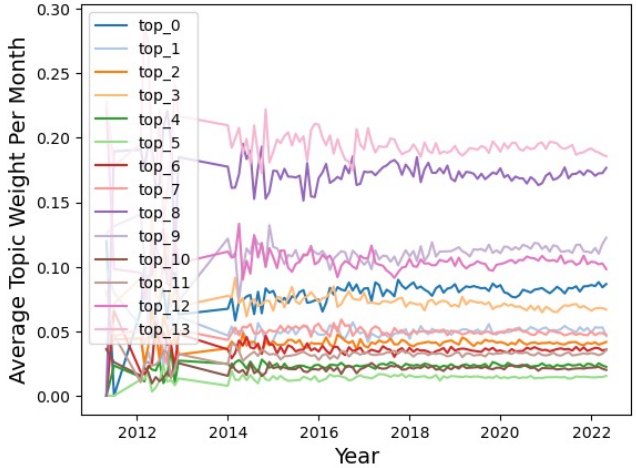
Do Topics Change over time?

Possibly, but this model doesn't capture it

Winner Takes All

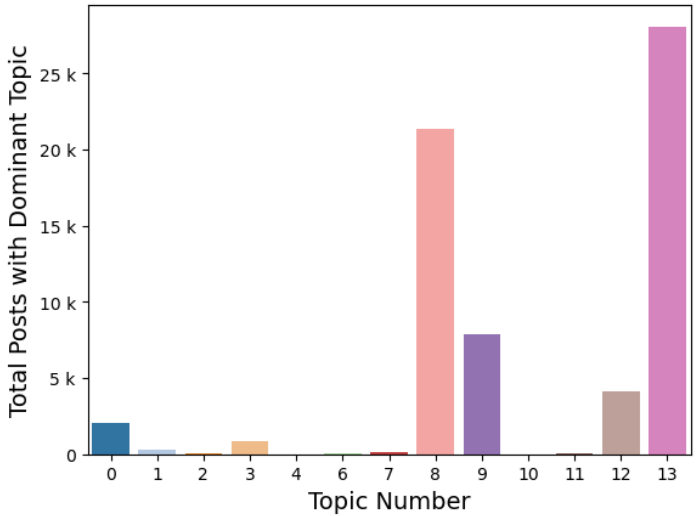


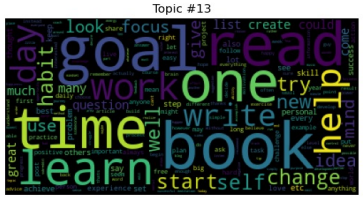
Sum The Corpus



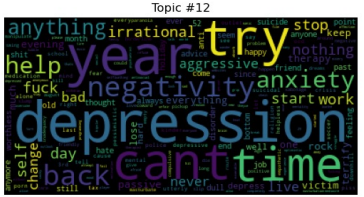
How are People Working on Themselves?

Winner Takes All

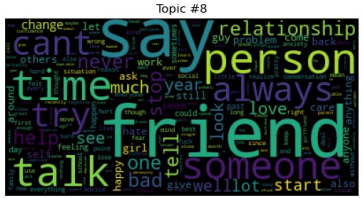




Learning/Reading

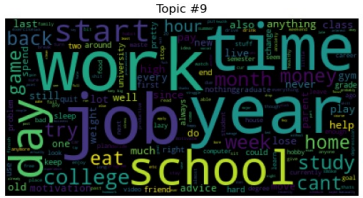


Mental Health

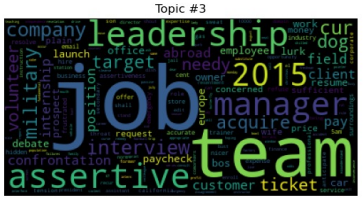


Relationships

Dating



Work/School



Career

Thank You!

Instructions for use (free users)

In order to use this template, you must credit **Slidesgo** by keeping the Thanks slide.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial purposes.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Delete the “Thanks” or “Credits” slide.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

https://slidesgo.com/faqs and https://slidesgo.com/slidesgo-school