Lab Note: The Advanced Question Query Generator

Team name: The Query Lab

Members: Mark Henri Mörsheim, Maryam El Ghadioui

In this micro shared task, we developed an Advanced Question Query Generator aimed at creating more realistic, question-based queries that reflect real user search behavior.
Our rule-based pipeline combines probabilistic selection with templated natural language generation. We begin by retrieving topic-specific queries via get_given_queries() from the iFind framework, using the last query as a topic seed. For each new query, the generator:

1. Selects a wh-word (e.g., how, what, why) based on weighted probabilities from real data.
2. Applies a predefined template containing placeholders for the wh-word and topic.
3. Adds punctuation (question mark, period, or none) based on real-world frequencies.

This is repeated 10 times to generate new queries appended to the original list.

Our original baseline generator simply prepended a random wh-word to an existing query and appended a question mark. While functional, it lacked the complexity and realism of actual user queries in exploratory search sessions. To improve, we analyzed real query clusters from the session group, revealing:

- Only **29%** of queries ended with a question mark.
- Almost **50%** had no punctuation at all.
- **72.6%** started with a wh-word (e.g., *how*, *what*, *why*).
- The most frequent start phrases were *how to*, *what is*, *how does*, etc.
- length averaged **79 characters** or **11.3 words** and most common length was **48 characters**.

We incorporated these findings into our generator through more diverse templates like:

- *"{w_word} are the main causes of {topic}"*

- *"{w_word} can governments do about {topic}"*

Punctuation was also probabilistically modeled: 29% "?", 22% ".", 49% none.

We attempted a more sophisticated syntactic approach using spaCy to extract sentence structures or POS patterns from the real queries, but this turned out to be noisy and less effective than expected.

This task highlighted the importance of behavior-aware simulation in interactive retrieval (IR). Future evaluation should move beyond static models and consider:

- **Behavioral realism**: Generators should mimic authentic patterns like reformulation and incomplete thoughts.

- **Dynamic intent modeling**: Evaluation should reflect evolving user needs over sessions.

- **Context-aware metrics**: Consider user history, diversity of queries, and exploration behavior.

Our experience shows fine-tuned adjustments rooted in real data, like realistic punctuation and templates, can significantly improve the quality of simulated user queries. Combining linguistic structure, statistical modeling, and user behavior mining could pave the way for more robust IR simulations in the future.