POLITECNICO
MILANO 1863

System and Methods for Unstructured Data Project: Amazon Reviews

Tommaso Capacci (10654230)

AY 2023/2024

Contents

1 Introduction

The dataset we've chosen to use represents a collection of reviews of products sold on Amazon. Datasets like this are usually used for **analytic purposes**, such as understanding the customers' needs and preferences, for **marketing purposes**, such as collecting the customers' opinions on a product, or for **sentiment analysis**, such as training a classifier to predict the sentiment of a review basing on its content, since information about the review's text is also provided.

In our case, we will consider the first task, since it can be reduced to the problem of writing a suited set of queries and can be also efficiently supported by the database technology we decided to use, which is MongoDB. We've decided to interpret the task of strategic analysis as the problem of extracting meaningful statistics from the dataset (such as the number of reviews per product, the average rating per product, the review with highest score, ...) which could be then included on the company's reports and used to make strategic decisions.

One of the main reasons behind our technical choice of using MongoDB as database technology is the fact that the dataset is in JSON format, which is natively supported by MongoDB. Moreover, it is composed of a collection of, possibly, nested documents, which is also a good fit for MongoDB, since it is a document-oriented database, meaning that it is specifically designed to store data as documents instead of relational tables.

2 Dataset

The dataset of our choice is managed by the University of California San Diego (UCSD) and is presented in this website. It is composed of multiple collections of documents, each one containing some reviews, registered between the 1996 and 2018, about a specific kind of products sold on Amazon. This fact makes the dataset suitable for our purposes, since it contains a lot of reviews but also would make it possible to filter the reviews by product category, allowing to eventually focus the analysis on a specific kind of goods.

In particular, in the website cited above, are provided 2 versions of the same datasets:

- *5-core*: this is a reduced version of the original, complete dataset, in which are included only the 5 most representative reviews per product
- *ratings only*: in this version of the dataset only information about the user, the product, the rating and the timestamp of pubblication time is kept for each review

While the second version of the dataset contains an higher amount of reviews, we decided to use the first one since it contained more interesting fields.

Among all the available collections, we've decided to use the one about **Digital Music** (which can be retrieved at this link) since this file includes a suitable number of datapoints (169.781) while still containing an interesting amount of attributes per document. Specifically, the documents inside the selected collection have the following shape:


```
{
    "image": ["https://images-na.ssl-images-amazon.com/images/..."],
    "overall": 5.0,
    "vote": "2",
    "verified": true,
    "reviewTime": "01 8, 2015",
    "reviewerID": "A36GE53TK8V94L",
    "asin": "B000T1EJ0W",
    "style": {"Format:": "MP3 Music"},
    "reviewerName": "MysticWolf229",
    "reviewText": "the theme song to the one and only movie that ...",
    "unixReviewTime": 1420675200
}
```

whose fields have the following meaning:

- **image**: an array of URLs pointing to the images of the product review;
- **overall**: the rating of the product, a float number going from 1 to 5;
- **vote**: the number of votes the review received, saved as a string;
- **verified**: a boolean value indicating whether the review has been verified or not;
- **reviewTime**: the date of the review, saved in RAW date format as a string;
- **reviewerID**: the alphanumeric ID of the user who wrote the review;
- **asin**: acronym of Amazon Standard Identification Number, is the alphanumeric ID that represents a specific product;
- **style**: a subdocument containing additional data about the version of the product. In the case of this collection it just contains information about the format of the product, whose possible values will be retrieved with a suited query;
- **reviewerName**: a string containing the name of the user who wrote the review;
- **reviewText**: a string containing the text of the review;
- **summary**: a string containing a summary of the review;
- **unixReviewTime**: the date of the review in Unix epoch time format.

3 Data Wrangling

After a quick inspection of the dataset we understood that the data entries it contained were mostly complete but, at the same time, we didn't have enough information to infer the value of missing attributes for incomplete documents.

One useful thing that could have been done in the scenario of a real analysis would be to find another dataset containing technical data about single products, together with their ASIN, and use this information to compute a join over this attribute and complete our dataset with this additional data: anyway, while in the presented setting this information would be very easy to be retrieved, this was not our case, since we weren't able to find any additional dataset with these characteristics and we couldn't afford (both from resources and time perspective) to perform a snapshot of the Amazon website through scalping.

The only interesting thing that is worth mentioning is the fact that we used a python script to extract a subset of the dataset:

```
import os

os.chdir("../Datasets")
json_file_path = os.path.join(os.getcwd(), "Digital_Music.json")
n = 150_000
suffix = str(n//1000) + "k"
output_file_path = json_file_path[:-5] + "_" + suffix + ".json"

with open(json_file_path, 'r') as input_file:
    with open(output_file_path, 'w') as output_file:
        for i in range(n):
            # BE SURE THAT EACH OBJECT OCCUPIES *EXACTLY* ONE LINE
            output_file.write(input_file.readline())
```

Specifically, as it can be seen, we used this script to round the number of documents to the nearest multiple of 50.000 by extracting the first lines of the original file. It is worth noticing that the core lines of this script can be easily modified to perform any kind of filtering over the JSON documents in this way:

```
import json

with open(json_file_path, 'r') as input_file:
    with open(output_file_path, 'w') as output_file:
        i = 1
        while(i <= n)
            # BE SURE THAT EACH OBJECT OCCUPIES *EXACTLY* ONE LINE
            line = input_file.readline()

            # Check if the line is not empty
            if line:
                json_object = json.loads(line)

                # Generic filter
                if filter:
                    output_file.write(line)
                    i += 1
```

Anyway we decided to not apply any further filtering since we wanted to keep as many attributes as possible in order to be more free in the writing of the queries, which will be presented in the next chapter.

4 Queries

```
db.reviews.aggregate([
    {"$unwind": "$style"},
    {"$group": {"_id": "$style"}}
])
```