Social Media Analysis Dashboard - System Design Explanation

# High-Level Goals

* 1. Ingest social media data (Reddit, Twitter, etc.)
* 2. Perform advanced text analysis (sentiment, topics, entities)
* 3. Visualize insights interactively
* 4. Be usable by non-technical users through a web UI (Streamlit)

# Architecture Overview

* Layered architecture from data ingestion to visualization and AI insights.
* Key components: Data loading, Preprocessing, Analysis Engines (Sentiment, NER, Topic Modeling), Visualization, AI Integration.

# Code Structure Breakdown

* main.py is the central controller.
* - setup\_nltk\_resources(): Downloads required NLTK data.
* - load\_spacy\_model\_from\_disk(): Downloads and loads spaCy model dynamically.
* - load\_data(): Loads and cleans the .jsonl data.
* - preprocess\_text(): Cleans and tokenizes the text data.

# Visual Tabs

* Overview: Summary metrics on posts, authors, subreddits
* Activity Trends: Time-series chart using Plotly
* Sentiment Analysis: VADER-based sentiment classification
* Top Entities: Frequent authors, subreddits, and words
* Word Cloud: Frequent terms visualized
* Author Network: Graph of authors and subreddits using NetworkX
* Topic Modeling: TF-IDF + NMF with spaCy-based topic naming
* AI Insights: GPT-generated qualitative summaries (optional)

# Topic Modeling Logic

* TF-IDF + NMF for topic extraction.
* spaCy NER used to name topics based on entities.
* Fallback to TF-IDF top keywords when no entities found.

# Network Graph

* Constructed using NetworkX.
* Visualized with Plotly (scatter and line for edges).
* Nodes: Authors and subreddits, edges: interactions.

# AI Insights (Optional)

* generate\_enhanced\_insights(): Statistical summaries.
* generate\_mock\_insights(): Placeholder for LLM-based insights.
* Can use OpenAI key via Streamlit secrets.

# Design Principles

* Modular functions for easy testing and debugging.
* Robust error handling with fallbacks.
* Caching with @st.cache for performance.
* User customization: date range, keyword filtering, topic renaming.

# Why This Design Works

* Scalable and performant with large datasets.
* User-friendly with Streamlit UI.
* Explainable and customizable analytics.
* Ready for future enhancements (e.g., more data sources or LLMs).