Fractal Hackathon : Q&A Engine

23rd Dec 2017 to 7th Jan 2018

Team no. 1 "NLPGiants"

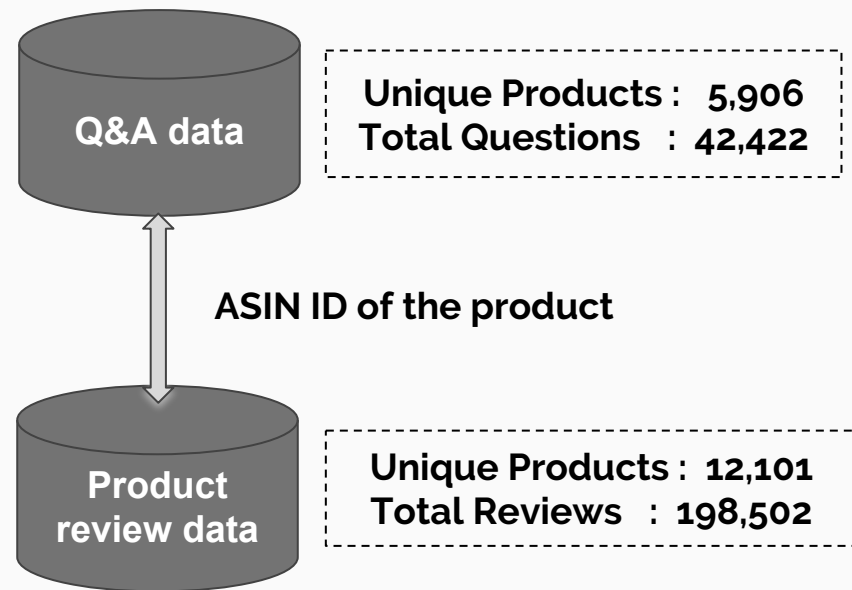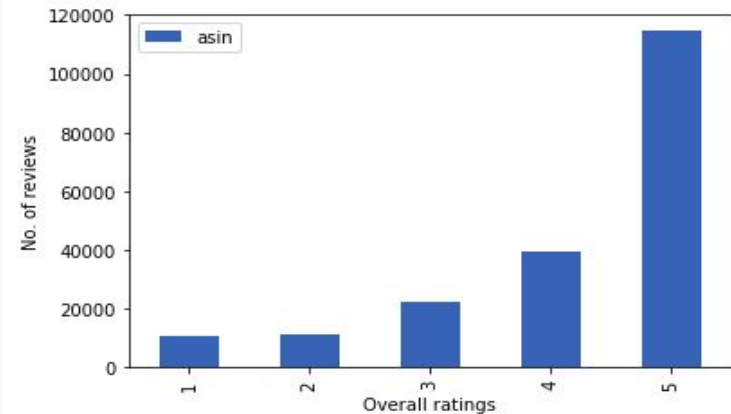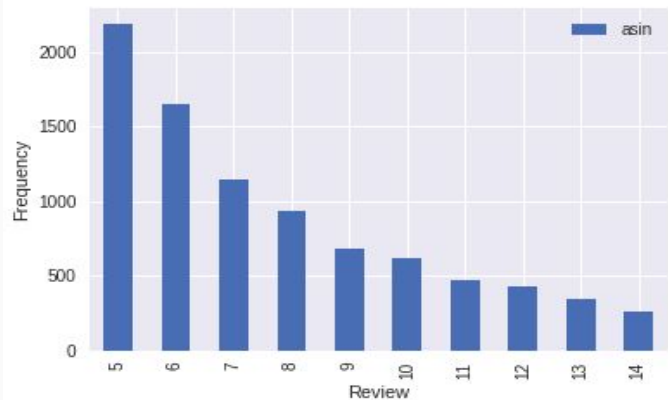- Vinod Boga
- Rohan Damodar
- Ankan Roy
- Nikhil Akki

Assumptions

- **User group** : Sales & Marketing team.

- **Objective** : To improve quality and sales of the product by analysing customer review sentiments.

- **Input mandate** : User needs to provide Product ID along with query.

- **Diversity** : Diversity in reviews are based on sentiments

Note - *It is assumed that the User knows the name of the product whose ID has been provided.*

Exploratory Data Analysis

Dataset Relationship and Facts

Unique Products : 5,906
Total Questions : 42,422

Q&A data

ASIN ID of the product

Product review data

Unique Products : 12,101
Total Reviews : 198,502

Sentiment Analysis : Textblob Vs. Vader

	textblob	vader	tricky_sentences
0	positive	negative	Sentiment analysis has never been good.
1	positive	positive	Sentiment analysis has never been this good!
2	neutral	negative	Most automated sentiment analysis tools are shit.
3	negative	positive	With VADER, sentiment analysis is the shit!
4	negative	negative	Other sentiment analysis tools can be quite bad.
5	negative	positive	On the other hand, VADER is quite bad ass!
6	positive	positive	Roger Dodger is one of the most compelling variations on this theme.
7	neutral	neutral	Roger Dodger is one of the least compelling variations on this theme.
8	neutral	positive	Roger Dodger is at least compelling as a variation on the theme.

Sentiment + Overall Product Rating

Problem : Vader's performance was not too good when run on reviews.

Solution : For a particular product, compared the review-sentiment returned by Vader against the overall product rating.

Result : Much more accurate sentiment analysis.

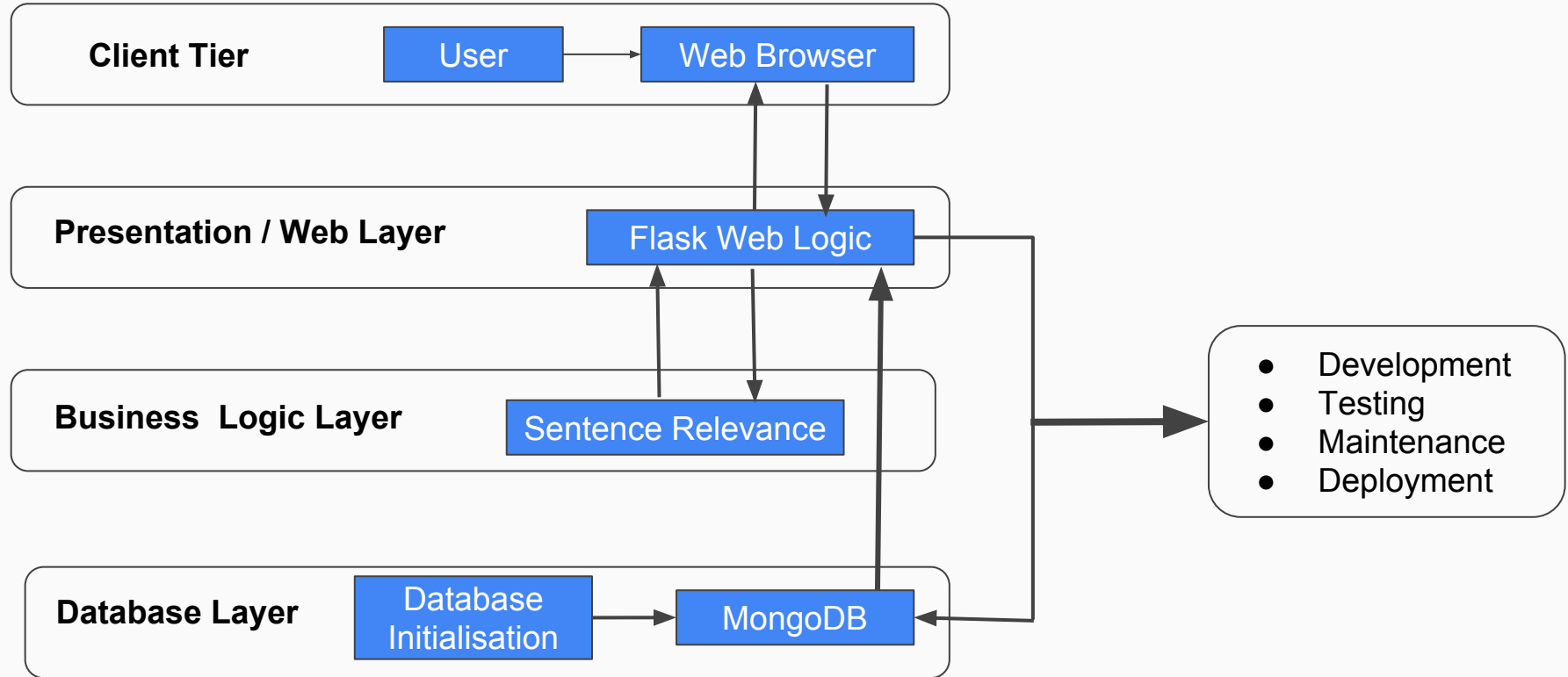| asin	overall	reviewText	VaderSentiments	VaderSentiments+OverallRating
7806397051	1	PLEASE DONT DO IT! this just rachett the palette size is not like in the picture, the colors go on sheer, slides on face like wax don't expect your makeup to stay put with this one. just spend the money for the good stuff. 2 thumbs down	positive	negative
9759091062	3	I have mixed feelings about this product. When I first started using it, I can see and feel the difference. It actually started to literally peel the dark skin away. There's a bit of a sting but that's how I know it was working. My skin was lighter in just a couple weeks! However, once I've stopped using it when I reached my desired skin color, it went back to being dark. I applied more and suddenly it was no longer working. It didn't peel or sting. I wonder what went wrong or if it has an e...	negative	positive
9759091062	1	Did nothing for my skin. Used as suggested and no signs of brightness. Wish it worked for me, but it didn't.	positive	negative
9759091062	1	This gell did nothing at all. I religiously put it on specific areas for weeks and nothing changed. DON'T WASTE YOUR TIME	positive	negative
9790794231	4	This is a very unique scent! It's deep and mysterious and light and floral at the same time. The deep purple bottle is so fitting for the scent- it's dusky, smoky (it doesn't smell like smoke, but it gives off that vibe), with a hint of feminine innocence. This is a hard scent to describe, one that invokes images rather than particular notes. I've not smelled anything like this anywhere else.	negative	positive
B00004TMFE	1	I have thin fine hair and I am always searching for the "miracle" to a beautiful head of hair. This is not it. It may be organic and have less surfactants but, it's the thinnest shampoo I've ever used. Can't get it from bottle to hand to head without losing it down the drain. Stupid purchase. Oh well, next time.	positive	negative
B00004TMFE	5	I bought this along with the shampoo because my previous shampoo was making my hair fall. After some months of usage I can't say for sure if it thickened my hair but at least it helped stop the loss. It leaves my hair very manageable as well. This conditioner is very thick and it didn't take a lot to detangle my hair. Will buy again.	negative	positive

Why not Word2Vec or Doc2Vec ?

- Both need larger and more vibrant corpus to train.

- Training on current dataset did not yield good results.

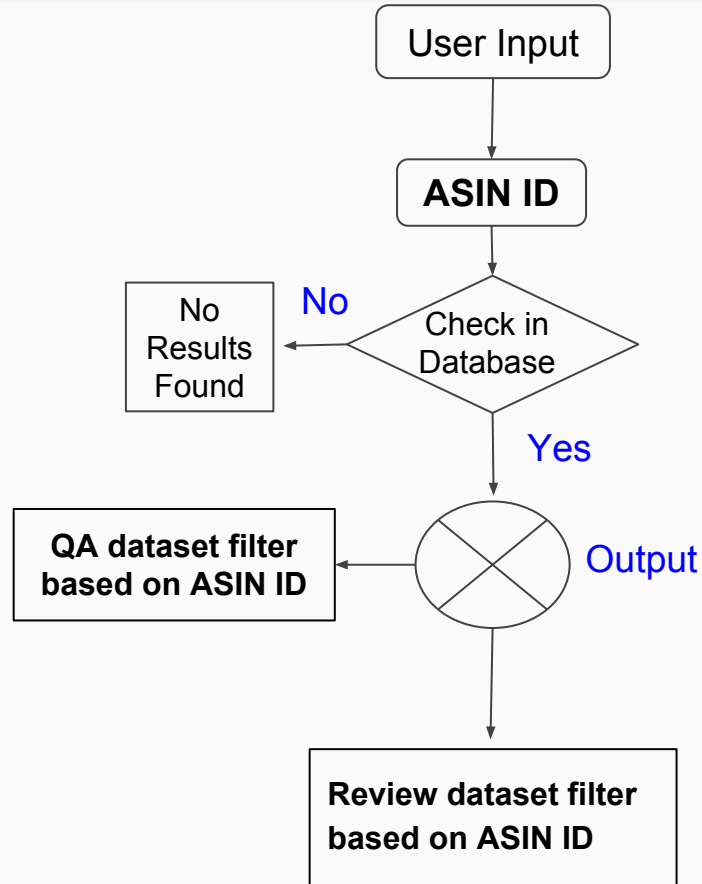- Found TF*IDF + BM25 to be more accurate.

Preprocessing

- **Sentiment :**
  - VaderSentiment analyzer
  - Analysis considering overall product rating - Review Dataset
- **Tokenization :**
  - nltk.word_tokenize
- **Stemming :**
  - Snowball Stemmer
- **Stopwords :**
  - nltk corpus english stopwords.
  - Less than two chars were discarded.
- **WH find :**
  - Performed POS tagging to find the WH words
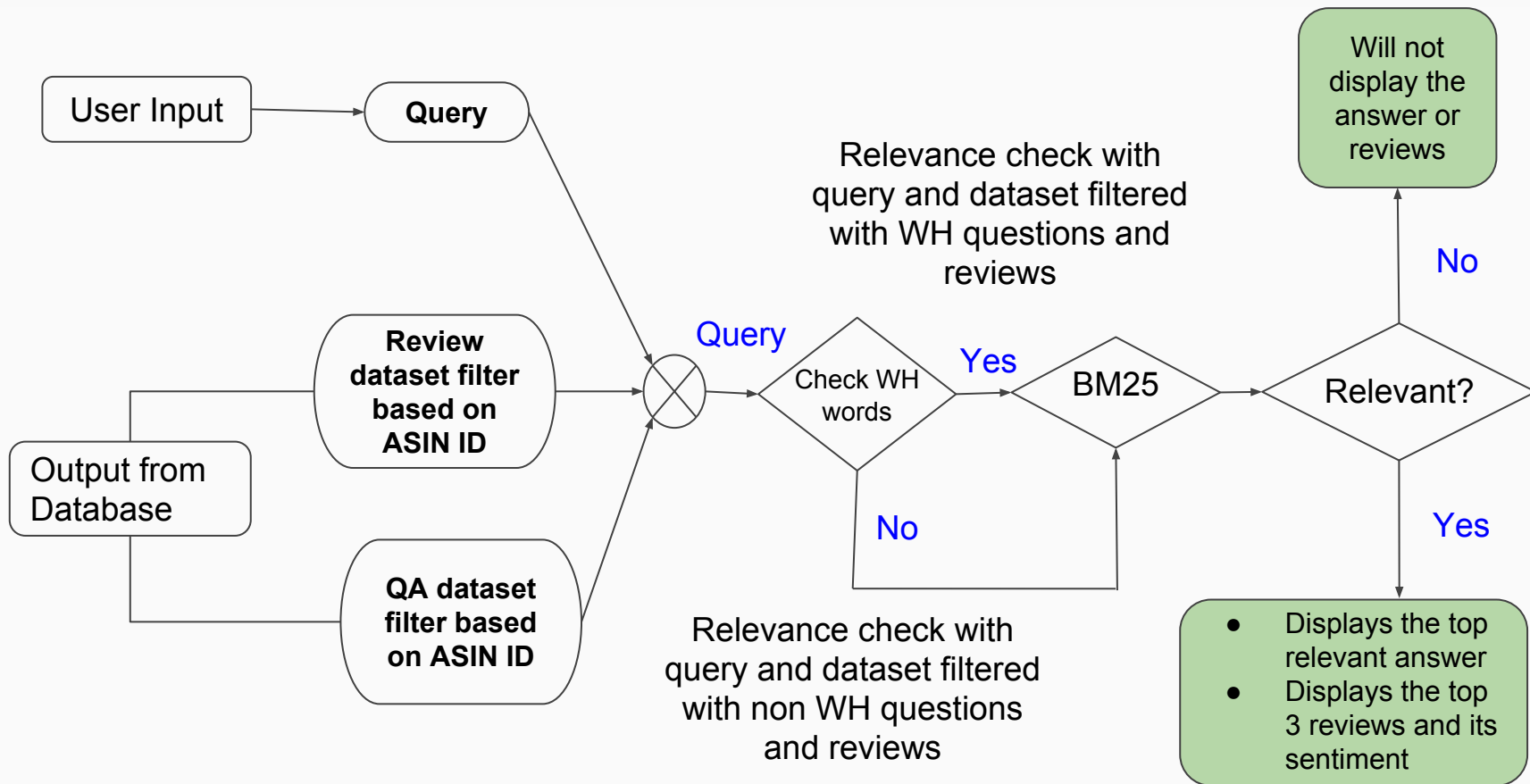
Architecture

Flow Chart - Filters dataset with the ASIN ID

User Input

ASIN ID

No Check in Database

No Results Found

Yes

Output

QA dataset filter based on ASIN ID

Review dataset filter based on ASIN ID

Flow Chart - Business Logic

User Input → **Query**

Output from Database

Review dataset filter based on ASIN ID

QA dataset filter based on ASIN ID

Query

Relevance check with query and dataset filtered with WH questions and reviews

Check WH words

Yes

BM25

Relevant?

Relevance check with query and dataset filtered with non WH questions and reviews

No

No

Will not display the answer or reviews

Yes

- Displays the top relevant answer
- Displays the top 3 reviews and its sentiment

Output Logic

Product	Has answer	Answer Sentiment	No. of complementary reviews	Output
	N	-	-	Top 3 relevant reviews
	Y	P/N	0	Top 3 relevant reviews
	Y	P/N	1	Most relevant review Second most relevant review Complementary third most relevant review
	Y	P/N	2	Most relevant review Complementary second most relevant review Complementary third most relevant review

N - Answers not present in database.
Y - Answers present in database
P - Positive sentiments
N - Negative sentiments
0/1/2 - Number of complementary reviews present in the database

Demo

1. How hard is it to install - B00004TUBL

2. can you fit make up brushes in the trays - 602260074X

3. is this good for me? - 7806397051

4. is it dual voltage 110v and 220V? - B00005O0MZ

Challenges

- Loading of Q&A (qa_Beauty.json) dataset in pandas dataframe

  - demjson.decode( ) + json.dumps( )

- Handling of the word, **"THAT"**.

- Loading Json file to MongoDB

Pitfalls

- Vader performs poorly with longer sentences/paragraphs

- Sentiment analysis using Overall product rating + Vader' might fail.

- End of the day, model does keyword matching.

Future Scope

- Helpfulness Feature : Least helpful reviews would be least relevant for an average user. Hence need to be removed from the dataset. (assumption)

- Topic Modeling for categorising product name from questions & reviews dataset.

Work Activities

Rohan Damodar

- Exploratory Data Analysis
- BM25 Algorithm
- Different Algorithm Analysis
- Business Logic Development

Vinod Boga

- Exploratory Data Analysis
- Architecture Design
- Review Sentiment Algorithm

Nikhil Akki

- Web & UI
- Database setup
- Testing & Presentation
- API Integration

Ankan Roy

- Exploratory Data Analysis
- Database Initialisation logic
- Data Preprocessing
- Business Logic Development

Thank You