**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 &amp; which the above process converts to amp. As a result, terms such as M&A and S&P appear as mampa and samp in this analysis.)

**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.

**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.

**Results & Discussion**

**Predictive Importance & Term Frequency**

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.

**Overall Query Success**

<table>
<thead>
<tr>
<th>Overall Query Success</th>
<th>Data Count</th>
<th>Read and Data</th>
<th>Data Count</th>
<th>Read and Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Export Views by Month**

<table>
<thead>
<tr>
<th>Month</th>
<th>Views</th>
<th>Month</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Status**

- Export of Views by Month

**Top Query Terms by Group**

<table>
<thead>
<tr>
<th>Top Query Terms</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Query Spy Exports**

- **Overall Query Success**
- **Export Views by Month**
- **Top Query Terms by Group**
- **Query Spy Exports**