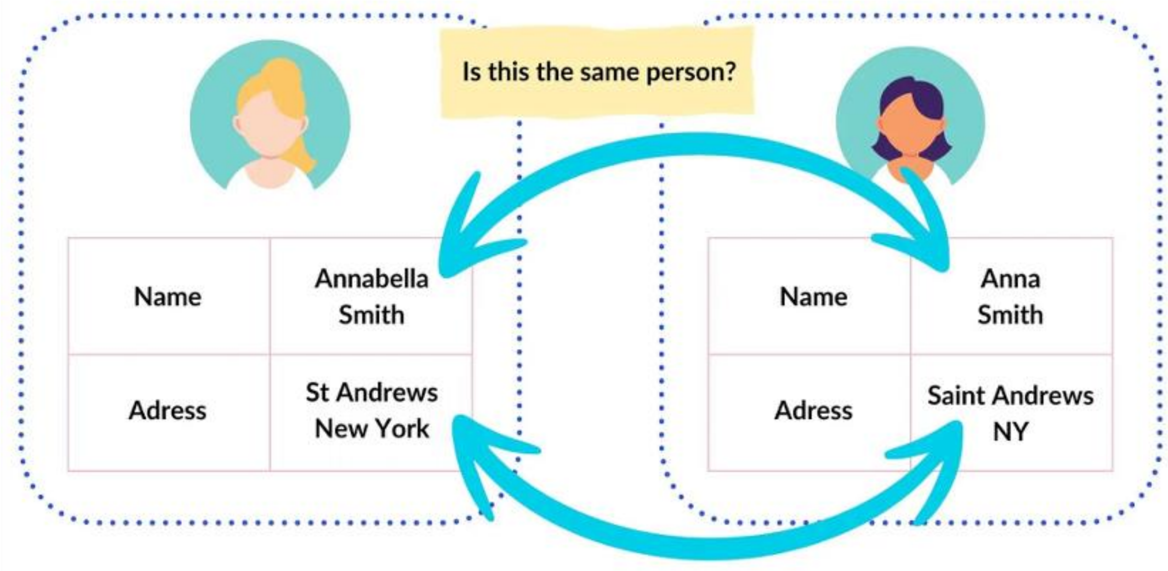
BrandNERD - An Extensive Brand Dataset and Analysis Pipeline for Name Entity Resolution

Nicholas Caporusso, Alina Campan,
Ayush Bhandari, Stephen Kroeger, and Sarita Gautam

School of Computing and Analytics
Northern Kentucky University

Named Entity Resolution (NER)

- Aggregating different names that refer to the same real-world entity

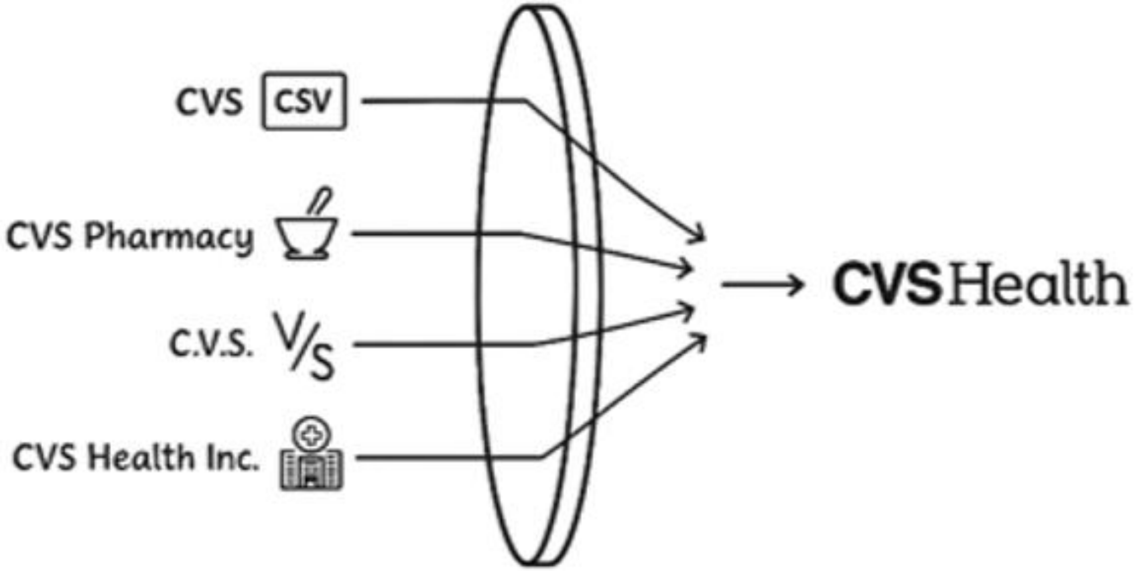


Name	Annabella Smith
Adress	St Andrews New York

Name	Anna Smith
Adress	Saint Andrews NY

Our specific case and problem

- Online auction platform
- ~40 million products
- 400,000+ raw brand names
- Unresolved names
  - misspelled, ambiguous, unverified brand names



State of the art (1/2): resources

Lack of datasets

- Few open-source datasets
- Limited number of entries
- Large but general datasets (USPTO)

Algorithms involve trade-offs

- String comparison (similarity)
- Context-aware or web-based
- Machine Learning and Deep Learning

State of the art (2/2): trade-offs

Automated processing

- Fast and scalable
- Inaccurate

- Issues:
  - Struggles with short, noisy text
  - Lacks semantic understanding
  - Easily broken by tiny differences (e.g. Dleta vs Delta)

Manual resolution

- Accurate
- Very slow

- Issues:
  - Not suitable for large datasets

Our work and contribution

A large dataset

- ~369,000 canonicalized brands
- ~60,000 validated brand names
- Separate stage-based datasets

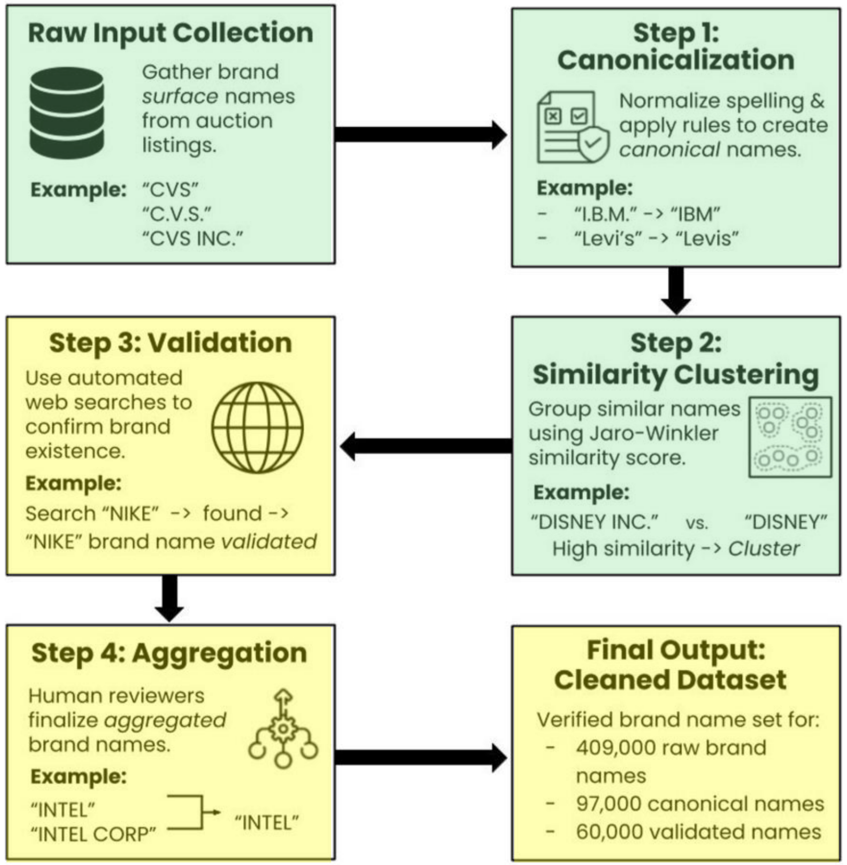
An extensible pipeline

- Separate processing modules
- Scalable approach
- Inspectable processing

An **open-source** resource for researchers and developers

Our processing pipeline

- Semi-automated processing
  - Rule-based normalization
  - String similarity scoring
  - Automated web search
  - Manual human review
- No external dependencies
- Additional context information
  - obtained through web search



Raw Input Collection

Gather brand *surface* names from auction listings.

Example: "CVS"
"C.V.S."
"CVS INC."

Step 1: Canonicalization

Normalize spelling & apply rules to create *canonical* names.

Example:
- "I.B.M." -> "IBM"
- "Levi's" -> "Levis"

Step 2: Similarity Clustering

Group similar names using Jaro-Winkler similarity score.

Example:
"DISNEY INC." vs. "DISNEY"
High similarity -> *Cluster*

Step 3: Validation

Use automated web searches to confirm brand existence.

Example:
Search "NIKE" -> found -> "NIKE" brand name *validated*

Step 4: Aggregation

Human reviewers finalize *aggregated* brand names.

Example:
"INTEL"
"INTEL CORP" -> "INTEL"

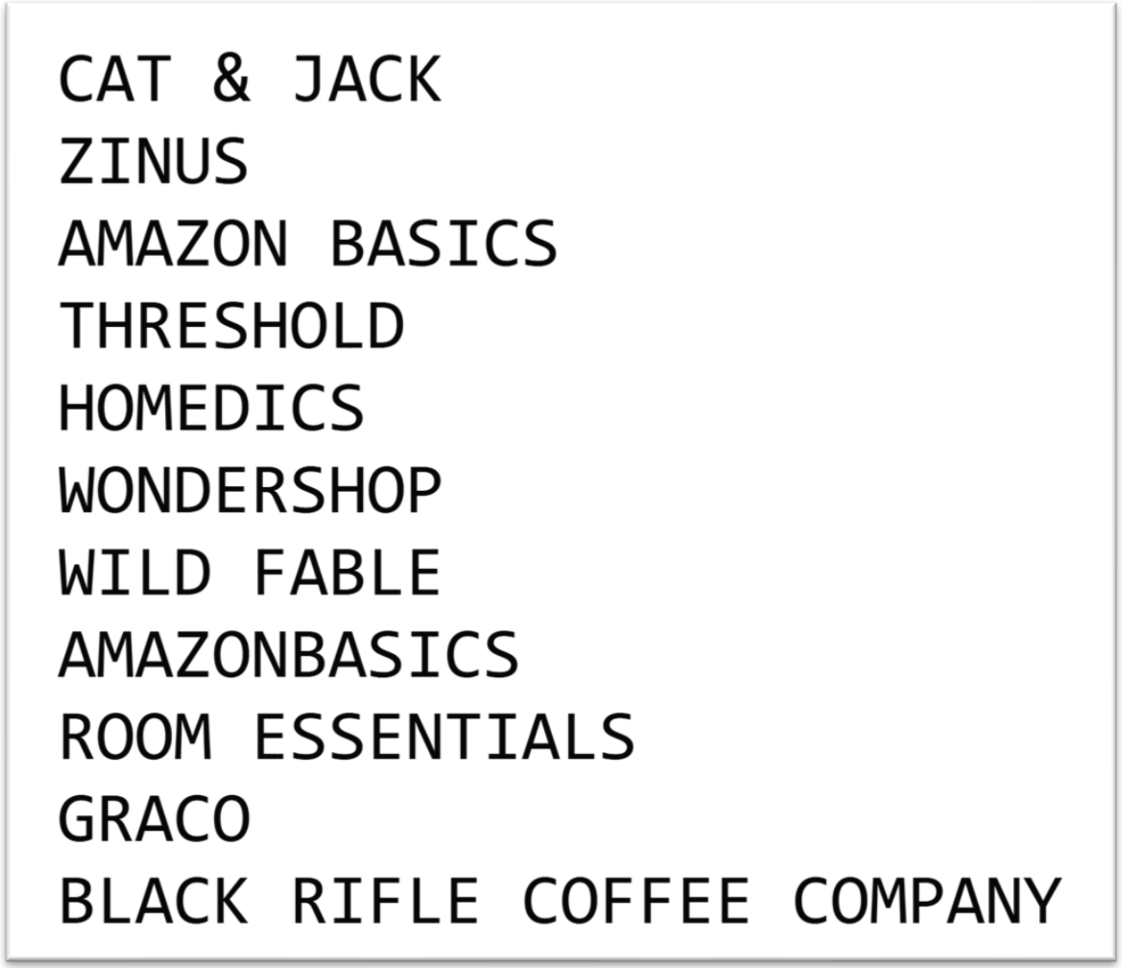
Final Output: Cleaned Dataset

Verified brand name set for:
- 409,000 raw brand names
- 97,000 canonical names
- 60,000 validated names

🟩 completed 🟨 ongoing

Step 1: Raw input collection

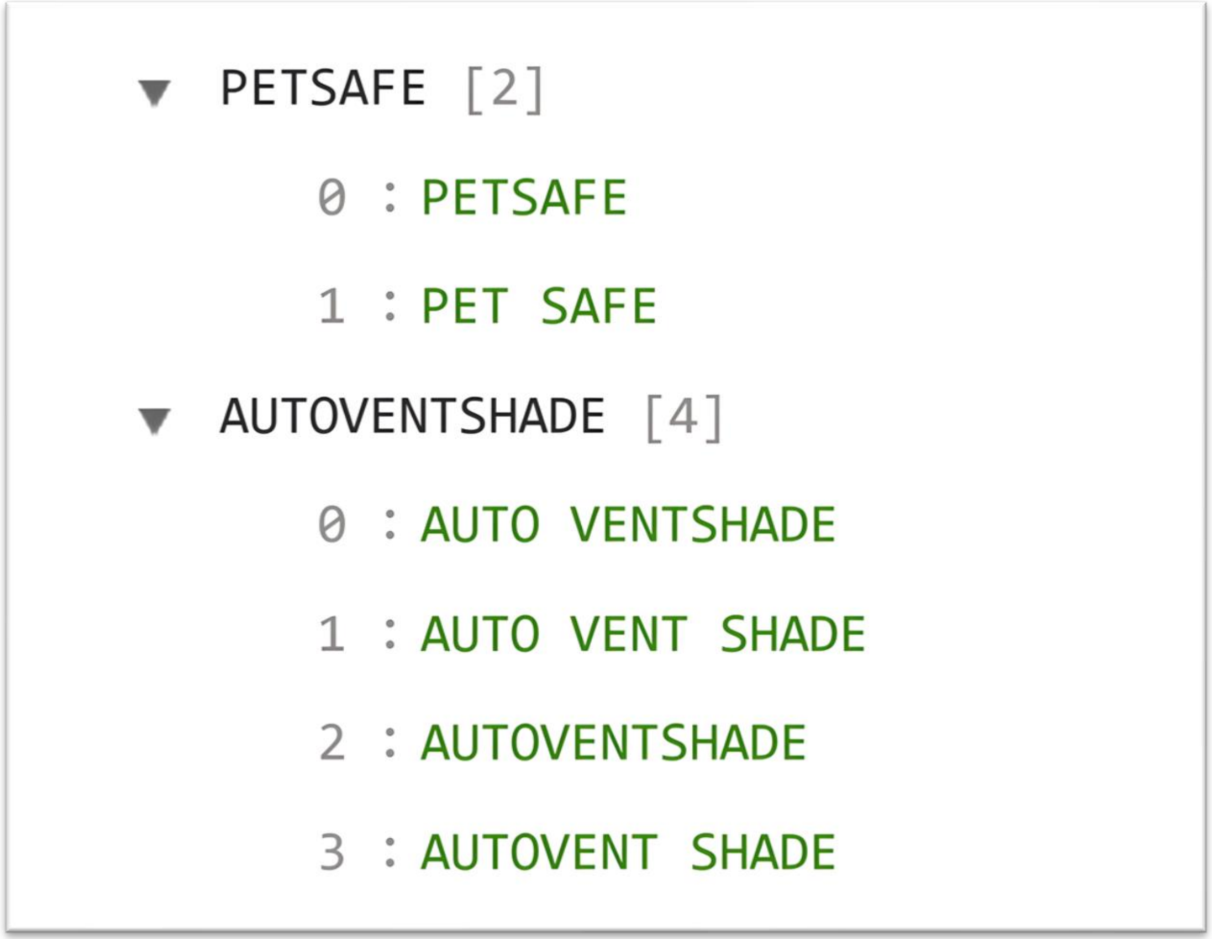
- Goal
  - Obtain a reliable set of brands
- Method
  - Scraped from online source
- Result
  - **394,542 raw brand names**
  - No context information
  - Need resolution to enable further processing



```
CAT & JACK
ZINUS
AMAZON BASICS
THRESHOLD
HOMEDICS
WONDERSHOP
WILD FABLE
AMAZONBASICS
ROOM ESSENTIALS
GRACO
BLACK RIFLE COFFEE COMPANY
```

Step 2: Canonicalization

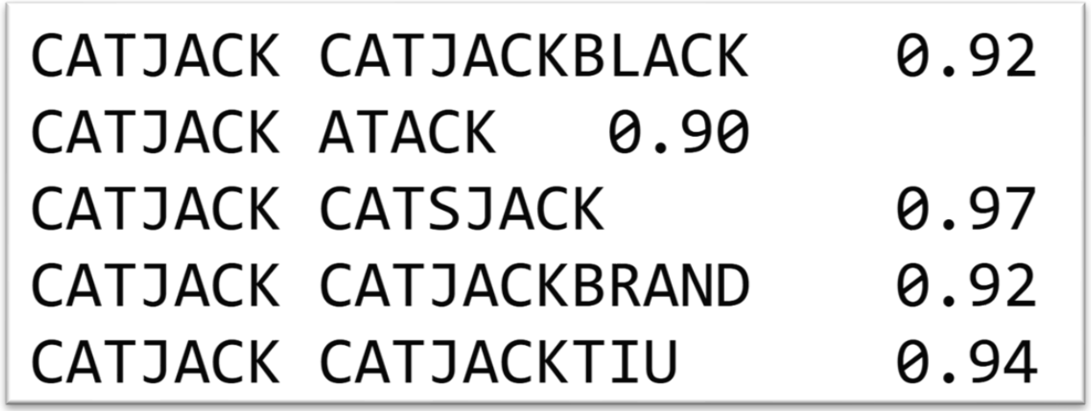
- Goal
  - Reconcile syntactically equivalent brand names
- Method
  - Two-stage transformation
    - Cleaning
    - Normalization
- Result
  - **368,703 canonical brands**
    - 93.45% of total



```
▼ PETSAFE [2]
      0 : PETSAFE
      1 : PET SAFE
▼ AUTOVENTSHADE [4]
      0 : AUTO VENTSHADE
      1 : AUTO VENT SHADE
      2 : AUTOVENTSHADE
      3 : AUTOVENT SHADE
```

Step 3: Similarity matching (1/2)

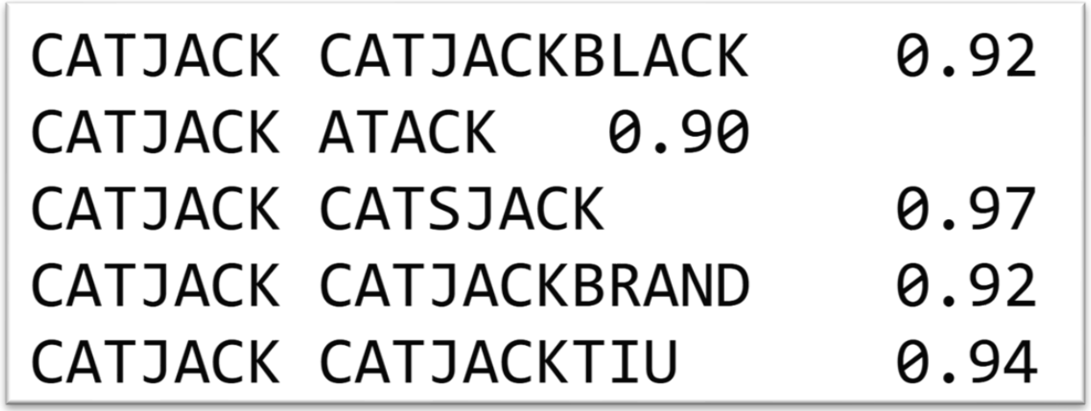
- Goal
  - find similar brand names
- Method
  - Experimented with various metrics
- Result
  - Jaro-Winkler has best accuracy



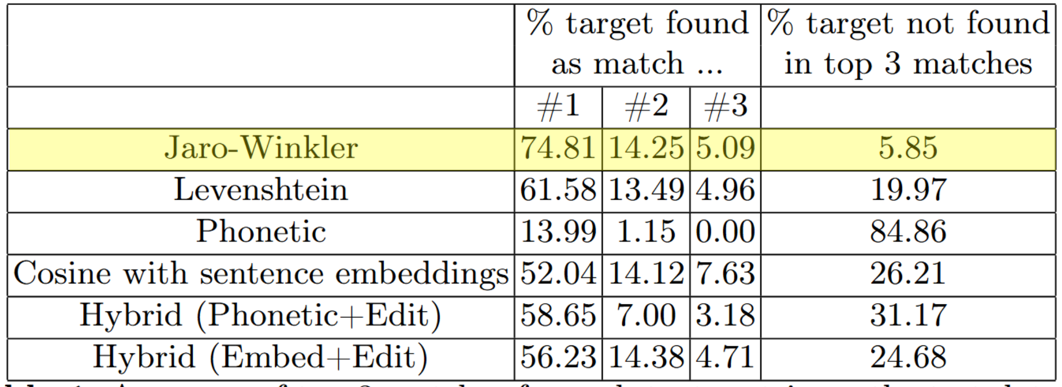
```
CATJACK CATJACKBLACK    0.92
CATJACK ATACK   0.90
CATJACK CATSJACK        0.97
CATJACK CATJACKBRAND    0.92
CATJACK CATJACKTIU      0.94
```

Step 3: Similarity matching (2/2)

- Goal
  - find similar brand names
- Method
  - Experimented with various metrics
- Result
  - Jaro-Winkler has best accuracy
  - **782,299 pairs with similarity >0.9**



```
CATJACK CATJACKBLACK     0.92
CATJACK ATACK    0.90
CATJACK CATSJACK         0.97
CATJACK CATJACKBRAND     0.92
CATJACK CATJACKTIU       0.94
```



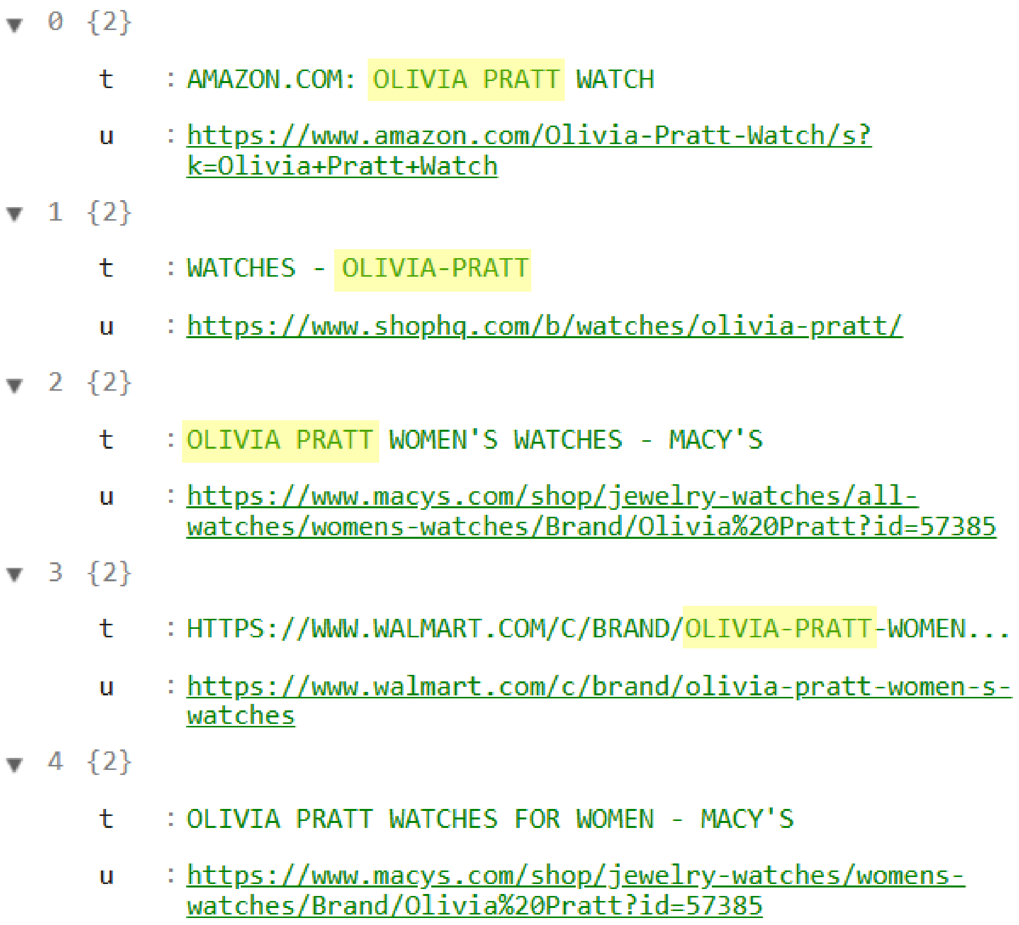
	% target found as match ...			% target not found in top 3 matches
	#1	#2	#3	
Jaro-Winkler	74.81	14.25	5.09	5.85
Levenshtein	61.58	13.49	4.96	19.97
Phonetic	13.99	1.15	0.00	84.86
Cosine with sentence embeddings	52.04	14.12	7.63	26.21
Hybrid (Phonetic+Edit)	58.65	7.00	3.18	31.17
Hybrid (Embed+Edit)	56.23	14.38	4.71	24.68

Table 1. Accuracy of top 3 matches for each text metric on the test dataset

Step 4: Validation (1/2)

- Goal
  - Verify that the brand exists
- Method
  - Web search via browser automation

BRAND NAME: OLIVIA PRATT

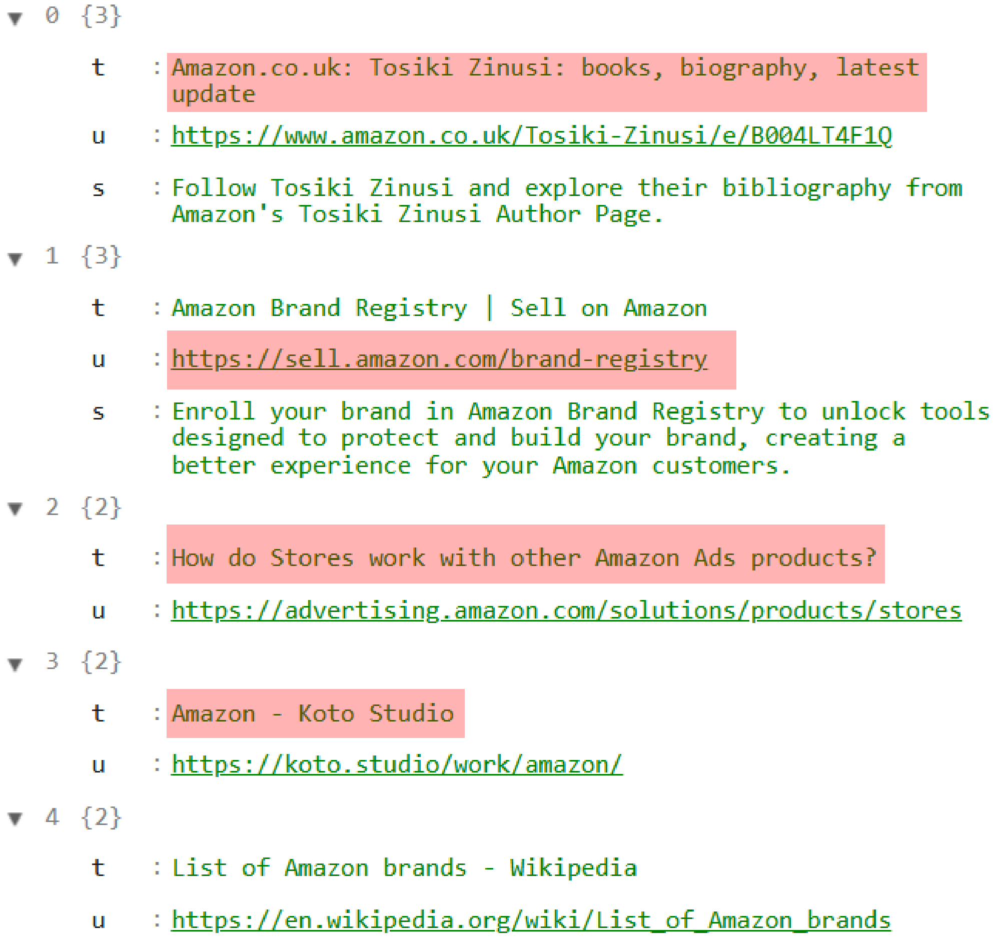


```
▼ 0 {2}
      t  : AMAZON.COM: OLIVIA PRATT WATCH
      u  : https://www.amazon.com/Olivia-Pratt-Watch/s?
           k=Olivia+Pratt+Watch
▼ 1 {2}
      t  : WATCHES - OLIVIA-PRATT
      u  : https://www.shophq.com/b/watches/olivia-pratt/
▼ 2 {2}
      t  : OLIVIA PRATT WOMEN'S WATCHES - MACY'S
      u  : https://www.macys.com/shop/jewelry-watches/all-
           watches/womens-watches/Brand/Olivia%20Pratt?id=57385
▼ 3 {2}
      t  : HTTPS://WWW.WALMART.COM/C/BRAND/OLIVIA-PRATT-WOMEN...
      u  : https://www.walmart.com/c/brand/olivia-pratt-women-s-
           watches
▼ 4 {2}
      t  : OLIVIA PRATT WATCHES FOR WOMEN - MACY'S
      u  : https://www.macys.com/shop/jewelry-watches/womens-
           watches/Brand/Olivia%20Pratt?id=57385
```

Step 4: Validation (2/2)

- Goal
  - Verify that the brand exists
- Method
  - Web search via browser automation
- Result
  - **309,581 brands searched**
    - 83% of canonical brands
  - **72,303 brands validated**
    - 20% of canonical brands
  - Process is ongoing

BRAND NAME: ZINUSI



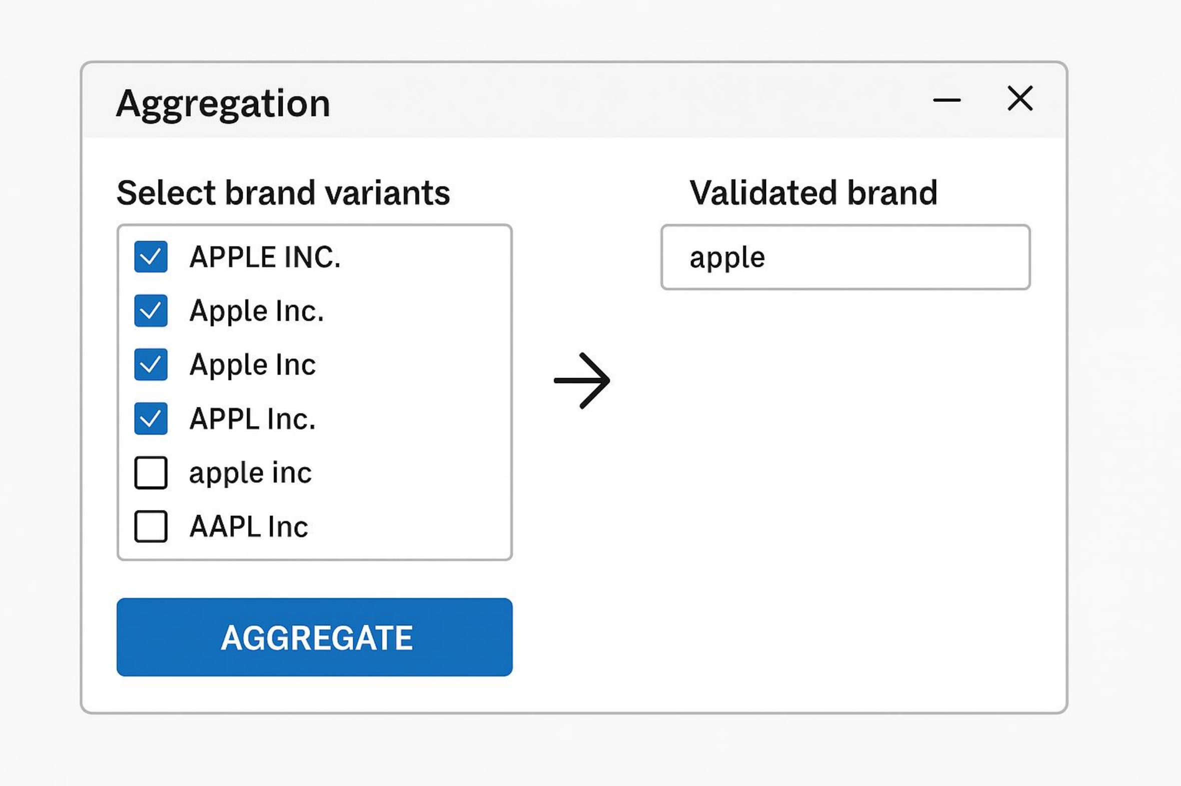
```
▼ 0 {3}
      t  : Amazon.co.uk: Tosiki Zinusi: books, biography, latest
           update
      u  : https://www.amazon.co.uk/Tosiki-Zinusi/e/B004LT4F1Q
      s  : Follow Tosiki Zinusi and explore their bibliography from
           Amazon's Tosiki Zinusi Author Page.
▼ 1 {3}
      t  : Amazon Brand Registry | Sell on Amazon
      u  : https://sell.amazon.com/brand-registry
      s  : Enroll your brand in Amazon Brand Registry to unlock tools
           designed to protect and build your brand, creating a
           better experience for your Amazon customers.
▼ 2 {2}
      t  : How do Stores work with other Amazon Ads products?
      u  : https://advertising.amazon.com/solutions/products/stores
▼ 3 {2}
      t  : Amazon - Koto Studio
      u  : https://koto.studio/work/amazon/
▼ 4 {2}
      t  : List of Amazon brands - Wikipedia
      u  : https://en.wikipedia.org/wiki/List_of_Amazon_brands
```

Step 5: Aggregation

- Goal
  1. Aggregate similar brands
  2. Resolve them to a canonical name

- Method
  - Web interface for manual review

- Result
  - **32,114 resolved brands**
    - 8.7% of canonical brands
  - Process is ongoing



Aggregation

Select brand variants

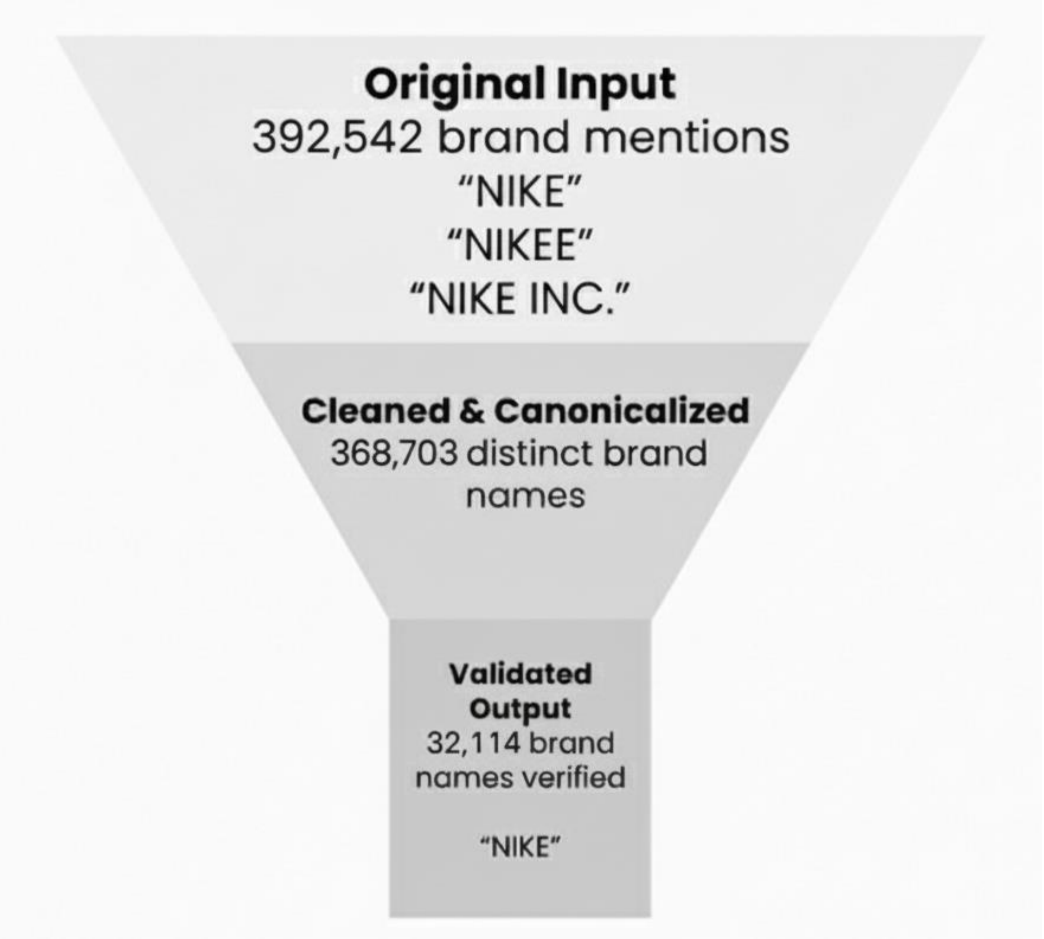
☑ APPLE INC.
☑ Apple Inc.
☑ Apple Inc
☑ APPL Inc.
□ apple inc
□ AAPL Inc

Validated brand

apple

Current status

- Completed 3/6 NER tasks
  - Currently working on validation, aggregation, and resolution
- Dataset released open-source
  - CC BY 4.0
  - https://bit.ly/3VCc2Sn
- Repository regularly updated



Original Input
392,542 brand mentions
"NIKE"
"NIKEE"
"NIKE INC."

Cleaned & Canonicalized
368,703 distinct brand names

Validated Output
32,114 brand names verified

"NIKE"

Future work

Dataset

- Continue validation, aggregation, and resolution
    - resolve 100% of brands
- Expand scope
    - to models and products
- Continue maintaining the repo

Pipeline

- Incorporate other methods
    - i.e., clustering, other similarity metrics, graph analysis
- Leverage context information
    - e.g., product description
- Explore collaborative options
    - e.g., crowdsourcing

Contacts

Contacts

- **Ayush Bhandari (presenting)**
  - bhandaria3@mymail.nku.edu
- **Nicholas Caporusso (PI)**
  - caporusson1@mymail.nku.edu
- **Alina Campan (PI)**
  - campana1@mymail.nku.edu

BrandNERD repository

- https://bit.ly/3VCc2Sn or
- https://github.com/NKU-HCI-lab/brandNERD-public