The Healthy Approach - Organic Data Enrichment Through Entity Extraction

3/31/21 12:28 PM

Speaker: Julia Naegu, Ian Bakst (Tamr)
Date: 03/31/2021

Unstructured data hiding in the wild

- Structured datasets often also contain unstructured attributes
- Example of Amazon catalog data containing package size information in the Product Description attribute

Product ID	Product Description	Price
B002BZOFKI	Mother Of The Bride Satin Button Party Accessory (1 count) (1/Pkg)	$5.4
B000MSFOJM	Instant Arches Women's 200 3 Pack Foam Arch Insoles	$26.95
B00A2AAUJE	Baseball Folded Thank You Notes (8 Pack)	$10.32
B000RNDA5C	Kiwi Liquid Instant Wax, 2.5 fl oz, Black	$6.53

- Example: Price standardization using size & quantity data

What is data enrichment?

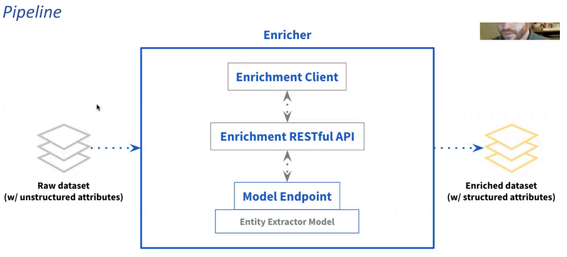
- Data enrichers usually apply external data to provide additional information to an existing dataset
- Example:
  - Enriching address datasets with geolocation data using Google Maps Geocoding API
  - 10550 S Sam Houston Pkwy W Houston TX 77071-3141 → **(29.64058 Lat , -95.53326 Long)**
  - 809 W Pasadena Fwy Pasadena TX 77506-1229 → **(29.70926 Lat, -95.21783 Long)**

From regular enrichment to organic enrichment

- Instead of relying on third-party enrichment sources, datasets can be enriched with attributes extracted from existing attributes
- Organic data enrichment increases the accessible information in a dataset by extracting structured attributes from unstructured ones

Product ID	Product Description	Price	Quantity	Unit of Measure
B002BZOFKI	Mother Of The Bride Satin Button Party Accessory (1 count) (1/Pkg)	$5.4	1	unit
B000MSFOJM	Instant Arches Women's 200 3 Pack Foam Arch Insoles	$26.95	3	unit
B00A2AAUJE	Baseball Folded Thank You Notes (8 Pack)	$10.32	8	unit
B000RNDA5C	Kiwi Liquid Instant Wax, 2.5 fl oz, Black	$6.53	2.5	fl z

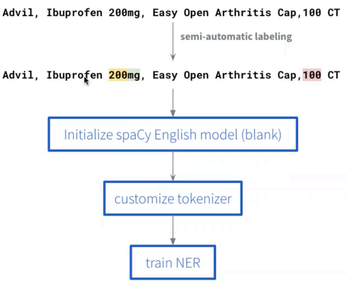
Pipeline



Model

Leveraging spaCy

- Provides pre-trained, language-specific NLP pipelines, for tokenizing, PoS tagging , NER, etc.



The Enricher

- Start with NLP engine (spaCy)
- Build functions/modules around the NLP to incorporate into workflows
- Build a **Flask App** which keeps track of metadata regarding the data and models
  - Flask is a lightweight WSGI web application framework
- Build **RESTful API** endpoints to run workflows with data/models on disk
  - A RESTful API is an architectural style for an application program interface (API) that uses HTTP requests to access and use data
- Deploy app in location where users can easily access

Architecture

- The enrichment client enriches data by hitting the APIs of the Enricher App
- The Enricher App controls the NLP models and their use and manipulation
- The models are trained on data and set up to enrich data with single API call

The Entity Extractor App

- Load labeled datasets
- Build models
- Tweak parameters
- Train models
- Generate results/analytics
- Enrich data