**Amazon Review Classification Using Clustering in PySpark**

A PROJECT REPORT

***Submitted by***

BL.EN.U4AIE20024 ------ Kanisettypalli Harshitha

BL.EN.U4AIE20026 ------ Kavya Duvvuri

BL.EN.U4AIE20031 ------ Kunisetty Jaswanth

**19AIE214 – Big Data Analytics**

**B.Tech. in Computer Science and Engineering (Artificial Intelligence)**

Logo

Description automatically generated with low confidence

AMRITA SCHOOL OF ENGINEERING, BANGALORE

AMRITA VISHWA VIDYAPEETHAM

BANGALORE 560 035

July – 2022

Table of Contents

[Amazon Reviews Dataset 3](#_Toc108539333)

[MeDAL Dataset 3](#_Toc108539334)

[Workflow 3](#_Toc108539335)

[Execution 4](#_Toc108539336)

[Changes made in the spark configuration file 5](#_Toc108539337)

[Conclusion 6](#_Toc108539338)

[References 6](#_Toc108539339)

# Amazon Reviews Dataset

Amazon is a multinational company that provides an online platform for shopping a wide range of products. Millions of Amazon products are bought every year and they receive mixed reviews. It is extremely necessary to identify the sentiment behind these reviews and modify the products accordingly. This helps the company in gaining more customers. Also, knowing the general rating of the previous reviews helps customers in making better choices when selecting their products.

Amazon reviews are large in number and come in real-time. Standalone systems with normal computational capabilities are not capable of processing data on such a large scale. To address this problem, in our project, we have made use of the concept of clustering in big data analytics and processed the amazon dataset on three different laptops.

# MeDAL Dataset

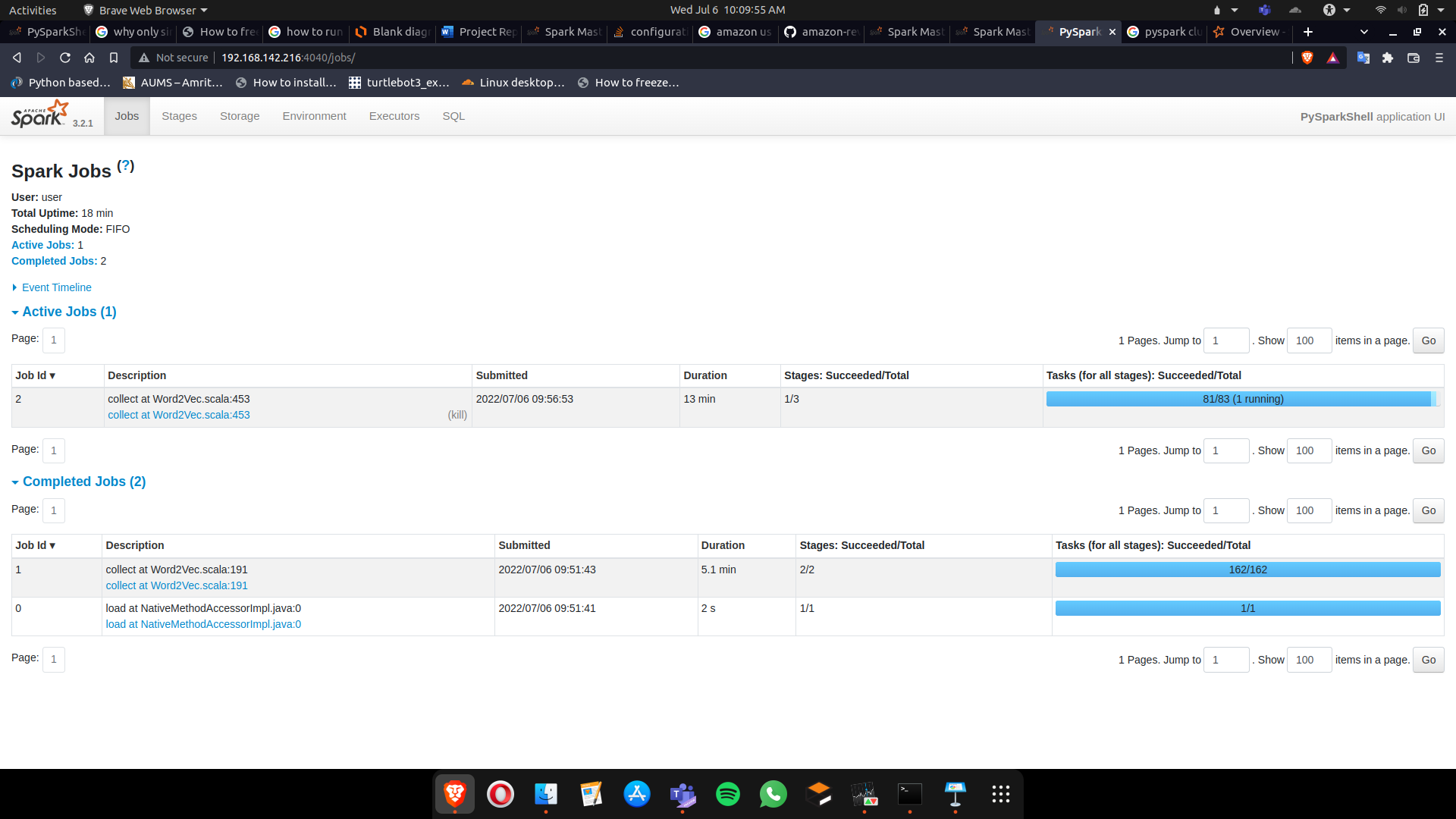
Abbreviation sense disambiguation is the logic that determines which of an abbreviation sense is the most relevant for a given instance of an abbreviation. In this project, medical abstracts have been taken up and the full form for a given medical abbreviation has been found. There can be multiple options available but selecting the best word is the key challenge of the problem. This dataset is 15 GB large, and it requires high computational strength of systems to be able to fit a large amount of data.

# Workflow

The workflow of the project involves multiple steps:

1. Downloading the dataset and placing the downloaded files in a directory together
2. Writing the code for preprocessing the dataset and obtaining the raw features required for classification
3. Try out multiple Machine Learning models to analyze and classify the data
4. Connect the laptops in a cluster using a hub or a switch and RJ45 cables
5. Place the code on all laptops in the same location
6. Start the master on one of the systems and slaves in the rest of the systems
7. Start the PySpark shell on the master and run the code in the shell
8. Note the time taken for processing in different cases and make observations

# Execution



A picture containing graphical user interface

Description automatically generated

## Changes made in the spark configuration file

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a video game

Description automatically generated

# Conclusion

* The amount of data generated is continuously increasing
* The necessity of processing data is increasing
* Healthcare sector development necessity increasing with population
* Big Data is an upcoming field that has great importance in healthcare

# References

<https://spark.apache.org/docs/2.1.0/ml-classification-regression.html>

<https://stackoverflow.com/questions/40981512/spark-cores-tasks-concurrency/41171632#41171632>

<https://stackoverflow.com/questions/34597923/how-to-set-the-number-of-partitions-nodes-when-importing-data-into-spark>

<https://spark.apache.org/docs/latest/>

<https://www.kaggle.com/datasets/xhlulu/medal-emnlp>

<https://www.kaggle.com/code/cynthiarempel/amazon-us-customer-reviews/data>