



GA DSIR-920

PROJECT 3: WEB APIS AND NLP

ALONZO MAYS
DATA SCIENTIST

AGENDA



01

DEFINE THE PROBLEM,
BACKGROUND INFO

02

GET THE DATA,
EXPLORE DATA

MODEL AND
EVALUATE DATA


03

04

CONCLUSION/RESULT

A dark, low-key photograph of three business professionals (two men and one woman) standing side-by-side, smiling and holding large, blank white rectangular signs. The image is dimly lit, with the subjects' faces and the signs being the primary light sources. The overall mood is professional and collaborative.

PROBLEM STATEMENT



The marketing department for a liberal arts college needs to increase student enrollment for their Psychology and Anthropology departments. These study disciplines share many similarities, therefore they need distinctive keywords to differentiate their marketing campaigns to increase enrollment.

As a data scientist for the college, I need to find the most predictive words/phrases that help classify psychology and anthropology using subreddit posts as a resource.


Background Information - Subreddits r/psychology and r/AskAnthropology

About Community

A Reddit community for sharing and discussing science-based psychological material.

1.0m
Members

152
Online


 Created Mar 8, 2008

About Community

Welcome to AskAnthropology

128k
Members

86
Online

 Created Mar 10, 2013



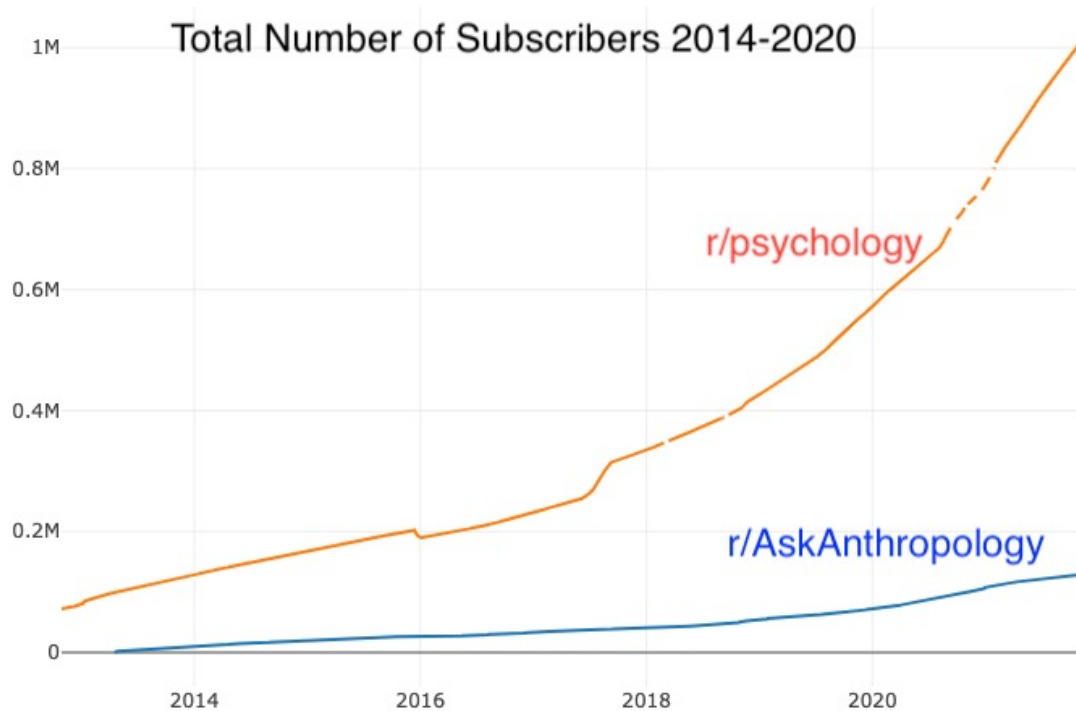
Psychology

r/psychology

AskAnthropology

r/AskAnthropology

Background Information - Subreddits r/psychology and r/AskAnthropology



Top r/AskAnthropology Post



Posted by u/arataumaihi 1 year ago

704



Any other anthropologists find this reddit a bit cringey sometimes?

Great to see people asking genuine questions, but if I see another post asking why X is better/more advanced/civilised than Y, or asking for evidence to support prejudicial worldviews, I'm going to cry.



92 Comments



Share



Save



Hide



Report

94% Upvoted

Top r/psychology Post



Posted by u/Mynameis___ 5 years ago

13.8k



Millennials Aren't Coddled - They Just Reject Abuse As A Management Tactic

canadianbusiness.com/blogs-...



1.1k Comments



Share



Save

...



A dark, low-key photograph of three business professionals (two men and one woman) standing side-by-side, each holding a white rectangular board. They are all smiling and looking towards the camera. The image is dimly lit, with the subjects' faces and the boards being the primary light sources. The text is overlaid in the center.

GETTING/EXPLORING THE DATA



Psychology

r/psychology

68,000 body comments initially scraped

The average amount of words in the body column is 50, and average character length in the body was 306

r/psychology

**The average words and character length are less than that of the anthropology subreddits

AskAnthropology

r/AskAnthropology

50,000 body comments initially scraped

The average amount of words in the body column is 81, and average character length in the body was 504

NLP PROCESS

The steps taken to go through the NLP

01

COMBINED DFS

- Dropped Unnamed columns and merged 2 DFs
- Total of 50_000 rows
- 25_000 from each subreddit

02


DEF FUNCTION

- Created a function that would CountVectorize columns and count the words and give a list of top 50 words

03

SPLIT DATA INTO X AND Y

- X was the body comments (set transformed using cv)
- y represented the target of r/anthropology and r/psychology

A dark, grayscale background image featuring three business professionals (two men and one woman) smiling and holding whiteboards. The image is dimly lit, with the subjects' faces and the whiteboards being the primary light sources. The whiteboards are blank and positioned in front of the individuals.

MODEL AND EVALUATE DATA

COMPARING MODELS

logreg

Logistic
RegressionCV
-CountVectorizer
-liblinear
-max_iter: 500
Train: 92.62%
Test: 86.46%

logreg

Logistic
RegressionCV
-TfidfVectorizer()
-liblinear

Train: 92.72%
Test: 86.99%

mnb

MultinomialNB
-CountVectorizer
-liblinear

Train: 87.11%
Test: 84.96%

mnb

MultinomialNB
-TfidfVectorizer

Train: 89.80%
Test: 86.11%

rf

RandomForest
-RegexTokenizer

Rf
cross_val_score: 80.7

Et
Cross_val_score: 82.5

BEST MODEL

Baseline score: 50%

Logistic Regression: train score: 92.7% | test score: 86.9%

*this score shows the high similarities of the two subreddits. Interesting, both disciplines are in the social sciences and there are a number of subdisciplines in psychology and anthropology that overlap (ie. Cultural psychology)



A background image showing three business professionals (two men and one woman) in business attire, each holding a whiteboard. The image is dark and serves as a backdrop for the text.

CONCLUSION/ RECOMMENDATION

RECOMMENDED KEYWORDS

Based on the findings from the data, the reddit posting show a close similarity with psychology and anthropology. The differences in keywords for psychology included: **study, message, feel, psychology, research**

Anthropology keywords include: **different, human, culture, lot(?), say, modern**



It is recommended to target ads for psychology using emotion words similar to 'feel'. Also to use keywords that appeal to an audience looking to do more research and study.

For Anthropology, try to highlight 'different' cultures, and use language that shows the progression of humans from ancient to modern.

Questions?

