INTISARI

ANALISIS KOMPARATIF PENGAMBILAN DATA *BATCH* VS. *STREAM* UNTUK *TOPIC MODELING* MENGGUNAKAN DATA DINAMIS REDDIT

Oleh

Veronika Regina Shanty
20/454543/PA/19574

Dengan pertumbuhan media sosial, analisis data real-time menjadi semakin penting. Penelitian ini membandingkan metode pemrosesan batch dan input data aliran untuk pemodelan topik menggunakan data yang bersumber dari subreddit 'r/beauty' Reddit. Studi ini mengkaji Latent Semantic Analysis (LSA) dan Latent Dirichlet Allocation (LDA) sebagai teknik pemodelan topik, dengan kedua proses tersebut dikemas melalui Docker untuk diterapkan dalam waktu 24 jam.

Evaluasi kedua pendekatan penyerapan ini berfokus pada indikator kinerja seperti latensi, akurasi model, dan pemanfaatan sumber daya. Hasilnya menekankan perbedaan antara strategi batch dan real-time. Penyerapan aliran menawarkan pemrosesan data yang lebih cepat dan wawasan langsung, sehingga cocok untuk lingkungan yang dinamis. Secara khusus, latensi rata-rata untuk LDA adalah 0,66 detik dan 0,25 detik untuk LSA dalam penyerapan aliran, dibandingkan dengan 1,40 detik untuk LDA dan 1,02 detik untuk LSA dalam penyerapan batch. Namun, penyerapan streaming juga menghabiskan bandwidth jaringan yang lebih tinggi, dengan LDA memerlukan 532,27 MB dan LSA memerlukan 277,79 MB dalam mode real-time dibandingkan 241,3 MB untuk LDA dan 60,4 MB untuk LSA dalam mode batch.

Di sisi lain, penyerapan batch memberikan stabilitas dan efisiensi dalam penggunaan jaringan. Hasilnya menunjukkan bahwa pengambilan data real-time menguntungkan situasi yang memerlukan pemrosesan data cepat. Sebaliknya, pengambilan data batch lebih cocok untuk situasi yang memprioritaskan kualitas model dan penggunaan jaringan yang lebih rendah.

Kata kunci: analisis social media, *natural language processing*, *topic modeling*, *big data*, *reddit*.

ABSTRACT

COMPARATIVE ANALYSIS OF BATCH VS. STREAM DATA INGESTION FOR TOPIC MODELING USING REDDIT DYNAMIC DATA

by

Veronika Regina Shanty
20/454543/PA/19574

Due to the growth of social media, analyzing real-time data is becoming increasingly crucial. This research contrasts batch processing and stream data input methods for topic modeling using data sourced from Reddit's 'r/beauty' subreddit. The study examines Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) as topic modeling techniques, with both processes containerized through Docker for deployment within 24 hours.

The evaluation of these two ingestion approaches focuses on performance indicators such as latency, model accuracy, and resource utilization. The results emphasize the differences between batch and real-time strategies. Stream ingestion offers faster data processing and immediate insights, making it well-suited for dynamic environments. Specifically, the average latency for LDA was 0.66 seconds and 0.25 seconds for LSA in stream ingestion, compared to 1.40 seconds for LDA and 1.02 seconds for LSA in batch ingestion. However, stream ingestion also consumes higher network bandwidth, with LDA requiring 532.27 MB and LSA needing 307.8 MB in real-time mode versus 241.3 MB for LDA and 60.4 MB for LSA in batch mode.

On the other hand, batch ingestion delivers stability and efficiency in network usage. The results indicate that real-time data intake benefits situations requiring rapid data processing. In contrast, batch data intake is better suited for situations prioritizing model quality and lower network usage.

Keywords: *social media analytics, natural language processing, topic modeling, data ingestion, reddit*