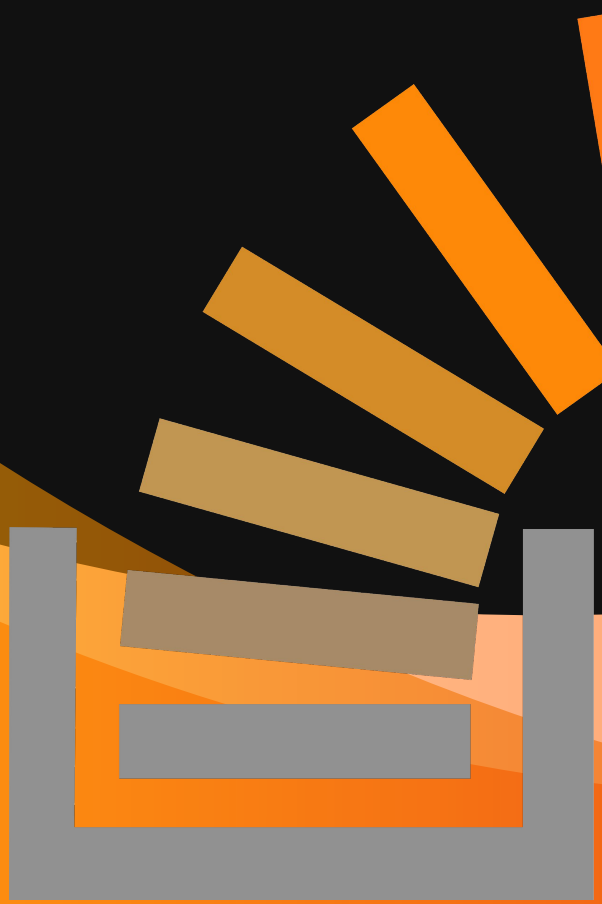
Will G. Nobles



# What makes a question useful?

## Topic Modeling of Stack Overflow Text

# Introduction

- **Motivation**: understand questions that are commonly the most difficult
- **Objective**: identify terms associated with high scoring posts
- **Goal**: determine topics associated with helpful questions

# Methodology

## Data

Stack Overflow dataset from BigQuery

## Metrics

Term frequency, topic probability

## Model

Latent Dirichlet Allocation with CountVectorizer

## Tools

Standard SQL, NLTK, Gensim, Seaborn, Matplotlib

*“can anyone please explain and please do give examples of other situations also i have seen large lines of code where its just hex numbers and never really understood why*

**Text from Topic 6 with > 97% probability**

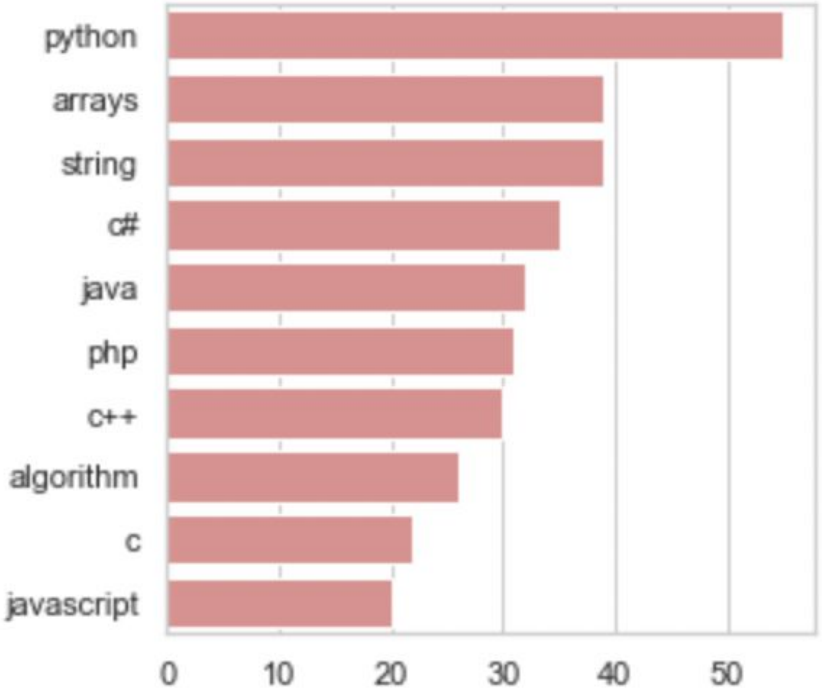
# Results

Keywords that define the dominant topic in question samples

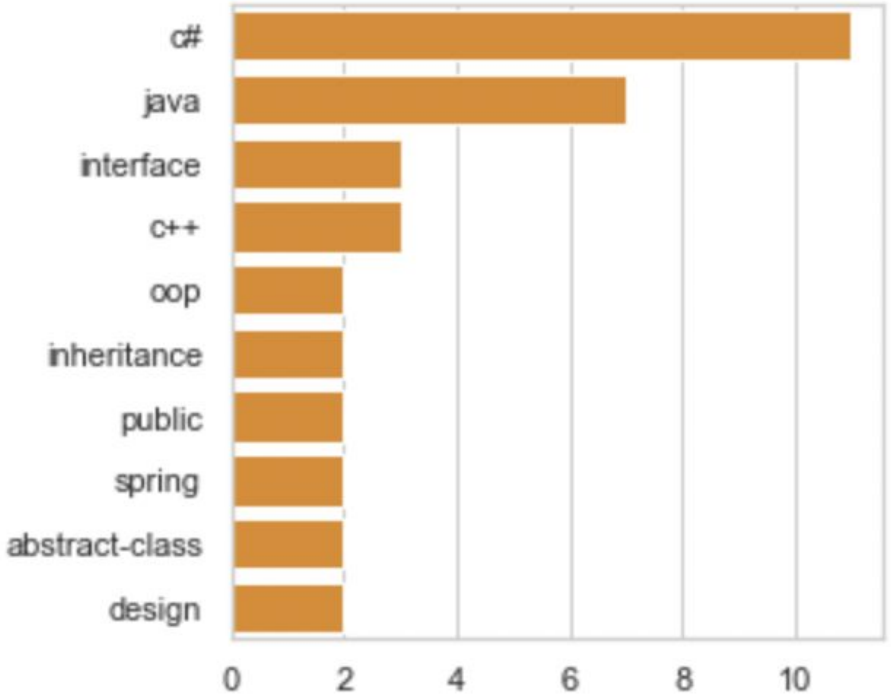
| Dominant Topic | Topic Percent Contribution | Keywords |
|---|---|---|
| **5** | **29** | **like, way, want, im, data, code, using, string...** |
| **16** | **34** | **new, object, var, null, test, code, error, exc...** |
| **1** | **37** | **warning, compiler, difference, reference, pers...** |

# Counts of Top 10 Tags per LDA Topic

## Topic 6



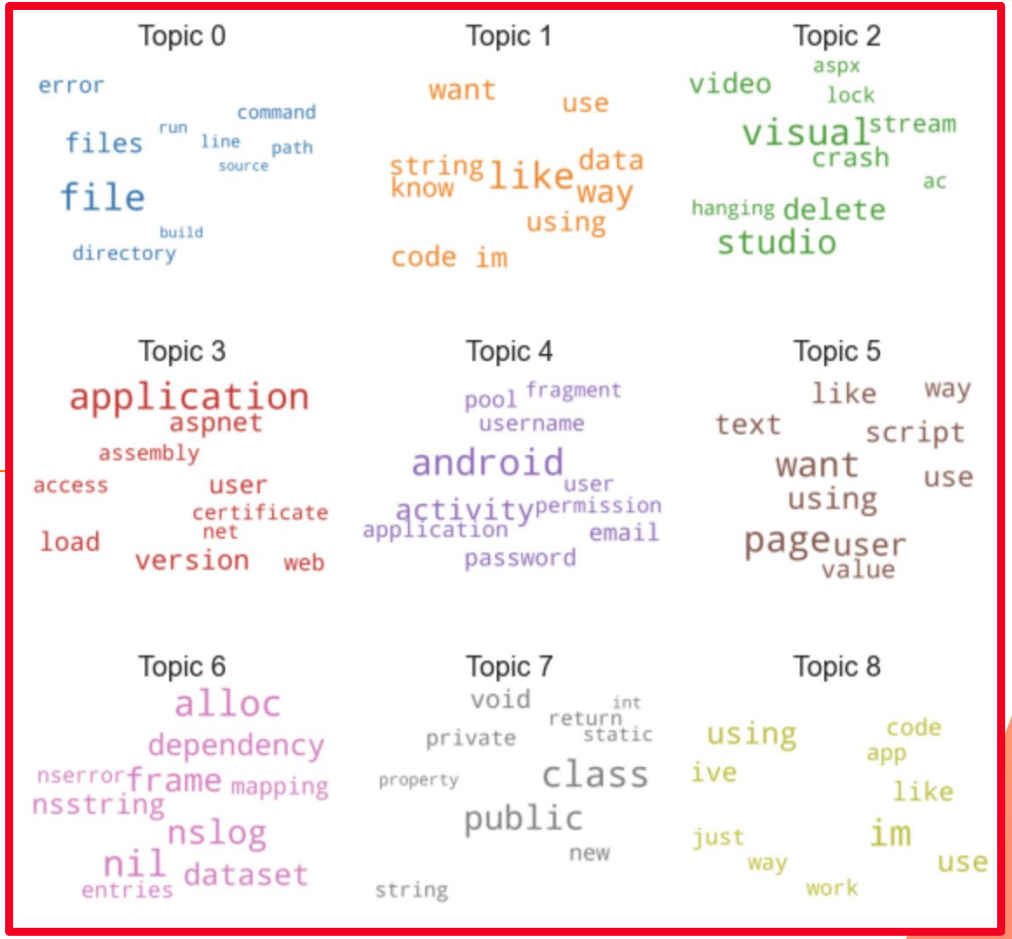

## Topic 8




Questions filtered for >80% probability

# Word Cloud

Top 10 Words in Each Topic



# Conclusions

## Recommendations

- Use the model to identify commonly misunderstood topics

## Impact

- Educators can prioritize
- Users can increase reputation
- Question posters can use helpful language

# Insights

**48%**

Average topic probability

**14,772**

Top scoring post

**2,141,036**

Post with highest number of views

**4,436**

Topic with highest number of associated questions

**14,772**

Top scoring post

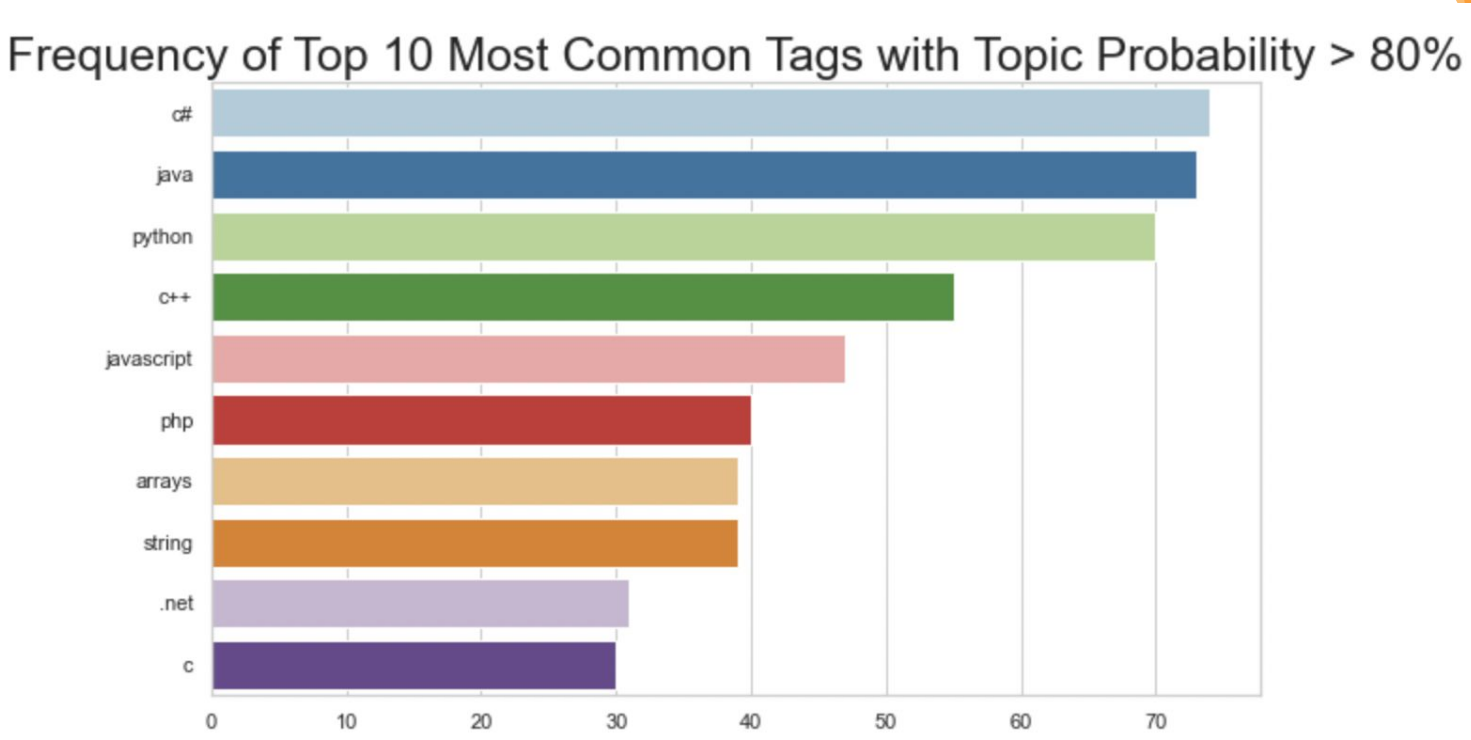
**2,141,036**

Post with highest number of views

# Future Work

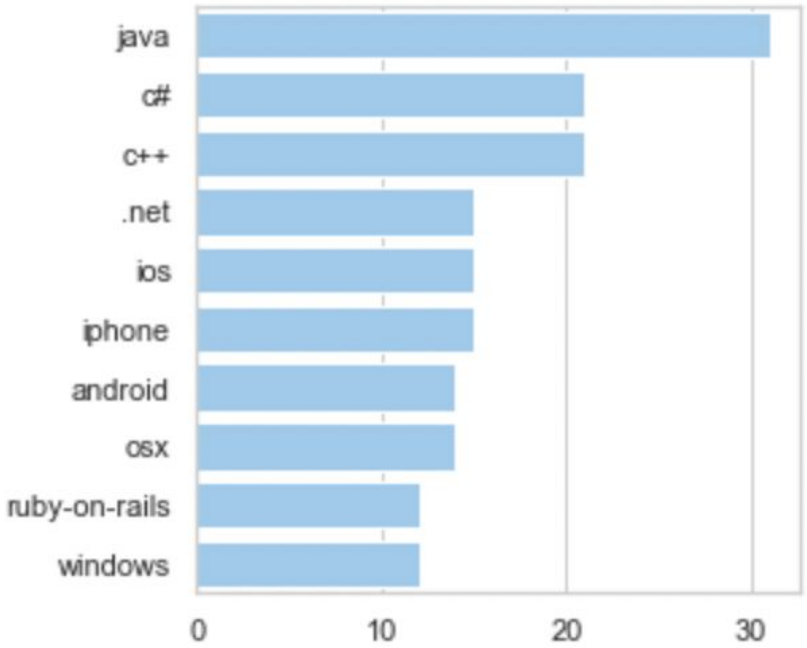
- Increase Stack Overflow domain knowledge on text that isn't normally defined as a word
- Try again with NMF to see if better results can be gained

# Appendix

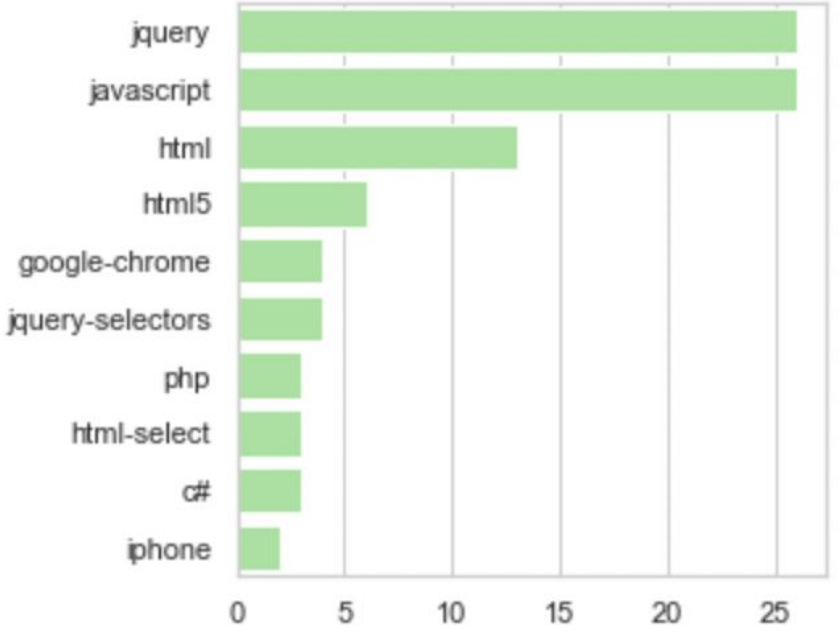

Frequency of Top 10 Most Common Tags with Topic Probability > 80%
c#
java
python
c++
javascript
php
arrays
string
.net
c
0
10
20
30
40
50
60
70

# Counts of Top 10 Tags per LDA Topic

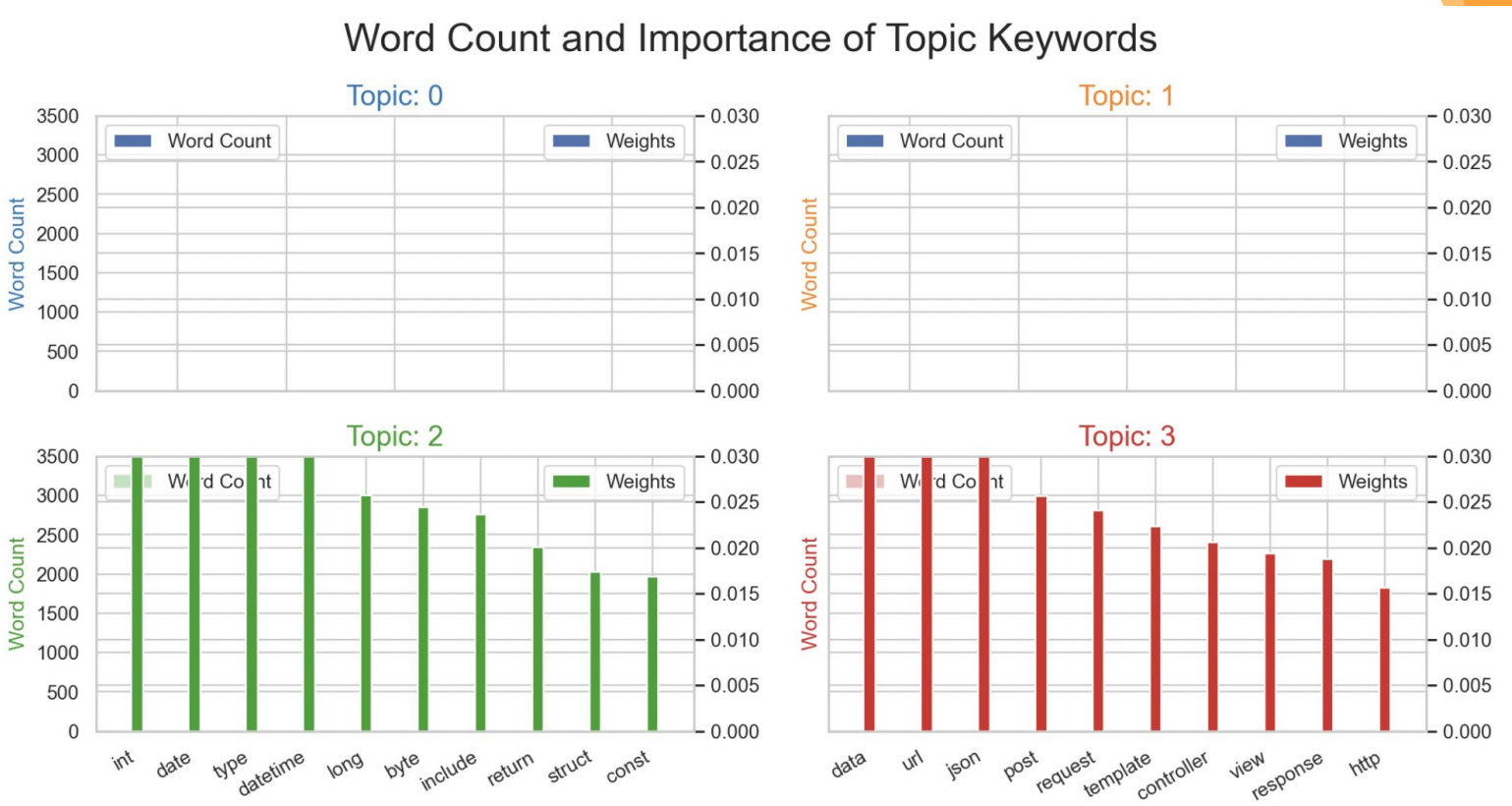
## Topic 19




## Topic 7




Questions filtered for >80% probability


Word Count and Importance of Topic Keywords
Topic: 0
Topic: 1
Topic: 2
Topic: 3
Word Count
Weights
int
date
type
datetime
long
byte
include
return
struct
const
data
url
json
post
request
template
controller
view
response
http