**TEAM 9 - FINAL REPORT**

**SCALABLE DATA ANALYTICS**

**TABLE OF CONTENTS**

1. **Abstract \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 3**
2. **Introduction \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**
3. **Motivation \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**
4. **Dataset Description \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**
5. **Data Pre-processing \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**
6. **Topic Modeling \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**
7. **Abstract**

The hospitality industry has been revolutionized with the advent of online platforms that allow customers to leave reviews of their experiences with restaurants. By analyzing these customer reviews, businesses can identify their strengths and weaknesses and make improvements accordingly. In this project, we present a novel approach to extract the highlight features of restaurants using topic modeling with PySpark.

Sushi has become an increasingly popular cuisine in recent years, with many restaurants offering their own unique takes on the classic dish. To help sushi lovers find the best sushi in their area, we present a data-driven approach to discovering and rating the best sushi in restaurants using Yelp reviews.

Our approach involves using topic modeling to extract the most frequently occurring topics related to sushi in Yelp reviews. We then use a rating system to score each restaurant based on the frequency and sentiment polarity of these topics in their reviews. By combining these scores with other features such as price range and location, we are able to identify the top-rated sushi restaurants in a given area.

To test our approach, we applied it to a dataset of Yelp reviews for sushi restaurants in the United States. Our analysis revealed the most commonly discussed topics in sushi reviews, as well as the highest-rated sushi restaurants in various cities.

Our results demonstrate the potential for using data-driven approaches to discover and rate the best sushi in restaurants. By analyzing large volumes of customer reviews, we can provide valuable insights into the strengths and weaknesses of different restaurants and help sushi lovers find the best dining experiences in their area.

Overall, our approach provides a powerful tool for restaurant owners and managers to improve their businesses by identifying areas that need improvement and promoting their strengths. It can also assist customers in making informed decisions about which service or product to purchase, ultimately improving customer satisfaction.

1. **Introduction**

In today's world, customer reviews are more important than ever before. With the rise of online review platforms such as Yelp, customers can easily share their experiences with others and businesses can gain valuable insights into their operations. The vast amount of data generated by customer reviews presents both an opportunity and a challenge for businesses. While this data can provide valuable insights into customer preferences and satisfaction levels, it can also be overwhelming to analyze and make sense of.

In this project, we present a data-driven approach to analyzing customer reviews using PySpark. Specifically, we focused on discovering and rating the best sushi restaurants in our target area using Yelp review data. We utilized two datasets: the Yelp Business dataset and the Yelp Reviews dataset. The Yelp Business dataset provided information about each business, such as its name, location, and description, while the Yelp Reviews dataset contained the actual reviews left by customers.

To begin our analysis, we used the Google SDK to push data to Google Cloud HDFS and create a cluster to run iPython Notebook. We then loaded a CSV file into a Spark DataFrame and filtered out the businesses that were relevant to our analysis. Specifically, we filtered for businesses that had sushi in their description, as we were interested in identifying the best sushi restaurants in our target area.

Once we had our filtered dataset, we performed text preprocessing tasks such as removing punctuation and stopwords on the Spark DataFrame. We used Spark ML's CountVectorizer to extract topics from each business's reviews. CountVectorizer is a powerful tool that converts a collection of text documents into a matrix of token counts, which can then be used to extract topics from the text.

After we had extracted topics from the reviews, we applied Spark UDFs, RDD operations, and higher-order functions to calculate popularity scores for each extracted topic. Popularity scores were calculated based on the frequency and polarity of the topics in the reviews. We then used Spark Streaming to label whether a review is positive or not using a simple if-else classification.

Overall, our approach provides a powerful tool for businesses to uncover insights from customer reviews and make data-driven decisions to improve their operations. By analyzing customer reviews, we can identify the strengths and weaknesses of each restaurant, and help owners and managers improve their business accordingly. Additionally, we assign a comparable normalized score to these highlights for the comparison of restaurants, making it easier for customers to decide which restaurant to visit when they're in the mood for sushi or any specific dish. All our analysis is based on customer reviews, making it an accurate and reliable representation of the restaurant's performance. Overall, this project has the potential to improve customer satisfaction, boost business success, and provide a valuable resource for sushi lovers.

1. **Motivation**

The restaurant industry is highly competitive, and customers often rely on online reviews to decide where to dine. However, with the abundance of information available on review sites, it can be overwhelming for customers to make a decision. Moreover, restaurants are always looking for ways to improve their business and gain an edge over their competitors.

This project aims to address these challenges by analyzing customer reviews of sushi restaurants on Yelp and using topic modeling and sentiment analysis techniques to identify the strengths and weaknesses of each restaurant. By doing so, we can help customers make more informed decisions and help restaurants improve their business by focusing on their strengths and addressing their weaknesses. Additionally, by rating the best sushi restaurants based on their reviews, we can provide a valuable resource for sushi lovers and help restaurants attract new customers. Overall, this project has the potential to improve customer satisfaction and boost the success of sushi restaurants.

1. **Dataset Description**

The Yelp dataset is a comprehensive collection of customer reviews, ratings, and business information for various categories of businesses, including restaurants, bars, and cafes, among others. The dataset contains over 8 million reviews from more than 200,000 businesses across many metropolitan areas in four countries. In this project, we focused on two specific datasets from Yelp - the Yelp Business dataset and the Yelp Review dataset.

The Yelp Business dataset includes information about businesses, such as their name, location, category, ratings, and review counts. We used this dataset to filter out the restaurants that offer sushi from the rest of the businesses. This dataset contains over 200,000 businesses across multiple categories and locations.

The Yelp Review dataset includes customer reviews for various businesses, including the filtered restaurants that offer sushi. This dataset contains over 8 million reviews and includes information about the user, the business, and the review itself. We used this dataset to analyze customer reviews for each restaurant, extract the relevant topics, and determine the sentiment of each review. The dataset provided us with a rich source of information to gain insights into customer preferences, highlight the strengths and weaknesses of each restaurant, and help businesses make data-driven decisions.