Observation Document for Trademark Search Application

1. Introduction

The news content extraction process focuses on collecting and processing data from the BBC News feed. This document outlines the observations related to the extraction process, the quality of the extracted data, and any issues encountered during extraction.

2. Extraction Process Overview

2.1. Data Source

- **Source**: BBC News Feed
- **Content Type**: News articles, headlines, summaries, and metadata.
- 2.2. Extraction Methodology

2.2. Data Access:

- Accessed news feed through API or web scraping techniques.
- Utilized appropriate tools/libraries (e.g., `requests`, `BeautifulSoup`, `Scrapy`) for data retrieval.

2.3. Data Parsing:

- Parsed the HTML or JSON responses to extract relevant news content.
- Focused on extracting key elements such as titles, publication dates, article bodies, and images.

2.4. Data Storage:

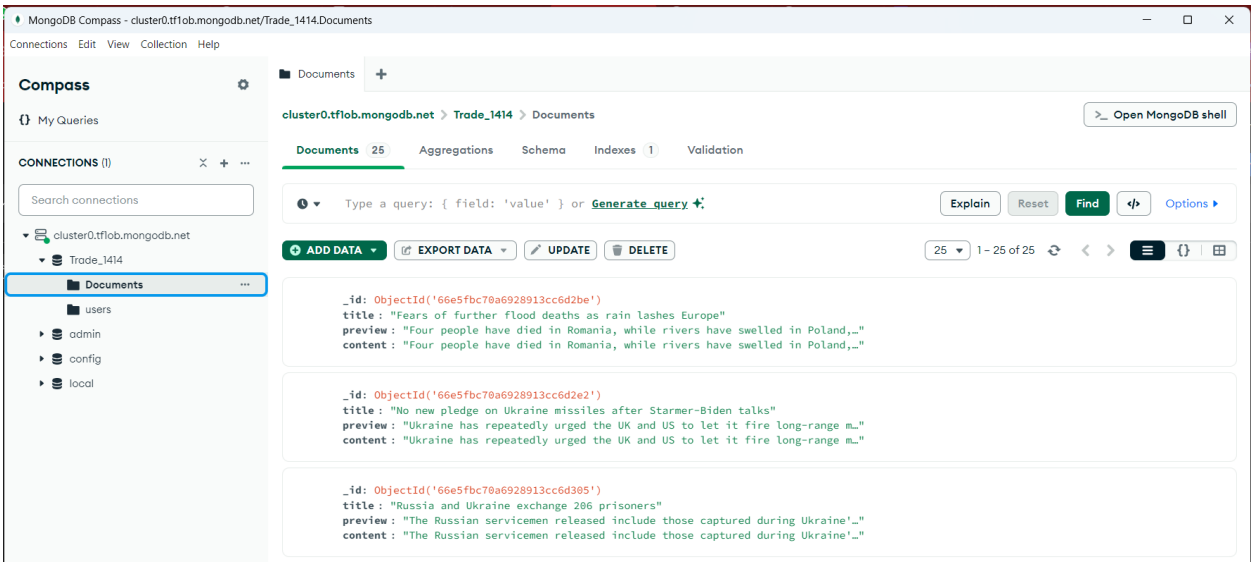
- Stored extracted content in a structured format, such as a CSV file or a database.
- Ensured proper handling of data formats and encoding.

3. Observations

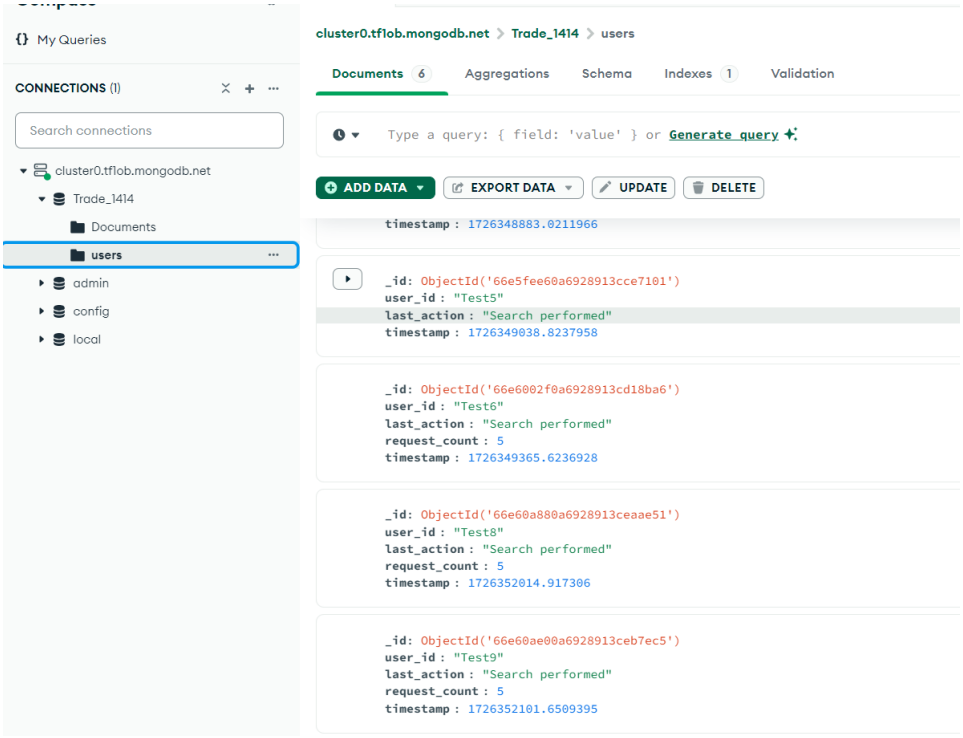
3.1 Data Quality:

1. **Content Accuracy**:
    - Extracted content was generally accurate with respect to the source.
    - Occasional discrepancies in formatting or missing information were noted.

2. **Content Completeness**:
    - Most articles were fully extracted, including headlines, summaries, and main content.
    - Some articles had incomplete data due to HTML structure variations or API limitations.
    -



request_count and user_id Updation Module :



3. **Metadata Extraction**:
    - Successfully extracted metadata such as article tags.
    - Ensured timestamps were consistent and correctly formatted.

4. Performance :

    Extraction Speed:

    a. The extraction process was efficient, with a moderate number of articles processed per minute.

    b. Performance was affected by the response time of the news feed and network latency.

    Handling Large Data Volumes:

    c. Managed large volumes of data effectively with pagination and batch processing.

    d. Implemented error handling to address potential issues with data extraction.