AskHistorian and Khan Academy: Project Historical Literacy

By: Blake Franey

Overview

- Objective
- Subreddits
- Modeling
- Analysis
- Conclusion

Objective

Historical literacy is becoming increasingly more apparent. A partnership with AskHistorians and Khan Academy that can make learning about history more accessible and enjoyable to anyone (even if they aren't students) would combat this trend.

This proposal is a springboard for the project, looking at some simple yet powerful classification models, Logistic Regression and Multinomial Naive Bayes, to infer popular topics and lingo used by laypeople and historians alike.

The purpose of these models was to dig into the correlated terms between AskHistorians and our sister subreddit History to get a feel for how history is talked about on the internet. Success was a combined consideration between accuracy, F1, and sensitivity.

The Subreddits

r/AskHistorians

- 986k subscribers
- A very actively moderated subreddit for people to ask real historians questions and get detailed answers.
- Maintain their own Twitter and podcast series.

r/history

- 14.1m subscribers
- A place for discussion about history from books and articles to obscure topics like discussing people's favorite figure of minor French nobility.
- Also actively moderated

Example Post from r/history

Did conquistadors like spicy food?

Discussion/Question

I've been living in Spain for two years, and since I've been here one thing I've been kind of shocked by is the utter blandness of the food people eat here. I've been utterly unable to find anything spicy, and it's made me think a lot about cuisine during the Spanish conquest of Latin America. It's become clear to me that the trademark spiciness we expect from Mexican cuisine *must* be native, and not coming from the Spanish who can't stand even a little bit of heat.

This makes me curious: are there any records of Spanish conquistadors or colonists going to Mexico and complaining about being unable to handle the cuisine? How did they acclimate themselves to the spiciness?

Example Post from r/AskHistorians

What was going on between the 1850s and 1880s that so many organized sports had their starts then?

Four of the five major modern sports saw their initial organization begin in the mid-to-late 19th century: Soccer started in England in 1863; baseball in New Jersey in 1846; American football was developed and codified between the 1870s and 1880s; hockey in Montréal in 1875.

Were there any social, environmental or economic changes going on to spur this growth? Or was it just random chance?

Methods Used for Preparation

Gathering/Cleaning the Data

- First, the data was scraped to obtain ~2000 posts for each subreddit
- Lots of unnecessary columns, only chose 7 to keep for the ‘unclean’ data frame. (only Title and Selftext made it into the model in the end)
- Combined title and selftext to create full text column.

Pre-processing

- Lemmatized the full text and removed stop words, ie common English words that wouldn’t be helpful for classification.
- Combined the cleaned text data from each subreddit into a single data frame.
- Vectorized the text to make it usable in an algorithm.

Modeling

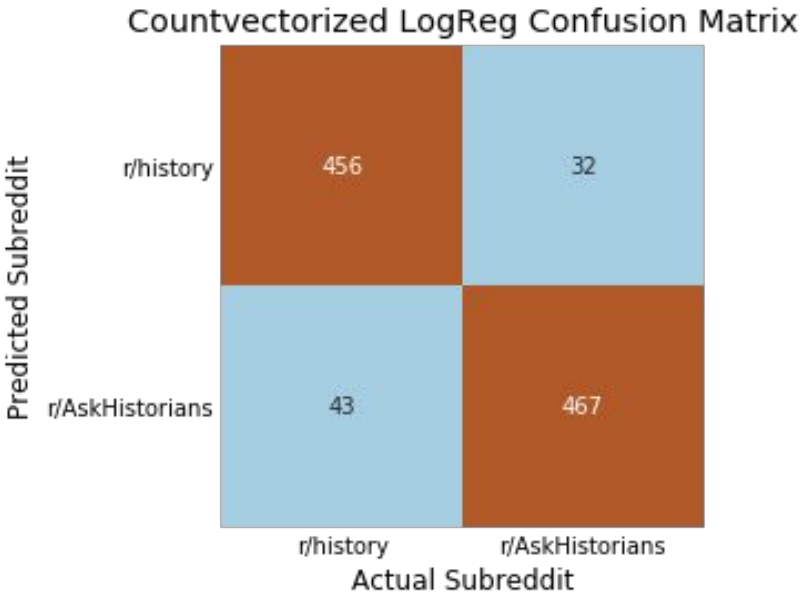
- The data was vectorized using the Countvectorizer and the Term Frequency–Inverse Document Frequency (tf-idf) methods..

- I compared a baseline score, Logistic Regression, and Multinomial Naive Bayes classifier.

- I also applied a Vader sentiment analysis to see how that affected the model.

Results

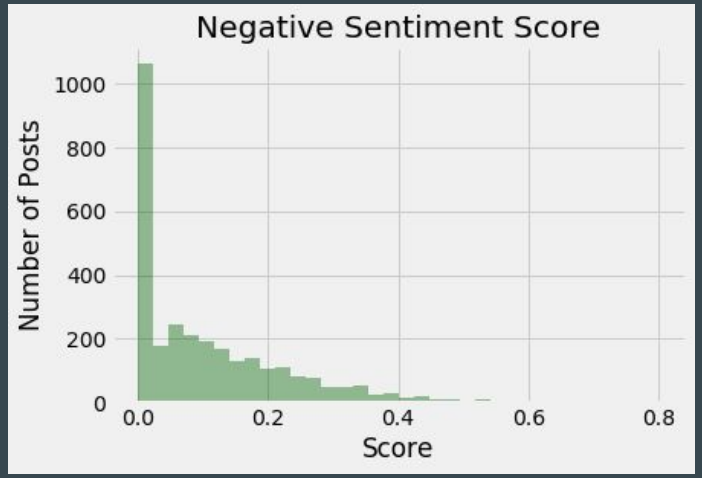
- Baseline was ~0.5 because each class was the same size. r/AskHistorians was the positive class.
- The countvectorizer gave the best results.
- Tf-idf and sentiment analysis gave slightly lower scores than the simple countvectorizer.

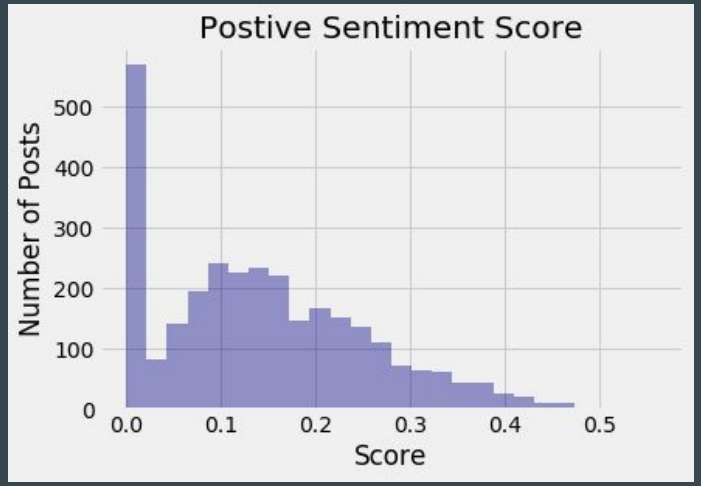


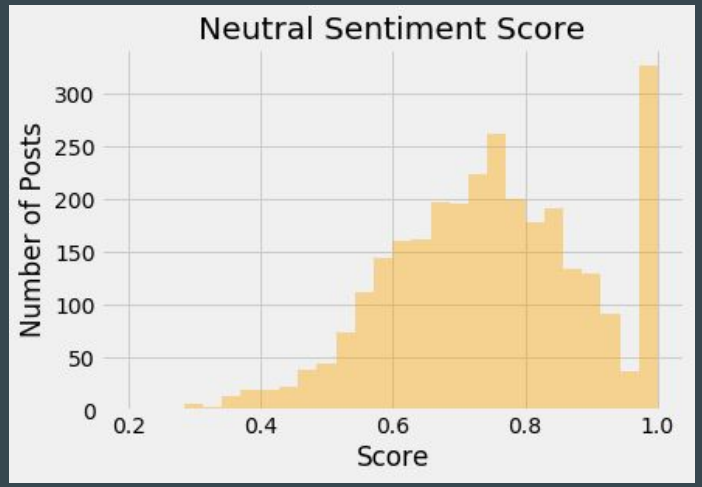
Specificity:
0.91

F1 Score:
0.926

Cross Validation Score (k=20):
0.905

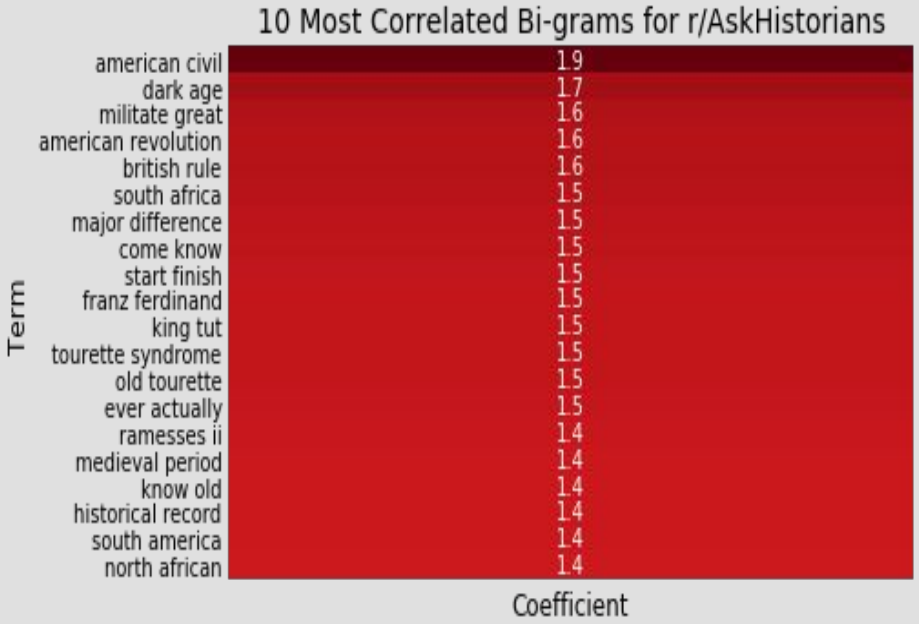

Negative Sentiment Score
Number of Posts
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
Score

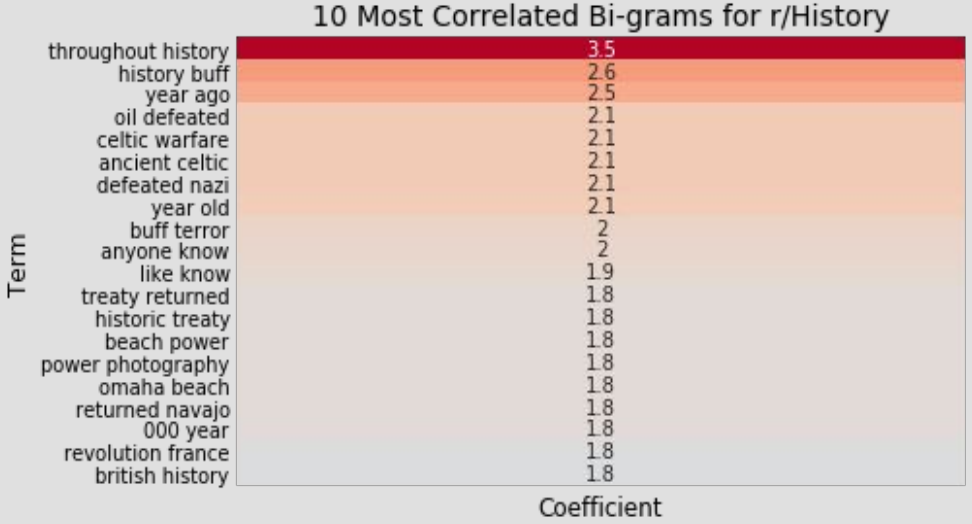

Postive Sentiment Score
Number of Posts
500
400
300
200
100
0
0.0
0.1
0.2
0.3
0.4
0.5
Score


Neutral Sentiment Score
Number of Posts
300
250
200
150
100
50
0
0.2
0.4
0.6
0.8
1.0
Score

Analysis

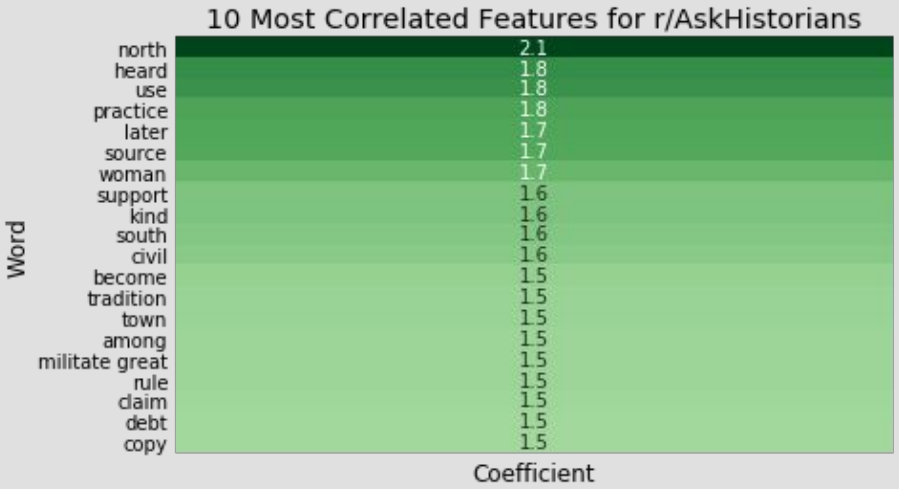
- Certain topics and time periods showed up more often.
- Diction between subreddits was also distinct in the most correlated terms.
- 2- and 3-grams didn't predict as well but gave better insight to the subreddits.


10 Most Correlated Bi-grams for r/AskHistorians
Term
american civil 1.9
dark age 1.7
militate great 1.6
american revolution 1.6
british rule 1.6
south africa 1.5
major difference 1.5
come know 1.5
start finish 1.5
franz ferdinand 1.5
king tut 1.5
tourette syndrome 1.5
old tourette 1.5
ever actually 1.5
ramesses ii 1.4
medieval period 1.4
know old 1.4
historical record 1.4
south america 1.4
north african 1.4
Coefficient


10 Most Correlated Bi-grams for r/History
Term
throughout history 3.5
history buff 2.6
year ago 2.5
oil defeated 2.1
celtic warfare 2.1
ancient celtic 2.1
defeated nazi 2.1
year old 2.1
buff terror 2
anyone know 2
like know 1.9
treaty returned 1.8
historic treaty 1.8
beach power 1.8
power photography 1.8
omaha beach 1.8
returned navajo 1.8
000 year 1.8
revolution france 1.8
british history 1.8
Coefficient

Interesting Words:

- North
- Jesus
- militate great
- woman



Interesting Topics:

- American Civil War
- Nudity become taboo
- Franz Ferdinand Assassination
- Military generals and feats.

Conclusion and the Future

1. There was a distinct difference in diction, topics, and phrasing between AskHistorians and History
2. Certain topics seem to be more popular.
3. Knowing popular topics can be a great way to people interested in our content. We will be able to collect more data to better guide what we offer