News Parsing System Optimization

Executive Summary

This document outlines critical improvements to our existing news parsing backend to optimize LLM-based address extraction. The focus areas include parallelizing the parsing infrastructure across multiple machines, implementing a multilabel classification system for accident categorization, and expanding our regional news sources to enhance coverage and accuracy.

1. Parallelization of News Parsing Infrastructure

Current Limitations

- Single-machine processing creates bottlenecks when handling high news volume
- Sequential parsing leads to delays in address extraction during peak news cycles
- Limited throughput affects our ability to process breaking news in real-time

Proposed Improvements

- **Distributed Processing Architecture**
  - Convert parsing pipeline to a distributed system using Apache Spark or Dask
  - Implement sharding strategy to distribute news articles across worker nodes
  - Design load balancing to ensure even utilization of processing resources
- **Optimized Workflow**
  - Segment news processing into discrete pipeline stages:
    - Ingestion → Preprocessing → Entity Recognition → Address Extraction → Validation
  - Allow parallel execution of each stage across different worker pools
  - Implement priority queues to expedite processing of high-value news sources
- **Performance Enhancements**
  - Add batch processing capability for archival content
  - Implement streaming processing for real-time news feeds
  - Create checkpointing system to recover from node failures

2. Multilabel Classification System for Accidents

Current Limitations

- Limited filtering capabilities for downstream analysis applications

Proposed Accident Classification System

- **Accident Type Labels**
    - Traffic (vehicle collisions, pedestrian incidents)
    - Workplace (industrial, construction, office)
    - Natural Disaster (floods, fires, earthquakes)
    - Public Space (mall, park, entertainment venue)
    - Residential (home accidents, building incidents)
    - Transportation (railway, aviation, maritime)
- **Severity Classification**
    - Fatal
    - Major Injury
    - Minor Injury
    - Property Damage Only
- **Scale Tags**
    - Individual
    - Small Group (2-10 people)
    - Large Group (11-50 people)
    - Mass Casualty (>50 people)
- **Response Entity Involvement**
    - Police
    - Fire Department
    - Emergency Medical Services
    - Hazmat
    - Military/National Guard
    - Coast Guard/Maritime Response
- **Implementation Strategy**
    - Fine-tune LLM specifically for accident context classification
    - Develop confidence scoring for each label assignment
    - Create annotation interface for correcting misclassifications
    - Implement hierarchical classification for nested categories

3. Regional News Source Expansion

Regional News Sources to Integrate

1. We need to add new sources for news extraction

Integration Requirements

- Develop custom scrapers and API connectors for each regional source
- Implement standardization process to normalize address formats across sources

- Create geolocation mapping to associate sources with geographic coverage areas
- Design quality scoring system to weight address reliability by source
- Build region-specific extraction rules to handle local address formats

4. Implementation Considerations

Key Performance Indicators

- **Processing Speed**: 3x improvement in articles processed per hour
- **Address Extraction**: 20% increase in validated addresses discovered
- **Classification Accuracy**: >85% precision and recall for accident labels
- **Source Coverage**: 60% increase in geographic diversity of regional news sources
- **System Reliability**: 99.9% uptime for the parsing infrastructure

Resource Requirements

- Additional compute nodes for distributed processing
- Storage expansion for caching and archives
- Regional-specific language models for improved extraction accuracy
- Training data for accident classification model development