ITEMMASTER - NATURAL LANGUAGE PROCESSING FOR MATCHING CUSTOMER COMPANY NAMES

Manqing Sun and William L. Guzmán Daugherty

Supervisor: Nick Kadochnikov

THE UNIVERSITY OF CHICAGO
GRAHAM SCHOOL
CONTINUING LIBERAL AND PROFESSIONAL STUDIES

ITEMMASTER
every brand. everywhere

Agenda

- Executive Summary
- Company Background
- Problem Statement
- Research Purpose
- Data
- Data Assumption and Transformation
- Data Exploratory
- Methodology
- Pseudocode
- Results
- Impact to ItemMaster's Business
- Acknowledgement
- Q&A

Executive Summary

- The objective of this project is to develop a Natural Language Processing (NLP) algorithm that can assist in finding the matching account in ItemMaster's Salesforce database for every product in a retailer's assortment sheet

- The Jaccard similarity index is used for measuring the similarity of a retailer's product name with the ItemMaster's product account name
- Before calculating the Jaccard similarity, the product names are pre-processed by using the NLP Toolkit (NLTK), Pandas frame, NumPy array, autocorrect (for spelling correction), re (regular expression), and ngrams, etc.

Executive Summary (cont.)

- Our algorithm is not only able to automate the manual work, but also able to find additional accounts and unique items that manual work cannot. An assessment performed by ItemMaster concluded that, based on the four sets of data they provided for our development purposes, our algorithm can add more than $2,000,000 dollars to ItemMaster's sales pipeline.

Company Background

- ItemMaster® is a **spinoff** of **Peapod**.com established in 2009. It provides product content solutions for brands, retailers, and consumers alike.
- ItemMaster enables partners to:
  - plan and merchandise
  - market across channels
  - build online eCommerce and media experiences
  - provide product content creation, management, and distribution
- ItemMaster's flexible product content management system is available to Consumer-Packaged Goods (**CPG**) brands, CPG retailers, and ecosystem partners or mobile applications that can benefit from the 100,000+ manufacturer products active in the rapidly growing ItemMaster platform.
- ItemMaster structures data for today's consumer trends and tomorrow's technologies.

Problem Statement

ITEMMASTER®
Every brand. Everywhere.

Problem Statement (Cont.)

Problem

- Grocery retailers sell products from thousands of manufactures; however, their assortment sheets are rarely standardized.
- There are often variations in spelling, punctuation, and Unicode characters in manufacturer and brand names across retailers and even within one retailer's data file.
- Mapping a retailer's product to ItemMaster's Salesforce database has been done manually.

Research Purpose

- The purpose of this project is to develop a Natural Language Processing (NLP) algorithm that can map the unstandardized products from retailer's assortment sheets to ItemMaster's standardized Salesforce database account.
- The deliverable of this project is to predict the matching account name, account ID, and parent account ID and give a similarity index value for this prediction in every unique combination of manufacturer and brand names.

Data – ItemMaster's data

Every Brand Every Where

ItemMaster helps manufacturers structure, manage, enhance, and distribute product content for a data driven world.

The standardized Salesforce database Account Names file format:

Variable	Definition	Other attributes (Format)
SFAccountName	Item Salesforce Account Name	Unique Key
SFAccountID	Item Salesforce Account ID	
SFParentAccountID	Item Salesforce Parent Account ID	

Example records in Account Names file:

SFAccountName	SFAccountID	SFParentAccountID
Gildan	001G000001H841ZIAR	001G000001XciJiIAJ
Doskocil Manufacturing Company	001G000001H841aIAB	
…	…	…

Data – The Retailer's Assortment Sheet

The unstandardized Retailer's Assortment Sheet format:

Variable	Definition
Manufacturer	Manufacture's name
Brand	Brand
Category A	Category A that the item belongs to
Category B	Category B that the item belongs to
Category C	Category C that the item belongs to
ItemDescription	Description of the item
ItemUPC	Item Universal Product Code (UPC), or barcode.
ShipItemID	Item Shipment ID
VerificationStatus	Item Verification Status
PublishedStatus	Item Published Status

Example records of retailer's unique manufacturer and brand names file:

	Manufacturer	Brand
1	IMPORT - PURCELL INTERNATIONA	
2	BLAZERS SPECIALTY FOODS	BLAZERS
3	GULF PRIDE ENTP INC	GULF PRIDE SELECT
4	1HARBOR SEAFOOD,INC	
5	LOREAL USA	SOFT SHEEN CARSON LETS JAM
6	BON SECOUR FISHERIES INC	NELSONS
7	1NORTH COAST	
8	Fisk Industries	.
9	SEA DELIGHT LLC	SEA DELIGHT
10		APPLE

Data Assumption and Transformation

Data Assumption:

- ItemMaster's Data Analytics team informed us that the ItemMaster's Saleforce database Account Names are created based on:
  - Product's Manufacturer Name
  - Brand name

Data Transformation:

- Based on the data assumption, an additional file that only contains the unique combination of the Manufacturer name and Brand name is created.
- This file is used to predict the matching standardized ItemMaster's Salesforce Account Name, Account ID, and Parent Account ID.

Data Exploratory – Explore unstandardized retailer's data

Example records of retailer's unique manufacturer and brand names file:

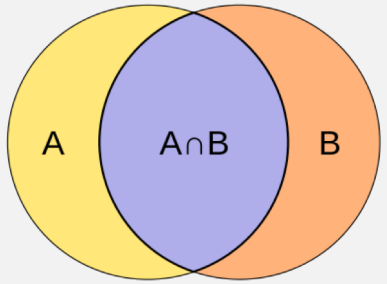
	Manufacturer	Brand
1	IMPORT - PURCELL INTERNATIONA	
2	BLAZERS SPECIALTY FOODS	BLAZERS
3	GULF PRIDE ENTP INC	GULF PRIDE SELECT
4	1HARBOR SEAFOOD,INC	
5	LOREAL USA	SOFT SHEEN CARSON LETS JAM
6	BON SECOUR FISHERIES INC	NELSONS
7	1NORTH COAST	
8	Fisk Industries	.
9	SEA DELIGHT LLC	SEA DELIGHT
10		APPLE

1. Misspelling of Manufacturer name (i.e. INTERNATIONA, LOREAL).
2. Name combines number and word together (i.e. 1HARBOR, 1NORTH)
3. Mixture of uppercase and lowercase names, etc.
4. Manufacturer is not null, but Brand is null or Brand will become null after filtering, i.e. the record 8.
5. Manufacturer is null, but Brand is not null
6. Both Manufacturer and Brand are not null.

Methodology - Jaccard Similarity versus Minimum edit distance

Jaccard Similarity:

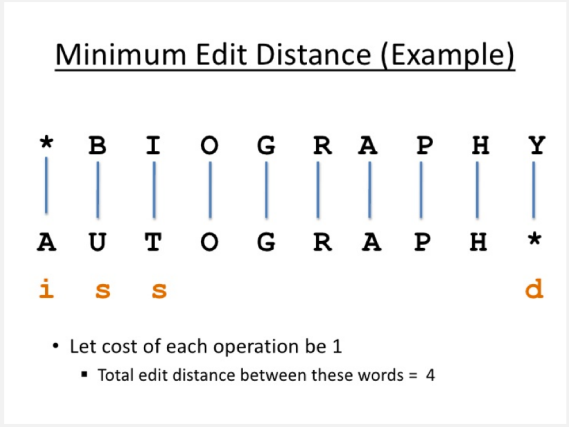
- Jaccard similarity is chosen for measuring the similarity of unstandardized names from retailer's assortment sheet to standardized names from ItemMaster's Salesforce database.
- The Jaccard similarity coefficient is defined as the size of the intersection divided by the size of the union of the sample sets.
- The Python **'Map a Lambda function to a list'** can get us the best matching account as well



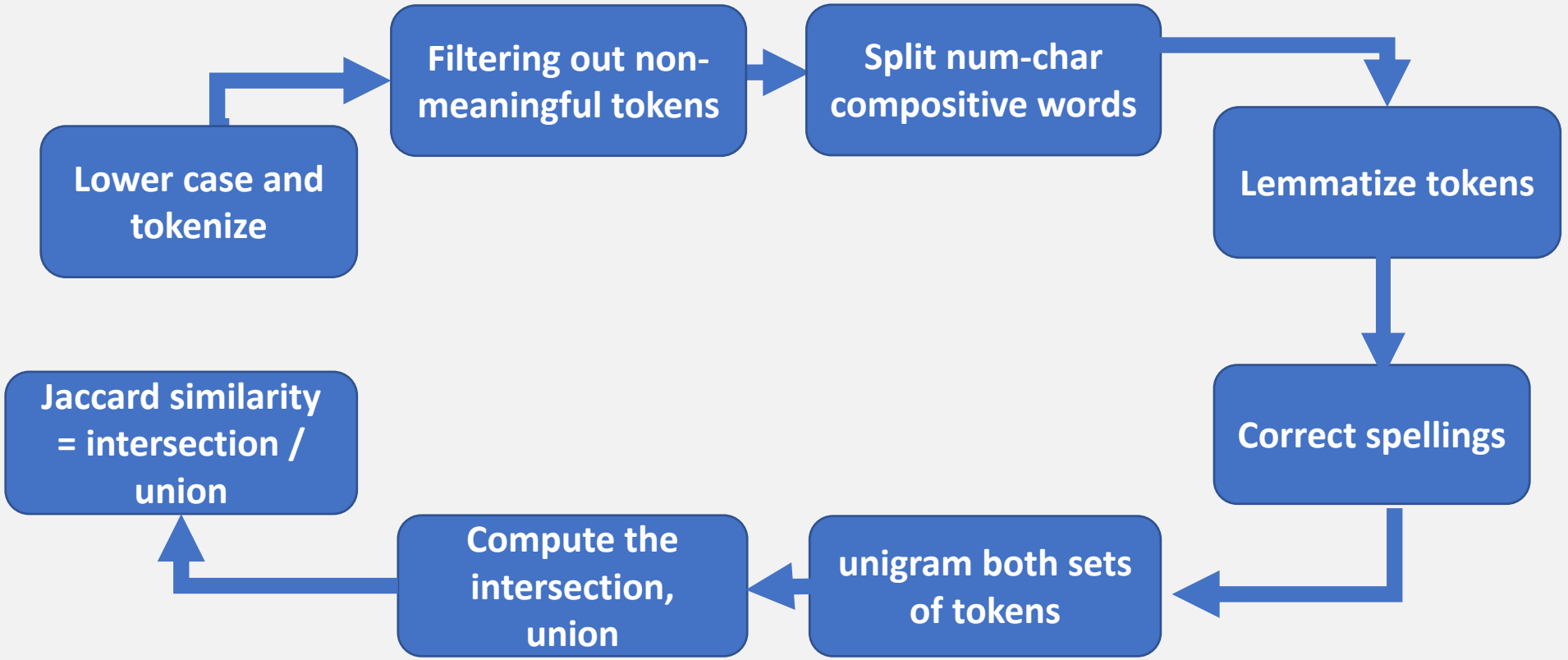
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Alternative Method:

- The minimum edit distance between two strings can be an alternative method. We chose Jaccard because its index value is easier to interpret and it gives satisfying results.



Methodology – Data Pre-processing steps



Note: For more detailed description of data pre-processing steps, please see appendix

Methodology – techniques for Data Preprocessing

NLTK:

- **NLTK** is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language.

RE:

- A ***regular expression*** is a special sequence of characters that helps you match or find other strings or sets of strings, using a specialized syntax held in a pattern.

Autocorrect spell:

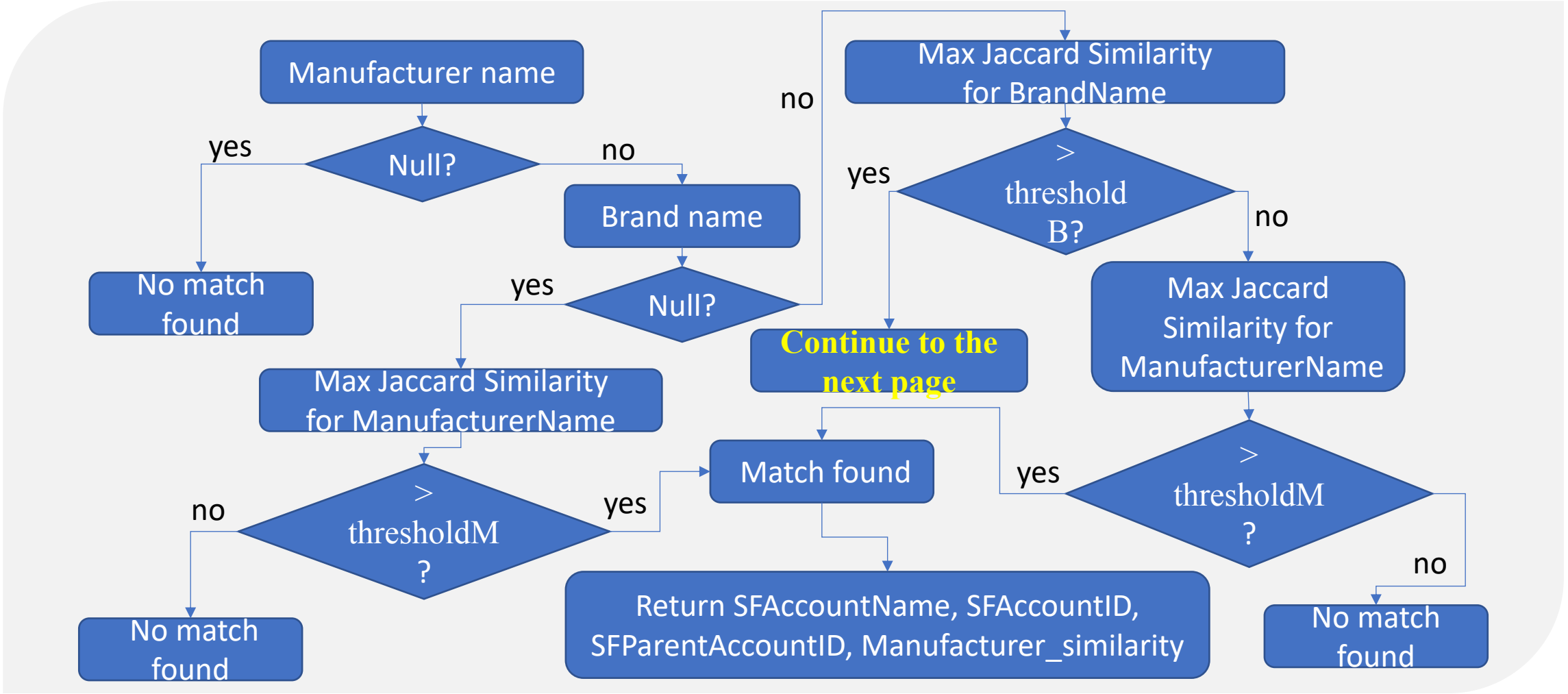
- We found **autocorrect spell** from **GitHub**. **GitHub** is where people build software. More than 27 million people use **GitHub** to discover, fork, and contribute to over 80 million projects.

Results – example records after data preprocessing

Manufacturer	Brand	SFAccountName	SFAccountID	SFParentAccountID	AcctName_ Similarity
IMPORT - PURCELL INTERNATIONA		Purcell International	001G000001H84CLIAZ		1
4 C FOODS CORP BR		4C Foods Corp.	001G000001H83vLIAR		0.75
1HARBOR SEAFOOD,INC		Harbor Seafood	001G000001Mj3AeIAJ		0.66667

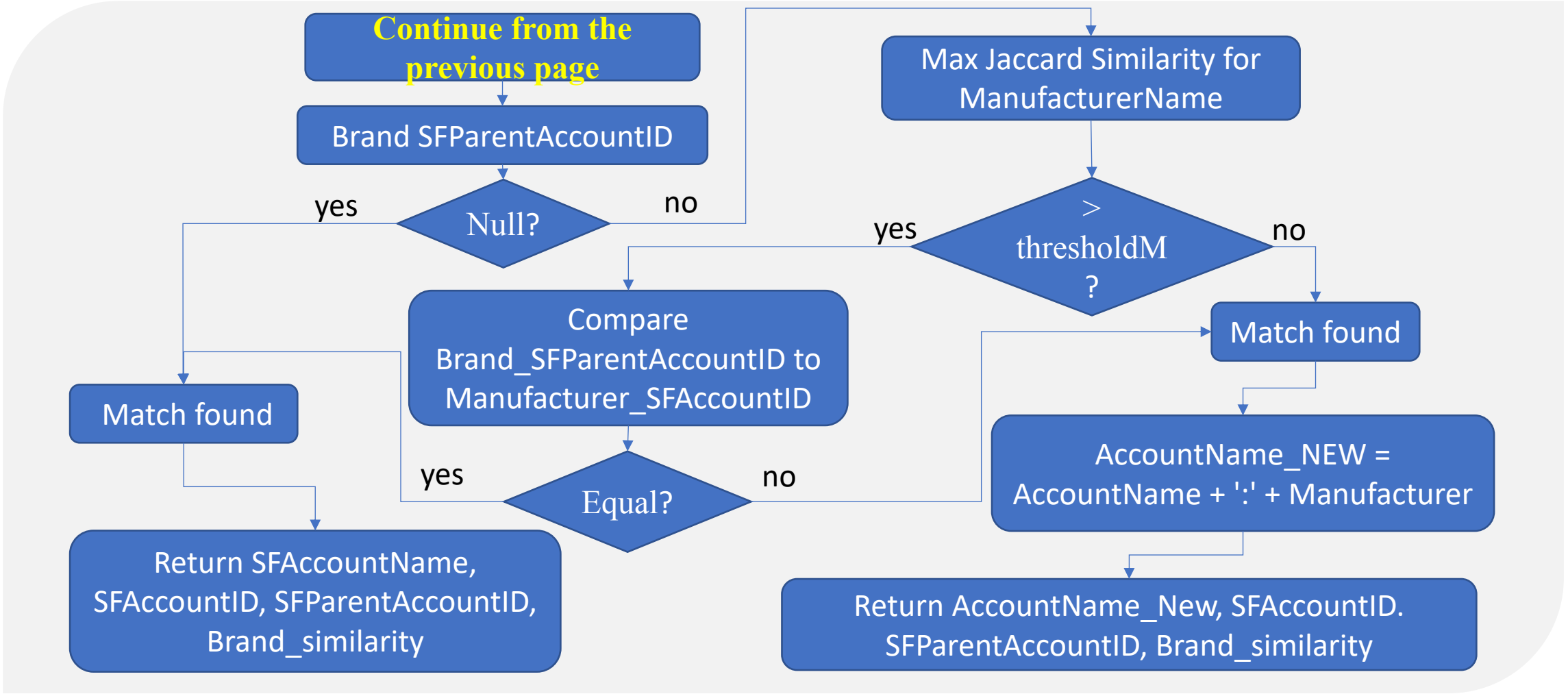
- The misspelled upper-case word ‘INTERNATIONA’ matched to ‘International’
- The num-char compositive word ‘4C’ matched to ‘4’ ‘C’
- The non-meaningful words ‘IMPORT’, ‘Corp’, ‘.’, ‘,’ and ‘INC’ were filtered out
- The num-char compositive word ‘1HARBOR’ splitted into ‘1’ and ‘HARBOR’ and then matched to ‘Harbor’

Pseudocode - Rules to determine the match being found (diagram)



Note: For more detailed description of this page, please see appendix

Pseudocode - Rules to determine the match being found (cont.)



Note: For more detailed description of this page, please see appendix

Results – example records for matching rules

Manufacturer	Brand	SFAccountName	SFAccountID	SFParentAccountID	AcctName_Similarity
BEAVER STREET FISHERIES INC	SEA BEST	Sea's Best	001G000002BUvsXIAT	001G000001ffUhIIAU	1
PEPSI LIPTON TEA PARTNERSHIP NORTH AMRCA	LIPTON PURE LEAF	Pepsico:PEPSI LIPTON TEA PARTNERSHIP NORTH AMRCA	0014A00002Hxs88QAB	001G000001H84BNIAZ	0.666666667
3M COMPANY	POST-IT	3M Homecare Division:3M COMPANY	001G0000025ZxY9IAK	001G000001IcYy4IAF	1

- First record, the match is based on 'Brand', and the parent accountId matches Manufacturer's accountId. The Jarccard similarity is 1.
- Second record, the matching account is also based on Brand; however, its parent accountId matched an account name as 'Pepsico', not the manufacturer name of 'PEPSI LIPTON TEA PARTNERSHIP NORTH AMRCA'. This tells us that our algorithm can find the match for newly merged customer company such as 'PEPSI LIPTON TEA PARTNERSHIP NORTH AMRCA'.

Results – example records of products per account from retailer

Manufacturer	Brand	Description	ItemUPC	Category A	Category B	Category C	Retailer	Private Label
IMPORT - PURCELL INTERNATIONA		WD MUSHROOMS STEMS/PIECE	21140216830	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD RED GRPFRUIT SECTIONS	21140218346	GROCERY	FRUIT CANNED	SPECIALTY FRUIT	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD ARTICHOKE HEARTS	21140218650	GROCERY	VEGETABLES - CANNED	PIMIENTOS/ONIONS/ARTICHOKES	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD MUSHROOMS BUTTONS	21140216854	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		SH SLICED MUSHROOMS	6.0788E+11	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		SH MUSHROOM STEMS & PCS	6.0788E+11	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		SH MUSHROOMS PCS&STEMS	6.0788E+11	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD MUSHROOMS SLICED/GLASS	21140216861	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD MUSHROOMS WHL /GLASS	21140216878	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD P/A TIDBITS IN JUICE	21140218421	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD MUSHROOMS SLICED	21140216847	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD SLICED PINEAPP N JUICE	21140218414	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD SLICED PINEAPP IN SYRP	21140218407	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD CRUSHED P/A IN JUICE	21140218438	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD MUSHROOMS SLICED	21140216892	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD PINEAPPCHNK /HVY SYRP	21140218391	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD CRUSHED P/A IN SYRUP	21140218445	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD CRUSHED P/A IN JUICE	21140218377	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD CHUNK P/A IN JUICE	21140218452	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		SH CHUNK PINEAPPLE IN JUICE	6.0788E+11	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD SLICED P/A IN JUICE	21140218360	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD MUSHROOMS STEMS/PIECES	21140216885	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		SH CRUSHED PINEAPL IN JUICE	6.0788E+11	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		SH SLICED PINEAPPLE IN JUICE	6.0788E+11	GROCERY	FRUIT CANNED	PINEAPPLE- CANNED	SEG	Y
IMPORT - PURCELL INTERNATIONA		SEG MUSHROOMS SLICED	38259107607	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		SH ARTICHOKE QUARTERS CAN	6.0788E+11	GROCERY	VEGETABLES - CANNED	PIMIENTOS/ONIONS/ARTICHOKES	SEG	Y
IMPORT - PURCELL INTERNATIONA		SEG MUSHROOMS STEMS/PIECE	38259107591	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD ARTICHOKE QUARTERS	21140016492	GROCERY	VEGETABLES - CANNED	PIMIENTOS/ONIONS/ARTICHOKES	SEG	Y
IMPORT - PURCELL INTERNATIONA		WD MUSHROOM WH BUTTON	21140216908	GROCERY	VEGETABLES - CANNED	MUSHROOMS CAN & GLASS	SEG	Y
IMPORT - PURCELL INTERNATIONA		SH ARTICHOKE HEARTS	6.0788E+11	GROCERY	VEGETABLES - CANNED	PIMIENTOS/ONIONS/ARTICHOKES	SEG	Y

Impact to ItemMaster's Business

Brian Cross via uchicagoedu.onmicrosoft.com Mar 29

to me, Cyril, Sanjay, William

Hi Manqing and William,

Here is a quick review of the impact on our organization.

HOURS SAVED:
It takes us an hour to do 100 accounts to manually identify the correct Salesforce Account
If you identify 1000 accounts, that saves us 10 hours of manual work
You were able to identify 3,057 additional accounts which translates to saving us 305 hours of manual work.

VALUE ADDED TO PIPELINE:
There is a sales pipeline value of $50 per item
If you identify the account for a brand with 5 items, it will add $250 to our sales pipeline
You were able to identify 41,734 additional unique items, which will add $2,086,600 to our sales pipeline.

I've also attached the review statistics that we were looking at on Tuesday. Let me know if you have any questions.

...

Conclusion

- Our algorithm is not only able to automate the manual work, but also able to find additional accounts and unique items that manual work cannot find. An assessment performed, based on four sets of data they provided for our development purposes alone, by ItemMaster concluded that our algorithm is able to
  - identify 3,057 additional accounts, which translates to saving ItemMaster 305 hours of manual work
  - plus 41,734 additional unique items, which will add $2,086,600 to ItemMaster's sales pipeline.
- This NLP algorithm we have developed for ItemMaster Inc. has reached or even exceeded our project client company's expectation.

Acknowledgement

- Our sincere thanks to Dr. Sema Barlas for her support, her advice along the way, and for making the NLP class available when we really needed it.
- Thanks to Nick Kadochnikov for teaching us these NLP techniques and all the advice he's given.
- Thanks to ItemMaster's analytics team for always being available whenever we had questions .

Appendix – Data pre-processing

Data pre-processing steps:

1. Using nltk library function nltk.word_**tokenize()** to tokenize both **lower cased** unstandardized and standardized names.
2. **Filtering out** non-meaningful tokens (words) from both sets of words
3. Using **re.match**() (regular expression match) to identify tokens consisting of both number and character (i.e. 4C), then to **split** these types of tokens into number string and character string (i.e. “4”, “C”).
4. Using nltk library function **nltk.WordNetLemmatizer**() to lemmatize each token in both sets of words.
5. Using autocorrect function spell() found from GitHub to autocorrect spelling errors in every token
6. Using nltk library function nltk.ngrams() to get the unigram of both sets of words.
7. Computing the intersection and union of the two sets of unigrams
8. Computing the Jaccard similarity = intersection / union

Appendix - Rules to determine the match being found (words)

Rules used to determine the match being found:

1. When a product's manufacturer name is null in a retailer's file, we say "no match found". The function call will give a return_values = ["No Match","", "", ""]
2. When the largest Jaccard similarity value among all matches found for a brand name in a retailer's file is larger than thresholdB and this matching account's parent account name matches the retailer's manufacturer name of the brand in the retail file, we say that the match is found! The function call will give a return_values = [AccountName_Brand, Brand_Account _SFAccountID, Brand_Account_SFParentAccountID, brand_similarity].
3. When the largest Jaccard similarity value among all matches found for a brand name in a retailer's file is larger than thresholdB and this matching account's parent account ID is 'nan', we say that the match is found! The function call will give a return_values = [AccountName_Brand, Brand_Account _SFAccountID, Brand_Account_SFParentAccountID, brand_similarity]. (cont.)

Appendix - Rules to determine the match being found (cont.)

Rules used to determine the match being found (cont.):

4. When the largest Jaccard similarity value among all matches found for a brand name in a retailer's file is larger than thresholdB and this matching account's parent account name doesn't match the retailer's manufacturer name, then according to Item Master's analytics team, this should also be considered a match. Additionally, a new record is created with AccountName_NEW = parent_SFAccountName:retailer_manufacture. The function call will give a return_values = [AccountName_NEW, Brand_Account _SFAccountID, Brand_Account_SFParentAccountID, brand_similarity].
5. When the largest Jaccard similarity value among all matches found for a brand name in a retailer's file is less than or equal to thresholdB, but the largest Jaccard similarity value among all matches found for the manufacturer name of the product is larger than thresholdM, we say that the match is found! The function call will give a return_values = [AccountName_Manufacturer, Manufacturer_Account _SFAccountID, Manufacturer_Account_SFParentAccountID, Manufacturer_similarity].


Thank You
Q&A