Predictive Typing

Ethan Passmore, Lane Phillips, Layne Bernardo, Roya Rashidi, Sarah Paracha

Computer Science & Computer Engineering

Problem:

Big data can cause difficulties in searching for relevant results instantaneously. Our capstone project is for the company Sorcero. Sorcero is a startup company with a focus on natural language processing solutions. Currently, Sorcero lacks phrase completion for users wishing to review and search their corpus of documents.

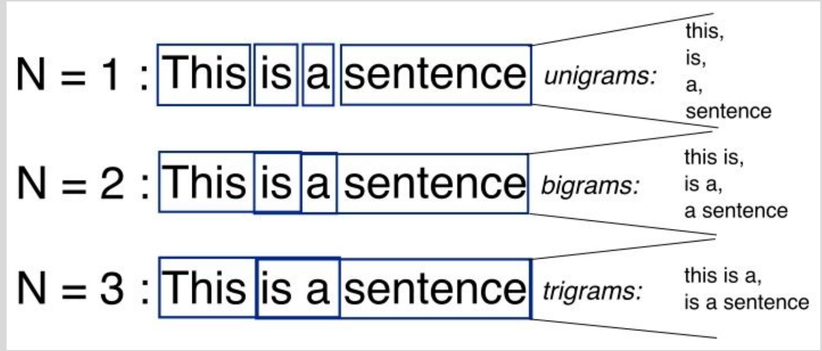
Objective:

To develop a reactive predictive search bar for Sorcero using n-grams and a scalable database

Design:

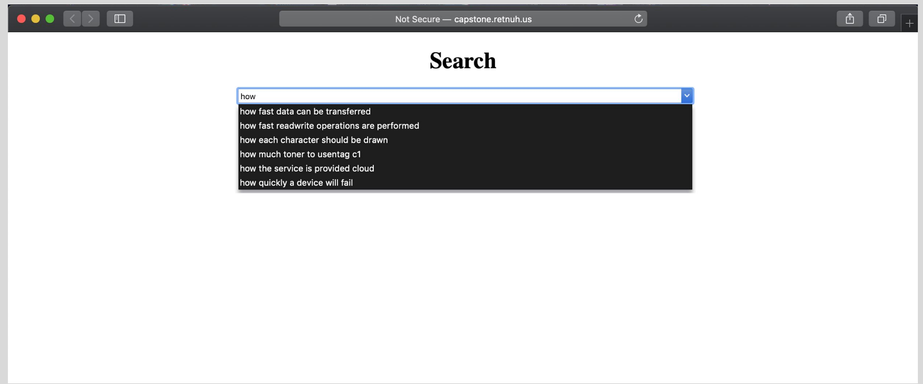
There are a few key pieces to the system. A document is submitted to the system where it gets parsed into phrases that are then submitted into a database consisting of n-grams to be used in the search. The search bar then accesses the databases and displays the most frequent matches for the n-grams.

N-Gram Example:



"N-Grams." *DeepAI*, 17 May 2019, deepai.org/machine-learning-glossary-and-terms/n-gram.

Results and Conclusion:



High-Level Architecture:

- Document Parser
- N-gram generation
- Hashing of n-grams into buckets
- Sorting buckets with highest frequency n-grams on top for efficient retrieval

The Way Forward:

Future modifications can be made to the program for a more streamlined user experience and better time-memory tradeoff. This may include creating nested alphabetical buckets to continually modify predictions in real time. Additionally, the scalability of the program may be improved to accommodate a widespread search amongst documents.

References:

capstone.retnuh.us
https://www.sorcero.com
https://capstone-csce.uark.edu/fall-spring-2019-2020/teams-11-17/team-12-predictive-typing/

Acknowledgements:

This project was an industry project proposed by Sorcero, a start-up based in Washington, DC.
Sorcero is a company specializing in NLP (Natural Language Processing) Solutions.