Comp 762—Assignment 2—Data integrity and consistency

Edward Newell

January 17, 2014

1 Overview

I performed some integrity checks on the categories and products, (working from those provided by Prof. Robillard)

I found that there are discrepencies between what is reported in the categories (`productsCount`, `testedProductsCount`, `ratedProductsCount`) and the actual counts in the product listings. While performing the checks, I found it more convenient to transform the categories listing into a flat format that provides easy lookups and traversal in both directions. I annotated this with the counts that I found, and will make the file available if others would find it useful.

With regard to the products, I tested the penetration of fields (e.g. `name`, `imageLarge`, `summary`). By penetration, I mean the percentage of products that contain the field. I also report various other stats by field, such as average length (for `lists`, `dicts`, or `strings`), average value (for numeric types) as well as the standard deviation of these. Types lists the absolute number of times the field was seen, broken down by what variable type it contained. There are 64 fields, and there are definitely some "gotcha's", where a field is just shy of 100% penetration. You can peruse the results here: `http://shpow.com/reports2html/?file=product_fields.json`. When the page loads, click "htmlify".

2 Details—Categories integrity check

I counted the number of products in the categories by iterating over all products, and keeping running totals associated to each category, based on the `category.id` field.

I similarly counted totals for rated products, by looking for the existence of the `ratings` field and checking that it contained at least one rating. For counting totals of tested products, I checked the `isTested` field.

The categories data provided from the API of course has its entries for `productsCount`, `ratedProductsCount`, `testedProductsCount`. There are many discrepencies, so I would like to submit that we use my annotated counts, respectiveld `count`, `count_rated`, `count_tested`, whenever counts are needed.

These counts are available in a flattenned version of the categories data, which I will make available. In this flattened verision, the field `downLevel` is eleminated, and the fields `parent` and `children` are added instead. These reference the parent and children by `id`s. The listing is structured as a dictionary, which all categories keyed by their `id` as a numeric string.

The categories Money, Food, and Baby & Kid are eliminated in this file along with their sub-categories recursively

3 Details—Products field penetration

By iterating over all products in the franchise-based listings provided by Prof. Robillard, I produced the union of all fields appearing in the products. Whenever a field was a `dict` type, I would recursively descend,

so that for example, the listing of fields contains `price`, `price.description`, `price.value`, and so on. For list types, no descent was made.

I have made available the full listing of stats for the 64 fields. in addition to penetration, I show counts for the variable type (e.g. in case fields are sometimes integer and sometimes a numeric string), lengths (average and standard deviation), and value (average and standard deviation). No lengths are reported for numeric types, and no value is reported for compound types (dicts and lists). For compound types, the length is the number of entries, and for strings it is the length of the string.

Counts do not include products from the eliminated categories. Averages and standard deviations are not affected by empty strings or missing fields, but are affected by empty compound types (lists or dicts).