


Project 3: Web API's & NLP



Marva Loyfer, DSI 321



Problem Statement

This project proposes a classification model that can more accurately identify the subreddit a post belongs to, based only on the title.



AskWomen: Questions about women's thoughts, lives, and experiences

r/AskWomen

[Join](#)

[Posts](#) [Rules](#) [FAQs](#)

[Hot](#) [New](#) [Top](#) ...



18



PINNED BY MODERATORS

Posted by u/AutoModerator 5 hours ago

Mate Free Monday

4 Comments Share Save ...

About Community

AskWomen: A subreddit dedicated to asking women questions about their thoughts, lives, and experiences; providing a place where all women can comfortably and candidly share their responses in a non-judgmental space. As part of our commitment to that mission, the AskWomen subreddit is curated to promote respectful and on-topic discussions, and not serve as a debate subreddit.

3.5m
Subscribers

3.3k
Viewing

Created Jul 17, 2010

r/AskWomen topics

[Ask](#) [Meta/Reddit](#)



We don't know, but we'll answer anyway

r/AskMen

[Join](#)

[Posts](#)

[Hot](#) [New](#) [Top](#) ...



146



PINNED BY MODERATORS

Posted by [Happy Little Vegemite](#) 10 days ago

FAQ Friday: Dating- What kind of things do you do to show interest, test the waters, or escalate things when it comes to dating?

168 Comments Share Save ...

About Community

We don't know, but we'll answer anyway

3.7m
Members

11.9k
Online

Created Aug 30, 2010

s?

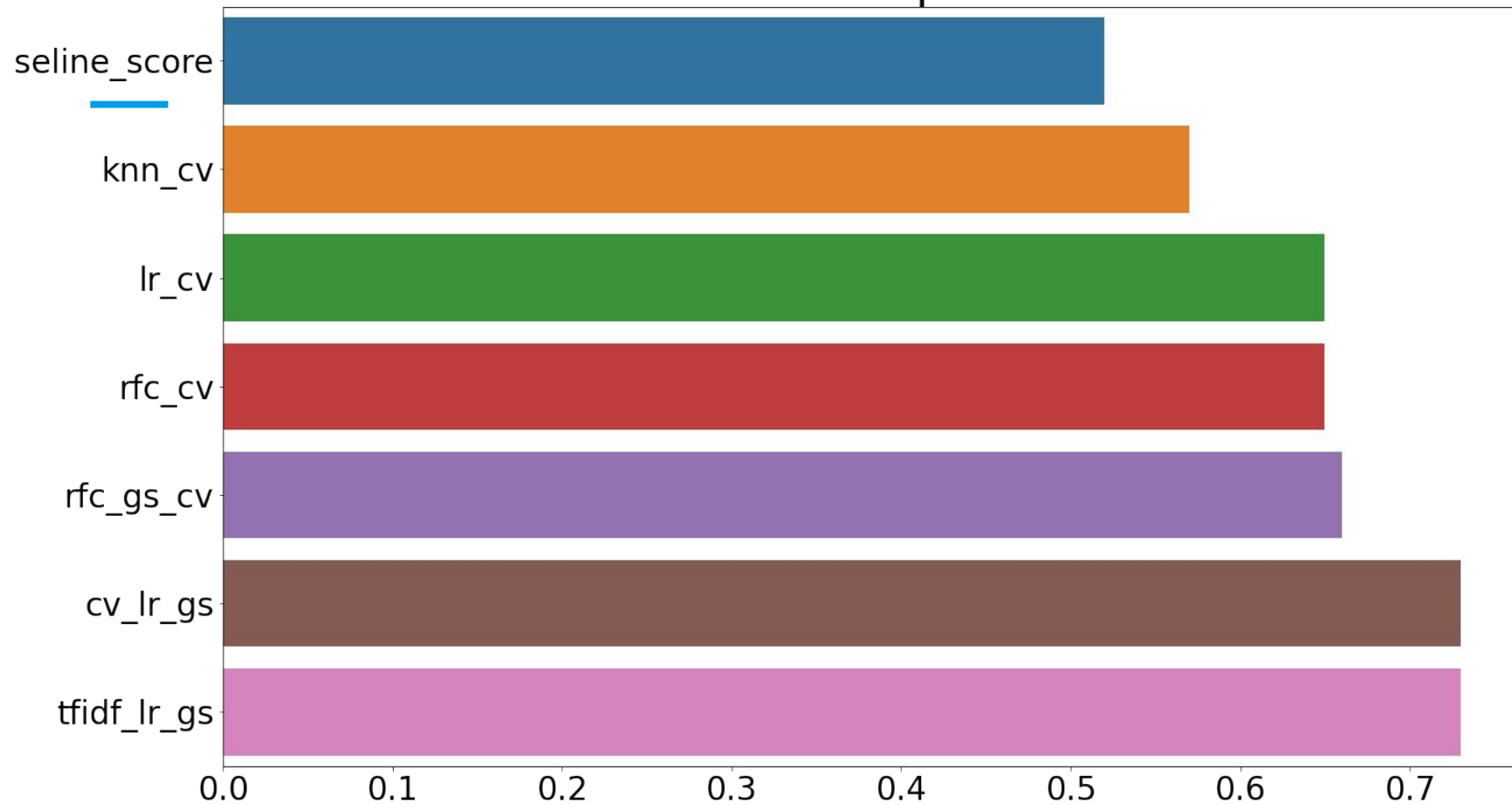
oman?

25: why?



33% 79°F Sunny

How did the models compare to the baseline?



Top 10 Predictive Words

stop_word = 'english'	stop_word = None
men	you
women	what
like	do
did	to
feel	how
woman	your
think	the
guy	is
girl	and
partner	of

Conclusion & Recommendations

Conclusions:

- All models beat the baseline score, though none achieved an accuracy score higher than 72%.
- There was no difference between two vectorizers (CountVectorizer, TfidfVectorizer).
- Keeping the stop words resulted in better scores

Recommendations:

- Using GridSearch improved the accuracy score; expanding the GridSearch might lead to a better result.
- Gradient Boosting might improve the model