Amazon Search Engine Blog

Fred Gigou, Idris Hanafi, James Rosenkoetter, Curtis Guo

(Embedded within each section) Data visualization (**10 points)**: Visuals are easy to read and understand. They are clearly connected to the main points in your report. Visualization choices are reasonable for your audience and data.

* Product Infrastructure (add Rekognition) (this visualization will be in the **project architecture**)
* Flowchart (this visualization will be in the **introduction**)
* Model Results (this visualization will be in the **interpretation section**)

# 1. Introduction

(max 1/2 pages - Fred/rest of team) (10 points): Your writing is concise, accurate and effective. You clearly state the value your project adds to data science, machine learning, or your specific domain of interest.

* Abstract
  + Data Science Questions
  + Business Impact
* The Hook
* Flowchart Visualization

# 2. Expertise

(Max 1 page) (5 points): Your report describes the state of the field currently, references relevant previous work, and identifies how your work fits in the broader context. Ethical considerations are stated in your report and mitigated appropriately in your work, as needed.

## 2.1. Similar/Relevant technology

(2 paragraph max)

* (Idris) Google Index,
* (Curtis) Amazon Search Engine paper,
* (Idris) Other Existing Search Engines
* (Idris) Average accuracy for search engines

## 2.2. What did we try to tackle?

(1 paragraph max - Curtis)

* Cold Start (anonymous user)

## 2.3. Ethical Considerations and Mitigation

Building data applications that interact with users and/or collect user data require data scientists to adhere to frameworks that ensure the highest ethical standards. The project does not collect any user data, and the user information provided in the Amazon dataset is anonymized. However, it is logical for project extensions to collect some user data. Collecting user data would require consent, clarity in what the user is agreeing to, transparency of what data is collected and why, controls on how the data is used and accessed, transparency to the user, and finally an understanding of the consequences about using the data and combining data sets. Some ideas in the current project that were discussed, but rejected, were IP address tracking and geolocation tracking. Geolocation tracking may increase the relevancy of search results, but at the cost of user privacy. These concerns were mitigated by not collecting any user data. Future extensions of the project that collect user data will need a clear privacy framework. Another area of possible concern is the use of the user reviews themselves. Potentially user reviews could be biased. The reviews could be generated in large part by one social-economic group, and thus not representative of users at large. In theory, this could result in recommending products to some groups of users that are sub-optimal. Future extensions of the project could mitigate this concern with the help of additional user data.

# 3. Data

(no points): (1 page max - Idris)

* (Idris) Description of the dataset
* (Idris) Why this dataset?
* (Idris) Challenges
* (Idris) Schema, Size, Storage, etc.
  + Head of dataset

# 4. Project Architecture

(no points): (max 1 page - Idris)

Describe the overview of the setup and the decisions.

* Project architecture visualization
* AWS
* DVC
* Git
* Colab
* Frontend/Backend

# 5. Data Science Methods

(**35 points)**: Your report **justifies** your methodological decisions and provides sufficient detail for another data scientist to understand your approach. Methods are carried out correctly and are well-matched to your project aims.

## 5.1. Pipeline (Idris - max 1 paragraph):

* Data Pipeline => Model Pipeline => Inferences

## 5.2. Data Engineering (Curtis/Idris/James - max 2 paragraphs):

* Transformation
* Cleaning
* Storage
* Feature Selections
* Vectorizing

## 5.3. Data Modeling (max 5 pages):

* Subcategory:
  + 1 visualization showing all of the Subcategory models
* Deployed Models:
  + Algorithm Overview: M1 (I think we are calling this M3) (LSI Bigram via sklearn) (James - ½ page)
  + Algorithm Overview: Tfidf vectorizer + cosine similarity: M2 and M2+ (Fred - ½ page)
* Other Models Assessed
  + Algorithm Overview: Word2vec (Fred - 2 lines)
  + Algorithm Overview: LSI Gensim + Unigram (Idris/James - 1 paragraph)
  + Algorithm Overview: Term Frequency Inverted Index (Curtis - 3 lines)
* Ranking Algorithm Model:
  + 1 visualization showing all of the Ranking Algorithms
  + Deployed Models:
    - R1 Baseline Algorithm Overview (Fred - ½ page)
    - R2 Product Context Algorithm Overview (Fred/James - ½ page)
      * Review quality dataset (James)
    - R3 User Review - Keyword Context Recommender System Algorithm Overview (Curtis - 1 page)

# 6. Interpretation of Results

(**10 points)**: It is not enough to simply show your results- you need to interpret them for the audience. You summarize and interpret your findings and connect them to broader impacts. If you found null results, you explore why this might be the case.

## 6.1. Subcategory Model:

### 6.1.1. Deployed Models:

* M1 (LSI Bigram via sklearn) (James - ½ page)
  + Result, Summary, and Challenges
* Tfidf vectorizer + cosine similarity: M2 and M2+ (Fred - ½ page)
  + Result, Summary, and Challenges

### 6.1.2. Other Models Assessed:

* Word2vec (Fred - 1 sentence)
  + Result, Summary, and Challenges
* LSI Gensim + Unigram (Idris - 2 sentence)
  + Result, Summary, and Challenges
* Term Frequency Inverted Index (Curtis - 2 sentence)
  + Result, Summary, and Challenges

## 6.2. Ranking Algorithm Model:

* R1 Baseline Algorithm (Fred - ½ page)
  + Result, Summary, and Challenges
* R2 Product Context (Fred - ½ page)
  + Result, Summary, and Challenges
* R3 User Review - Keyword (Curtis - ½ page)
  + Result, Summary, and Challenges

# 7. Conclusions/Final Thoughts/Extensions

(**no points)**  (everyone - max 2 paragraphs)

# 8. Statement of Work

(**no points)**  (everyone)

Please include a statement of work in your final project. The teaching team may award bonus points to individuals with exceptional contributions.

# 9. References

(**no points)**  (everyone)

# 10. Appendix

(**no points)**  (everyone)