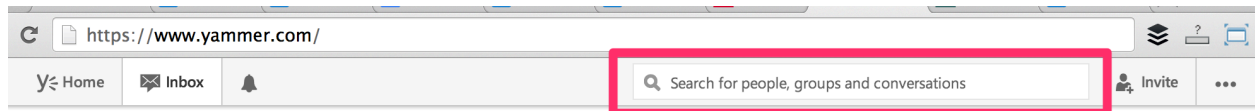Project: Understanding Search Functionality | SQL Analytics Training - Mode
Yammer has a search bar feature that helps users look for people, groups, and conversation text.



The functionality is akin to the search mechanism employed by Google. As users begin to type within the search field, an autocomplete system offers suggestions that are pertinent to the entered text, streamlining the process for users to access content without the need to fully articulate their search query. This action triggers a logging process known as **search_autocomplete.**

Furthermore, users have the option to select from among the top ten recommended queries, which are organized according to their relevance, through a dropdown menu. Simply initiating a search and arriving at the results display triggers a log entry under an event titled **search_run**. Subsequently, interactions with these recommendations are tracked under events designated as **search_click_X**, where "**X**" denotes a numerical range from 1 to 10, indicating the sequence of the chosen result. For instance, selecting the foremost suggestion from the dropdown menu would be recorded as **search_click_1**.
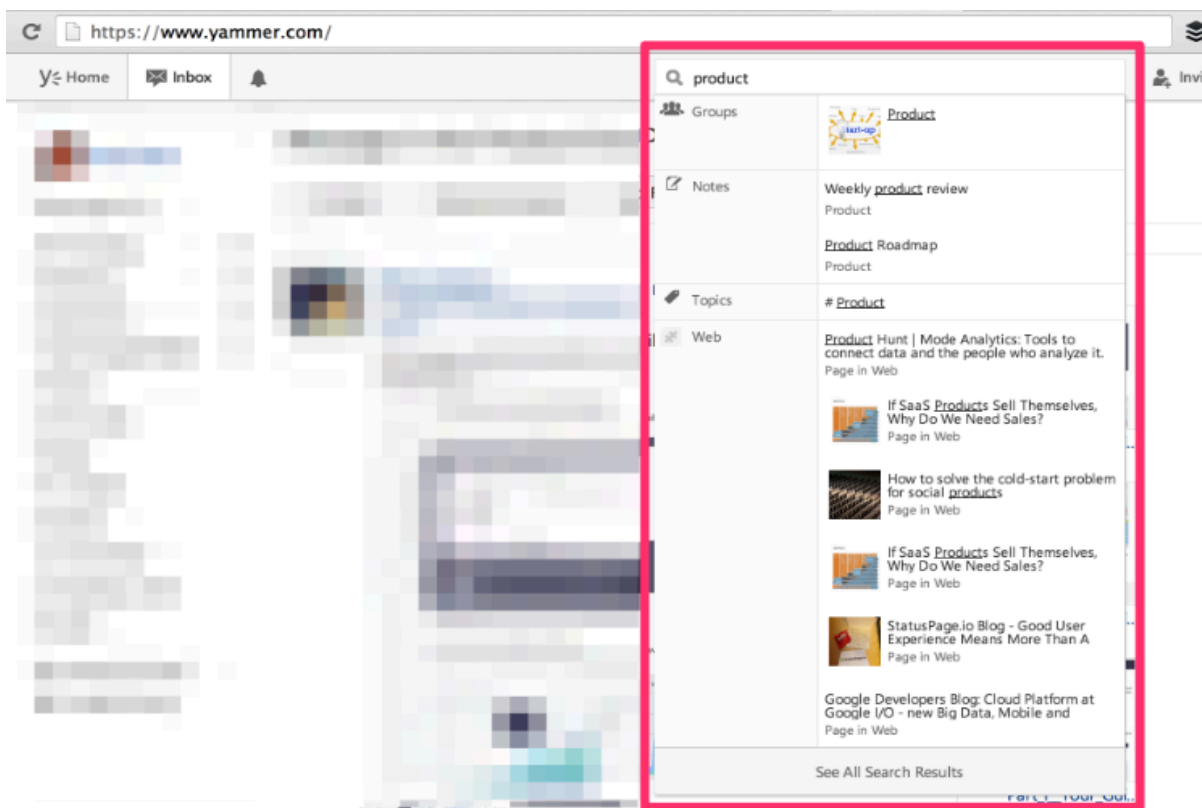
The goal of this project was to **determine whether the search query should be modified, and if so, how?**

---

Question 1: Could there be instances where a user, fluent in a less commonly spoken language, submits a search query but refrains from clicking on any results due to the language discrepancy in the search outcomes?

To investigate whether language barriers impact user engagement with search results on the platform, my approach involves understanding the relationship between the number of search queries initiated and the engagement with the search results. The hypothesis is that if users frequently initiate searches but seldom interact with the search results (e.g., by clicking on items from the dropdown menu), this could indicate a language mismatch between the users' queries and the language of the search results.

```sql
SELECT u.language,
  COUNT (DISTINCT u.user_id) AS users,
  COUNT (CASE WHEN e.event_name IS NULL THEN 1 END) AS
uses_who_never_engaged_in_application,
  COUNT (DISTINCT e.user_id) AS users_who_engaged_in_app,
  COUNT (CASE WHEN e.event_name = 'search_run' THEN 1 END) AS
searches_ran,
  COUNT (CASE WHEN e.event_name ilike 'search_click_%' THEN 1
END) AS search_results_that_users_selected
FROM tutorial.yammer_users u
LEFT JOIN tutorial.yammer_events e
ON u.user_id = e.user_id
GROUP BY 1
ORDER BY 3
```

See **optimizing-search-image1** for output**.**

```sql
SELECT u.language,
  COUNT (DISTINCT u.user_id) AS users,
  COUNT (CASE WHEN e.event_name IS NULL THEN 1 END) AS
uses_who_never_engaged_in_application,
  COUNT (DISTINCT e.user_id) AS users_who_engaged_in_app,
  COUNT (CASE WHEN e.event_name = 'search_run' THEN 1 END) AS searches_ran,
  COUNT (CASE WHEN e.event_name = 'search_autocomplete' THEN 1 END) AS
```

```
autocomplete_results_selected
FROM tutorial.yammer_users u
LEFT JOIN tutorial.yammer_events e
ON u.user_id = e.user_id
GROUP BY 1
ORDER BY 3
```
See **optimizing-search-image2** for output**.**

By left joining the **yammer_users** table with the **yammer_events** table on **user_id**, I can track user engagement at a granular level across different languages.

The preliminary results indicate no significant disparity between the volume of searches initiated and the frequency of selecting search results from the dropdown. This observation suggests that search results are, for the most part, in alignment with the users' language preferences. While it's natural for the number of selections to be lower than the number of searches—since not every search yields actionable results—the crucial insight is the absence of a stark drop-off that would suggest a language barrier issue.

However, a notable finding is the considerable proportion of users who have never engaged with key features of Yammer, such as reviewing the inbox, liking messages, or visiting the homepage. This lack of engagement points towards a broader issue beyond the search functionality: **user unfamiliarity with the platform's features.**

To address this, introducing an optional tutorial that showcases Yammer's features could enhance user engagement. Such a tutorial would guide new users through the platform, highlighting how to effectively use search among other functionalities, potentially increasing overall activity and satisfaction with the platform.

---

## Question 2: Are there regions where Yammer's server connectivity issues hinder users from accessing search results after submitting queries?

While I lack direct access to server performance data, an alternative approach to infer potential connectivity issues is by analyzing fluctuations in search query volumes by region on a week-to-week basis. A substantial drop in search queries from one week to the next could indicate possible accessibility or server connectivity problems affecting user experience. To explore this hypothesis, the following SQL query examines weekly search activity across different locations, comparing the volume of searches executed one week to those in the preceding week:

```
SELECT
   location,
```

```
   DATE_TRUNC('week', occurred_at) AS week,
   COUNT(CASE WHEN event_name = 'search_run' THEN 1 END) AS searches_ran,
   LAG(COUNT(*)) OVER (PARTITION BY location ORDER BY DATE_TRUNC('week',
occurred_at)) AS previous_week_searches,
   COUNT(CASE WHEN event_name = 'search_run' THEN 1 END) - LAG(COUNT(*))
OVER (PARTITION BY location ORDER BY DATE_TRUNC('week', occurred_at)) AS
difference_in_searches
FROM tutorial.yammer_events
WHERE event_name = 'search_run'
GROUP BY location, DATE_TRUNC('week', occurred_at)
ORDER BY 5
```

| | location | week | searches_ran | previous_week_searches | difference_in_searches |
|---|---|---|---|---|---|
| 1 | United States | 2014-07-07 00:00:00 | 146 | 248 | -102 |
| 2 | Brazil | 2014-07-14 00:00:00 | 14 | 80 | -66 |
| 3 | United States | 2014-08-04 00:00:00 | 192 | 255 | -63 |
| 4 | United States | 2014-05-12 00:00:00 | 168 | 224 | -56 |
| 5 | United States | 2014-06-16 00:00:00 | 210 | 265 | -55 |
| 6 | Japan | 2014-06-09 00:00:00 | 52 | 102 | -50 |
| 7 | France | 2014-07-07 00:00:00 | 28 | 70 | -42 |
| 8 | Japan | 2014-08-11 00:00:00 | 49 | 88 | -39 |
| 9 | Italy | 2014-06-16 00:00:00 | 3 | 42 | -39 |
| 10 | Russia | 2014-08-04 00:00:00 | 9 | 45 | -36 |
| 11 | Australia | 2014-07-14 00:00:00 | 3 | 38 | -35 |
| 12 | Russia | 2014-08-18 00:00:00 | 15 | 50 | -35 |
| 13 | Japan | 2014-07-07 00:00:00 | 49 | 84 | -35 |
| 14 | Germany | 2014-05-12 00:00:00 | 27 | 61 | -34 |
| 15 | United Kingdo... | 2014-07-07 00:00:00 | 28 | 61 | -33 |
| 16 | Colombia | 2014-07-28 00:00:00 | 10 | 42 | -32 |
| 17 | Indonesia | 2014-06-30 00:00:00 | 11 | 41 | -30 |
| 18 | France | 2014-08-04 00:00:00 | 25 | 55 | -30 |
| 19 | Russia | 2014-06-16 00:00:00 | 21 | 49 | -28 |
| 20 | Japan | 2014-05-26 00:00:00 | 42 | 70 | -28 |
| 21 | Sweden | 2014-06-16 00:00:00 | 2 | 30 | -28 |
| 22 | United Kingdo... | 2014-07-28 00:00:00 | 21 | 48 | -27 |
| 23 | Venezuela | 2014-07-28 00:00:00 | 5 | 32 | -27 |
| 24 | Germany | 2014-06-09 00:00:00 | 36 | 62 | -26 |
| 25 | Egypt | 2014-06-23 00:00:00 | 3 | 29 | -26 |
| 26 | India | 2014-06-02 00:00:00 | 13 | 39 | -26 |

This analysis aims to pinpoint regions where there might be significant week-to-week variances in search request volumes, potentially highlighting areas with unstable server connectivity or accessibility issues. Despite observing fluctuations in search activity across different countries,

the variations are not markedly drastic enough to conclusively attribute them to server outages or severe connectivity disruptions.

---

## Question 3: When a user types in a search query, how many other queries are entered through the search box in a 5 minute window?

Yammer wants its users to take advantage of the search feature, but users shouldn't be accessing it too often. Otherwise, it means that none of the search results are useful.

```sql
SELECT
  DATE_TRUNC('week', sub.occurred_at) AS week,
  COUNT(*) AS searches_started,
  COUNT(CASE WHEN sub.next_search_time < sub.occurred_at + INTERVAL '5 minutes' THEN 1 END) AS searches_within_5_minutes
FROM (
  SELECT
    user_id,
    occurred_at,
    LEAD(occurred_at) OVER (PARTITION BY user_id ORDER BY occurred_at) AS next_search_time
  FROM
    tutorial.yammer_events
  WHERE
    event_name = 'search_run'
) sub
GROUP BY 1
```

| | week | searches_started | within_5min_follow_ups |
|---|---|---|---|
| 1 | 2014-04-28 00:00:00 | 358 | 270 |
| 2 | 2014-05-05 00:00:00 | 753 | 577 |
| 3 | 2014-05-12 00:00:00 | 640 | 480 |
| 4 | 2014-05-19 00:00:00 | 744 | 566 |
| 5 | 2014-05-26 00:00:00 | 741 | 571 |
| 6 | 2014-06-02 00:00:00 | 780 | 601 |
| 7 | 2014-06-09 00:00:00 | 771 | 589 |
| 8 | 2014-06-16 00:00:00 | 715 | 540 |
| 9 | 2014-06-23 00:00:00 | 744 | 566 |
| 10 | 2014-06-30 00:00:00 | 859 | 672 |
| 11 | 2014-07-07 00:00:00 | 779 | 602 |
| 12 | 2014-07-14 00:00:00 | 809 | 612 |
| 13 | 2014-07-21 00:00:00 | 829 | 634 |
| 14 | 2014-07-28 00:00:00 | 954 | 733 |
| 15 | 2014-08-04 00:00:00 | 717 | 532 |
| 16 | 2014-08-11 00:00:00 | 639 | 482 |
| 17 | 2014-08-18 00:00:00 | 545 | 408 |
| 18 | 2014-08-25 00:00:00 | 642 | 482 |

Based on the provided SQL script and the analysis of search query data for Yammer, I can draw several insights and consider implications for improving the user experience regarding the search feature.

**High Frequency of Follow-up Searches:** The fact that 270 follow-up searches were conducted, after 358 search initial queries, suggests a high level of user engagement with the search functionality. However, it also indicates potential issues with the search results' relevancy or completeness.

**Search Efficiency:** An efficient search tool should help users find the desired information with minimal effort. Right now, the autocomplete tool fills in that void. With this feature, the number of search queries submitted within 5 minutes of the original query was halved.

```sql
SELECT
  DATE_TRUNC('week', sub.occurred_at) AS week,
  COUNT(*) AS searches_started,
  COUNT(CASE WHEN sub.next_search_time < sub.occurred_at + INTERVAL '5
minutes' THEN 1 END) AS searches_within_5_minutes
FROM (
  SELECT
    user_id,
    occurred_at,
```

```
    LEAD(occurred_at) OVER (PARTITION BY user_id ORDER BY occurred_at) AS
next_search_time
   FROM
     tutorial.yammer_events
   WHERE
     event_name = 'search_autocomplete'
) sub
GROUP BY 1
```

| | week | searches_started | searches_within_5_minutes |
|---|---|---|---|
| 1 | 2014-04-28 00:00:00 | 424 | 207 |
| 2 | 2014-05-05 00:00:00 | 841 | 402 |
| 3 | 2014-05-12 00:00:00 | 910 | 437 |
| 4 | 2014-05-19 00:00:00 | 958 | 475 |
| 5 | 2014-05-26 00:00:00 | 908 | 443 |
| 6 | 2014-06-02 00:00:00 | 947 | 455 |
| 7 | 2014-06-09 00:00:00 | 962 | 463 |
| 8 | 2014-06-16 00:00:00 | 1031 | 514 |
| 9 | 2014-06-23 00:00:00 | 978 | 458 |
| 10 | 2014-06-30 00:00:00 | 1056 | 539 |
| 11 | 2014-07-07 00:00:00 | 1118 | 537 |
| 12 | 2014-07-14 00:00:00 | 1159 | 542 |
| 13 | 2014-07-21 00:00:00 | 1167 | 571 |
| 14 | 2014-07-28 00:00:00 | 1231 | 599 |
| 15 | 2014-08-04 00:00:00 | 1113 | 551 |
| 16 | 2014-08-11 00:00:00 | 898 | 410 |
| 17 | 2014-08-18 00:00:00 | 1028 | 508 |
| 18 | 2014-08-25 00:00:00 | 1091 | 551 |

**Implications for User Experience:** When a large proportion of users perform another search shortly after their initial one, it could imply that they did not find what they were looking for on their first attempt. This leads me to my next question.

## Question 4: What is the selection rate for the first 3 "relevant" search results versus the latter 7 within five minutes of initiating a search query?

The assumption is that the most pertinent search results are displayed at the beginning of the dropdown list, with the initial three being the most likely to be clicked. To investigate this:

```
SELECT
```

```sql
    DATE_TRUNC('week', sub.occurred_at) AS week,
    COUNT (CASE WHEN sub.event_name = 'search_run' THEN 1 END) AS
searches_ran,
    COUNT (CASE WHEN sub.event_name IN ('search_click_result_1',
'search_click_result_2', 'search_click_result_3') THEN 1 ELSE NULL END) AS
top_3_seletected,
    COUNT (CASE WHEN sub.event_name IN ('search_click_result_4',
'search_click_result_5', 'search_click_result_6',
    'search_click_result_7', 'search_click_result_8',
'search_click_result_9', 'search_click_result_10') THEN 1 ELSE NULL END) AS
remaining_7_selected
FROM (
  SELECT
    user_id,
    occurred_at,
    event_name,
    LEAD(occurred_at) OVER (PARTITION BY user_id ORDER BY occurred_at) AS
next_search_time
  FROM  tutorial.yammer_events
  WHERE event_name = 'search_run' OR event_name ilike
'search_click_result_%'
) sub
GROUP BY 1
```

| | week | searches_ran | top_3_seletected | remaining_7_selected |
|---|---|---|---|---|
| 1 | 2014-04-28 00:00:00 | 358 | 141 | 202 |
| 2 | 2014-05-05 00:00:00 | 753 | 226 | 342 |
| 3 | 2014-05-12 00:00:00 | 640 | 194 | 270 |
| 4 | 2014-05-19 00:00:00 | 744 | 250 | 363 |
| 5 | 2014-05-26 00:00:00 | 741 | 234 | 355 |
| 6 | 2014-06-02 00:00:00 | 780 | 267 | 368 |
| 7 | 2014-06-09 00:00:00 | 771 | 198 | 271 |
| 8 | 2014-06-16 00:00:00 | 715 | 261 | 326 |
| 9 | 2014-06-23 00:00:00 | 744 | 246 | 346 |
| 10 | 2014-06-30 00:00:00 | 859 | 269 | 397 |
| 11 | 2014-07-07 00:00:00 | 779 | 257 | 368 |
| 12 | 2014-07-14 00:00:00 | 809 | 270 | 347 |
| 13 | 2014-07-21 00:00:00 | 829 | 242 | 360 |
| 14 | 2014-07-28 00:00:00 | 954 | 269 | 381 |
| 15 | 2014-08-04 00:00:00 | 717 | 216 | 329 |
| 16 | 2014-08-11 00:00:00 | 639 | 194 | 267 |
| 17 | 2014-08-18 00:00:00 | 545 | 139 | 204 |
| 18 | 2014-08-25 00:00:00 | 642 | 173 | 230 |

Analysis reveals that contrary to expectations, the initial three search results do not dominate user selections. Surprisingly, the latter seven options receive more attention, indicating that users may not always find the most relevant results at the top of the list. Occasionally, users do not select any of the presented options. This suggests a need to reevaluate the algorithm that ranks search results in the dropdown menu, ensuring it aligns more closely with user preferences and search intents.

## Question 5: What is the frequency of users re-selecting search results within a 5-minute interval?

My objective was to assess the feasibility of Yammer utilizing a caching mechanism for both the search queries entered by users and the subsequent results they click on. While determining the specific queries and clicked results without explicit data is unfeasible due to the lack of data, the potential benefits of implementing such a caching strategy should be considered.

## Conclusions and Recommendations for Application Improvement:
The comprehensive analysis of Yammer's search functionality reveals insightful trends and user behaviors that underscore several opportunities for enhancing the overall user experience within the platform. Drawing on the findings from the inquiries and data analysis, I propose targeted recommendations aimed at refining Yammer's search feature and, by extension, its application as a whole.

1. **Enhance Search Relevance and Efficiency:**

- The high frequency of follow-up searches within a 5-minute window suggests that users often do not find what they are looking for on their first attempt. To address this, Yammer should invest in optimizing its search algorithms to improve the relevance of search results. Enhancing the autocomplete functionality with more accurate predictions could significantly reduce the need for subsequent searches, thereby improving user satisfaction and efficiency.

2. **Introduce Multilingual Support:**
   - While the analysis did not indicate a stark language barrier issue, incorporating multilingual support for search queries and results can cater to a wider user base. This involves optimizing the search engine to better understand and deliver results in the users' preferred languages, reducing any friction for non-English speakers.

3. **Optimize the Ranking of Search Results:**
   - Given that users often click on results beyond the top three recommendations, it's crucial to reassess the criteria for ranking search results. Yammer should consider leveraging user engagement data to refine how results are prioritized, ensuring that the most relevant and useful content is easily accessible.

4. **Foster User Familiarity with Platform Features:**
   - The notable proportion of users who have never engaged with key features of Yammer highlights a gap in user onboarding and familiarity. Implementing interactive tutorials or guided tours can significantly enhance user engagement by demonstrating the value and functionalities of the platform, including effective search practices.

5. **Investigate and Address Potential Connectivity Issues:**
   - While the fluctuation in search query volumes did not conclusively point to server connectivity issues, it remains important for Yammer to monitor and ensure robust server performance across all regions. Proactively identifying and resolving any potential connectivity problems can prevent disruptions in user access to search results and other features.

6. **Explore Caching Strategies for Search:**
   - Although specific queries and clicked results were not analyzed, the concept of caching search queries and results warrants further exploration. By caching frequently searched terms and their corresponding results, Yammer can expedite search response times and enhance the overall user experience.

7. **User Feedback Loop:**
   - Establish a continuous feedback mechanism to gather direct input from users about their search experiences. This can inform ongoing improvements and adjustments to the search functionality, ensuring it meets user needs and expectations.