Issues Faced & Solutions – NewsScope AI (Global News Tracker)

Problem 1:

HTML Content Containing Extra Tags

- **Issue:**
  RSS feeds often contain , ,
, or embedded HTML that pollutes the article text. Without cleaning, duplicate detection and summarization fail.

- **Solution:**
  Implemented clean_text() using **BeautifulSoup + regex** to strip HTML tags and normalize whitespace before fingerprinting and storing in DB.

Problem 2:

Duplicate Articles in Database

- **Issue:**
  Same news article fetched multiple times (different RSS feeds, slight formatting changes).

- **Solution:**

    - Added **SHA-256 fingerprinting** of cleaned text (fingerprint_text()).

    - Dedupe logic (dedupe_new_articles()) removes duplicates based on fingerprint before committing to DB.

Problem 3:

SQLite Database Lock Errors

Issue:
When background scheduler (apscheduler) and Streamlit UI both accessed SQLite simultaneously, errors like:

sqlite3.OperationalError: database is locked

- **Solution:**
    - Enabled check_same_thread=False in SQLAlchemy engine for SQLite.
    - Used **session per request** pattern (SessionLocal()) and ensured sessions are closed after use (db.close()).

Problem 4:

Scheduler Not Running with Streamlit

- **Issue:**
  APScheduler jobs didn't run when Streamlit was launched with streamlit run.
- **Solution:**
    - Started scheduler inside a separate thread/process.
    - Used start_scheduler() at app startup to ensure periodic RSS fetches run in background.

Problem 5:

RSS Feed Rate Limits / Errors

- **Issue:**
  Fetching too frequently caused HTTP 429 (Too Many Requests) or feed parsing errors.
- **Solution:**
    - Limited jobs to every **15 minutes** (sched.add_job(..., 'interval', minutes=15)).
    - Added error handling + retry logic inside fetch_google_news_rss().

Problem 6:

Missing or Corrupt Articles

- **Issue:**
  Some feeds lacked content, title, or published_at, breaking ingestion.

- **Solution:**
    - Added field validation + default fallbacks (datetime.utcnow() for missing timestamps, "Unknown Source" for missing publisher).
    - Skipped incomplete articles gracefully.

Problem 7:

Performance – Slow Queries in UI

- **Issue:**
  Loading 100s of articles with .all() made Streamlit slow.
- **Solution:**
    - Limited results (limit(50) for recent fetch).
    - Added DB indexes (fingerprint, published_at) for faster lookups.

Problem 8:

Incorrect Module Imports

Issue:
Relative imports like from .db import init_db caused errors when running Streamlit:

ImportError: attempted relative import with no known parent package

- **Solution:**
    - Restructured project as a package (app/ folder with __init__.py).
    - Used absolute imports (from app.db import init_db).