Azure Lab 4: Text Feature Engineering on Azure

Student Name: Shaima Mohammed
Student ID: 60104699
Course: DSAI 3202 Parallel & Distributed Comp
Instructor: Dr. Oussama Djedidi
Institution: University of Doha for Science & Technology
Date: 14 November 2025

🔍 Objective

The goal of this lab was to perform **text feature engineering** on the Goodreads reviews dataset using Azure Databricks.
We aimed to extract and combine various text-based features such as **sentiment scores**, **review length**, and **TF-IDF**, and then merge them into a single Gold-layer dataset ready for modeling.

⚙️ Steps Performed

Step 1 – Data Preparation

- Loaded the curated Goodreads reviews (in Delta format) from the **Gold layer**.
- The dataset included `review_id`, `book_id`, `user_id`, `title`, and `review_text`, which served as the base for feature extraction.

Step 6.5 – Sentiment & Length Features

- Used **NLTK VADER Sentiment Analyzer** to calculate emotional tone for each review.
- Created a PySpark UDF to generate:
  - `sentiment_pos` – positive score
  - `sentiment_neg` – negative score
  - `sentiment_neu` – neutral score
  - `sentiment_compound` – overall polarity (-1 to +1)
- Added text-length metrics:

    - `review_length_chars` – character count
    - `review_length_words` – word count
- Verified output and saved it to:
- `/gold/sentiment_length_features/`

Step 7 – Combine All Features

- Loaded **TF-IDF Gold dataset** from `/gold/text_features_tfidf/`.
- Joined it with the curated dataset (including sentiment & length features) using `review_id`.
- Renamed TF-IDF columns for clarity: `tfidf_score`, `idf_weight`, `term_frequency`, `document_frequency`.

Bonus Feature Engineering

Added three extra features to enrich the dataset:

1. **`sentiment_label`** – categorizes each review as Positive, Neutral, or Negative based on the compound score.
2. **`review_density`** – ratio of characters to words (measures text compactness).
3. **`exclamation_count`** – counts how many "!" are in the review (used as an emotion indicator).

Step 8 – Final Save and Validation

- Printed and checked the final schema.
- Saved the merged dataset to:
- `/gold/features_v2/`
- Displayed a preview showing successful integration of sentiment, TF-IDF, and bonus features.

🧠 Results & Observations

- The final Gold dataset combines **semantic**, **structural**, and **statistical** text information.
- It is now ready for machine-learning tasks such as review classification or rating prediction.
- The bonus features improve both interpretability and potential model accuracy.

✅ Conclusion

This lab demonstrated how to use Azure Databricks for end-to-end text feature engineering.
We successfully created sentiment, length, and TF-IDF features, merged them into a final Gold dataset, and added three custom bonus features.
All steps were executed without errors, and the output was validated and saved for future machine-learning model development.