Final Capstone Project Proposal

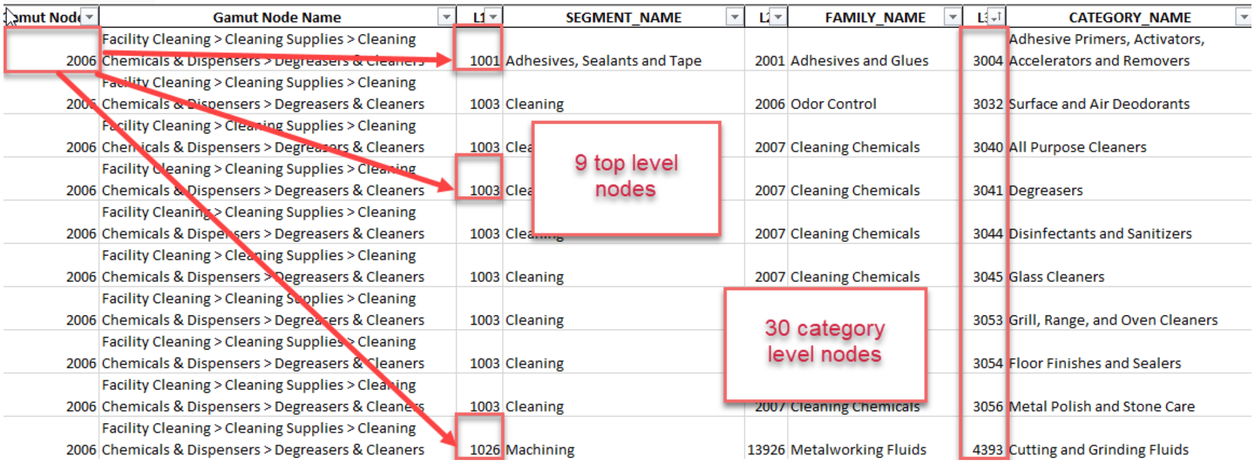
Project Title: Taxonomic Attribute classification using attribute values and definitions

Proposed By: Colette Gabriel

What is the problem you are attempting to solve?

The company I work for has two legacy software systems and is moving forward to one combined system that will encompass all of the products it sells. Legacy system one contains about 1.8 million products and legacy system two contains nearly half a million. Nearly all of the products in system two are also contained in system one, but the taxonomy and attribute values between the two systems do not match up.

As shown here, a single node (2006) in system two maps to 30 terminal nodes in system one. So the first problem to solve was one more of data engineering. I had to build a pipeline to extract data from both systems (system one in Teradata, system two using Postgres SQL) and interleave the data so that attributes and values can be compared.



Gamut Node	Gamut Node Name	L1	SEGMENT_NAME	L2	FAMILY_NAME	L3	CATEGORY_NAME
2006	Facility Cleaning > Cleaning Supplies > Cleaning Chemicals & Dispensers > Degreasers & Cleaners	1001	Adhesives, Sealants and Tape	2001	Adhesives and Glues	3004	Adhesive Primers, Activators, Accelerators and Removers
2006	Facility Cleaning > Cleaning Supplies > Cleaning Chemicals & Dispensers > Degreasers & Cleaners	1003	Cleaning	2006	Odor Control	3032	Surface and Air Deodorants
2006	Facility Cleaning > Cleaning Supplies > Cleaning Chemicals & Dispensers > Degreasers & Cleaners	1003	Clea	2007	Cleaning Chemicals	3040	All Purpose Cleaners
2006	Facility Cleaning > Cleaning Supplies > Cleaning Chemicals & Dispensers > Degreasers & Cleaners	1003	Clea	2007	Cleaning Chemicals	3041	Degreasers
2006	Facility Cleaning > Cleaning Supplies > Cleaning Chemicals & Dispensers > Degreasers & Cleaners	1003	Cleaning	2007	Cleaning Chemicals	3044	Disinfectants and Sanitizers
2006	Facility Cleaning > Cleaning Supplies > Cleaning Chemicals & Dispensers > Degreasers & Cleaners	1003	Cleaning	2007	Cleaning Chemicals	3045	Glass Cleaners
2006	Facility Cleaning > Cleaning Supplies > Cleaning Chemicals & Dispensers > Degreasers & Cleaners	1003	Cleaning			3053	Grill, Range, and Oven Cleaners
2006	Facility Cleaning > Cleaning Supplies > Cleaning Chemicals & Dispensers > Degreasers & Cleaners	1003	Cleaning			3054	Floor Finishes and Sealers
2006	Facility Cleaning > Cleaning Supplies > Cleaning Chemicals & Dispensers > Degreasers & Cleaners	1003	Cleaning	2007	Cleaning Chemicals	3056	Metal Polish and Stone Care
2006	Facility Cleaning > Cleaning Supplies > Cleaning Chemicals & Dispensers > Degreasers & Cleaners	1026	Machining	13926	Metalworking Fluids	4393	Cutting and Grinding Fluids

9 top level nodes

30 category level nodes

The goal of the project is to build a scaffold to map attributes from system one into attributes of system two where there is enough similarity between concepts in each system. The first pass was simple to match up attributes that are named the same (ex. color in system one = color in system two).

The more sophisticated method is to use the attribute values and definitions stored in the two systems to predict when attributes are similar enough for mapping. For example, sometimes “overall height” maps to “depth” depending on how values were captured. “Current” could map to “amps”, “fragrance” to “scent”, “commercial/residential” to “application”, etc.

By performing TF-IDF on three columns and then clustering the results, my goal is to provide guidance to predict when attributes in each system are close enough in concept to be mapped. This data will be verified by third party reviewers, but I can cut down their work time and hopefully boost accuracy by recommending potential mapping that might not be obvious at first glance.

How is your solution valuable?

The company goal is to merchandise 500,000 products in the new system by the end of year, so timing is very important for phase one and this work can help speed the third party review. For the second phase of merchandising the remaining 1.2 million SKUs, this work can help save money and time by making the third party review process more efficient.

I also hope to be able to suggest attribute mappings that might not be obvious through manual review but can be teased out through regression/clustering methods.

What is your data source and how will you access it?

The data source consists of the taxonomic structure of two different data systems. System one data is pulled from Teradata (aka Business Objects), a data access platform. System two data is pulled using a Postgres backend. Both are accessed using SQL queries integrated into Python code I’ve written.

What techniques from the course do you anticipate using?

I plan to use natural language processing, specifically using my own TF-IDF vectors or using Word-2-Vec or Gensim to build an attribute/value vocabulary, then applying logistic regression for supervised learning on about 3,000 mappings that have already been manually completed. Text classification clustering methods can help with unsupervised classification for the remaining data.

Based on the accuracy of the results, I plan to set a threshold level and include the recommended mapping as “Potential Match” in the column that is being reviewed by the third party. Currently the only values store there are “Grainger only” (system one), “Gamut only” (system two), and “Match”, where attribute names match up exactly. By making predictions and including high value potential matches, I can speed up the review process.

The final step will be to make this pipeline available to the taxonomy team, so that they don't have to come to me to create these attribute mappings.

What do you anticipate to be the biggest challenge you'll face?

The biggest challenge was building the pipeline and smoothly integrating the two data sources accurately. There are more than 7,800 nodes in system one and nearly 11,800 nodes in system two, and figuring out an efficient way to download and accurately integrate this data has taken more than a month of work.

Speeding the entire process up will be an ongoing challenge, but I was already able to improve speed by up to 40% by building up a dictionary of dataframes to prevent multiple calls to the same node.

The next challenge is to utilize the values and value frequency in a meaningful way to generate accurate predictions.