NATURAL LANGUAGE PROCESSING

—

REDDIT

Brianna Sanzone

Natural Language Processing

Reddit

Goal

Develop a classification model that can accurately predict which subreddit a given post came from based on language recognition

Problem Statement

Can we predict which subreddit a given post came from based off the language used.

r/AskScience

"Ask a science question, get a science answer."

- Biology
- Psychology
- Physics
- Medicine

r/longevity

"Reasons to hope to see the age of 100 and beyond"

- Regenerative Medicine
- Biomedical rejuvenation
- Stem cell Therapy
- Gene Therapy

OBTAINING THE DATA

- Web Scraping from Reddit API

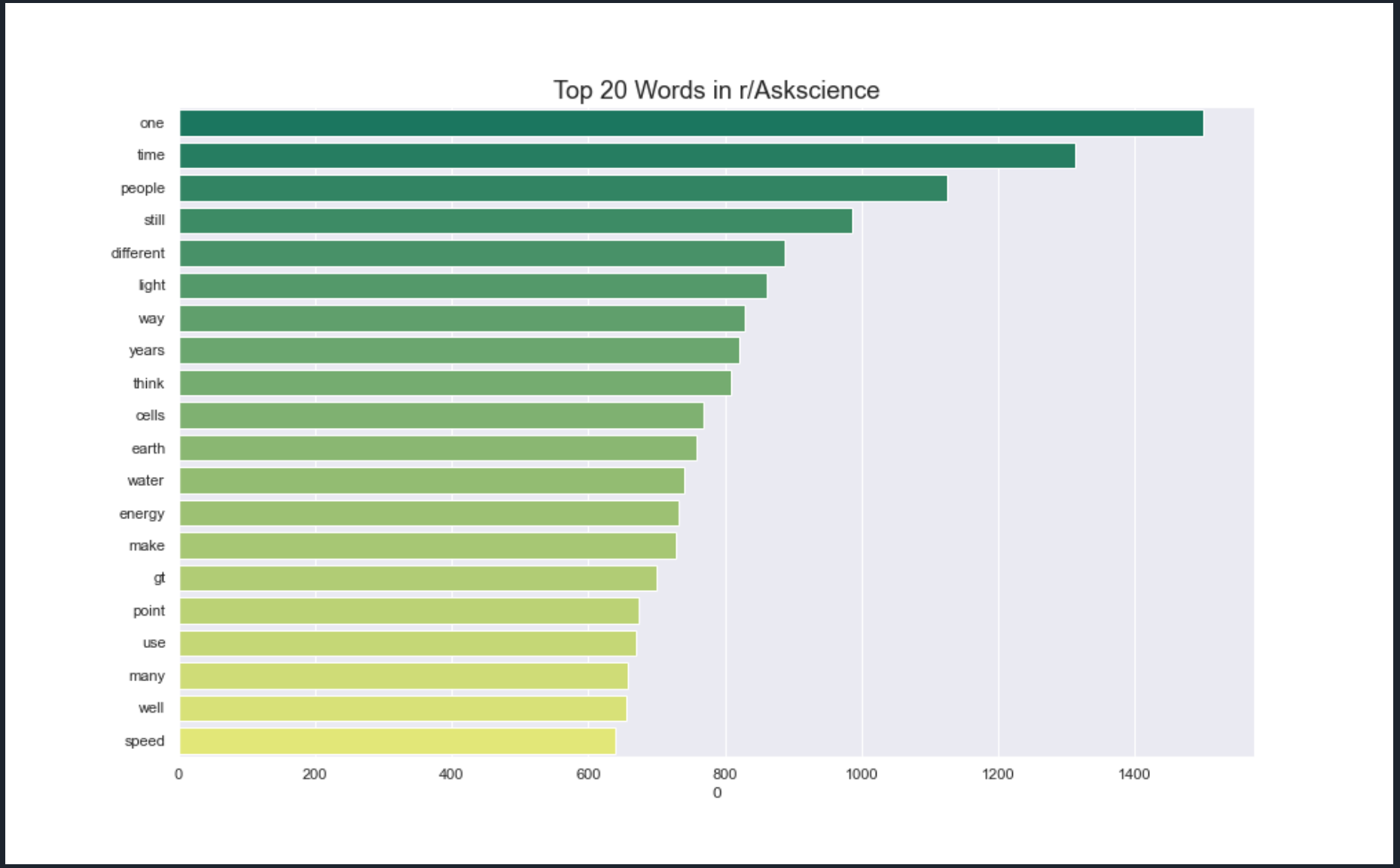
- 53,000 Total Posts

  ONLY 300 posts with usable text

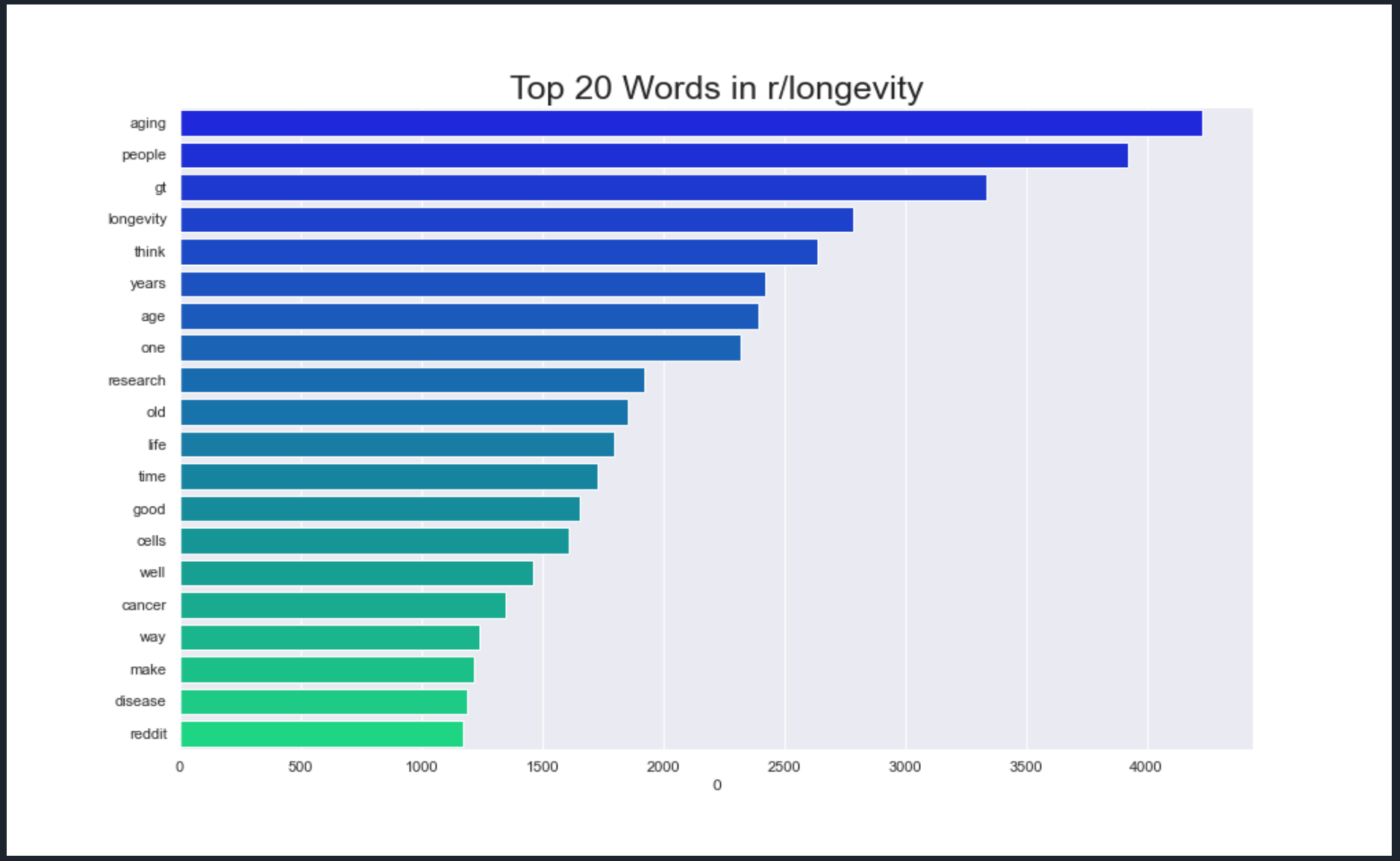
- 51,000 Total Comments

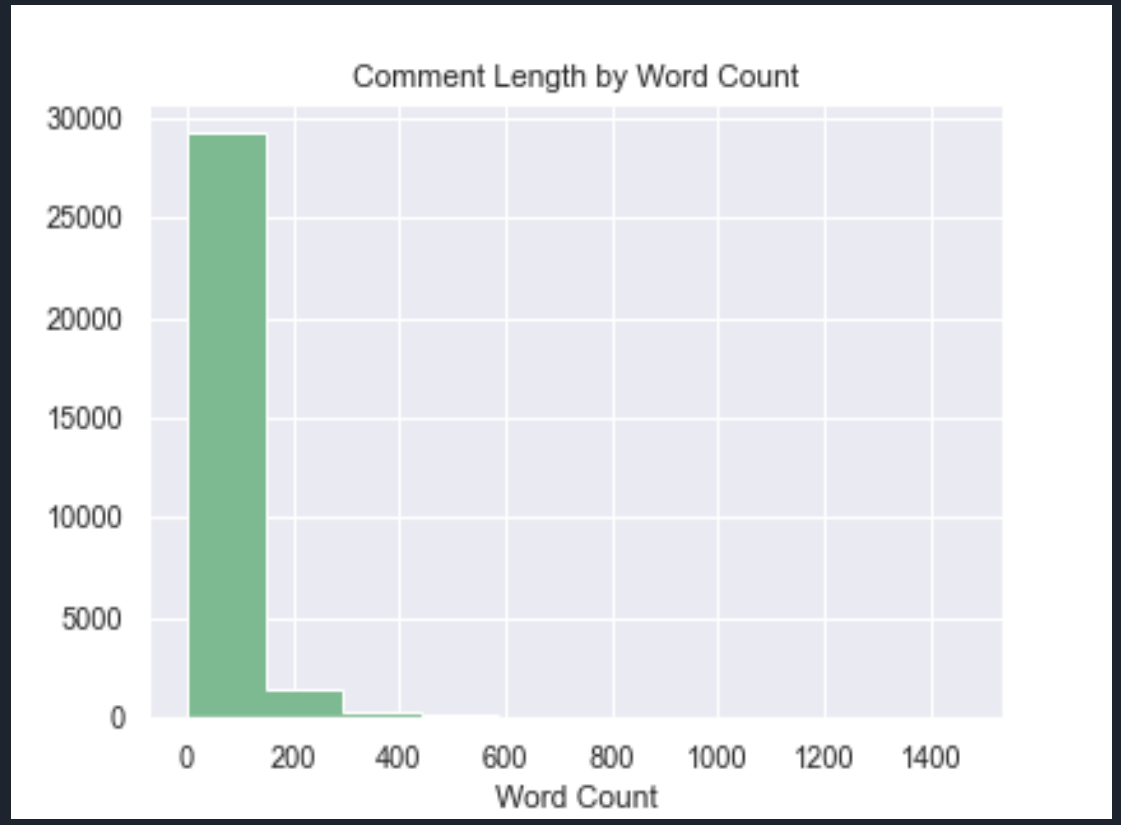
  31, 000 posts with usable text

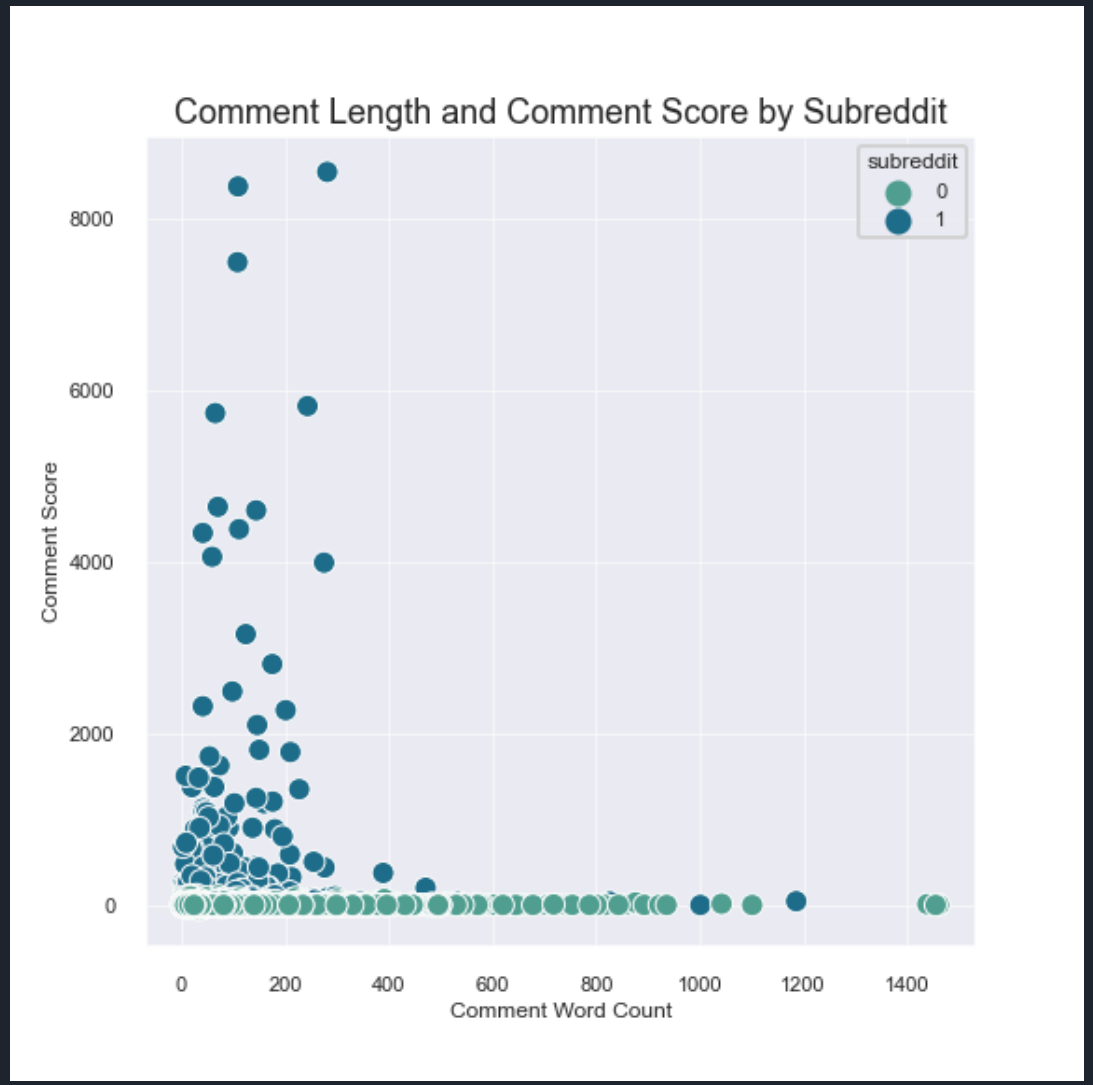
r/Askscience



r/longevity


Top 20 Words in r/longevity
aging
people
gt
longevity
think
years
age
one
research
old
life
time
good
cells
well
cancer
way
make
disease
reddit
0
500
1000
1500
2000
2500
3000
3500
4000
0


Comment Length by Word Count
30000
25000
20000
15000
10000
5000
0
0
200
400
600
800
1000
1200
1400
Word Count


Comment Length and Comment Score by Subreddit
subreddit
0
1
8000
6000
4000
2000
0
Comment Score
0
200
400
600
800
1000
1200
1400
Comment Word Count

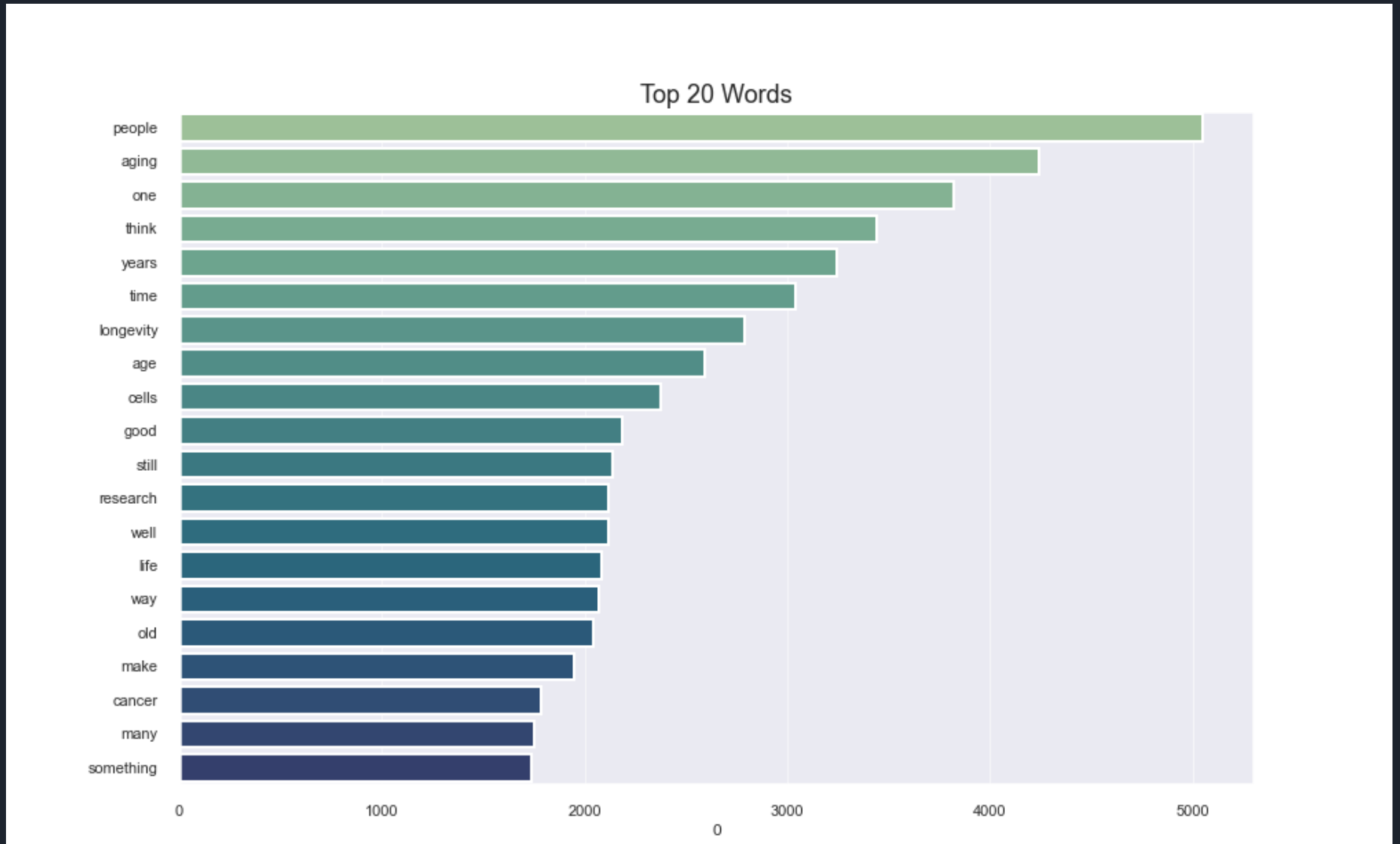
Transform

- CountVectorizer
- TF-IDF Vectorizer

Model

- Logistic Regression
- Random Forest
- DecisionTreeClassifier
- AdaBoostClassifier
- GradientBoostingClassifier

TOP 20 MOST FREQUENT WORDS



BEST MODEL:

TF-IDF CLASSIFIER & LOGISTIC REGRESSION

- Training Score: 89%
- Accuracy Score: 87%

CONCLUSIONS & THINGS TO CONSIDER

ASKSCIENCE

General Science related key words such as light, energy, earth, cells

LONGEVITY

More tailored towards ageing, fasting cancer, cells, research

DELETED and REMOVED text