SIGMOD Programming Contest 2020

Sainyam Galhotra
Advised by: Barna Saha
University of Massachusetts Amherst
sainyam@cs.umass.edu

UMassAmherst
College of Information & Computer Sciences

Problem Statement:

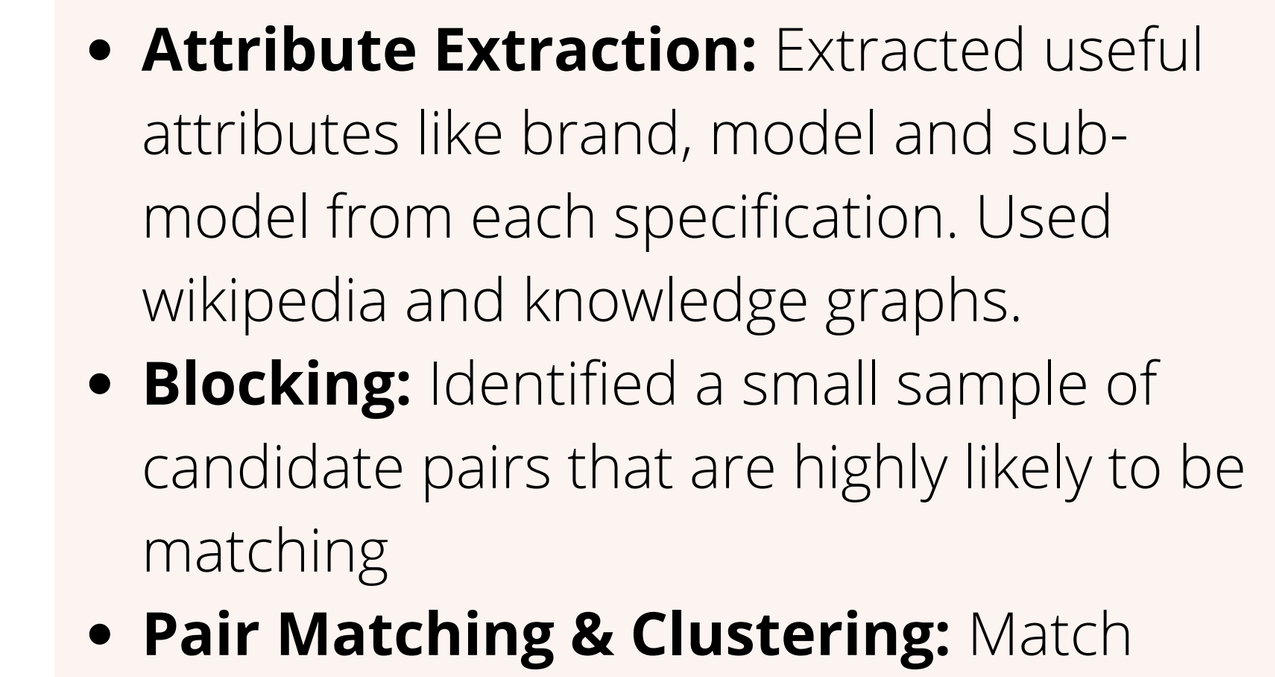
Given a collection of camera specifications from various e-commerce websites, identify all pairs of specs that refer to same entity

Number of products: Around 30K

System Architecture

Four components

- **Preprocessing and schema mapping:** Removed unwanted characters, identified useful attributes and mapped schema across different sources.
- **Attribute Extraction:** Extracted useful attributes like brand, model and sub-model from each specification. Used wikipedia and knowledge graphs.
- **Blocking:** Identified a small sample of candidate pairs that are highly likely to be matching
- **Pair Matching & Clustering:** Match candidate pairs and construct cluster of records that refer to same entity.



Blocking:

- Token based blocking
  - Considered unigram tokens along with bi-grams and tri-grams
- Block Weight estimation:
  - Calculated block weights using the validation dataset. These weights helped us shortlist high quality blocks

- **Observation**: Most of the high weight bi-gram and tri-gram block tokens were brand, model, sub-model of the considered products
- This was especially helpful to identify that canon eos 5d mark ii is different from canon eos 5d mark ii

Pair Matching & Clustering

- Rule based matching:
  - Products with same brand and model information are labelled as matches
- Performed transitive closure to handle inconsistencies

Results

- Precision: **0.99**
- Recall: **0.99**
- Running time: less than 25 seconds*

*Implementation in Python, tested on a laptop with 16GB RAM.