# SEBI Data Pipeline & MongoDB Access

## 1. Data Extraction Summary

The SEBI data pipeline follows a multi-step process to extract and structure enforcement action data:  
  
1. Extracted the base API endpoint listing enforcement actions.  
2. From that API, parsed paginated links to get full coverage.  
3. Extracted individual PDF links from each page.  
4. Downloaded all PDFs using requests and stored PDF IDs.  
5. Used SpaCy NER and Flair NER to extract named entities like organization names, dates, and legal terminology.  
6. Applied sentiment analysis using both a heuristic-based scoring approach and SpaCy sentiment models.  
7. Cleaned and stored the final structured data as a CSV.

## 2. MongoDB Atlas Access

The cleaned SEBI data is hosted on MongoDB Atlas and can be accessed as follows:

📌 Connection URI:

mongodb+srv://guest:guest\_pass@intellewings.enj2zjz.mongodb.net/?retryWrites=true&w=majority&appName=Intellewings

📂 Access Code:

from pymongo import MongoClient  
  
# Connect to MongoDB Atlas (guest access)  
client = MongoClient("mongodb+srv://guest:guest\_pass@intellewings.enj2zjz.mongodb.net/?retryWrites=true&w=majority&appName=Intellewings")  
  
# Select the database and collection  
db = client["sebi\_data"]  
collection = db["data"]  
  
# Fetch and print all documents  
docs = collection.find()  
for doc in docs:  
 print(doc)