# Digging into the FAQs: analyzing user queries to inform service and system development

Ask for a logo!

Marcella E. Barnhart and Kevin A. Thomas
Lippincott Library of the Wharton School, University of Pennsylvania

Ask for a logo!

## Introduction

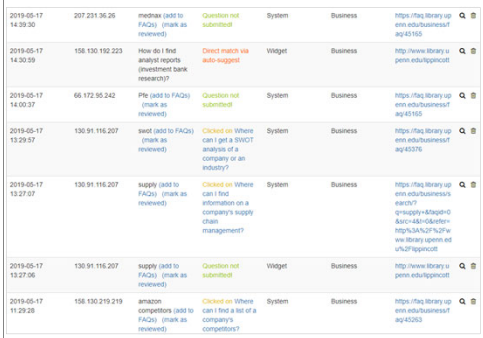
The Lippincott Library's Business FAQ, a knowledge base of 300+ questions that receives over 100,000 views a year, is an integral part of Lippincott's reference service offering. This poster reports on an analysis of user query data from the Business FAQ. User queries vary widely in their structure from long phrases like *'trust building strategies in inter-organizational negotiations'* to company names like *Nike*. Using machine learning techniques to analyze the data, we address a range of questions including:

- What types of queries are more likely to be successful as defined by the user selecting a question from the existing knowledge base?
- What user behaviors are common for successful versus unsuccessful queries?
- Does query complexity impact the success or failure of the match?
- Is there content we could add to the FAQ to improve the users' success?

## Data

Business FAQ content is currently hosted on Springshare's LibAnswers platform. The Query Spy function captures the exact query entered, as well as whether the query led to the questioner selecting an answer from within the system.

- LibAnswers maintains 6 months of query data. Data used in this analysis were pulled from 8/31/2018 through 3/28/2019.
- To focus on patron behavior, Lippincott's business librarians identified IP addresses of machines they typically operate, and data were filtered to remove all searches from those addresses.



The **Status** field reflects the user's next step as:

- **Clicked on question** when the user navigated to a FAQ on the Search Results page
- **Match via auto-suggest** when the user navigated to an FAQ in the suggested results before submitting the query
- **Not Submitted** when the patron did not select a result,
- **Question Submitted** when the user chose the *Submit Your Question* option.

This analysis considers a query successful if the patron navigated to an FAQ (either of the first two **Status** values) and unsuccessful if the patron submitted a query but did not identify a result (either of the last two **Status** values).

## Methods

**N-gram Conversion**

Queries were converted, via a series of steps, into terms consisting of 1 to 3 consecutive words as follows:

1. Remove numbers
2. Remove punctuation (including apostrophes and dashes)
3. Convert all letters to lower case
4. Extract one-word, two-word, and three-word terms (spaces replaced by underscore)
5. Remove stop words identified by the Snowball stemmer project as implemented in the R package SnowballC (Bouchet-Valat 2019)
6. Perform stemming according to algorithms from the Snowball Stemming Library as implemented in the R package SnowballC (Bouchet-Valat 2019)
7. Remove one-letter terms

(Note that Query Spy exports the ampersand character as *&*, which the above process converts to *amp*. As a result, terms such as *M&A* and *S&P* appear as *mampa* and *sampp* in this analysis.)

## Methods

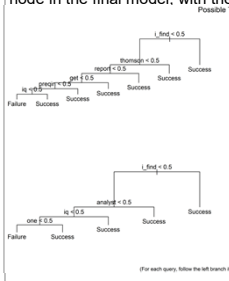
**k-Means Clustering**

k-means analysis groups similar items, in this case terms from related queries. Five of the top query terms by group are shown below.

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| am_i | a_compani | financi | estat | analyst_reports_invest |
| am_i_hav | aapl | financial_timindustri | | bank |
| difficulti | access_cb | time | real | bank_research |
| difficulty_us | access_cb_insight | | real_estdo_i | |
| difficulty_using_thomson | access_to_cas | | | do_i_find |
| having_difficulti | account | | | find_analyst |
| having_difficulty_us | acquisit | | | find_analyst_report |
| i_hav | activ | | | how_do |
| i_having_difficulti | advertis | | | how_do_i |
| one | affili | | | i_find_analyst |
| thomson | alum | | | invest |
| thomson_on | america | | | investment_bank |
| use | american | | | investment_bank_research |
| using_thomson | amp | | | reports_invest |
| using_thomson_on | analysi | | | reports_investment_bank |
| why_am | analyst | | | research |
| why_am_i | and_acquisit | | | |
| | and_report | | | |
| | annual | | | |
| | annual_report | | | |
| | appl | | | |

**Random Forest Analysis**

In a random forest model, an algorithm randomly generates a large number of decision trees, each of which predicts an outcome. In this case, the trees used the presence or absence of groups of query terms to predict the success or failure of a query. For this analysis, 6376 total query records were divided into 5100 records used for training the model and 1276 records for testing the trained model's performance. Multiple variations on random forest and extra-random forest models were trained and tested for comparison. The ultimate model, chosen for its stability and test accuracy across both failed and successful searches, is a random forest consisting of 350 trees, validated through repeated cross-validation of four folds and four repetitions. Between 7 and 12 randomly-selected terms were available for splitting at each tree node in the final model, with the optimal model allowing 12 terms.




Term importance is measured by the difference in gini coefficient, which measures how dissimilar data are, with versus without the given term. Terms with higher gini importance scores have more impact on results.

**Search Behavior & Follow-up Searches**

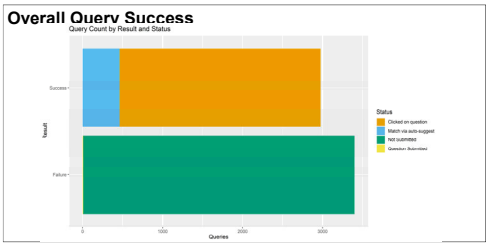
Queries do not always stand alone, but sometimes form part of a larger, iterative process. This research defines a search as a series of one or more queries from the same IP address that occurred within 30 minutes of another query. Although SpringShare does not provide information that follows a patron's entire search process, we can use a query's referring URL to identify when a search originated from a given FAQ. Comparing these instances to the quantity of visits to the same FAQ during the same timeframe (*Export Views by Month* report), we can arrive at a follow-up search rate.

**Regression**

Regression analysis can help indicate which factors contribute toward or detract from query success. This analysis uses a logistic regression model to estimate the relative odds of query success vs. failure under a given set of conditions. For search queries, which tend to be brief and need not be formatted as a sentence, typical text complexity measures may not be highly valid. See Liu, Croft, et al for research assessing the validity of several approaches, which found a substantially valid hybrid approach.

## Results

**Overall Query Success**

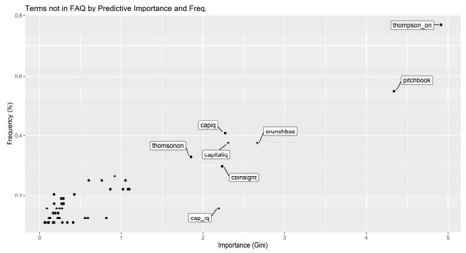



## Results & Discussion

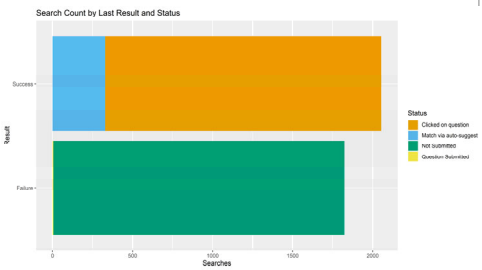
**Predictive Importance & Term Frequency**

| Term | Importance (Gini) |
|---|---|
| analyst | 38.98 |
| report | 33.27 |
| euromonitor | 30.28 |
| industri | 30.08 |
| economist | 29.09 |
| analyst_report | 26.94 |
| preqin | 25.85 |
| thomson | 25.44 |
| one | 18.52 |
| i_find | 18.41 |
| gartner | 17.37 |
| iq | 16.89 |
| factiva | 16.31 |
| proquest | 15.25 |
| ers | 15.04 |
| capital_iq | 15.00 |
| can_i | 14.31 |
| find | 13.96 |
| statista | 13.71 |
| where_can | 12.92 |
| research | 12.49 |
| bloomberg | 11.64 |
| can_i_find | 11.51 |
| capit | 11.10 |
| thomson_on | 10.70 |
| where_can_i | 9.42 |
| can | 9.07 |

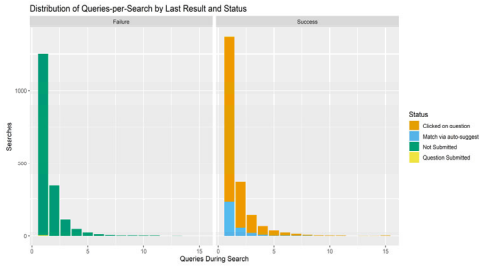
Terms that are important to and frequent in queries, yet not present in any FAQs present a clear first target for expanding FAQ content.




**Search behavior**




53% of searches ultimate leading the patron to view an FAQ, an improvement over queries considered individually. Searches that ended in success tended to include more queries. While the difference accounts, on average, for only a fraction of an additional query per search, there is a strong indication that the difference is more than coincidence. Could patrons be struggling for the specific words or spellings needed to return a relevant result? We might consider follow-up searches an indicator of the referring FAQ's failure to address the patron's query and explore potential improvements for FAQs with both high traffic and high follow-up search rates.




**Regression**

The generalized linear model used attempts to predict a query's success based on its word count. Applied to a test sample, this model demonstrates 52.5% accuracy; it is better than a coin toss, but only slightly.