Predicting the Success of Answers on Stack Overflow

Spencer Haupert, Frank Ockerman, Selena Scott, John Tsirigotis

12/17/2021

Introduction

For this project, we are exploring a dataset of all the questions and answers from StackOverflow, focusing on R as the programming language of interest (i.e., the questions that contain an R tag) [1]. Our primary aim is to identify factors that influence an answer's score and its likelihood of acceptance.

The data include questions asked on Stack Overflow between 2008 and 2017 and are limited to questions that weren't closed or deleted. Within this dataset, there are three files available that contain information on the questions, answers, and tags specific to each post. The Tags file includes the tags used on each question besides the R tag. The Questions file includes the user ID, title, full body of text, date the question was posted, and score (the difference between the number of upvotes and downvotes). The Answers file includes the user ID, full body of text, date the answer was posted, score, and whether or not the answer was accepted (selected as the best solution to the question).

Methods

We adopted a technique known as topic modeling in order to work with the text from answers in the data. Topic modeling is a statistical method of identifying the underlying topics within a corpus (i.e., set of written texts). The process of topic modeling ultimately derives a set of K topics from the corpus, where each topic consists of a set of terms that define that topic [2]. Using Latent Dirichlet Allocation (LDA) to carry out this process, the model output consists of a topic's proportional contribution to a document (i.e., body of text of an answer), as well as the proportional contribution of a term to a topic.

To perform topic modeling, it was first necessary to create a cleaned corpus by reducing words to their root form and removing capitalization, punctuation, numbers, and common stopwords in the English language. Additionally, we only included words that were in at least 2% and at most 80% of the documents. After this process of forming a cleaned corpus, there were some documents that were left without any words and these documents were necessarily removed prior to inputting the document-term matrix into the LDA model.

The LDA model also requires that the number of K topics is specified. Traditionally, an arbitrary value for the number of topics within a corpus is K = 20. There is also a procedure of identifying the ideal value of K by fitting several iterations of models with different values of K and comparing the coherence score of the topics in each model [3]. Coherence score measures the extent to which words within a topic are related to each other, meaning a high coherence score suggests that the derived topics consist of highly correlated words. This procedure of determining the value K through coherence score was computationally intensive and required the use of the department cluster.

Results

As mentioned above, high coherence scores indicate a better fit of the LDA model. Figure 1 shows the coherence scores of models with different numbers of K topics, where models with over 12 topics have the highest coherence scores. However, the effort to name topics when K was set to 20, 15, or 10 by looking at the top 15 terms per topic suggested that a lower value of K was more appropriate. Therefore, we ran an LDA model with K = 8 and found that the topics were not overlapping excessively at this value of K, meaning the topics could be reasonably named based on the top 15 terms per topic.

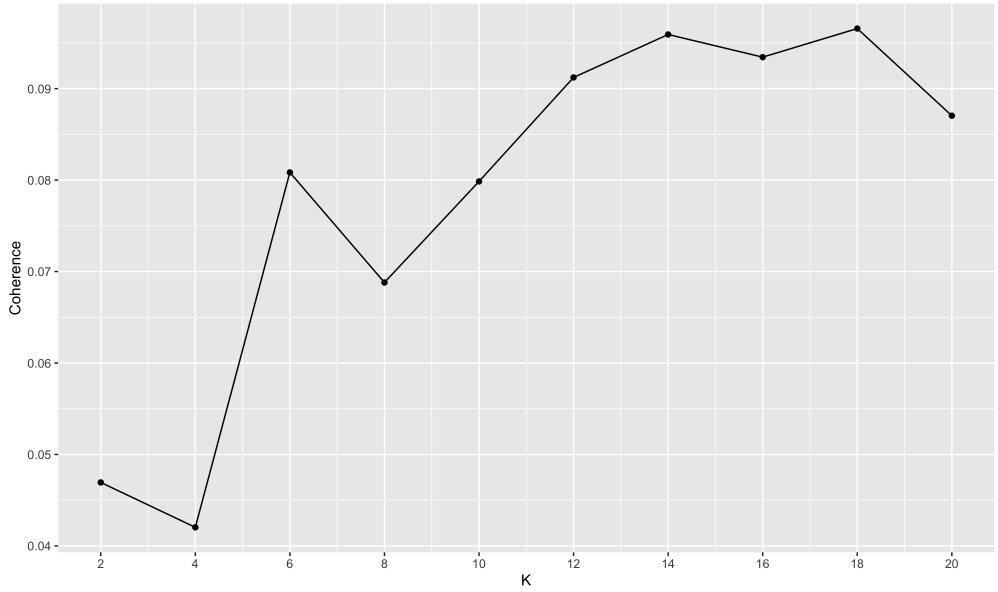

Figure 1: Best LDA Model by Coherence Score

Based on this LDA model, there were eight topics identified from the corpus of answers. Table 1 shows the names we gave those topics and the top 15 terms belonging to those topics. Figure 2 visualizes how each answer is made up of those topics with varying proportions for a random subset of 50 answers.

Table 1: Top 15 Terms Per Topic

Topic	Terms
function definitions and calls	function code call return object argument the error result method pass defin loop oper assign
dates and times	data time group date frame start sampl year format end day count creat dat long
installing and loading packages	packag work file problem test run version read instal user check comment sourc load find
vectors	true fals vector functionx element length null sep you this here solut result fun collaps
strings	you class output if charact string match option replac remov numer convert input this or
lists and variables	list variabl set type case factor model make order level answer question dont your correct
plots and figures	plot imag line descript add label point size text here color set you fill chang
data frames and matrices	column valu row name datafram number matrix col creat index tabl select subset dataset max

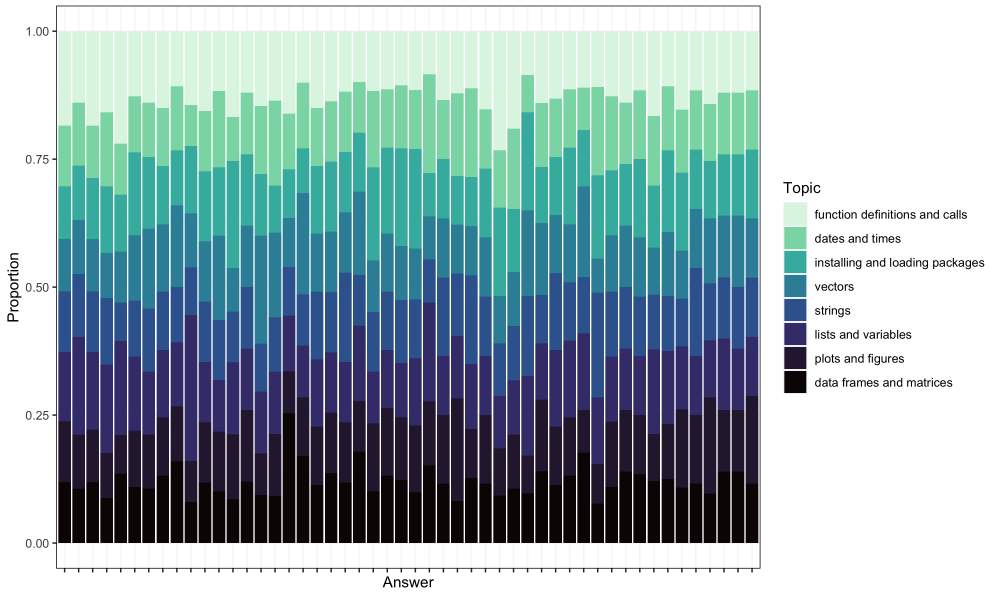

Figure 2: Topics of 50 Answers on Stack Overflow

Discussion

Contributions

- Spencer: feature engineering
- Selena: topic modeling, introduction, parts of methods and results
- Frank: sentiment analysis, regression analysis, diagnostics
- John: parts of methods and discussion

References

1. Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning, 113-120.
2. "R Questions from Stack Overflow." Accessed November 17, 2021. https://kaggle.com/stackoverflow/rquestions.
3. Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 399-408.