PES UNIVERSITY
CELEBRATING 50 YEARS

Algorithms for Information Retrieval and Intelligence Web Project - Review 1

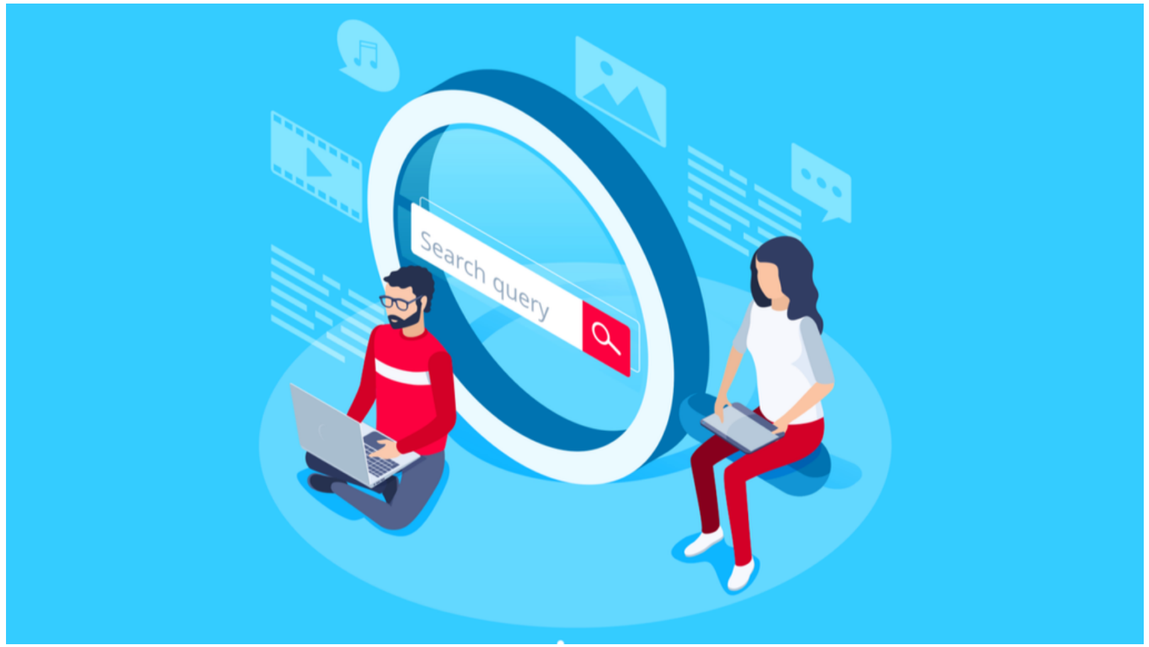
Query Autocompletion using Trie and LSTM model

PES1UG21CS152 : Charutha Rajeesh
PES1UG21CS416 : Pragya Srivastava

INTRODUCTION

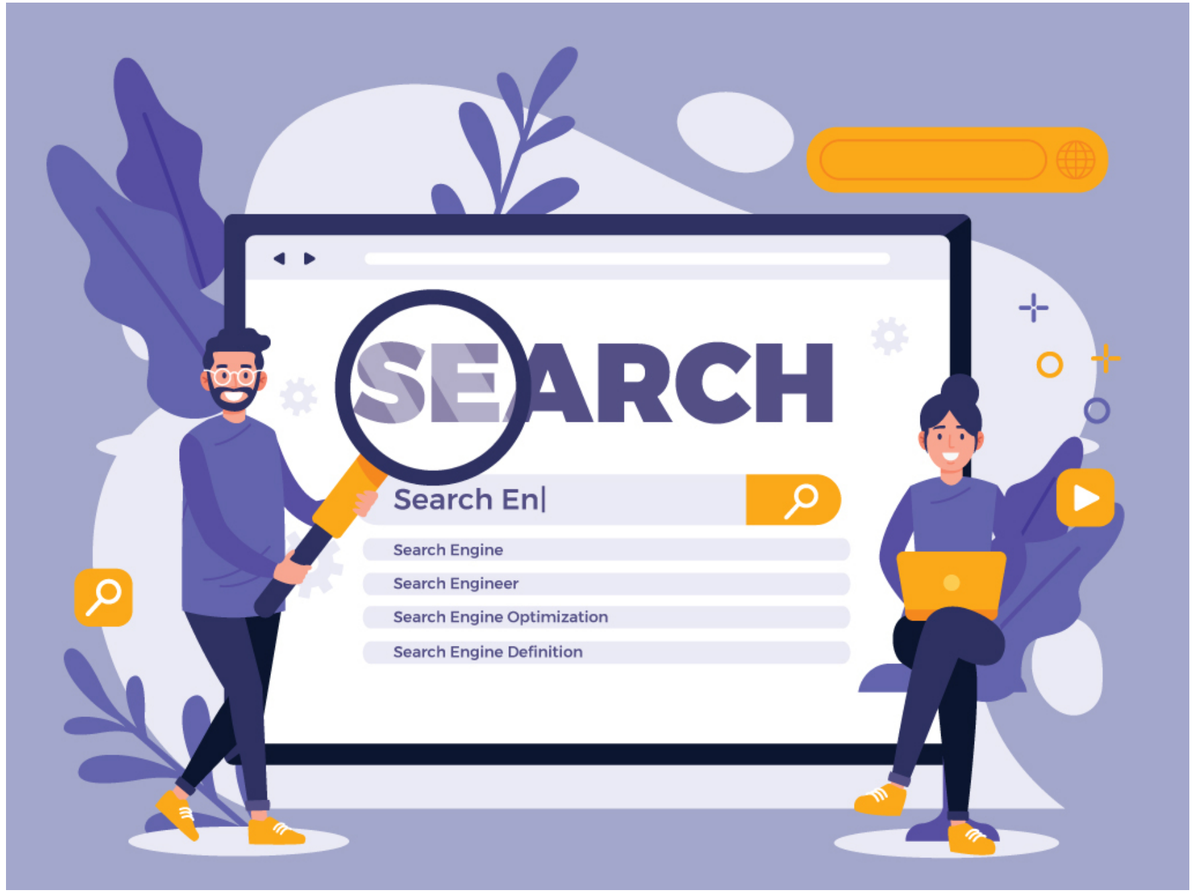
In the domain of information retrieval, users often face challenges when formulating search queries, especially when dealing with long or complex terms.

Query autocompletion aims to assist users by **predicting** and **suggesting completions** for **partially entered queries** in real-time, enhancing the search experience and improving query formulation efficiency.



PROBLEM STATEMENT

Design and implement an **autocomplete system** to efficiently **predict and suggest completions for partially entered queries** in real-time.



SUPPORTING LITERATURE

Early Research:

- Initial focus on rule-based or statistical methods.
- Common techniques include prefix matching, trie data structures, and n-gram models.

Integration of IR Techniques:

- Adoption of inverted index structures, TF-IDF weighting, and relevance ranking algorithms.

Machine Learning Approaches:

- Research on Markov models, decision trees, and support vector machines for suggestion generation.

Neural Language Models:

- Adoption of RNNs, LSTMs, BERT, and GPT for capturing linguistic patterns and generating relevant suggestions.

Generative Models:

- Transformer-based architectures gain traction for generating completions by learning from large text corpora and capturing semantic relationships.

SUPPORTING LITERATURE

1. QueryBlazer: Efficient Query Autocompletion Framework:
   Generative query completion system using an n-gram language model

2. Efficient and Effective Query Auto-Completion:
   The solution is based on the combination of an inverted index with succinct data structures

1. Efficient query autocompletion with edit distance-based error tolerance:
   Neighborhood generation-based method to process error-tolerant query autocompletion maintaining only a small set of active nodes, thus saving both space and time to process the query.

METHODOLOGY

I

- **Data Collection and Preprocessing:**
  - Preprocess the data by tokenizing it into words or subwords, removing stop words, punctuation, and special characters, and applying techniques like stemming or lemmatization for normalization.

- **Building a Trie Data Structure for word completion:**
  - Constructing a trie data structure to efficiently store and retrieve completions for partially entered prefixes.

- **Training an LSTM RNN for sentence completion :**

  a.Training an LSTM model, such on the preprocessed text data. Fine-tuning the model on the specific task of query autocompletion.

METHODOLOGY

- **Combining Suggestions:**
  - Combining the completions obtained from the trie structure with those generated by the LSTM model to form a hybrid set of autocomplete suggestions.
  - rank and filter the hybrid suggestions based on relevance scores, user preferences, or other criteria.

- **User Interface Integration(demo):**
  - Integrating the hybrid autocompletion functionality into user interfaces
  - Provide a mechanism for users to input queries and display the hybrid suggestions in real-time as they type.

DATASET

The SIGIR 2017 data set includes files for four data:

- **background** (6,773,535 queries; used for training our neural language model; queries that appear less than three times are removed to filter out noisy queries.)
- **training** (75,198 prefixes; used for training LambdaMART.)
- **validation** (30,993 prefixes)
- **test** (32,559 prefixes; it is written as "32,044" in the paper by mistake.)

In each of training, validation, and test data, there are only two columns: a **prefix** and the **corresponding query**.

(or) subset of wikipedia dataset in word document format consisting of paragraphs on various topics.

REFERENCES

https://dl.acm.org/doi/pdf/10.1145/3437963.3441725

https://link.springer.com/article/10.1007/s00778-019-00595-4

https://dl.acm.org/doi/abs/10.1145/3397271.3401432

https://sites.google.com/site/daehpark/Resources/data-set-for-query-auto-completion-sigir-2017

THANK YOU

TEAM 5