Modeling Pipeline Overview - JSON to Transformer Prep

1. JSON Flattening

Raw race data in JSON format is loaded and flattened into tabular structure.

- *flatten_day.py logic for flattening a single JSON*

- *flatten_day_batch.py batch processor for all .json files in raw/ directory*

2. Initial DataFrame Cleaning

Cleans data types, converts strings to categories, parses dates.

- *clean_flattened_df.py applies dtype optimization and cleanup*

3. Embedding Index Creation

Adds categorical embedding indices for transformer ingestion.

- *add_embedding_indices.py label-encodes fields like 'country', 'going', 'sex'*

4. NLP & Regex Phrase Feature Extraction

Embeds text from 'comment' and 'spotlight'; extracts domain-specific binary NLP features.

- *process_text_fields.py merges text embedding + phrase detection*

- *extract_race_phrases.py regex-based binary flags (also internal to process_text_fields)*

5. Embedding Save & Load

Exports embedded vectors to .npz for efficient model input; supports validation on load.

- *embedding_io.py save/load .npz with schema validation*

6. Final Assembly for Modeling

Combines all processed features into a DataFrame for model ingestion.

- *save_model_ready_df.py entry point to clean, embed, and save final model input DataFrame*