Solution Presentation - Company Classifier

Generated: 2025-10-19 17:11

Solution overview

This solution assigns one or more insurance taxonomy labels to each company in the input list. It combines two matching signals so labels are both meaningful and precise.

What I implemented (choices made)

1) Data cleaning and unification of tags and text fields.
2) Field consolidation so each decision uses full company context.
3) Semantic matching using sentence embeddings to capture meaning.
4) Lexical matching using TF-IDF to capture exact word matches.
5) A balanced combination of semantic and lexical scores for final ranking.
6) Returning the top 3 labels per company, with simple safeguards to avoid poor matches.
7) Per-label diagnostics (counts and average scores) to assess label quality.

Results from the latest run

Rows: 9492
Coverage: 100.00%
Avg labels/company: 3.00
Scores — mean: 0.79, median: 0.78, std: 0.11

Top labels (by frequency)

Label	Count	Avg score
Agricultural Equipment Services	726	0.79
Food Processing Services	636	0.80
Accessory Manufacturing	485	0.76
Sheet Metal Services	420	0.82
Travel Services	396	0.83
Plastic Manufacturing	395	0.80