IT University of Copenhagen
Large Scale Data Analysis

Scalable Data Processing

Yelp Reviews and Authenticity

March 10, 2025

1 Introduction

Unlocking Insights from big data

In this assignment, we embark on an exciting expedition into the realm of Big Data, focusing on the exploration of large datasets[1], and extracting meaningful insights from these. You will learn to use some tools built to deal with large datasets in a distributed manner. The task will require a blend of your technical expertise, analytical acumen, and creative thinking to navigate through JSON files efficiently and extract actionable insights. You will harness the power of distributed computing (by using a Spark cluster) for proving hypotheses, uncovering hidden insights, and even predicting trends in data. So, fasten your seatbelts and prepare to embark on a data-driven adventure that will challenge your skills, expand your horizons, and empower you to harness the true potential of Big Data.

Authenticity Study

In 2019, an interesting study looked into the use of the term "authenticity" in Yelp reviews and how it was used to signal different characteristics from popular cuisines. The author claims that, for example, in Chinese and Mexican restaurants, the term "authenticity" was used to describe typically negative things like "dirty", "kitsch" or "cheap", where in Italian and French restaurants was used to signal high quality and delicious details. Thus, the author concludes that the use of the term "builds an authenticity trap where reviews reinforce harmful stereotypes".

The study was done by manually reading and classifying 20,000 reviews for restaurants in New York City. The dataset you will be using is vastly larger (6,685,900 reviews) and much more geographically diverse. We want to test the author's conclusions with this richer dataset, but reviewing the data manually would be just unfeasible. Luckily, you have access to distributed computing and great skills.

[1] The datasets are not huge, but big enough that working with them on a normal laptop would quickly become inefficient.

2 The Data

The data you will use is from the YELP ACADEMIC DATASET (documentation), a dataset of reviews and tips given by yelp users to business. The version of the dataset we are using takes up around 8 GB uncompressed and includes more than 6 million reviews.

The data is saved as JSON-files, which have been uploaded to UCloud. UCloud is a digital research environment hosted at The University of Southern Denmark (SDU), but accessible for all Danish Universities via WAYF. It provides data handling and analysis powered by a supercomputer that can handle computations and programs too time-demanding to run on a personal computer as, for example, access to a HPC environment and other computing environments for software development, data engineering, big data analytics, machine learning, and artificial intelligence. UCloud also provides cloud data storage, allowing users to analyze and share data.

The JSON-files are:

```
FULL PATH NAME                                                  SIZE
/datasets/yelp/yelp_academic_dataset_business.json            131.9 M
/datasets/yelp/yelp_academic_dataset_checkin.json             389.9 M
/datasets/yelp/yelp_academic_dataset_review.json               5.0 G
/datasets/yelp/yelp_academic_dataset_tip.json                  233.2 M
/datasets/yelp/yelp_academic_dataset_user.json                 2.3 G
```

Reading files in Spark is done with the `spark.read`-module, which has methods for different file formats. You will have a template file to start working with and some basic instructions for reading the files and looking at the data. Spark can read the files from the drive and convert them to in-memory DataFrames.[2]

3 Requirements and Hand-in

The assignment consists of three sections: In the first, you will create specific queries, to get familiarized with the data. In the second part you will have more freedom, and the objective is to try to answer high-level questions by querying the data and extracting insights. In the final part of the assignment, you will do some feature engineering and machine learning to arrive to some predictions.

You need to compile everything into a pdf report where you give your solution and explain your decisions, along with a discussion of your approach to solving the tasks. Feel free to use screenshots from part of your code and write in your report what you want feedback on.

The maximum length of the report should be 6 pages, excluding figures.

3.1 Specific DataFrame Queries

Formulate the following queries using Spark DataFrames.

1. Find the total number of reviews for all businesses.
2. Find all businesses that have received 5 stars (on average: Use business.stars column) and that have been reviewed by 500 or more users. The output should be in the form of a DataFrame of (name, stars, review count).
3. Find influencers who have written more than 1000 reviews. The output should be in the form of a Spark Table/DataFrame of user id.
4. Find the businesses names that have been reviewed by more than 5 influencer users. You can use a view created from your answer to Q3.
5. Find an ordered list of users based on the average star counts they have given in all their reviews.

[2] Not the same as Pandas DataFrames, but inspired by it.

Note: **do not use SQL queries, i.e., the `spark.sql` method. For example do not write:**

```
business.sql('SELECT * FROM ... WHERE ...')
```

Instead use the "pandas-style":

```
business.select('*').filter('...')
```

However, you are allowed to import functions from `pyspark.functions.sql`.

3.2 Authenticity Study

The next questions should be answered using statistics found in the data. You are free to use the `spark.sql` method for this section.

3.2.1 Data Exploration

Look in the data for the use of "authenticity language", as defined in the Eater New York article. These queries should include (but not be limited to) the following questions:

- What is the percentage of reviews that contain a variant of the word "authentic"?
- How many reviews contain the string "legitimate" grouped by type of cuisine?
  Note: The data unfortunately doesn't have a "cuisine type" column you can group by. So you need to be creative to answer this question
- Is there a difference in the amount of authenticity language used in the different areas? (e.g., by state, north/south, urban/rural)
  Note: As part of answering this question, you could compute the full cube or rollup combining the location of the business and whether the review contains authenticity language, and use this to aggregate their counts per state and city.

Explain your exploratory approach. Explain the queries you formulate, the results you get and how they inform your further exploration.

3.2.2 Hypothesis Testing

The hypothesis proposed in the article is that authenticity language is used to describe different characteristics for different cuisines (and by extension, makes it harder for non-white restaurant owners to enter the higher-end restaurant market).

- Can you identify a difference in the relationship between authenticity language[3] and typically negative words [4], in restaurants serving South American or Asian cuisine compared to restaurants serving European cuisine? And to what degree?

Explain your approach, assumptions, and results. Explain how your queries actually answer the question. Think critically about your approach, its strengths, and limitations.

Note: You might have to be creative in using the categories feature/column for this task. Think about how they structured the data, perhaps adding new columns could help.

3.3 Rating Prediction

We now want to build an ML model that predicts the rating of a certain restaurant given a user review. For this task, you should use the MLlib library. Built on top of Spark, MLlib is a scalable machine learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, and underlying optimization primitives.

Your code should contain a set of (pre)processing steps, such as, but not limited to:

[3] words such as "authentic" or "legitimate" or their derived forms
[4] like "dirty", "kitsch", "cheap", "rude", "simple" or similar

- splitting the dataset into training and test sets,
- extracting features from the reviews (think what features could potentially influence the ratings),
- training your model on the training set (e.g., SVM),
- and evaluating your model on the test set (with your metric of choice, e.g. MSE).

Explain your choices in each of the above steps. With which task (regression, classification, clustering, etc.) did you decide to model your problem? Explore the possibility of using authenticity language and location as features in your model. Does this improve the quality of your model?

Recall to check MLLib documentation, and ask the TAs whenever in doubt.

3.4 Suggested reading and useful links

- Spark 3.3.1 Quick Start (Remember to choose the python tab)
- Spark SQL Programming Guide (Much more detail)
- Architecture Overview of Cluster-based Spark
- The PySpark API Documentation for version 3.3.1 (Remember to always use the documentation for this version). Especially these modules will be useful to you:
  - DataFrame
  - Column
  - Spark SQL Functions
- MLlib: Apache Spark's scalable machine learning library
- A Tale of Three Apache Spark APIs: RDD vs. Dataframes and Datasets