Food Fact Search Engine

Jiayi Wang

Background and Use Case

93%

Consumers feel compelled to eat healthy at least some of the time

64%

Consumers consider healthfulness as an important influencing factor while purchasing food

1/4

Consumers actively seek health benefits from food

~30%

Grocery consumers in the US and Europe are health-conscious

Problem

- Rising concerns of health and well-being
- Health Consciousness, under the impact of the pandemic
- Consumers tend to look beyond the brands and shiny labels for keeping a balanced lifestyle and energy level

Use Case

This food search engine is designed to support individual users to explore detailed behind-scene facts of certain food products that match their criteria.

Users will be able to:

- Search a **food keyword** to retrieve its **nutritional elements**
- Search **food product** under certain **category**
- Sort **food** based on **nutrition score** or **certain nutrition**
- Search **food** in a specific **calorie range**

Data Source

Data Source

- Retrieved from **Open Food Facts database**:https://world.openfoodfacts.org/data
- Open, free, crowdsourced
- Data model: relational
- Updated to 2021, update expected in 2022

Data Exploration

Information on food products

- ~130k products in US
- 186 variables including brand, manufacturing location, category, list of ingredients, nutritional information, nutrition grade, etc,.
- **26 key variables** are extracted to build the search engine

Data Procurement

- Pandas package to import data in **CSV file**

```
df.columns
```

```
Index(['code', 'last_modified_datetime', 'product_name', 'nutriscore_grade',
       'brands', 'brand_owner', 'origins_en', 'ingredients_text', 'allergens',
       'additives_en', 'pnns_groups_2', 'energy_kcal_100g', 'fat_100g',
       'saturated_fat_100g', 'trans_fat_100g', 'cholesterol_100g',
       'carbohydrates_100g', 'sugars_100g', 'proteins_100g', 'salt_100g',
       'sodium_100g', 'caffeine_100g', 'calcium_100g', 'iron_100g',
       'zinc_100g', 'fiber_100g'],
      dtype='object')
```

Design Choices /3 Tech

Back-end

MongoDB

- MongoDB is used for storing **documents of full data** about important food facts with each within 16M in PyMongo
- But, MongoDB has **flexibility and efficiency** when importing JSON files (which is more compatible with our food data as it has list in certain columns, faster and easier than PostgreSQL)
- Can upload to MongoDB Atlas for cloud sync and allows multiple kinds of connection with Elasticsearch via reliable replicas to protect data and retrieve changes by Watch methods

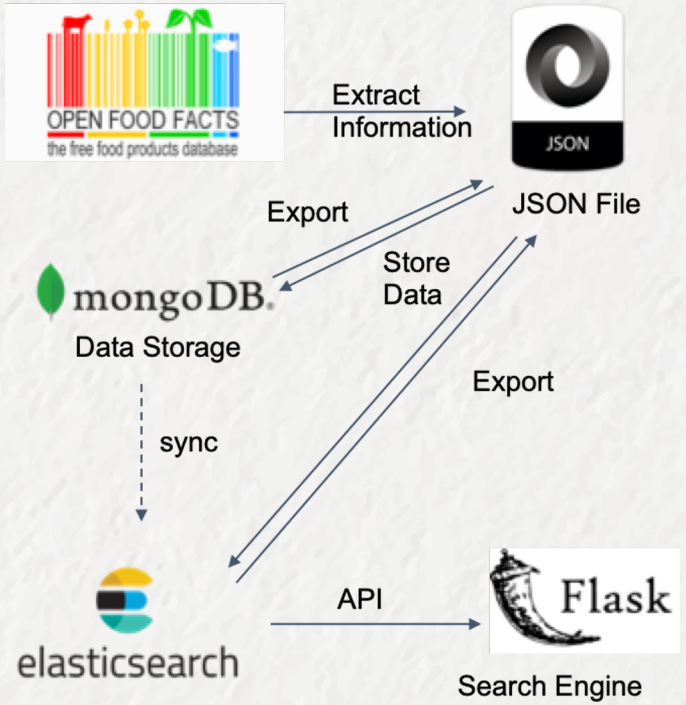
Elasticsearch

- Elasticsearch is a **distributed, scalable, real-time** data analysis engine which can be used for both **open search** and **full-text search**
- Elasticsearch allows combination of queries like "match", "range", and "sort", enabling users to search by food keyword, category, certain calories range and sort the results by certain nutrition
- Our food data is **unstructured,** it's better to use Elasticsearch with emphasis on speed, scale, and relevance. Elasticsearch allows search, analyze and explore **large amounts** of data, can be **synced** with MongoDB in the future (through connector) for data input, and connect to web interface for interaction for output return

Front-end

Flask

- Flask with be used as interactive **web-based interface**, which is a micro web framework used for food fact searching without requiring particular tools or libraries
- Highly **customized** and suitable for technical experimentation for different food searching queries
- **Easy setup and debugging**, flexible and fast to deploy as microservice for food fact searching
- We connect it to Elasticsearch database to form an API system for future commercialized expansion



MongoDB & Elasticsearch

MongoDB

```
1   _id : ObjectId("6260cb275a1c23b044b87b8e")                ObjectId
2   code : 20043131                                          Int32
3   last_modified_dateti… : "2018-10-17T10:18:25Z"           String
4   product_name : "Sliced Plain Bagel"                      String
5   nutriscore_grade : "c"                                   String
6   brands : "Fresh & Easy"                                  String
7   brand_owner : null                                       Null
8   origins_en : null                                        Null
9   › ingredients_text : Array                               Array
10  › allergens : Array                                      Array
11  additives_en : null                                      Null
12  pnns_groups_2 : "Bread"                                  String
13  energy_kcal_100g : 365                                   Int32
14  fat_100g : 1.18                                          Double
15  saturated_fat_100g : 0                                   Int32
16  trans_fat_100g : 0                                       Int32
17  cholesterol_100g : 0                                     Int32
18  carbohydrates_100g : 71.76                               Double
19  sugars_100g : 4.71                                       Double
20  proteins_100g : 12.94                                    Double
21  salt_100g : 1.19634                                      Double
22  sodium_100g : 0.478536                                   Double
23  caffeine_100g : null                                     Null
24  calcium_100g : 0.071                                     Double
25  iron_100g : 0.00424                                      Double
26  zinc_100g : null                                         Null
27  fiber_100g : 2.4                                         Double
```

```
ingredients_text: Array
  0: "Unbleached enriched flour (wheat flour"
  1: " barley malt flour"
  2: " niacin"
  3: " reduced iron"
  4: " thiamin mononitrate"
  5: " riboflavin"
  6: " folic acid)"
  7: " water"
  8: " sugar"
  9: " contains 2% or less of: wheat gluten"
  10: " wheat flour"
  11: " salt"
  12: " yeast"
  13: " distilled vinegar"
  14: " cultured corn syrup solids"
  15: " cul"
allergens: Array
  0: "gluten"
```

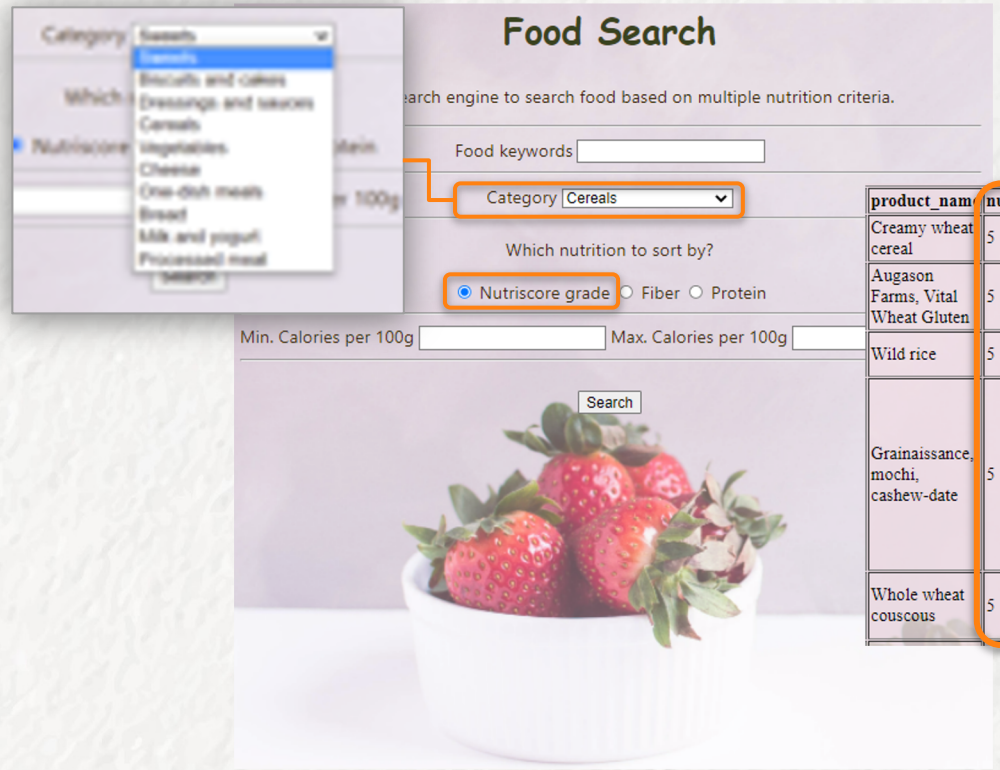
Elasticsearch

```
res =  es.search(index = "food",
                 body = {
                    "sort":{"nutriscore_grade": {"order": "desc"}},
                    "query":{
                        "bool":{
                            "must":[
                                {"match" : {"product_name" : "chocolate"}},
                                {"match" : {"pnns_groups_2" : "Sweets"}},
                                {"range": {"energy_kcal_100g": {"gte": 50}}},
                                {"range": {"energy_kcal_100g": {"lte": 200}}}
                            ]
                        }
                    }
                 })
result = [entry['_source'] for entry in res['hits']['hits']]
result
```

```
[{'code': 9542015350.0,
  'last_modified_datetime': '2020-07-28T17:17:26Z',
  'product_name': 'Classic recipe milk chocolate pretzel bar',
  'nutriscore_grade': 3,
  'brands': 'Lindt,  Lindt & Sprungli (Usa) Inc.',
  'brand_owner': 'Lindt & Sprungli (Schweiz) AG',
  'origins_en': None,
  'ingredients_text': ['Sugar',
   ' cocoa butter',
   ' milk',
   ' pretzel pieces [corn starch',
   ' palm oil',
   ' potato starch',
   ' sea salt',
   ' sugar',
   ' cellulose gum',
   ' soya lecithin',
   ' leavening (sodium bicarbonate',
   ' sodium acid pyrophosphate)',
```

Flask Demo

Use case example 1:
search for foods with **highest nutrition grade** in certain **category**

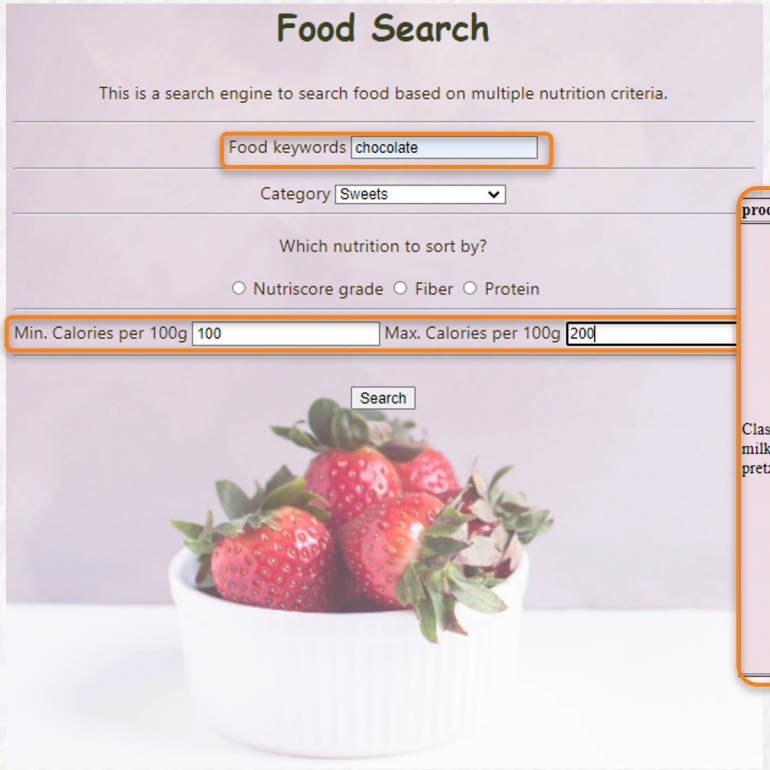


Cereals Sort by Nutrition Grade

product_nam	nutriscore_grade	brands	brand_owner	origins_en	ingredients_text	allergens	additives_e	pnns_groups_2
Creamy wheat cereal	5	None	Blue Chip Group Inc.	None	['Wheat.']	['gluten']	None	Cereals
Augason Farms, Vital Wheat Gluten	5	Blue Chip Group	Blue Chip Group Inc.	None	['Vital wheat gluten.']	None	None	Cereals
Wild rice	5	Trader Joe's	None	None	['California wild rice.']	None	None	Cereals
Grainaissance, mochi, cashew-date	5	Grainaissance	GRAINAISSANCE	None	['Organic sweet brown rice', ' filtered water', ' dates', ' cashews', ' cinnamon', ' natural vanilla flavor', ' and sea salt.']	None	None	Cereals
Whole wheat couscous	5	Trader Joe's	Y I, Inc.	None	['whole durum wheat semolina', '']	None	None	Cereals

Flask Demo

Use case example 2:
search for food with certain **keyword** and **calories range**



Chocolate with Calories from 100 to 200 Kcal

product_name	nutriscore_grade	brands	brand_owner	origins_en	ingredients_text	allergens	additives_en	pnns_groups_2	energy_kcal_100g
Classic recipe milk chocolate pretzel bar		Lindt, Lindt & Sprungli (Usa) Inc.	Lindt & Sprungli (Schweiz) AG	None	['Sugar', ' cocoa butter', ' milk', ' pretzel pieces [corn starch', ' palm oil', ' potato starch', ' sea salt', ' sugar', ' cellulose gum', ' soya lecithin', ' leavening (sodium bicarbonate', ' sodium acid pyrophosphate)', ' yeast extract]', ' chocolate', ' skim milk', ' soya lecithin (emulsifier)', ' barley malt powder', ' sea salt', ' artificial flavor.']	None	['E322 - Lecithins', 'E322i - Lecithin', 'E450 - Diphosphates', 'E450i - Disodium diphosphate', 'E466 - Sodium carboxy methyl cellulose', 'E500 - Sodium carbonates', 'E500ii - Sodium hydrogen carbonate']	Sweets	170.0

Data Governance & Future Concerns

Data Governance

Governance Responsibility

- **Governance Council :** responsible for data ownership, compliance with data policies and data governance process and procedures.
- **Data Stewards** : responsible for data content, context, and associated business rules , which are what is stored in a data field.
- **Data Custodians** : responsible for the safe custody, transport, storage of the data and implementation of business rules ,which are technical environment and database structure.

Data License

- **Open Database License** : The Open Food Facts database is available under the ODbl. Rights granted by ODbl Licensor **include commercial use** and the right to sublicense the work.
- **Database Contents License** : The individual contents of the database are available under the DbCL. Rights granted by DbCL Licensor **include commercial use** and the right to sublicense the work.
- **Creative Commons Attribution ShareAlike licence** : Products images are available under the CC BY-SA. Rights granted by CC BY-SA Licensor include sharing and adapting data in **commercial use**, but they must license their new creations under the same terms.

Further Concerns

Scaling up

Potential Problems:

- **Search problem**: Increase difficulty of finding the information needed. The bigger the data set, the more intensive it is to find what we're looking for.
- **Concurrency problem**: Increase difficulty of making data available to several people and programs simultaneously, which may japodize availability of data.
- **Consistency problem**: Increase difficulty of dealing with constantly updated data whose updates need to be reflected to the people and programs using that data, which may japodize integrity of data.
- **Speed problem**: Increase difficulty of handling more requests or transactions within specific time.

Solutions:

- Organizie data in alphabetical order
- Create replication
- Publish change logs

Database update

The category selections in our search engine options is based on the popularity (top 10) of the food category in existing database. If we update database, the popularity level may change which means we need monitor the database and **revise search engine** constantly with database update.

Cost Analysis

Cost Analysis

Cost Detials

The openness and completeness of the data means there is **no cost to acquire the data** .

Monthly costs for MangoDB and Elasticsearch are calculated like:

- **MangoDB Cost per month** = Base price per hour (based on storage size) * 24hr * 30day
- **Elasticsearch Cost per month** = $0.10 * storage size * 24hr * 30day

Monthly costs for **Website Hosting Service** is $250

Storage Size	MongoDB Atlas	Elasticsearch	Heroku Website Hosting Services	Total cost per month
10 GB	$57	$30	$250	$337
20 GB	$144	$60	$250	$454
40 GB	$389	$120	$250	$759
80 GB	$749	$240	$250	$1,239
160 GB	$1,440	$480	$250	$2,170

Reference

- MangoDB Pricing:

https://www.mongodb.com/pricing

- Elasticsearch Pricing:

https://coralogix.com/blog/elasticsearch-pricing-is-aws-really-cost-effective/

- Heroku website hosting services Pricing:

https://www.heroku.com/pricing

Quantitative & Qualitative Success Metrics

		Industry Benchmark
Quantitative	Visitor Volume	The Number of Visitors Come to Food Search Engine
	Search Rankings	The Website Ranks for The Relevant Keywords Excluding the Paid Ads
	Bounce rate	The Percentage of Visitors Who Leave The Site without Taking Any Action **< 40-60%**
	Average Session Duration	The Average Time User Spends on The Page of The Food Search Engines **> 3 Mins**
	PageSpeed	Response Speed of Searching Food Information < **5.4 Secs**
	Coverage	Food Facts Database Coverage Size and Update Frequency **[Weekly]**
	Relevance of Search	Recall: Number of Relevant Documents Retrieved/ Total Number of Relevant Documents
		Precision: Number of Relevant Documents Retrieved/ Total Number of Documents Retrieved
		Comprehensiveness: The Scope of Relevant Documents
Qualitative	Functionality	Implement Basic Functions of Search Engine Boolean Logic and Scope Limiting Abilities
	User Satisfaction	Users' Satisfaction with our food Search System (Questionnaire to Obtain Feedback)

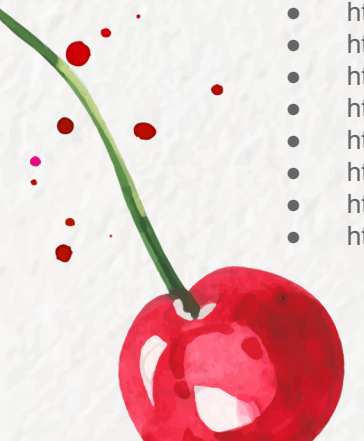
Conclusion and Recommendations

We extracted and cleaned valuable information from raw food fact dataset, loaded them into MongoDB. We further developed a food search engine based on Elasticsearch and Flask that allows multiple search criteria on food facts

Recommendation

1. **To scale up on cloud:**
    - From **Localhost** to **Cloud-based**: Using **MongoDB atlas** upgraded subscription accounts to access a virtual private cloud database with great level of security including IP whitelisting and encryption
    - **Sync with Elasticsearch**: Using MongoDB **replica** and MongoDB **connector** to automatically sync data from MongoDB to Elasticsearch for future data import and storage.
2. **To scale up user base:** expand to Asian/European users by including food data from countries other than US and supporting multiple languages
3. **Develop commercialized use case:** combine with the recommendation system and explore commercial applications
4. **Improvement based on user feedback**
    - Add functionality of website (e.g. more searching criteria)
    - Improve the database coverage by adding other data sources of food nutrition and update data regularly to ensure data integrity
    - Optimize search accuracy and improve page response speed

Thank you!
Questions?



Reference:

- https://world.openfoodfacts.org/data
- https://www.mongodb.com/
- https://opendatacommons.org/licenses/odbl/1-0/
- https://opendatacommons.org/licenses/dbcl/1-0/
- https://creativecommons.org/licenses/by-sa/3.0/deed.en
- https://www.renolon.com/health-conscious-consumer-statistics/
- https://www.bluecorona.com/blog/how-fast-should-website-be/
- https://www.klipfolio.com/metrics/marketing/average-time-on-page
- https://www.conductor.com/learning-center/organic-website-traffic-industry-benchmarks-2022/
- https://www.analyticsvidhya.com/blog/2021/06/part-20-step-by-step-guide-to-master-nlp-information-retrieval/