Group 7 Blog Post

Kacey Cleveland, Nicholas Larsen, Chi Zhang, Vincent Lu

Introduction and Main Goal

The Stack Exchange network is a popular online conglomeration of 133 question-and-answer websites, with each website devoted to a specific field of interest within the academic and social sciences. Hosting over 100 million users each month, Stack Exchange websites provide platforms for users of all backgrounds to ask field-specific questions and gain constructive feedback from experts in both professional and academic spheres. As these websites grow in popularity, so does the amount of data containing user information. This data is freely available on an online archive hosted by the Stack Exchange Data Dump. Our motivating questions are directed at both how these online communities grow and evolve over time, and how we could visualize these findings. To answer these questions, we focused on the Stack Exchange categories devoted to the fields of chemistry, economics, engineering, data science, computer science, and biology.

Our team has conducted an in-depth analysis of over 170,000 instances of user-contributed content from the Stack Exchange network, with the main goal of seeking to understand and summarize user behavior over time by modeling our select Stack Exchange websites as temporal networks and using natural language processing and machine learning to analyze user posts. We have succeeded in identifying category-specific syntax by analyzing post content and have fitted over a dozen classification models to datasets containing tens of millions of entries. Co-currently, we have also developed an app that models each category as a dynamic network, allowing users to compute basic network statistics and track user activity over time. This blog post begins with a showcase and description of the Shiny app, followed by a “big picture” analysis of user-behavior over time using our network models, and concludes with a summary of our natural language processing and machine learning results.

Shiny Application

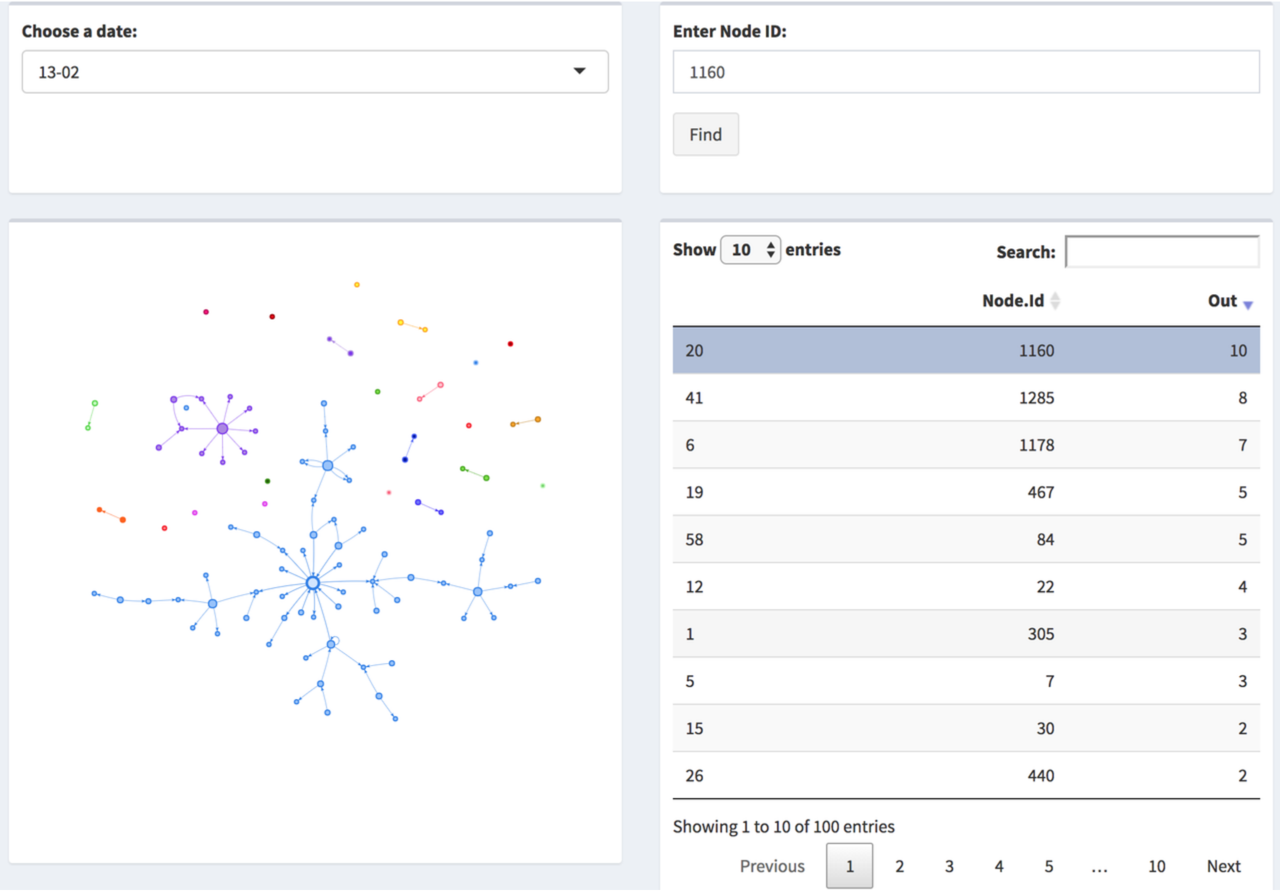
The Shiny application is the main visual aid for our project in addition to the machine learning analysis. We have published a online version of our application that can be found here: https://stor320stackexchange.shinyapps.io/deliver/

Instructions for App Usage:

First, select the category under “Stack Exchange Data” in the left sidebar menu. Next, choose how you wish to summarize the data, either as a dynamic network (“Networks”) or as an interactive wordcloud (“Word cloud”) generated by user posts content.

The ”Networks” Option:

The network panel has five components: “Choose a date”, “Enter Node ID”, the network graph visual, the corresponding data frame, and the “Activity Graph” (not shown in below screenshot). Date selection allows users to visualize the network of user-posts interactions at intervals of one month. The node ID input allows app users to track a specific Stack Exchange user in the network over time; it can be achieved by either typing in a number or selecting an observation in the generated data frame (lower right-hand corner). The network graph is an interactive model with each node colored by its cluster and marked with a node ID and an “out” degree. “Out” degree refers to the number of outward directed edges attached to the node. Physically, these represent the number of responses a given user wrote to other users’ posts. A user of our application can zoom in and zoom out to get a comfortable view of the graph, as well as click and drag nodes. The data frame can be used to search for highly active users by sorting by the out degrees. At the bottom of the screen, we have provided an “activity graph”, which summarizes the number of posts over the entire timeline of the category’s existence. A user of our app can use this chart as a reference for certain months of interest. For example, suppose for the Chemistry data, one was interested in the month corresponding to the highest peak in volume of posts. By looking at the activity graph and identifying the month “2017-03”, one can now select the corresponding date in the “choose a date” box and view the network for that month.



Choose a date: 13-02

Enter Node ID: 1160

Show 10 entries

Search:

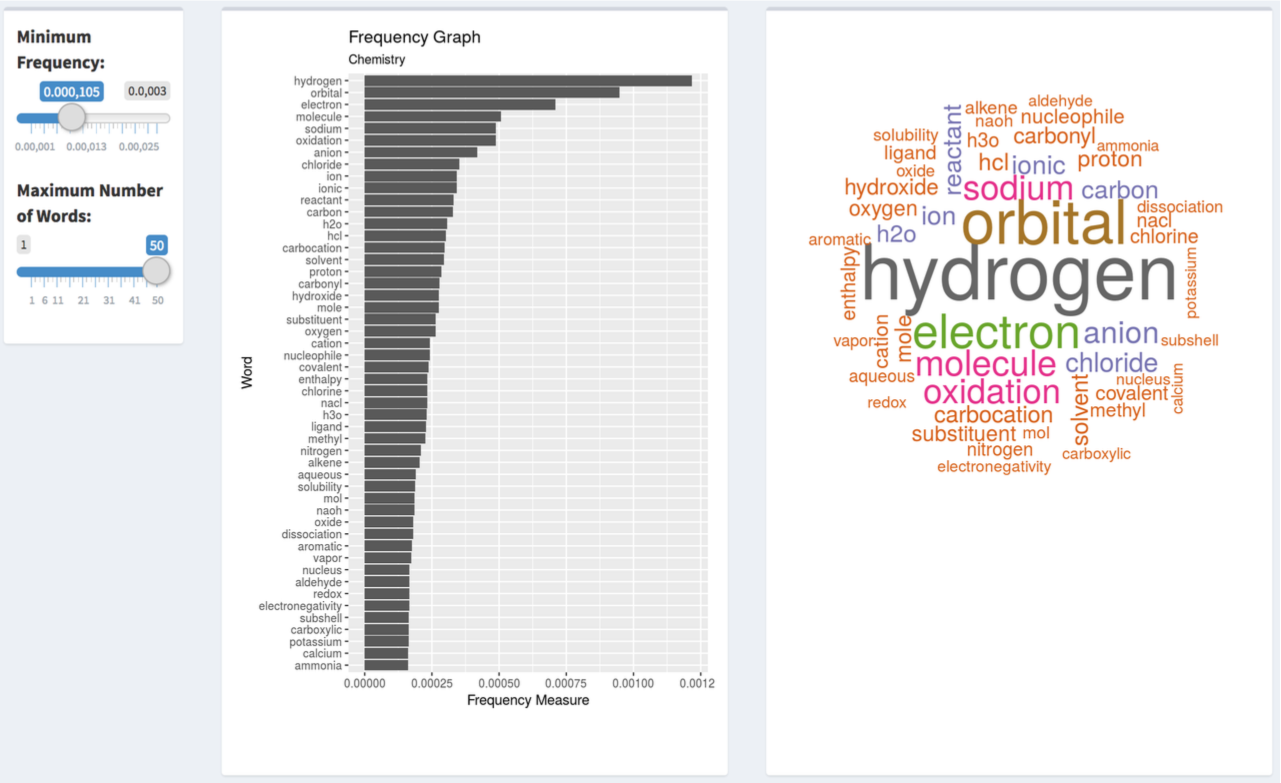
	Node.Id	Out
20	1160	10
41	1285	8
6	1178	7
19	467	5
58	84	5
12	22	4
1	305	3
5	7	3
15	30	2
26	440	2

Showing 1 to 10 of 100 entries

The network panel

The ”Word cloud” Option:

The word cloud panel consists of three boxes. By changing the two sliders in the leftmost box (“Minimum Frequency” and “Maximum Number of Words”), the frequency graph and the word cloud graph update accordingly. The “Minimum Frequency” slider filters out the words across all posts contained in the category that occur with a frequency below the value specified in the slider. The words surviving the filtering process are presented in “Frequency Graph.” The maximum number of words constrains the number of words shown in the word cloud.



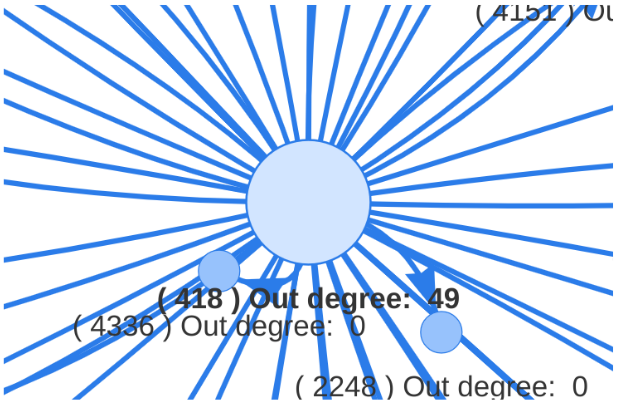
Minimum Frequency: 0.000,105

Maximum Number of Words: 50

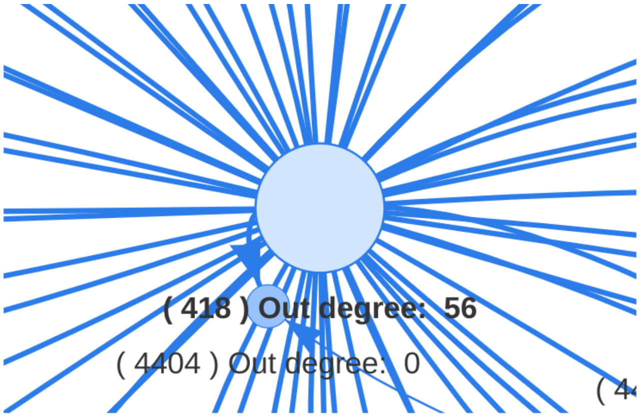
The Chemistry word cloud panel

Modeling and Visualizing Temporal Networks

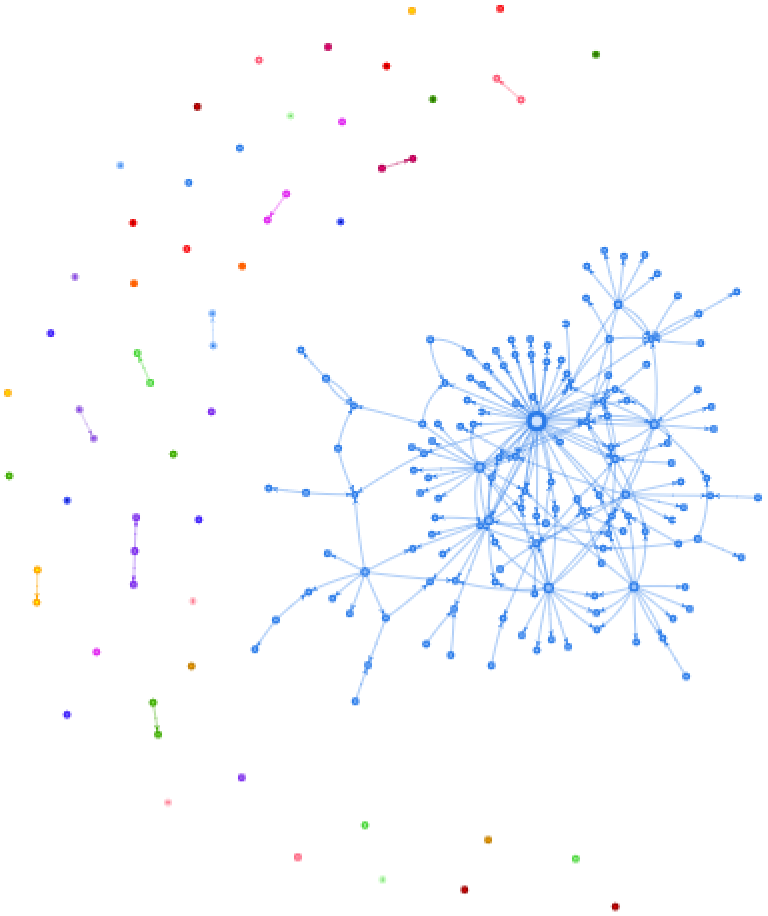
The highlight of our Shiny application is how it models post activity in the Stack Exchange communities for chemistry, economics, engineering, data science, computer science, and biology as dynamically evolving networks and visualizes the result. Using frameworks for graph visualization and network analysis, we modeled each of the six Stack Exchange categories as interactive graphs that includes features such as node location and node tracking over time. Nodes represent particular Stack Exchange users on a particular month. Edges are drawn between users if one comments or replies to another post. With this modeling scheme, we are able to simultaneously track user activity from month to month and visualize the networks as they evolve over time. Below is an image of one of the top active Chemistry users (represented as User 418) at the beginning of 2014 between January and February. Underneath shows User 418's relative position in the entire network for January 2014.



User 418 on January 2014



User 418 on February 2014



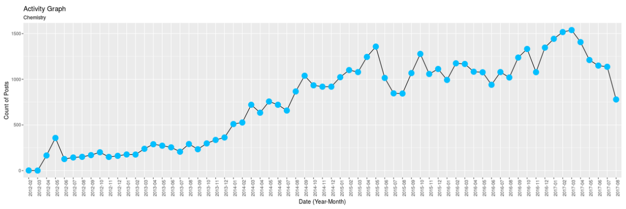
The highlighted node is user 418

In the above graph, we define three sets of node-groups. There is the main group (blue) that comprises the majority of the Chemistry network at this time and includes the most active users and popular posts. The smaller bodies of nodes represent users who commented on less or more "niche" popular posts. Edge-less nodes represent posts that attracted zero comments or answers.

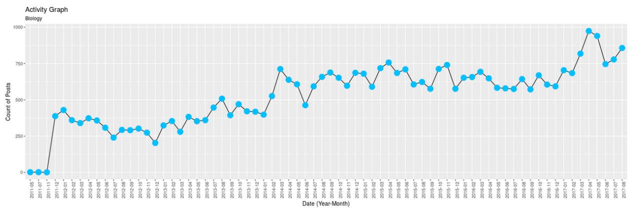
We observed this pattern of three groups for each of our six categories across the entire timeline. Our best guess as to why this pattern exists is that the large group reflects how sub-topics are often inspired by other issues. As users contribute answers, other questions can arise, spawning entirely new threads or creating "chains" of Q&A's as seen in the above figure. To return to the figure where user 418 is highlighted, you can observe the "chains" as nodes connected to user 418 in a straight line. Some of these chains spawn new threads of conversation, represented by "star-bursts" of node-edge patterns. The smaller groups perhaps represent users who contribute only on a semi-active basis, or perhaps comment frequently but only for certain sub-fields or "niche issues." A "niche issue" could be specific area of expertise or academic problem. Users with the same or similar knowledge-bases would be likely to comment, however, because the post is so obscure, the chains of comments and answers is not strong enough to connect to the larger group. Edge-less nodes are more difficult to interpret; they could represent areas where very few people have the background to answer, or they may simply be questions that have not received an answer at that particular time.

"Big Picture" Thoughts

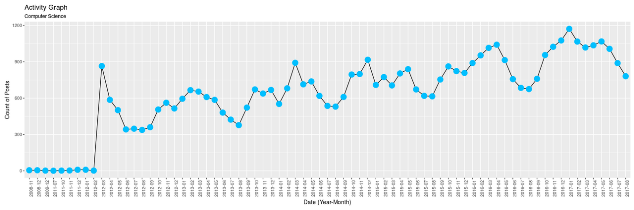
Here, we take a step back from our month-to-month analysis frame to consider the "big picture" of user activity in terms of post volume over all time. To do this, we include plots for each category that show number of posts per month since the categories' creation.



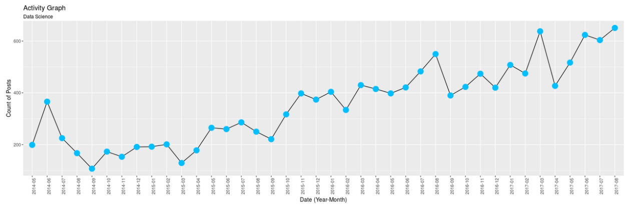
Chemistry



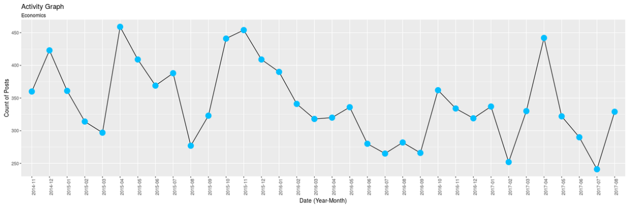
Biology



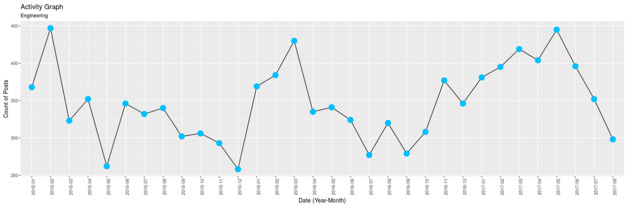
Computer Science



Data Science



Economics



Engineering

Each category exhibits a unique pattern of evolution with respect to monthly volume of posts. We can see that for Chemistry, Biology, Computer Science, and Data Science there is almost consistent positive growth occurring at different rates. We can see that Data Science, although a relatively recent category with a first recording date of May 2014, shows sign of huge growth where between January 2015 and January of 2017 where it more then doubled in terms of post volume (from 200 posts to 400+ posts). The more "elderly" categories (Chemistry, Biology, and Computer Science) have slower growth rates in terms of percentage growth, however all exhibit consistent positive growth as well. Economics and Engineering appear to have frequent variance in terms amounts of activity from month to month and do not show the same consistent positive growth as seen in the other categories.

It is interesting to note when periods of great increase or decrease in post volumes occur. Consider the Computer Science category; we can see during the Summer months that post volume noticeably dips. Our group hypothesizes that perhaps a large group of the Computer Science community are academics and students that take a leave of absence due to breaks during the Summer.

The reader should also note that not all of the respective websites for these categories were created at the same time (for example, the Computer Science website began in 2008, whereas the Data Science website was not created until 2014). As a result, we can at least partially attribute the size of the categories, with respect to post volume, to lifespan. Consider the table below:

Table 1:

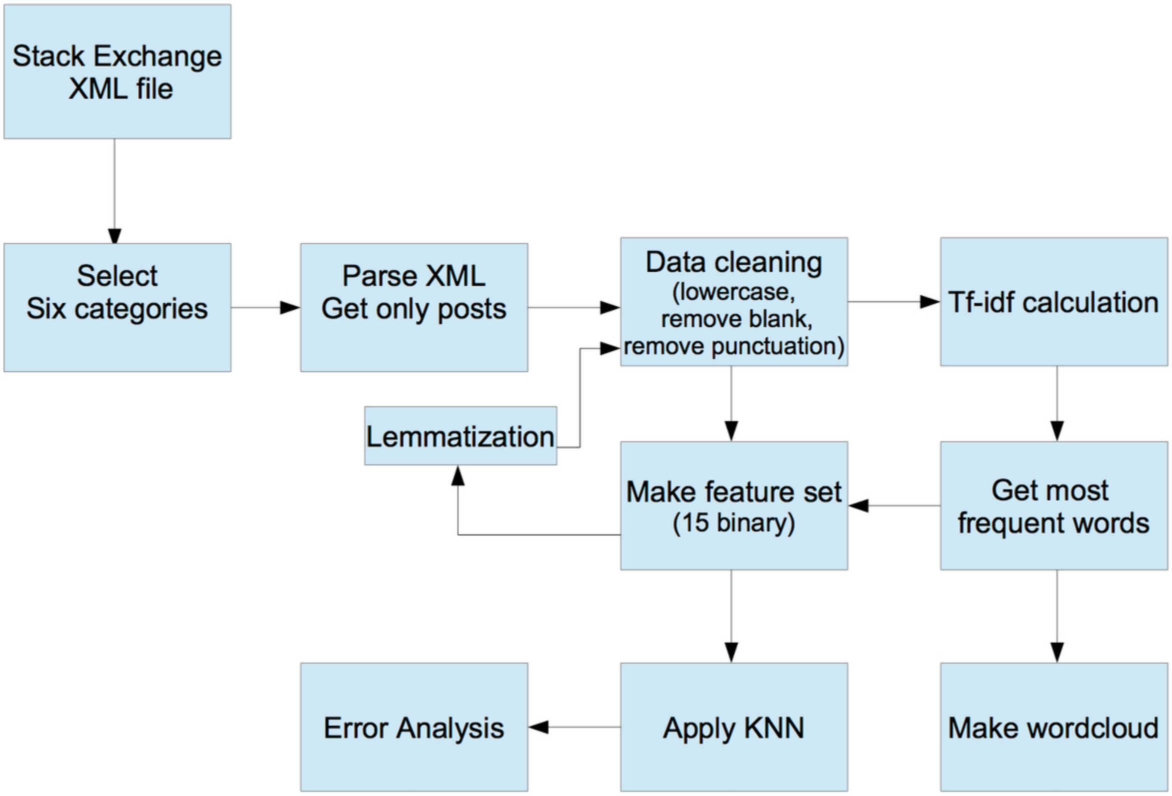
Computer Science ("*cs*", 2008)	47,812
Biology ("*bi*", 2011)	37,703
Chemistry ("*ch*", 2012)	50,547
Data Science ("*da*", 2014)	14,159

Economics (“*ec*”, 2014)	11,570
Engineering (“*en*”, 2015)	11,033

Each entry on the left denotes category, abbreviation, and “date of birth”; the right-hand side contains total posts over all time. While the trend is not necessarily linear (Chemistry is not the oldest category but boasts the largest number of posts), there is a clear decrease for post volumes in categories started after 2013.

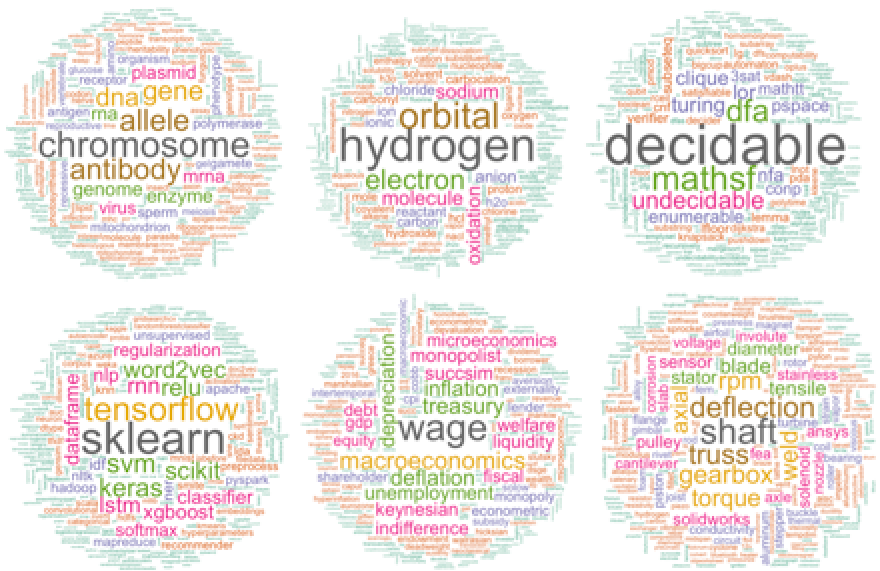
Natural Language Processing and Machine Learning

Using the contents of the over 170,000 user-contributed posts from our six Stack Exchange websites (see above table), we used natural language processing methods to identify words that are both relevant and unique to their host-websites, allowing us to clearly differentiate between websites simply by analyzing users’ vocabulary. We then used this information to build a suite of binary k-Nearest Neighbors classifiers to see if we could predict a given post’s host-website. The following is a diagram that outlines our process:



Natural Language Processing: Results

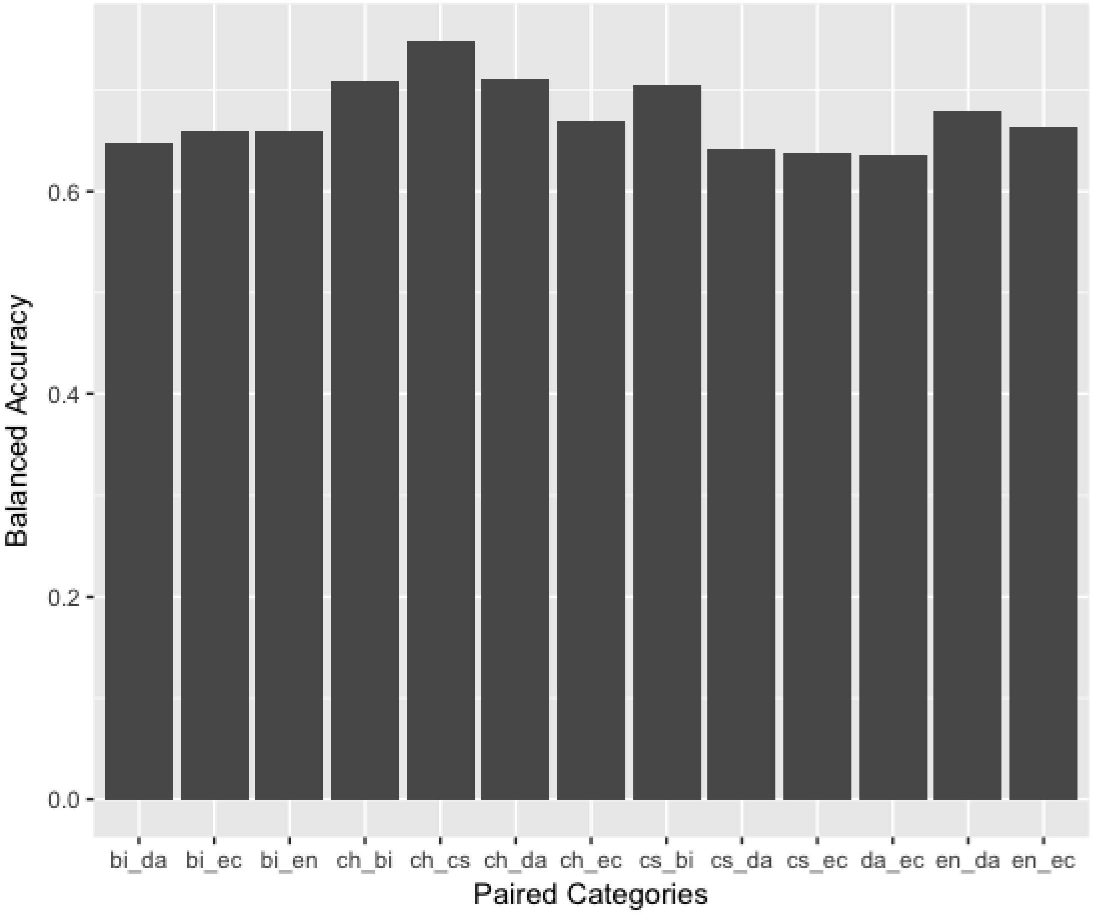
Our results from our text-analysis of the user posts were quite strong. We were able to successfully identify high-frequency words in each category that were not only unique to each category, but also were highly reflective of the respective category. For the purposes of this blog post, we will not delve into the technical aspects of how we parsed the posts (see Process Book). The following set of wordclouds summarizes our results quite nicely, and can also be generated in real-time by the Shiny app. The size of each word corresponds to the unique “importance scores” we assigned. Notice how it is very easy to guess which category these words come from.



K-Nearest Neighbors: Results

For each category, we extracted 500 words with the top “importance scores”, calling them *keywords*. For a given pairing of posts, we counted the number of times each keyword appeared in each post, resulting in a 1x1000 feature vector of integers per post. We combined the results into a dataframe and labeled each frequency vector by the relevant category. With our method of feature selection, the number of columns in the feature set grows quickly with the number of categories, which can result in datasets too large to perform any significant calculations in a relatively short amount of time. For this reason, we chose to fit binary classifiers to each pairing of the six category. For each category pairing, we divided the feature data into 70/30 training and test sets and fit k-NN models for k = 5, 7, and 9 to the training set using 3-fold cross validation.

As discussed in the *“Big Picture” Thoughts* section, the number of posts per category varied largely. As a result, our models were often biased in favor of whichever category contained the larger number of posts. To correct this, we computed the balanced accuracy scores of each model’s prediction on the held-out test set. The results are summarized here:



Reference "Table 1" for abbreviations meanings.

Across all models, the balanced prediction scores vary 63-74%. Some of these prediction errors likely resulted from flaws intrinsic to how we defined our feature set. In some cases, posts of either category did not contain any of the top 500 keywords. If enough of these “zero-posts” belong to a certain category, then the k-NN algorithm may incorrectly classify any post close in Euclidean distance to these zero-posts.

Summary

We have shown that modeling the Stack Exchange websites as dynamic networks provides nuanced insights into how users of each category interact, as well as allowing for big picture analysis of how these categories behave at certain times of the year and of overall growth trends. Our work has been finalized in the form of a Shiny application, which permits users to freely explore these networks.

We have also successfully parsed all posts in each category and identified key words that can be used to identify the category of a given post. To show this, we trained 15 binary k-NN classifiers on data that uses these important words with modest success. We hope to improve our classification approach by adapting our method of computing frequency vectors to account for “zero-posts” cases and perhaps use more sophisticated models such as Random Forests.