Insights for Curriculum Development: Identifying Emerging Data Science Topics through Analysis of Q&A Communities

Habib Karbasian

George Mason University

Fairfax, Virginia

hkarbasi@gmu.edu

Aditya Johri

George Mason University

Fairfax, Virginia

johri@gmu.edu

Abstract

Updating curricula in new computer science domains is a critical challenge faced by many instructors and programs. In this paper we present an approach for identifying emerging topics and issues in Data Science by using Question and Answer (Q&A) sites as a resource. Q&A sites provide a useful online platform for discussion of topics and through the sharing of information they become a valuable corpus of knowledge.We applied latent Dirichlet allocation (LDA), a statistical topic modeling technique, to analyze data science related threads from from two popular Q&A communities "Stack Exchange and Reddit". We uncovered both important topics as well as useful examples that can be incorporated into teaching. In addition to technical topics, our analysis also identified topics related to professional development. We believe that approaches such as these are critical in order to update curriculum and bridge the workplace-school divide in teaching of newer topics such as data science. Given the pace of technical development and frequent changes in the field, this is an inventive and effective method to keep teaching up to date. We also discuss the limitations of this approach whereby topics of importance such as data ethics are largely missing from online discussions.

CCS Concepts

• Information systems→Document topic models; Data cleaning;

• Applied computing→Document analysis.

Keywords

online Q&A platforms, text mining, topic modeling, StackExchange, Reddit, curriculum development

ACM Reference Format:

Habib Karbasian and Aditya Johri. 2020. Insights for Curriculum Development: Identifying Emerging Data Science Topics through Analysis of Q&A Communities. In The 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20), March 11–14, 2020, Portland, OR, USA. ACM, New

York, NY, USA, 7 pages. https://doi.org/10.1145/3328778.3366817

SIGCSE '20, March 11–14, 2020, Portland, OR, USA

ACM ISBN 978-1-4503-6793-6/20/03. . . $15.00
https://doi.org/10.1145/3328778.3366817

1 Introduction

Data Science has emerged as an important area within Computer Science (CS) at both the undergraduate and graduate levels but as with any new domain, developing relevant curricula for the field has required significant effort. The ACM Education Council has set up a Data Science Task Force to articulate the role of computing discipline-specific contributions to this emerging field [14]. The National Science Foundation (NSF) sponsored a workshop [11], the EDISON Data Science Project [2] launched in 2015 with the purpose of "accelerating the creation of the Data Science profession" [24], and the Park City Math Institute brought together an interdisciplinary faculty group to devise curriculum guidelines for undergraduate programs in data science. Finally, National Academies of Sciences, Engineering, and Medicine released an extensive report on Data Sciences for Undergraduates [30].

Although these efforts are commendable, as the ACM task force has noted, a particular challenge of developing Data Science curricula is that the field is inherently interdisciplinary, requires balancing computing, statistics, and domain knowledge, and as with most computing related fields, keeping up with advances in the field is difficult and changes in curriculum are common and often an ongoing exercise. The need for regular updates does not imply that programs need to jump on bandwagons with every change [21] but a timely response to emerging areas is still essential for preparing the future workforce [27]. Given that the practice of curriculum development through the use of committees is comprehensive but often slow [1], there is need for instructors and programs to have a more dynamic view of changes in the field [36] which allows it to broaden the curricula as needed [31]. There is also a need to find specific problems and their solutions, scenarios and cases, that can be integrated into the curricula [26].

In this paper we present a data mining based approach for curriculum improvement that relies on drawing insights from online communities, especially question and answering (Q&A) communities, for keeping track of topics, tools, and techniques that are relevant to Data Science. In addition to theoretical and mathematical issues, which form the core of the curricula, there is a need to address pragmatic issues such as tools [8] which requires following a functional approach [13] to includes real world examples and scenarios [17].

2 PriorWork

A review of recent work on Data Science education shows that a number of topics, including modeling, optimization, and data lifecycle, are important for the field and so are professional skills such as communication and teamwork [35]. The other area that is of importance is data ethics [34]. Given this prior work that has synthesized important topics, one of the ways in which we can gauge the usefulness of the results is to compare with what has already been synthesized but also look at the value of novelty in our findings. Furthermore, we also want to examine if using more than one resource or platform adds value to our approach especially in terms of uncovering different topics.

Although many sources for information about topics are available, Q&A communities are a valuable resource for learning about different topics as they are used by thousands of professionals in the field as well as by newcomers to the topics. They provide easy access to experts [22] who can scaffold newcomers' learning [19]. They have high quality information [38], response rates are fast and they are up-to-date with new information[23], they contain useful examples of code [29], and they are largely publicly available. Overall, they have changed the landscape for information sharing and knowledge building [39] across a range of topics [40] and have been found to be useful for teaching [18] and for better understanding a topic[28]. Their application for curricula update and verification so far has not been investigated and that is one of the aims of this paper. An inclusive approach to curricula development and enhancement is crucial to capture and expand the diversity of ideas in the field[33].

3 Research Study

We use our data mining techniques, in particular, topic modeling, as a way to identify relevant topics [7] and the research question guiding the work was: What are the main discussion topics in Stack Exchange (data science) and Reddit (data science subcommunities)?

We first provide an explanation of our datasets followed by description of the methods and analysis.

3.1 Datasets

3.1.1 Stack Exchange: Data Science

Stack Exchange is an online platform that hosts a variety of Q&A forums including one on data science. The platforms features the ability for users to ask new questions and answer existing questions, as well as to “vote” questions and answers up or down, based on the perceived value of the post. Users of Stack Exchange can earn reputation points and “badges” through various activities. Stack Exchange makes its data publicly available in XML format under the Creative Commons license [5]. The dataset is divided into five XML documents: badges.xml, comments.xml, posts.xml, users.xml and votes.xml. For our purposes, we use posts.xml and comments.xml, which contain the actual text content of the posts and the comments, as well as the view count, favorite count, post type, creation date, and ID of the user who created each post and comment. The dataset spans for four years , beginning of 2015 until 2019 spanning 48 months. The dataset has 57,075 posts and comments: 27,249 (47.7%) posts including questions and answers and 29,826 (52.3%) comments.

3.1.2 Reddit: Data Science Subreddits

Reddit is a community-driven platform for submitting, commenting and rating links and text posts. It brands itself as a social news website where registered users submit content in the form of links or text posts. Content entries, submissions, are organized by areas of interest or subcommunities called subreddits, such as politics, programming, science.

We used the data dump provided here [4] under public licence which was collected originally from Reddit’s official API [3] for submissions and comments. For the purpose of this work we decided to filter the entire dataset to these 9 data science related subreddits: 1-DataIsBeautiful, 2-MachineLearning, 3-DataScience, 4-LearnMachineLearning, 5-Analytics, 6-MLQuestions, 7-BigData, 8-DeepLearning and 9-DataMining. To put them in the same time line as Stack Exchange, we limited the data for the recent four years, January 2015 until December 2018, 48 months. The dataset contains 3,559,702 submissions and comments: 226,134 (6.4%) submissions

Table 1: Submissions and comments for each subreddit

Ranking	Subreddit	Submissions+Comments (%)	
		Original	Preprocessed
1	DataIsBeautiful	2,975,912 (83.60)	1,781,973 (80.52)
2	MachineLearning	307,210 (8.63)	222,370 (10.05)
3	DataScience	154,149 (4.33)	116,328 (5.26)
4	LearnMachineLearning	40,642 (1.14)	31,088 (1.40)
5	Analytics	27,794 (0.78)	21,428 (0.97)
6	MLQuestions	20,946 (0.59)	17,432 (0.79)
7	BigData	18,435 (0.52)	11,843 (0.54)
8	DeepLearning	11,262 (0.32)	8,074 (0.36)
9	DataMining	3,352 (0.09)	2,669 (0.12)

and 3,333,568 (93.6%) comments. The breakdown of submissions and comments for each subreddit is shown in table 1 under original column.

3.2 Methodology and Analysis

3.2.1 Preprocessing Textual content was extracted from posts and comments by removing any code snippets, HTML tags, URLs, hashtags, by applying the Porter stemming algorithm[20], and by removing common English-language stop words. To increase the quality of text analysis 2-grams (equivalently, bi-grams) were used in the model [37]. Lemmatization was applied to identify intention in a part of speech and meaning of a word in a sentence. We used adjective, adverb, noun and verb as accepted parts of speech. To remove less frequent words from the sentences we set the minimum threshold as 10 for each word and 60% of the documents for maximum threshold. Finally, we removed sentences with less than 5 words to help with topic modeling (discussed later). After preprocessing, the dataset contained 26,856 (52.3%) posts and 24,152 (47.7%) comments for Stack Exchange dataset and 137,060 (6.2%) submissions and 2,076,145 (93.8%) comments for Reddit dataset. The breakdown of the submissions and comments for subreddits is shown in table 1 under preprocessed column.

3.2.2 Topic Modeling For topic modeling we applied latent Dirichlet allocation (LDA)[10] using the MALLET version 2.0.8 [25], an implementation of the Gibbs sampling algorithm [15]. LDA has previously been used for similar purpose, most notably by [32] to infer topics within different courses. In this approach the number of topics, denoted by K, is a user-specified parameter and there is no single value of K that is appropriate in all situations and all datasets ([41] and [16]). The coherence score provides a rough estimate of the quality of the model and was used to decide the numbers of topics for each dataset. This method clusters semantically related keywords in K groups which together reveal the nature, or concept, of the topic. For ease of readability, we have manually provided a short label for each topic. We chose labels based on the top words in the topics and by examining a sample of posts that contain the topics.

3.2.3 Metric In this study, we set σ defined in [10] to 0.10, which we found to remove noisy topic memberships while still allowing only the dominant topics to be present in each document. Then for each document, we normalized the weight of topics to be 1.

3.2.4 Topic Share We used the overall share of a topic z_k across all posts defined in [9]. The topic share metric measures the proportion of documents that contain the topic z_k. For example, if a topic has a topic share metric of 10%, then 10% of all texts contain this topic. The topic share metric allows us to measure the relative popularity of a topic across the entire corpus.

4 Results

We analyzed both datasets using a range of topics (from 2 to 100) and chose the highest coherence score as the basis for our optimal model for each dataset. The final model was trained for 1000 iterations. We uncovered 32 and 62 topics for Stack Exchange and Reddit respectively. Topics on Reddit were of wider variety as compared to Stack Exchange and are proabably an artifact of how the two platforms are moderated differently. The posting guidelines in Reddit are flexible but StackExchange enforces strict rules and off-topic postings are disallowed. The only non-data science topic among 32 topics in the StackExchange dataset was Q&A Guidelines which had 10.56% topic share; it was the second most dominant topic which was discarded from our analysis (Table 2). In Reddit, we removed several topics from Reddit (such as “US election”, “entertainment industry”, “sports”, “climate change”) from our analysis as by analyzing few samples from each topic, we realized that the discussions were not data science related and mostly personal opinion exchange with no data related substance. Therefore, we found only 19 topics relevant to data science out of 62 topics (Table 3). Overall, the 31 data science related topics out of 32 in StackExchange (data science community) have 89.45% of total posts and comments. The 19 data science related topics out of 62 in the Reddit (9 data science subreddits) have 27.37% of total submissions and comments. Tables 2 and 3 show 31 topics and 19 topics for Stack Exchange and Reddit dataset. The topics are sorted in the descending order in terms of topic share.

We now present and discuss a subset of the topics common in both datasets along with representative examples for each. Through comparison across bothwe can see their importance aswell as range. Our goal is to illustrate the usefulness of looking at Q&A forums as a way to keep current with topics.We also want to show that using different platforms for analysis is useful given both the similarities and variations across platforms.

4.1 Neural Network & Deep Learning

We find multiple topics across these platforms discussing “neural network” and specifically “deep learning”:

(1) Stack Exchange: Neural Network (Layer Structure, Activation Func), Deep Learning (GAN-CNN), Deep Learning (RNN-LSTM)

Table 2: 31 Data Science related Topics in Stack Exchange

Topics (SubTopics) %

(Top LDA Keywords)

1 Problem Formulation 14.6%
problem,data,good,make,tri,work,gener,case,approach,model

2 Model Selection (Cross Validation) 5.91%
model,data,training,test,train,set,accuracy,dataset,valid,crossvalidation

3 Code Debugging 4.66%
code,tri,function,error,keras,model,python,tensorflow,work,run

4 Preprocessing (Pandas, Data Manipulation) 4.44%
column,row,data,valu,tabl,list,dataframe,index,creat,function

5 Classification (Algorithm Selection) 4.01%
data,label,model,classification,class,problem,classifi,labels,features,classifier

6 Deep Learning (GAN-CNN) 3.86%
image,images,network,cnn,input,layer,filter,pixel,size,object

7 Nueral Network (Layer Structure, Activation Func) 3.76%
layer,input,output,network,weight,neuron,neuralnetwork,activ,valu,function

8 Big-Data processing (Hadoop, Spark, NoSQL) 3.71%
data,python,r,spark,languag,packag,databas,tool,queri,write

9 Readings (ML) 3.66%
paper,find,deeplearning,machinelearning,read,methods,book,learning,algorithms

10 NLP (BOW, Word2vec) 3.31%
word,document,word,sentenc,text,vector,topic,word2vec,model,term

11 Optimization (Neural Network, SGD) 3.1%
weight,loss,training,gradient,error,learningrate,optim,chang,iter,valu

12 Feature Engineering (RF, DT) 3.04%
features,tree,featur,model,split,import,xgboost,decisiontre,randomforest,algorithm

13 Job/Education Advice 3%
datascience,machinelearning,work,project,cours,datascientist,compani,learn,field

14 Model Selection (Performance Evaluation) 2.96%
predict,valu,class,score,model,probability,accuracy,posit,threshold,metric

15 Regression/Correlation 2.84%
model,variabl,regression,linear,linearregression,valu,correlation,fit,coeffici,predict

16 Preprocessing (Categorical Encoding, Missing Data) 2.82%
valu,data,variabl,features,featur,categori,encoding,categor,attribut,column

17 Clustering 2.78%
cluster,clustering,point,distance,clusters,data,kmeans,similar,find,similarity

18 Temporal Analysis (Prediction, TimeSeries) 2.78%
data,time,day,predict,timeseries,model,event,month,year,hour

19 Visualization (Plotting) 2.65%
plot,data,point,valu,scale,line,normal,rang,show,exampl

20 NLP (Text Extraction, Scraping) 2.38%
text,extract,data,find,tag,exampl,document,search,match,dataset

21 Recommender System 2.33%
user,product,item,custom,rate,recommend,base,data,purchas,time

22 Libraries Installation 2.27%
file,run,memori,instal,gpu,orange,tri,load,comput,save

23 Classification (Imbalanced- MultiClass) 2.19%
class,data,sampl,svm,weight,tri,balanc,classifi,dataset,classifier

24 Dimensionality Reduction 2.09%
vector,matrix,dimens,data,pca,features,space,transform,origin,compon

25 Deep Learning (RNN-LSTM) 2%
input,lstm,sequence,output,model,rnn,network,timestep,sequenc,predict

26 Statistical Tests 1.84%
sampl,test,random,data,valu,number,differ,distribution,estim,error

27 Math discussion in DS (Formula) 1.57%
c,r,sum,frac,equat,function,text,theta,fracparti,alpha

28 Reinforcement Learning 1.53%
action,state,reward,valu,agent,polici,game,move,player,reinforcementlearning

29 Outlier in TimeSeries 1.48%
data,timeseries,detect,outlier,signal,time,frequenc,sensor,pattern,outlier

30 Generative Models (PGM-GAN-MLE) 1.33%
model,distribution,probability,estim,give,observ,px,distribut,function,gaussian

31 Social Network Modeling 1.11%
graph,node,patient,age,person,citi,countri,peopl,edg,data

Table 3: 19 Data Science related Topics in Reddit

	Topics (SubTopics)	%	(Top LDA Keywords)
1	Visualization (Graph, Colors)	10.05%	data,graph,color,line,make,show,map,chart,visual,point
2	Job/Education Advice	9.06%	data,job,work,datascienc,scientist,experi,compani,program,skill,interview
3	Statistical Analysis (Mean, Median, STD)	8.95%	number,data,averag,rate,high,year,popul,show,total,time
4	Google Analytics	6.58%	googl,page,site,user,search,data,facebook,websit,track,analyt
5	Deep Learning (CNN, GAN, LSTM)	6.28%	imag,train,layer,network,input,model,output,gener,weight,tri
6	Model Selection (Cross Validation)	5.85%	model,data,train,predict,featur,set,test,dataset,good,problem
7	Readings (Intro to DS)	5.63%	cour,learn,machinlearn,book,math,ml,good,start,python,understand
8	Programming Languages	5.48%	data,python,r,code,languag,tool,work,sql,learn,databas
9	Deep Learning (TensorFlow, Performance)	4.79%	gpu,tensorflow,run,model,comput,work,code,train,librari,kera
10	Readings (AI)	4.79%	research,ai,paper,human,work,ml,field,machinlearn,scienc,comput
11	Q&A in ML	4.71%	project,work,code,interest,find,idea,tri,good,write,make
12	Visualization (Links)	4.49%	data,sourc,tool,visual,www,creat,excel,make,map,tableau
13	Readings (ML)	4.46%	read,someth,good,start,point,tri,understand,find,work,make
14	Readings (NN ,RL)	4.38%	learn,paper,model,train,algorithm,neuralnetwork,network,gener,optim,work
15	Statistical analysis (Correlation, Causation)	4.13%	differ,group,thing,correl,peopl,point,factor,gener,level,effect
16	Math Discussion in DS (Explanation)	3.91%	distribut,function,valu,point,probabl,model,sampl,number,weight,gener
17	Data (External Links)	3.11%	www,org,html,pdf,jpg,png,c,imag,file,googl
18	Logic in Game	2.19%	game,win,play,move,number,time,chanc,random,bet,odd
19	Preprocessing (Pandas, Data Manipulation)	1.17%	data,valu,column,cluster,code,file,import,c,function,return

(2) Reddit: Deep Learning (CNN, GAN, LSTM), Deep Learning (TensorFlow, Performance)

“Neural network” is one of the popular classification algorithms that is widely used in the industry due to its high accuracy and scalablity. Recently a new version of it has revolutionized the practicality of such algorithm called “deep learning”. It has different varitions such as “Convolutional Neural Network (CNN)”, “Generative Adverserial Network (GAN)”, “Recurrent Neural Network (RNN)”, “Long Short-Term Memory (LSTM)”. You will see one example of post/submission/comment for each topic in the Table 4.

4.2 Model Selection

Model selection is one of the important topics in data science where it helps to find a better candidate among others in terms of a prespecified performance metric. As you can see, this topic has been broken down in different subtopics in both platforms:

(1) Stack Exchange: Model Selection (Cross Validation), Model

Selection (Performance Evaluation)

(2) Reddit: Model Selection (Cross Validation)

Table 4: Neural Network/Deep Learning Examples

StackExchange

"How to train neural network that has different kind of layers - If we have MLP then we can easily compute the gradient for each parameters, by computing the gradient recursively begin with the last layer of the network, —truncated— .

Topic: Neural Network (Layer Structure, Activation Func)

Howto use GAN for unsupervised feature extraction from images? I have understood how GAN works while two networks (generative and discriminative) compete with each other. —truncated—

Topic: Deep Learning (GAN-CNN)

"How do you get an RNN to learn in real time? <p>In order to train a recurrent neural network, you have to unfold it say 50 times and treat it like a chain of RNN cells. —truncated—

Topic: Deep Learning (RNN-LSTM)

Reddit

Convolutional to LSTM in Tensorflow? Hi there, does anyone have a good example of how to handle timeseries and convolutional networks? Essentially I am looking for the equivalent of Keras' 'TimeDistributed' wrapper —truncated—.

Topic: Deep Learning (CNN, GAN, LSTM)

Any legacy deep learning environment for a Windows computer with a NVS 300 GPU and CUDA 6.5? This is my main workstation at work. I found out that the NVS 300 cards aren't supported by CUDA beyond 6.5. —truncated—

Topic: Deep Learning (Tensorflow, Performance)

Table 5: Model Selection Examples

StackExchange

"Why not train the final model on the entire data after doing hyper-paramaeter tuning basis test data and model selection basis validation data? By entire data I mean train + test + validation. —truncated—

Topic: Model Selection (Cross Validation)

I am confused. You are doing a Logistic Regression and using r2_score to quantify the quality of your prediction? Logistic Regression is for binary classification, —truncated—.

Topic: Model Selection (Performance)

Reddit

Cross-Validation and Feature Selection I have about 150 samples 1000 features (ranked by their importance by Relieff score). My question is, what would be the best approach to: choose the hyper parameters, choose the optimal number of features to use and report the accuracy of my model using SVM and kNN —truncated—

Topic: Model Selection (Cross Validation)

You will see one example for each topic in the Table 5. This shows that "cross validation (CV)" is the most important part of this topic as it has a dedicated topic in both platforms.

4.3 Visualization

"Visualization" is one of the last steps in a data science project where you need to present your results in an understandable and easy way to your targeted audience. It needs to be more efficient and have important information about the point being discussed. This topic has been a point of interest in both platforms as well but in different ways:

(1) Stack Exchange: Visualization (Plotting)

(2) Reddit: Visualization (Graph, Colors), Visualization (Links)

Table 6: Visualization Examples

StackExchange

I don't understand what the point of the first plot of each pair is. You are plotting your count data in blue and a density in green - the density will integrate to 1 over (-Inf, +Inf) so it will be indistinguishable from a flat green line on the scale of counts.

Topic: Visualization (Plotting)

Reddit

Different shades of a single colour is actually an excellent way to present continuous data series though... It is readable by the colourblind, it allows you to spot tiny changes in values, —truncated—

Topic: Visualization (Graph, Color)

I manualy copied the data from [my youtube history](https://www.youtube.com/feed/history) and pasted it to a google sheets document. To visualize the data I also used google sheets. I used Photoshop Elements 13 to put the charts together. Interactive versions of the used charts and two more charts: —Links to charts—"

Topic: Visualization (Links)

In Table 6, you will see one example for each topic.

4.4 Math/Statistics

"Mathematics" and "statistics" are two fundamental components of data science where math helps with theorem proof and validity of the hypothesis as opposed to statistics where it deals with uncertainty and builds models for practical purposes. Hence, there are different angles about discussing these two topics in the two platforms. In Stack Exchange, the content is presented in a more technical theme like math formula and proof as opposed to Reddit where people are more interested in the basics and explanatory version of math/stats:

(1) Stack Exchange: Math discussion in DS (Formula), Statistical Tests

(2) Reddit: Math discussion in DS (Explanation), Statistical analysis (Correlation, Causation), Statistical analysis (Mean, Median, Sd)

In Table 7, you will see one example for each topic.

4.5 Job/Education Advice

Due to the rise in the need of data science in industry, people are trying to adjust their education to this new trend. So there are two types of topics being discussed: 1-what are the related courses or path to the education to data science? 2- How is the career change toward data science?

In Table 8, you will see one example for job/education topics.

5 Discussion

Our analysis of Stack Exchange and Reddit sheds light on the important topics that people have been talking about in the past and current data science industry. The topics reflect the multi-disciplinary nature of the field where computer science, mathematics/statistics and business work closely. Our findings support prior work that has identified topics at a higher level by providing examples of specific problems and issues within those topics to support learning. The platforms also provide real world examples and worked problems to assist with instruction. Similar to what has been reported earlier, our findings show that techniques such as “neural network/deep

Table 7: Math/Statistics Examples

StackExchange

"Yes, this is guaranteed by the Moore, Aronszajn theorem.

$K(x, y) = e - \|x - y\|$ is a positive definite kernel. This means it

is a symmetric function satisfying Í$_{ni}$

, $_{j=1}$ $c_i c_j K(x_i, x_j) \geq 0$ for all

$n \in N$, all $x_1, \cdot \cdot \cdot, x_n \in R_n$, and all $c_1, \cdot \cdot \cdot, c_n \in R$. —truncated—

Topic: Math discussion in DS (Formula),

The t-test has many assumptions. That dataset violates several of them: 1) Data should be sufficiently large (>30 independent points). 2) Data should be approximately normally distributed Given that the assumptions are violated, you can not expect to valid results.

Topic: Statistical Tests

Reddit

"**Hyperbolic function** In mathematics, hyperbolic functions are analogs of the ordinary trigonometric, or circular functions. The basic hyperbolic functions are the hyperbolic sine ""sinh"" (or), and the hyperbolic cosine ""cosh"" (), —truncated—

Topic: Math discussion in DS (Explanation),

AFAIK correlation implies causation in the sense that there must be *some* causal structure that relates the two correlated variables. It may not always be possible or practical to discover that structure, but your explanation is the kind of thing I’m talking about

Topic: Statistical analysis (Correlation, Causation)

They’re two completely different measures of central tendency. The median is less affected by outliers because if you increase the value of the highest value in the set, it changes the average but not the median.

Topic: Statistical analysis ((Mean, Median, Sd))

Table 8: Job/Education Advice Examples

StackExchange

Mathematics major for data science - So I’m a recent transfer 2nd year student from Computer Science major to Mathematics major. Though I do have a bit of an issue here. I can choose between the applied mathematics, pure mathematics and statistics concentrations. —truncated-

Topic: Education Advice

Reddit

Data Engineer aspiring to be a data scientist evaluating my career path. I currently work as a data engineer for a large company using not so interesting tech. Think SQL, SSIS, etc… I do a substantial amount of on my own python programming —truncated—

Topic: Job Advice

learning” are a prominent topic and so are “math / statistics” showing how important and fundamental they are to the understanding of data science.

5.1 Dominance of Deep Learning

Of all the major themes apparent in the technical topics,we find that “neural network/deep learning” discussions are the most popular overall in both platforms (9.62% topics share in Stack Exchange and 11.07% in Reddit) closely to “math/statistics” discussions. This is because of the fact that “neural network/deep learning” discussions alone are distributed among 5 topics across these two platforms:

(1) Stack Exchange: Neural Network (Layer Structure, Activation Func), Deep Learning (GAN-CNN), Deep Learning (RNN-LSTM)

(2) Reddit: Deep Learning (CNN, GAN, LSTM), Deep Learning (TensorFlow, Performance)

These topics are completely dedicated to “neural network/deep learning” concept. But there are other topics that also talk about them:

(1) Stack Exchange: Optimization (Neural Network, SGD), Generative Models (PGM-GAN-MLE)

(2) Reddit: Readings (NN ,RL)

These discussions boil down to a classification problem where the goal is to find a decision boundary to be able to classify the inputs. This process is usually done through formulating the problem as an “optimization” task. That is why most of model-based classification algorithms are translated to “optimization” tasks. One of the methods solving “optimization” problems is Stochastic Gradient Discent (SGD). The “generative models” topic talks about the models such as Probabilistic Generative Models (PGM), Generative Adversial Network (GAN), Maximum Likelihood Estimation (MLE). “GAN” is another “deep learning” algorithm that is generative and popular. Finally “neural network” also shows up in “reading” section along with “reinforcement learning”. As both are two algorithms solved through “optimization”. So these topics show the importance of “deep learning” along with “optimization” as a solving tool in data science.

5.2 Comparison of Platforms

Due to the strict rules enforced upon the content of “Stack Exchange”, it is expected to be dominated by data science related topics with technical theme but we found that “Job/Education Advice” is popular and here people discuss about their workplace environment, the current job market and about their career choices and education path to data science. On Reddit, an interesting set of topics is the “Readings” topic, where usually people direct others to the resources such as links, books, papers or dataset and it can be viewed as a good repository for users from basic to advanced level to have access to the more comprehensive source of knowledge. These topics in “Reddit” are: Readings (Intro to DS), Readings (AI), Readings (ML), Readings (NN ,RL) and Q&A in ML.
The tone and content of “Stack Exchange” is more technical and algorithmic than those of “Reddit”. The exception is the topic of “Math/Statistics” are more found in “Reddit” than “Stack Exchange” even though they are the fundamental building blocks of data science. It shows that people on “Reddit” are at the basic level as opposed to users in “Stack Exchange”. That also partially explains why the tone of explaining them is more descriptive in “Reddit” than technical in “Stack Exchange” where “formula” has their own separate topic, Math discussion in DS (Formula).“Reddit” is more focused on data analytics as opposed to “Stack Exchange” is geared toward data science. The first three dominant topics of each platform, excluding Job/Education Advice, prove the point that in “Reddit”, Visualization, Statistical Analysis and Google Analytics are more about analytics where as Problem Formulation, Model Selection and Code Debugging are elements of building models based on the given data and that is one of data scientist’s responsibilities. The prevalence of “Job/Education Advice” shows that people have found “Reddit” an easier platform to exchange and share experience and ask their questions.

5.3 Comparison with Curricula Guidelines

Our findings show that the topics discussed in the two communities mirror the topics that are deemed important in the Data Science curricula guidelines we reviewed [1]. Overall, the information gained from the communities is good representation of the topics needed to gain expertise. There were some differences though. The curricula framework focuses on process skills more than the topics discussed in the communities and it also includes communication and teamwork skills guidance which is not discussed in the communities.
Therefore, overall, we see the information of the communities does

a good job of assisting with specific technical topics and problems that go with them rather than providing a well structured way to learn Data Science. This we believe still remains the purview of the formal education system. More critical, from the perspective of preparing the future workforce, is the lack of coverage on the platforms of the topic of data ethics. Most recent studies and report have argued that data ethics is a topic of extreme importance within data science and therefore it needs to be covered. Most curricula guidelines now include specific topics that need to be covered within the area including privacy and algorithmic bias but they are not discussed on the platforms.

5.4 Limitations

There are several limitations to this work including the data processing barrier and the technical know-how needed for data analysis. Our sample selection was selective and more research is needed to examine the usefulness of other sites as different sites support different kinds of information sharing [6] and interaction leading to differences in knowledge sharing [12]. Furthermore, we realize that although our work is important in terms of evaluating a resource that can assist with keeping up with new topics and tools in Data Science, for this information to actually get used there has to be an easy to use interface that can not only allow for topics to be seen but examples and questions to be selected from the online platform.

6 Conclusion

In this paper, we present an approach for uncovering Data Science topics in two popular Q&A platforms, “Stack Exchange” and “Reddit”. Our methodology is based on LDA, a widely-applied statistical topic model, which discovers topics from the textual. We used this method to uncover the topics, which allows us to understand the popular discussions in both platforms. Our analysis provides an insight which platform is better suited for which purpose. It also shows the current and dominant topics and concerns of the community. Our methodology can also be applied to other online platforms but different kinds such as Twitter, Instagram, web portals, blogs, and forums and compare the type of topics and messaging across these platforms and determine which one is suited for which purpose. Therefore, it provides a framework to stay updated with the current need of industry and academia. One of the simplest ways in which these platforms can be used by instructors is through their own participation on them even if that is lurking and monitoring to know what’s going on. There is also the possibility to integrate participation on these platforms within courses, similar to using GitHub for instance.

Acknowledgments

This work is supported in part by NSF Grants DUE-1712129 & DUE-1707837. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] 2013. Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science. (Jan 2013). https://doi.org/10.1145/2534860

[2] 2018. The EDISON Data Science Competence Framework. http://edisonproject.eu/edison/edison-data-science-framework-edsf

[3] 2019. Reddit API. http://www.reddit.com/dev/api

[4] 2019. Reddit Data Dump Website. http://files.pushshift.io/reddit/

[5] 2019. StackExchange Data Dump Website. https://archive.org/download/stackexchange

[6] Saif Ahmed, Seungwon Yang, and Aditya Johri. 2015. Does online q&a activity vary based on topic: A comparison of technical and non-technical stack exchange forums. In Proceedings of the Second (2015) ACM Conference on Learning@ Scale. ACM, 393–398.

[7] Robert Ball, Linda Duhadway, Kyle Feuz, Joshua Jensen, Brian Rague, and Drew Weidman. 2019. Applying Machine Learning to Improve Curriculum Design. In Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19). ACM, New York, NY, USA, 787–793. https://doi.org/10.1145/3287324.3287430

[8] Austin Cory Bart, Dennis Kafura, Clifford A. Shaffer, and Eli Tilevich. 2018. Reconciling the Promise and Pragmatics of Enhancing Computing Pedagogy with Data Science. In Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE '18). ACM, New York, NY, USA, 1029–1034. https://doi.org/10.1145/3159450.3159465

[9] Anton Barua, Stephen W Thomas, and Ahmed E Hassan. 2014. What are developers talking about? an analysis of topics and trends in stack overflow. Empirical Software Engineering 19, 3 (2014), 619–654.

[10] DavidMBlei, AndrewY Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research 3, Jan (2003), 993–1022.

[11] Boots Cassel and Heikki Topi. 2015. Strengthening data science education through collaboration. InWorkshop on Data Science EducationWorkshop Report, Vol. 7. 27.

[12] Bushra Chowdhury, Aditya Johri, Dennis Kafura, and Vinod Lohani. 2019. Be Constructive: Learning Computational Thinking Using Scratch™ Online Community. In International Conference on Web-Based Learning. Springer, 49–60.

[13] Sarah Dahlby Albright, Titus H. Klinge, and Samuel A. Rebelsky. 2018. A Functional Approach to Data Science in CS1. In Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE '18). ACM, New York, NY, USA, 1035–1040. https://doi.org/10.1145/3159450.3159550

[14] Andrea Danyluk, Paul Leidig, Lillian Cassel, and Christian Servin. 2019. ACM Task Force on Data Science Education: Draft Report and Opportunity for Feedback. In Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19). ACM, New York, NY, USA, 496–497. https://doi.org/10.1145/3287324.3287522

[15] Stuart Geman and Donald Geman. 1987. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In Readings in computer vision. Elsevier, 564–584.

[16] Scott Grant and James R Cordy. 2010. Estimating the optimal number of latent

concepts in source code analysis. In 2010 10th IEEE Working Conference on Source Code Analysis and Manipulation. IEEE, 65–74.

[17] Mark E. Hoffman, Paul V. Anderson, and Magnus Gustafsson. 2014. Workplace Scenarios to Integrate Communication Skills and Content: A Case Study. In Proceedings of the 45th ACM Technical Symposium on Computer Science Education (SIGCSE '14). ACM, New York, NY, USA, 349–354. https://doi.org/10.1145/2538862.2538916

[18] Courtney Hsing and Vanessa Gennarelli. 2019. Using GitHub in the Classroom Predicts Student Learning Outcomes and Classroom Experiences: Findings from a Survey of Students and Teachers. In Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19). ACM, New York, NY, USA, 672–678. https://doi.org/10.1145/3287324.3287460

[19] Aditya Johri and Seungwon Yang. 2017. Scaffolded help for learning: How experts collaboratively support newcomer participation in online communities. In Proceedings of the 8th International Conference on Communities and Technologies. ACM, 149–158.

[20] Karen Sparck Jones. 1997. Readings in information retrieval. Morgan Kaufmann.

[21] David G. Kay. 1996. Bandwagons Considered Harmful, or the Past As Prologue in Curriculum Change. SIGCSE Bull. 28, 4 (Dec. 1996), 55–58. https://doi.org/10.1145/242649.242666

[22] Xiaomo Liu, G Alan Wang, Aditya Johri, Mi Zhou, and Weiguo Fan. 2014. Harnessing global expertise: A comparative study of expertise profiling methods for online communities. Information Systems Frontiers 16, 4 (2014), 715–727.

[23] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design lessons from the fastest q&a site in the west. In Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2857–2866.

[24] Andrea Manieri, Steve Brewer, Ruben Riestra, Yuri Demchenko, Matthias Hemmje, Tomasz Wiktorski, Tiziana Ferrari, and Jeremy Frey. 2015. Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists. In 2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom). IEEE, 588–593.

[25] Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. (2002).

[26] A. McGettrick, M. D. Theys, D. L. Soldan, and P. K. Srimani. 2003. Computer engineering curriculum in the new millennium. IEEE Transactions on Education 46, 4 (Nov 2003), 456–462. https://doi.org/10.1109/TE.2003.818755

[27] M. Milosz and E. Lukasik. 2015. Reengineering of computer science curriculum according to technology changes and market needs. In 2015 IEEE Global Engineering Education Conference (EDUCON). 689–693. https://doi.org/10.1109/EDUCON.2015.7096044

[28] Sukanya Kannan Moudgalya, Kathryn M. Rich, Aman Yadav, and Matthew J. Koehler. 2019. Computer Science Educators Stack Exchange: Perceptions of Equity and Gender Diversity in Computer Science. In Proceedings of the 50th ACM Technical Symposium on Computer Science Education (SIGCSE '19). ACM, New York, NY, USA, 1197–1203. https://doi.org/10.1145/3287324.3287365

[29] Seyed Mehdi Nasehi, Jonathan Sillito, Frank Maurer, and Chris Burns. 2012. What makes a good code example?: A study of programming Q&A in StackOverflow. In 2012 28th IEEE International Conference on Software Maintenance (ICSM). IEEE, 25–34.

[30] Engineering National Academies of Sciences, Medicine, et al. 2018. Data science for undergraduates: Opportunities and options. National Academies Press.

[31] James Robergé and C. R. Carlson. 1997. Broadening the Computer Science Curriculum. In Proceedings of the Twenty-eighth SIGCSE Technical Symposium on Computer Science Education (SIGCSE '97). ACM, New York, NY, USA, 320–324. https://doi.org/10.1145/268084.268206

[32] Jean Michel Rouly, Huzefa Rangwala, and Aditya Johri. 2015. What Are We Teaching?: Automated Evaluation of CS Curricula Content Using Topic Modeling. In Proceedings of the Eleventh Annual International Conference on International Computing Education Research (ICER '15). ACM, 189–197.

[33] Mehran Sahami, Alex Aiken, and Julie Zelenski. 2010. Expanding the frontiers of computer science: designing a curriculum to reflect a diverse field. In Proceedings of the 41st ACM technical symposium on Computer science education. ACM, 47–51.

[34] Jeffrey S. Saltz, Neil I. Dewar, and Robert Heckman. 2018. Key Concepts for a Data Science Ethics Curriculum. In Proceedings of the 49th ACM Technical Symposium on Computer Science Education (SIGCSE '18). ACM, New York, NY, USA, 952–957. https://doi.org/10.1145/3159450.3159483

[35] Chris B. Simmons and Lakisha L. Simmons. 2010. Gaps in the Computer Science Curriculum: An Exploratory Study of Industry Professionals. J. Comput. Sci. Coll. 25, 5 (May 2010), 60–65. http://dl.acm.org/citation.cfm?id=1747137.1747147

[36] G. Subrahmanyam. 2009. A Dynamic Framework for Software Engineering Education Curriculum to Reduce the Gap between the Software Organizations and Software Educational Institutions. In 2009 22nd Conference on Software Engineering Education and Training. 248–254. https://doi.org/10.1109/CSEET.2009.8

[37] Chade-Meng Tan, Yuan-Fang Wang, and Chan-Do Lee. 2002. The use of bigrams to enhance text categorization. Information processing & management 38, 4 (2002), 529–546.

[38] Hon Jie Teo and Aditya Johri. 2014. Fast, functional, and fitting: expert response dynamics and response quality in an online newcomer help forum. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 332–341.

[39] Hon Jie Teo, Aditya Johri, and Vinod Lohani. 2017. Analytics and patterns of knowledge creation: Experts at work in an online engineering community. Computers & Education 112 (2017), 18–36.

[40] Bogdan Vasilescu, Alexander Serebrenik, Prem Devanbu, and Vladimir Filkov. 2014. How social Q&A sites are changing knowledge sharing in open source software communities. In Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 342–354.

[41] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In Proceedings of the 26th annual international conference on machine learning. ACM, 1105–1112.