Vizziest

https://www.ischool.berkeley.edu/projects/2019/vizziest-making-visualization-easiest-everyone

W210.6 Capstone Project

Chi Iong Ansjory, Jeff Braun, Grace Lin, Yang Yang Qian
[ansjory, jbraun, yqlin, yangyang.qian]@ischool.berkeley.edu

Vizziest

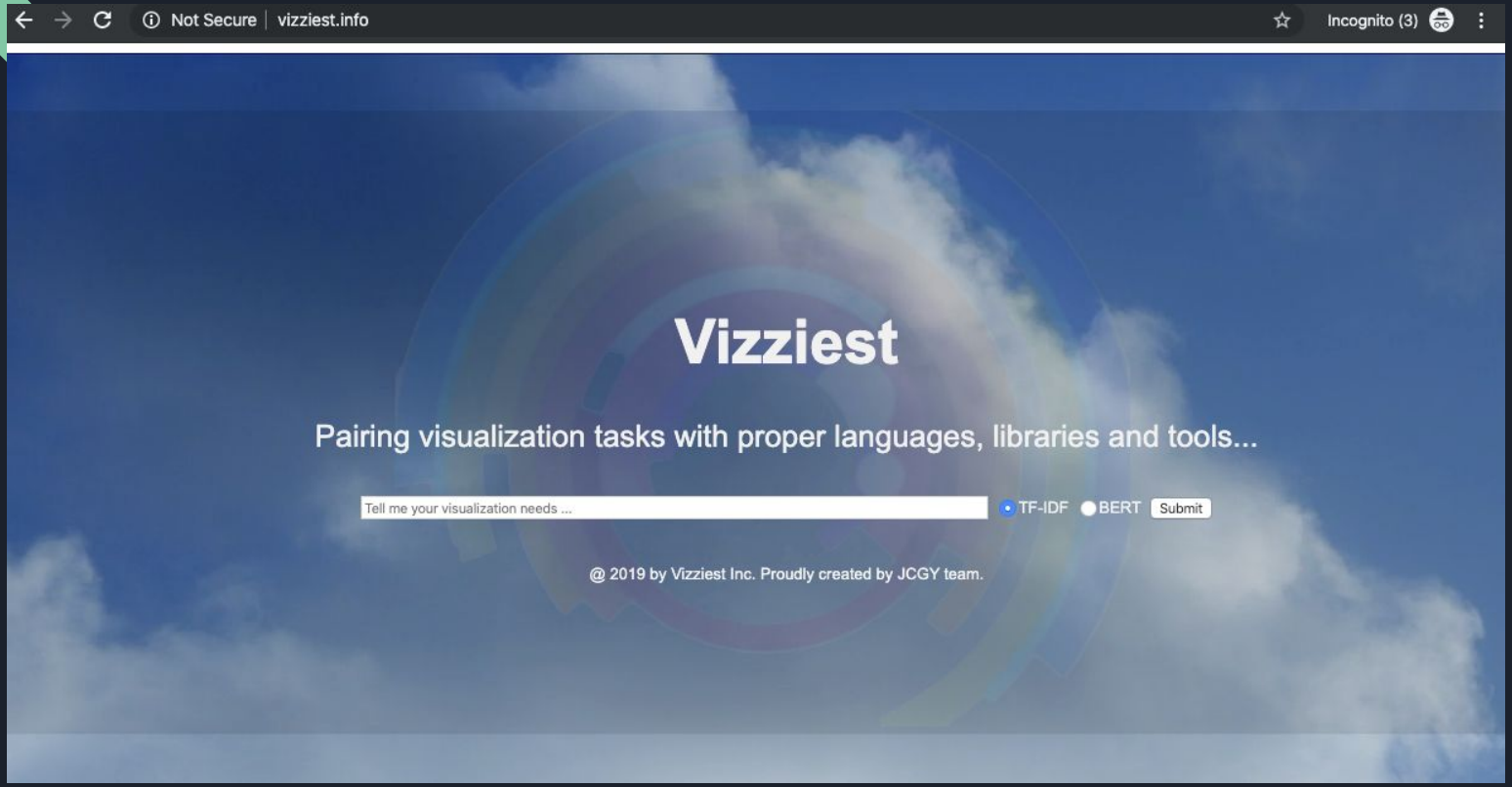
- **Our Mission**
  - **Vizziest** takes the time, frustration, and guesswork out of finding actionable guidance for creating the data visualization that best meets the user's requirements.

- **Customer Pain Points**
  - **20 hours** spent per year researching solutions

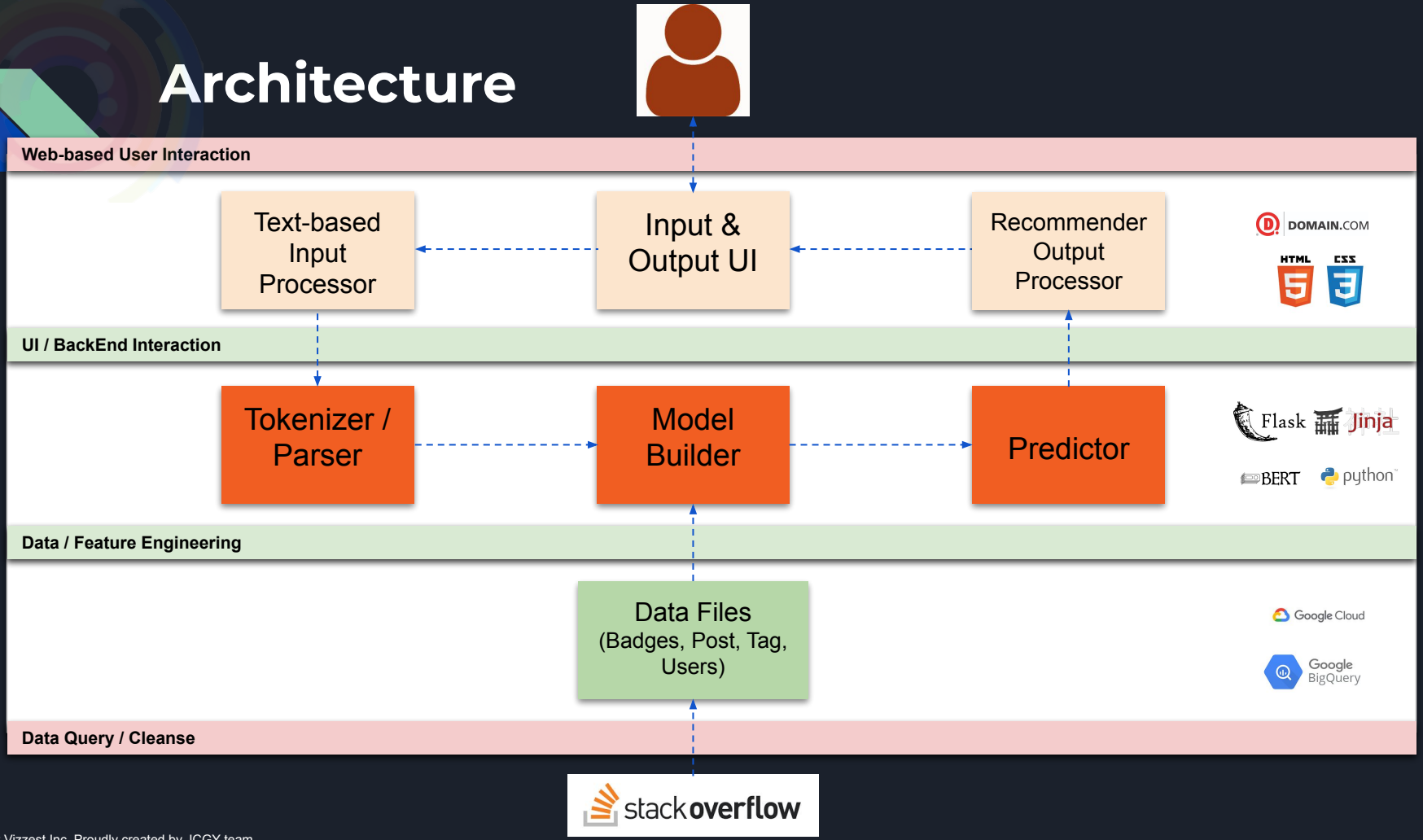
- **Our Concept**
  - NLP models + filtered public data = a knowledge tool for creating visualizations

- **Future Potential**
  - Create similar specialized knowledge tools for other problem domains

Demo (http://vizziest.info)



Architecture



How Vizziest works

- **User keys in a question**

- **Approach A**
  - **Various embedding approaches to vectorize questions:**
    - TF-IDF
    - Word to Vector
    - Sentence to Vector
  - **Logistic Regression to recommend questions and answers**

- **Approach B**
  - **BERT to directly recommend answers**
    - Next sentence prediction

Approach A: Models and Analysis

- **Baseline**
  - Top answer received from using the Stack Overflow search bar (Elasticsearch)

- **TF-IDF**
  - Uni-gram, bi-gram and tri-gram
  - Text frequency to determine importance
  - Pros and Cons: fast but no linguistic contexts

- **Word to Vector and Sentence to Vector**
  - A pre-trained model based on Google News
  - Neural networks trained to predict words from their neighbors
  - Pros and Cons: linguistic contexts but limited to Google News data and longer time

Approach A: Model Evaluation

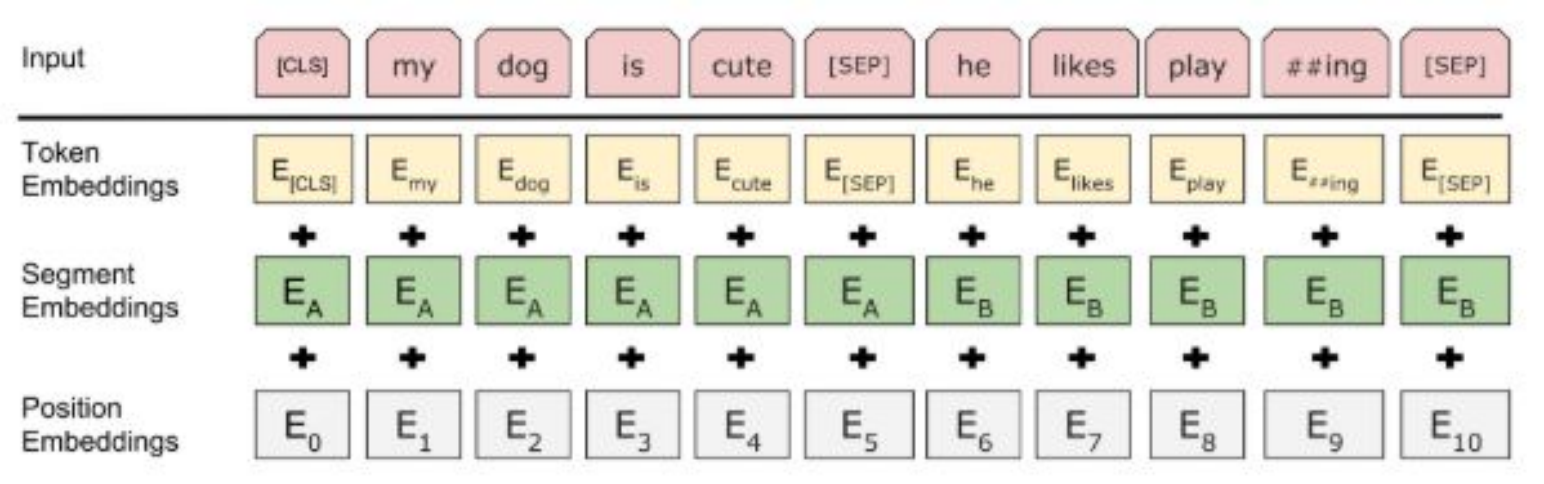
- **Method**
  - Surveyed 105 people through Mechanical Turk
  - Compared best recommended questions from Stack Overflow, TF-IDF, and Word / Sentence to Vector embeddings

- **Result**
  - 52% of respondents chose TF-IDF uni-gram
  - 31% of respondents chose Stack Overflow search engine
  - 10% of respondents chose word to vector
  - 7% of respondents chose sentence to vector

Approach B: BERT Model and Analysis

- **BERT, Bidirectional Encoder Representations from Transformers**
  - Aims for better semantic understanding of search terms
  - Transfer learning
- **We used a pretrain BERT model from Hugging Face (pytorch)**
  - We need two sentences, or “spans” for the NSP task
  - User’s input becomes one of the spans
  - Stackoverflow accepted answer bodies the second span
- **Assumption**
  - The question should be semantically closer to the accepted answer body than other bodies
  - This turned out to be an incorrect assumption

Approach B: BERT Next Sentence Prediction (NSP)



- BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings
- Image from “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”

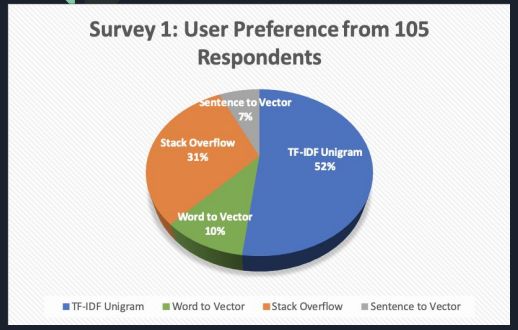
Approach B: Model Evaluation

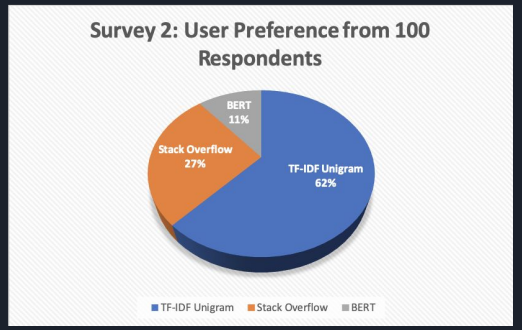
- **Method**
  - Surveyed 100 people through Mechanical Turk
  - Compared best recommended answers from Stack Overflow, TF-IDF, and BERT
- **Result**
  - 62% of respondents chose TF-IDF uni-gram
  - 27% of respondents chose Stack Overflow search engine
  - 11% of respondents chose BERT

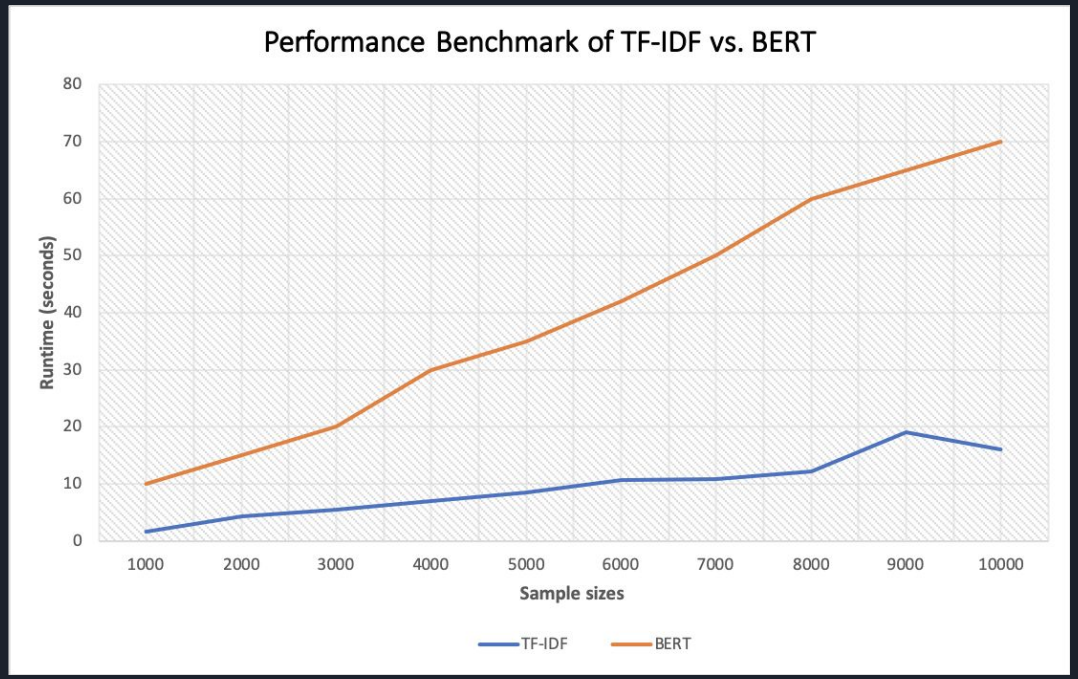
Approach B: Evaluation Takeaways

- **Performance**
  - Survey respondents tended to prefer results from TF-IDF vs BERT
- **Response Time**
  - Limited to 1 GPU, 11k question-answer pairs
  - Scales linearly with number of samples
- **Overall**
  - For search application, simple TF-IDF model offers better performance at lower cost than BERT

User Preference and Performance Benchmark







Lessons Learned

- Careful filtering to create corpus is essential
- Leverage cloud offerings
- Open-ended search box alone may not be enough to capture user context
- Stack Overflow language is very different from normal speech and newspaper articles

Future Enhancements

- Add viz-relevant data from other sources
- Create manually-labelled “ground truth” dataset
- Try ensemble models: TF-IDF + BERT Q&A
- Add user feedback functionality

Q & A

Vizziest® = Making **Vis**ualization Ea**siest** for Everyone

Try it out!

http://vizziest.info

Capstone Project Gallery:

https://www.ischool.berkeley.edu/projects/2019/vizziest-making-visualization-easiest-everyone