# 

# **Introduction to Dataset - rephrase**

This dataset represents a selection from Yelp’s extensive collection of businesses, reviews, and user information. It was initially compiled for the Yelp Dataset Challenge, an opportunity for students to explore and analyze Yelp data, and to share their findings.

From the original pool of 6,990,280 reviews, we have extracted a subset of 100,000 rows. This latest dataset encompasses details on businesses across eight major metropolitan areas in the USA and Canada.

The dataset contains 21 columns with various types of data:

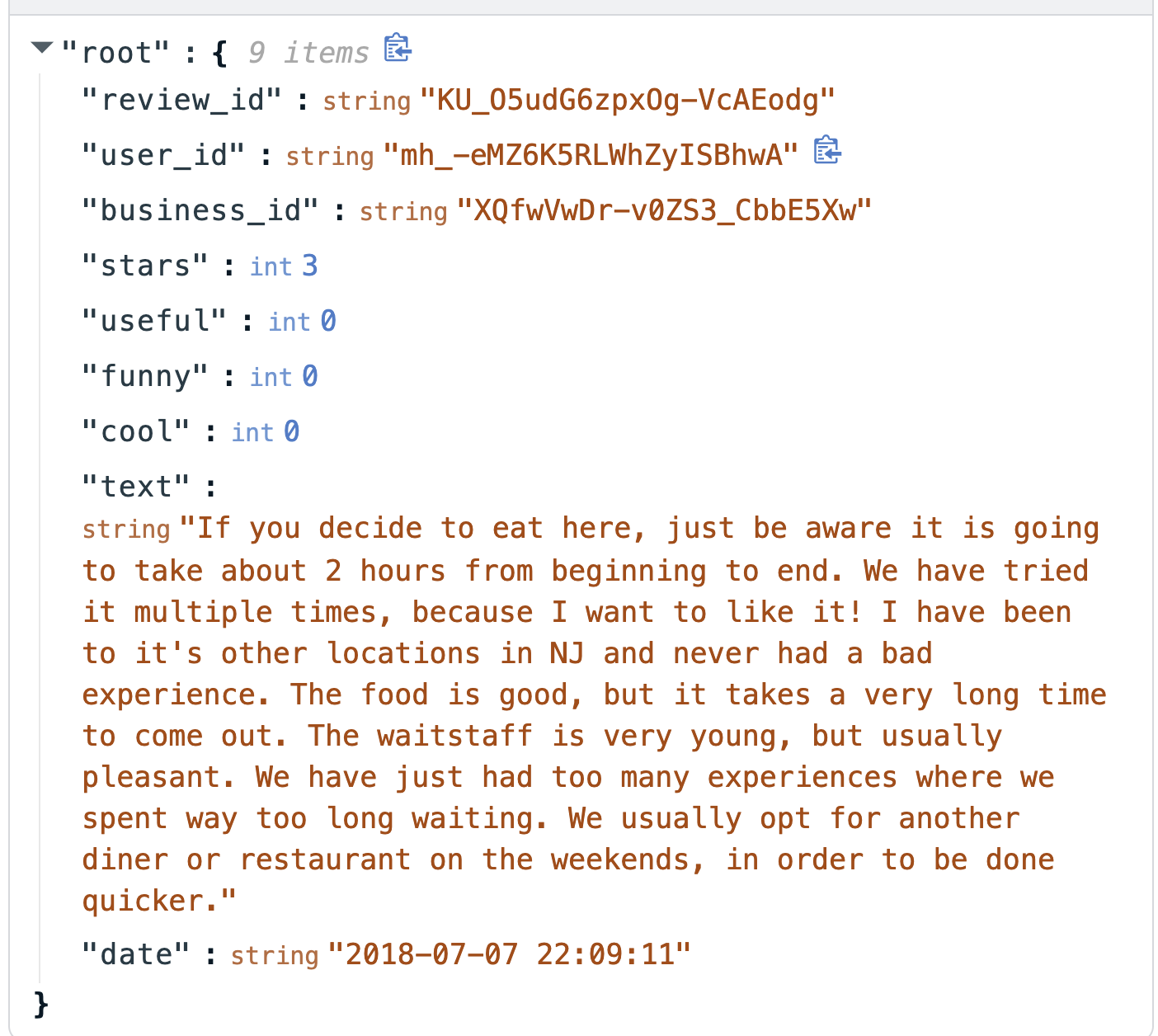
* **name:** 100,000 entries (object type) - Contains the name of the business.
* **address:** 99,180 entries (object type) - Provides the street address of the business. Missing values indicate data unavailability.
* **city, state, postal\_code:** 100,000 entries each (object type) - These columns list the city, state, and postal code where each business is located, respectively.
* l**atitude, longitude:** 100,000 entries each (float64 type) - Geographical coordinates of the business, useful for mapping and location-based analysis.
* **stars\_x (business rating):** 100,000 entries (float64 type) - The user rating for the business on a 1 to 5 scale.
* **review\_count:** 100,000 entries (int64 type) - The total number of reviews that the business has received.
* **is\_open:** 100,000 entries (int64 type) - A binary indicator (1 for open, 0 for closed) showing whether the business was open or closed at the time of data collection.
* **attributes:** 98,710 entries (object type) - Describes various attributes of the business, such as 'has WiFi', 'parking available', etc. Missing entries suggest incomplete data.
* **categories:** 100,000 entries (object type) - Lists categories or types of services the business offers, such as 'Restaurant', 'Beauty & Spas', etc.
* **hours:** 94,500 entries (object type) - Operating hours of the business on different days of the week. Missing values indicate that operating hours are not provided.
* **review\_id, user\_id:** 100,000 entries each (object type) - Unique identifiers for each review and user respectively.
* **stars\_y (review rating):** 100,000 entries (int64 type) - Ratings given by users in their reviews on a 1 to 5 scale.
* **useful, funny, cool (review votes)**: 100,000 entries each (int64 type) - Counts of votes for each review indicating how many other users found the review useful, funny, or cool.
* **text (review content):** 100,000 entries (object type) - The text content of each user review.
* **date:** 100,000 entries (float64 type) - The date when each review was posted, formatted as a floating-point number representing the time.

**Datasource:**

<https://www.kaggle.com/datasets/omkarsabnis/yelp-reviews-dataset>

**References:**

* <https://www.analyticsvidhya.com/blog/2022/10/using-text-mining-on-reviews-data-to-generate-business-insights/>
* <https://medium.com/ml2vec/training-and-tuning-a-lstm-to-classify-yelp-reviews-4d37b8aa2e91>



**Our work breakdown structure**

1. Come up with topic and outline of our project
2. Convert our outline in google colab
3. Perform coding and implement our projects
4. Review our results together
5. Make a presentation slide
6. Rehearse our slides

**We decided to choose the Yelp dataset to work on this project. Below is our proposal, and we would like to ask for your advice before we get started with our project.**

1. **Data collection**
   * Convert JSON to CSV - we use this yelp dataset, and we collected 100,000 rows.
2. [**Exploratory Data Analysis**](https://towardsdatascience.com/text-mining-and-sentiment-analysis-for-yelp-reviews-of-a-burger-chain-6d3bcfcab17b) 
   * Create a dataset summary
   * **Bar chart for top three business categories**
     + **Bars, Breakfast & Brunch, Restaurants, Barbeque, Cajun/Creole, Cafes, Cocktail Bars, Nightlife:** This category has the highest number of reviews, totaling approximately 8.93 million. The category includes a variety of dining and nightlife options, suggesting a high level of customer engagement and interest in these services.
     + **Live/Raw Food, Restaurants, Seafood, Beer Bar, Beer, Wine & Spirits, Bars, Food, Nightlife:** The second category has around 8.43 million reviews, which also indicates a strong consumer presence in the food and beverage sector, with a particular emphasis on seafood and establishments serving alcohol.
     + **Venues & Event Spaces, Performing Arts, Arts & Entertainment, Hotels & Travel, Food, Convenience Stores, American (New), Beauty & Spas, Restaurants, Museums, Event Planning & Services, Hotels, Cinema, Resorts, Day Spas:** The third category, with approximately 4.21 million reviews, is the most diverse. It covers a wide range of services from event spaces and hotels to arts, entertainment, and various types of food services.
   * **Rating Distribution across Yelp Reviews**
     + The majority of Yelp reviews tend to be **positive, centering around 4 stars.**
     + The presence of outliers on the lower end in the box plot could indicate particularly bad experiences or potentially harsher critics, while outliers on the higher end could represent exceptionally good experiences or less critical reviewers.
   * **Sentiment and Star Rating Distribution**
     + The **high number of positive reviews** could suggest that customers who have had good experiences are more likely to leave a review, or that businesses generally meet or exceed customer expectations.
     + The **relatively low number of neutral reviews** may imply that customers are often inclined to have clear perceptions, either good or bad, about their experiences.
     + The **quantity of negative reviews,** while the smallest, is critical for businesses to analyze in-depth to identify recurring issues and areas for service improvement.
   * Line chart for Best Restaurants based on average star rating
   * **Line chart for Year trend of count of reviews**
     + The line graph shows **a clear upward trend** in the number of reviews **from the year 2006 to around 2014,** suggesting increasing usage of Yelp and greater customer engagement over these years.
     + **After 2014,** there appears to be a plateau and **slight fluctuations in the number of reviews,** which could indicate a stabilization in Yelp's user activity or possibly increased competition from other review platforms.
3. **Text Mining:** 
   * **Step 1: Preprocessing**
     + Removing special characters and numbers that do not contribute to textual meaning.
     + Converting all text to lowercase to maintain consistency.
     + Eliminating stop words (common words like 'and', 'the', etc., that do not contribute much meaning).
     + Stemming or lemmatization, which simplifies words to their base or root form.
     + **Visualization:** Use word clouds or frequency bar charts to visualize the most common words and identify patterns.
   * **Step 2: Most Common Words Analysis**
     + **Tokenization and Bag-of-word: click here to view** [**resource**](https://urytrayudu1.medium.com/sentiment-analysis-for-yelp-review-classification-54b65c09ff7b)
       - **Tokenization:** Split the text of reviews into individual terms or words.
       - **Bag-of-Words (BoW):** Create a BoW model that represents the text as a collection of individual words along with their frequency counts.
         1. **N-gram: 2 or 3 words**
         2. In order to better capture the context, we can consider pairs or triples of words that appear next to each other and they can also give us more useful information.
   * **Step 3: Deep Dive into Positive Sentiments - click here to view** [**resource**](https://www.analyticsvidhya.com/blog/2022/10/using-text-mining-on-reviews-data-to-generate-business-insights/)
     + **Why do people love to go to this place?** - try to understand why people love the services.
     + **Word Embeddings:** Apply Word2Vec to transform words into vectors so that semantically similar words are closer in the vector space.
     + **Similarity Analysis:** Identify clusters or groups of words that are often mentioned together in positive reviews.
     + **Contextual Relationships:** Analyze the context around frequent positive words to understand the reasons behind positive sentiments.
   * **Topic Modeling - LDA:** figure out what topics are being discussed in the reviews and how the words relate to those topics. - **click here to view** [**resource**](https://www.analyticsvidhya.com/blog/2022/10/using-text-mining-on-reviews-data-to-generate-business-insights/)
     + **Choosing the Number of Topics:**
       - Use coherence score evaluation to determine the most suitable number of topics.
       - Start with a range (e.g., 5-15 topics) and select the number with the highest coherence score.
     + **Running LDA:**
       - Apply LDA to the preprocessed text data.
       - Train the model and examine the top words in each topic to understand their themes.

**Research questions and Select models**

* + **How have customer sentiments about a dining experience evolved over time? What specific aspects of the service drive these sentiments? Does it change based on reviews and restaurants?**
    - Sentiment Analysis
      * **Model?** the process of understanding the opinions of people about a subject.
        1. **Bag of word and supervised learning**
        2. **Supervised Learning Classifiers** - use historical data with known sentiment to predict the sentiment of a new piece of text (and we will build and compare models using metrics to see what models are working well.

**Logistic regression**

**Random forest**

**KNN**

**Naive**

**Use classification metric: F1-score to compare the model’s results.**

* + - * **Recurrent neural networks**
      * **Transfer learning (Roberta model)**

**====**