**Overview**

The objective is to match entries in two datasets - one containing a set of products and another containing a set of listings from a 3rd party retailer. These two datasets are contained in files named Products.txt and Listings.txt. Each entry is stored on its own line as a JSON element. Below are examples of two matching entries from each dataset.

From Products.txt:

{"product\_name":"Casio-EX-H30","manufacturer":"Casio","model":"EX-H30","family":"Exilim","announced-date":"2011-01-04T19:00:00.000-05:00"}

From Listings.txt:

{"title":"Casio Exilim EX-H30 Digitalkamera (16 Megapixel, 12,5-fach opt. Zoom, 7,6 cm (3 Zoll) Display, Akku für bis zu 1.000 Fotos, bildstabilisiert) schwarz","manufacturer":"CASIO","currency":"EUR","price":"193.50"}

Each entry in the products dataset can match to multiple entries in the listings set. Additionally, each listing can only match a single product. The desired output is each item in the products dataset paired with all matching items in the listings dataset, as shown below.

{  
"product\_name": String  
"listings": Array[Listing]  
}

**Hypothesis**

As both of the datasets contain a "manufacturer" parameter, the primary focus will be on grouping both of the datasets based on this common parameter. Both datasets will be segregated into matching groups based on their manufacturer value. This will serve both to filter out nonmatches as well as prevent wasted comparisons testing products against listings from other manufacturers.

Once the datasets have been grouped based on their manufacturer, the listing will be matched to products based on the product model and family using a deterministic methodology. Deterministic matching is predicted to be the ideal method due to the low number of comparable parameters and the high variance in listing formatting. The criteria for matching and testing will be further detailed in the Methodology section.

**Methedology**

The manufacturer values of a listing and product are considered to match if, ignoring capitalization, the entireity of the product manufacturer string can be found in the listing manufacturer string.

The method for matching model and family values is similar, however the "model" and "family" parameters of each product are compared to the "title" parameter of each listing. A listing is deemed to match a product if the manufacturer, model, and family match. The only exception is for products that only contain a manufacturer and model. In these cases, a match will be judged based on manufacturer and model alone.

Due to the high number of products and listings, manually verifying the matches is not feasible. While a considerable number can be verified in this way, other methods are required to provide more concrete, objective results. As each listing is guaranteed to match one product at the most, matching errors can be detected by examining the results for listings that match more than one product. By combining these two methods, the accuracy of the results can be approximated.