Recognizing Quantity Names for Tabular Data

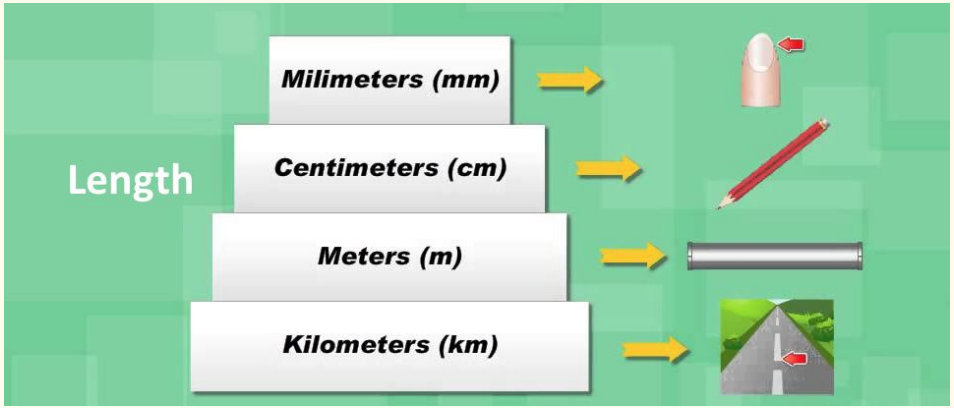
Yang Yi, Zhiyu Chen, Jeff Heflin, Brian D. Davison

Dept. of Computer Science and Engineering
Lehigh University

LEHIGH UNIVERSITY

What is a Quantity Name?

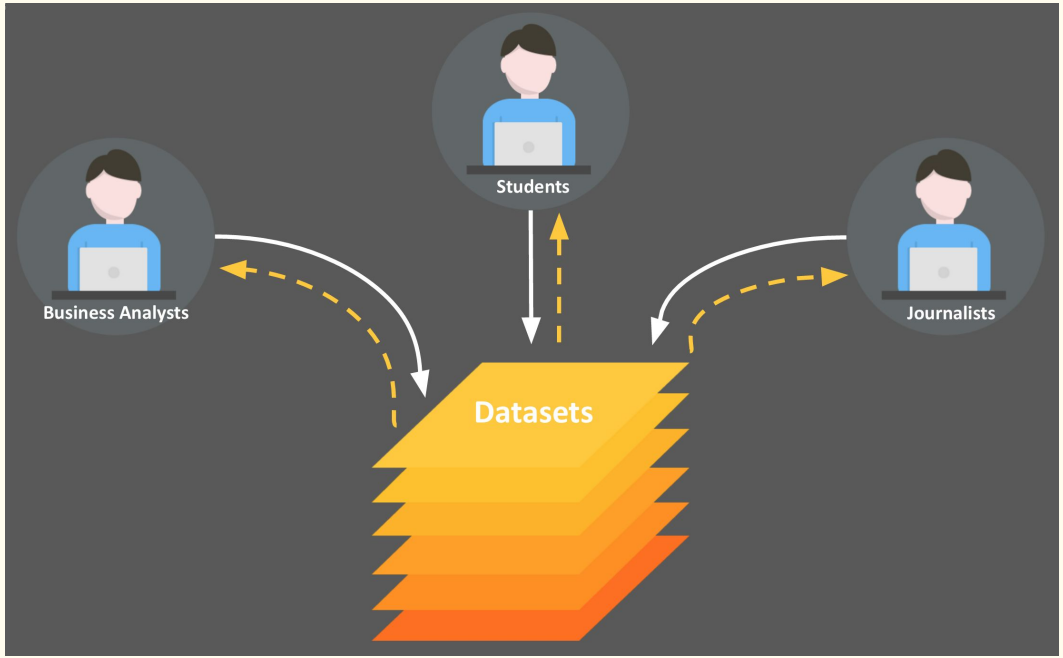
- A **quantity name** (also known as a **quantity kind**) is a kind of quantity that can be measured using defined and unrestricted units of measurement[1].



[1] Reference: https://docs.nomagic.com/display/SYSMLP183/Quantity+Kind

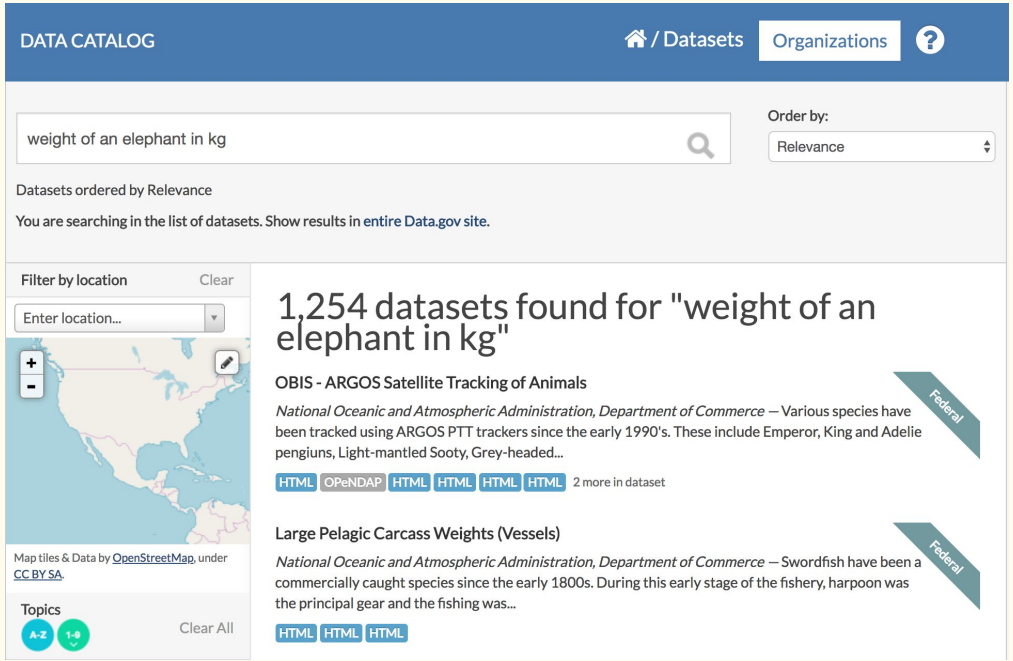
Motivation

- People in many roles are capturing, storing, and analyzing datasets



Motivation

- Inferring quantity names to improve the ability to match datasets



LEHIGH UNIVERSITY

Motivation

- Quantity names provide for a broader search scope than simply units

Objective

Design and implement a model to ***recognize*** and ***recommend*** **quantity names** for numeric columns based on features extracted from column name and column content.

LEHIGH UNIVERSITY

Objective

Length	Time	Weight	Percent	Currency
Elevation, ft	**duration_seconds**	**CO2 (tons)**	**Confidence_limit_High**	**Total income (dollars in millions)**
1155	30.24	26601.04	23.6	342.1
0	30.56	29448.39	35	2279.1
203	247.52	9932.26	38	3995.9
204	97.34	15689.41	15.4	5978.8
204	30.11	23015.94	7.4	8431.3
...	...	...	...	...
1074	36.76	7324.18	41.3	20034.5
1100	49.52	0	57	28997
1354	81.23	0	57.2	134038.4
1090	198.53	928126.66	22.7	230468.1
1090	49.82	0	87.3	

LEHIGH UNIVERSITY

Related Work

- Thomas et al. (ADCS 2015) & Au et al. (ADCS 2016)
  - infer data types: Strings, Numbers, Boolean values, Dates, and place names
- Sarawagi et al. (KDD 2014)
  - unit extraction in queries on web tables: unit extractors to pull units from column names
- Valera et al. (ICML 2017)
  - discover statistical types of variables in a dataset
- DasSarma et al. (SIGMOD 2012) & Wick: et al. (KDD 2008)
  - find related tables by computing schema similarity
- Ratinov et al. (WI 2004)
  - expand the abbreviation in schemas (units are typically stored in abbreviated form)

Dataset

- Terms associated with five common quantity names
- Another class label “other”
  - Other quantity names, compound units, not quantity names, ...

Quantity Name	Units	Abbreviation	Context
Length	meter, mile, inch, feet	m, mi, in, ft	height, width
Time	second, minute, hour	sec, s, min, hr, hrs	duration
Percent	percentage	%	accuracy
Currency	dollar, euro, pound	USD, $, EUR, GBP	amount, cost
Weight	gram, kilogram, pound, ounce, ton	g, kg, lb, oz, t	

in parentheses
Perimeter (m)

after “in”
Dist. from Coop in miles

after a dash or underscore
segment_length_ft

tie with context terms
time seconds

Dataset

Extract Data from *data.gov* and give ID → Retain numeric columns only → Label column with 0-5 → Remove duplicate column names within the same dataset

Quantity Name	# of Instances
Length	896
Time	352
Percent	1031
Currency	875
Weight	233
Total	**3387**

LEHIGH UNIVERSITY

Features

Built From	ID	Type	Feature
Column Content	1	Real with length 1	Maximum value
	2	Real with length 1	Minimum value
	3	Real with length 1	Average value
	4	Real with length 1	Range value (maximum - minimum)
	5	Integer with length 1	Length of the maximum value (when expressed as a string)
Column Name	6	Integer with 1ength 1	Number of words
	7	Integer with length 1	Number of characters
	8	Array of 5 booleans	Presence of quantity-specific terms for each quantity name

Column Name: Canopy Height in meters

...(other feature names)	Match with Length	Match with Time	Match with Percent	Match with Currency	Match with Weight
...(other features)	1	0	0	0	0

Column Name: Trip duration

...(other feature names)	Match with Length	Match with Time	Match with Percent	Match with Currency	Match with Weight
...(other features)	0	1	0	0	0

10-Fold Cross Validation

- Upsample classes to handle imbalanced class distribution

Quantity Name	# of Instances
Length	896
Time	352
Percent	1031
Currency	875
Weight	233
Total	**3387**

LEHIGH UNIVERSITY

10-Fold Cross Validation

- Upsample classes to handle imbalanced class distribution

Quantity Name	Training Part	Testing Part
Length	807	89
Time	317	35
Percent	928	103
Currency	788	87
Weight	210	23
None	∞	∞

10-Fold Cross Validation

- Upsample classes to handle imbalanced class distribution

Quantity Name	Training Part	Testing Part
Length	807	89
Time	317	35
Percent	928	103
Currency	788	87
Weight	210	23
None	∞	∞

Quantity Name	Training Part
Length	928
Time	928
Percent	928
Currency	928
Weight	928
None	928
Total	**5568**

10-Fold Cross Validation

- Upsample classes to handle imbalanced class distribution

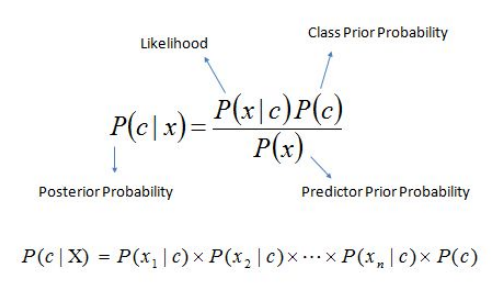
Quantity Name	Training Part	Testing Part
Length	807	89
Time	317	35
Percent	928	103
Currency	788	87
Weight	210	23
None	∞	∞

Quantity Name	Training Part	Testing Part
Length	928	89
Time	928	35
Percent	928	103
Currency	928	87
Weight	928	23
None	928	103
Total	**5568**	**440**

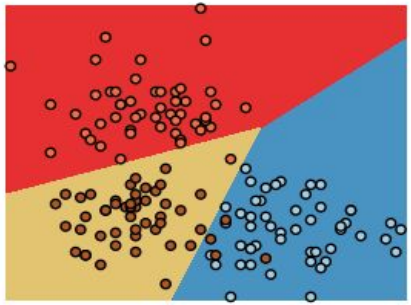
Remain unchanged

Classification Models

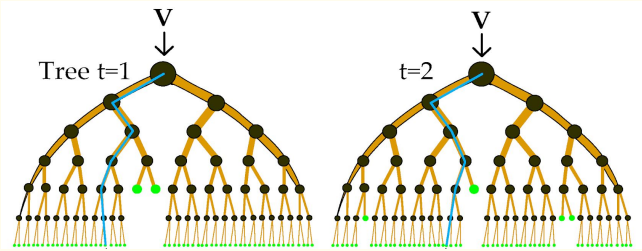
Naive Bayes for multivariate Bernoulli models
Accuracy: 77.3%



SVM with a linear kernel (LinearSVC)
Accuracy: 48.7%



Random forest ☆☆☆
Accuracy: 89.5% (with 200 trees and max depth 200)



LEHIGH UNIVERSITY

Evaluation

- Overall Accuracy: 89.5%
- Confusion Matrix showing counts

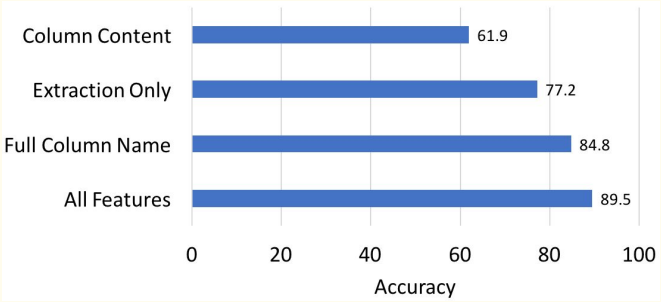
Example mistakes:

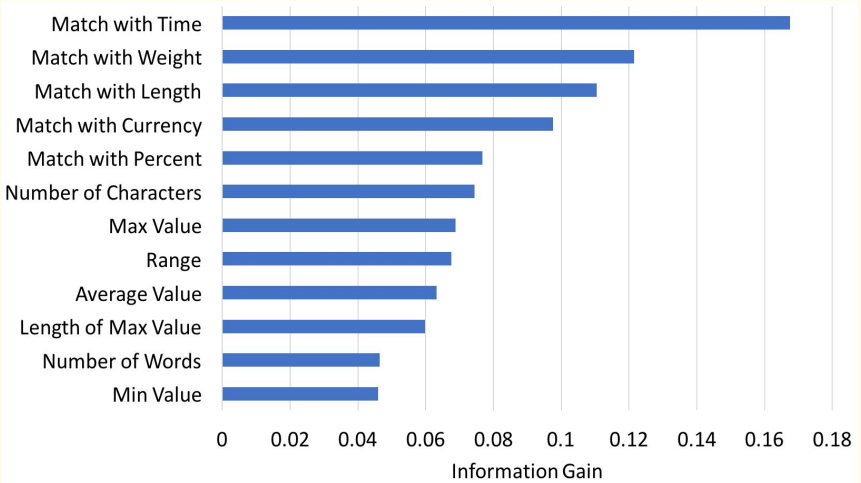
sqmile → Length
toe(s) → Time
Refunds - Individual Income Tax → Percent
Number of glass doors → Currency
Number of Boats → Weight
Steel (lbs) → Other

	Predicted Class					
Actual	Length	Time	Percent	Currency	Weight	Other
Length	86	0	0	0	0	3
Time	0	34	0	0	0	1
Percent	0	0	100	0	0	3
Currency	0	0	1	76	1	9
Weight	0	0	0	0	20	3
Other	1	1	12	4	1	84

Evaluation

- Expands on the rule-based approach presented by Sarawagi et al.
  - apply and expand their rule-based extractor: accuracy of 77.2%.
  - also predict for columns that quantity names and units that are not provided
- Feature Importance

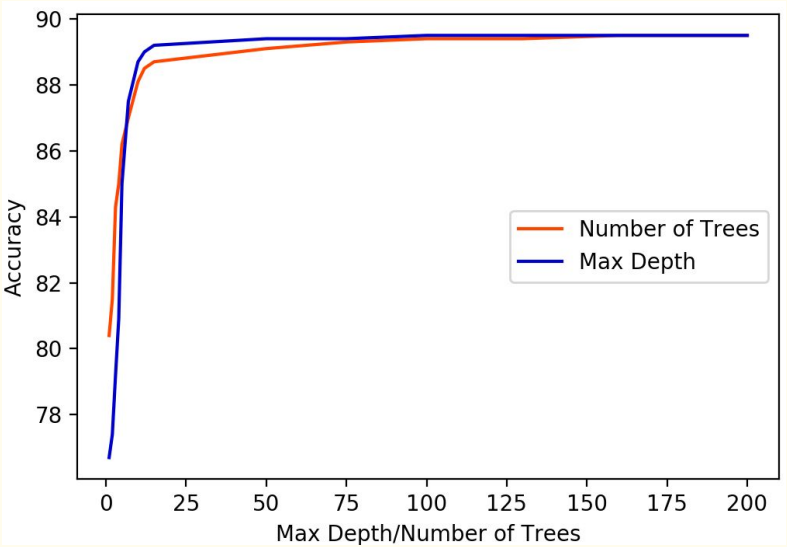




LEHIGH UNIVERSITY

Evaluation - Sensitivity Analysis

- Random forest model
  - Number of trees
  - Max depth
  - Other Parameters
    - use default in scikit-learn



LEHIGH UNIVERSITY

Why it Works Well

- Most important: presence of quantity-specific terms
- Focusing on five quantity names
  - Many more possible; likely more difficult
- Column content features perform unexpectedly well
  - Over 60% accuracy (when used alone)

Future Work

- Explore the metadata and description of datasets
  - Some datasets have attribute information provided in the description or an explicit data dictionary in a separate file
- Expand this work to actual units
- Expand the table containing terms associated with quantity names

LEHIGH UNIVERSITY

Thank you!

References

- [1] Au, V., Thomas, P., Jayasinghe, G.K.: Query-biased summaries for tabular data. In: Proc. 21st Australasian Document Computing Symp. ADCS '16 (2016) 69-72
- [2] Chen, Z., Jia, H., Heflin, J., Davison, B.D.: Generating schema labels through dataset content analysis. In: Companion Proceedings of The Web Conference. WWW '18 (2018) 1515-1522
- [3] Das Sarma, A., Fang, L., Gupta, N., Halevy, A., Lee, H., Wu, F., Xin, R., Yu, C.: Finding related tables. In: Proc. ACM SIGMOD Int'l Conf. on Management of Data. SIGMOD '12 (2012) 817-828
- [4] Ratinov, L., Gudes, E.: Abbreviation expansion in schema matching and web integration. In: Proc. IEEE/WIC/ACM Int'l Conf. on Web Intelligence. WI '04 (2004) 485-489
- [5] Sarawagi, S., Chakrabarti, S.: Open-domain quantity queries on web tables: Annotation, response, and consensus models. In: Proc. 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. KDD '14 (2014) 711-720
- [6] Thomas, P., Omari, R., Rowlands, T.: Towards searching amongst tables. In: Proc. 20th Australasian Document Computing Symp. ADCS '15 (2015) 8:1-8:4
- [7] Valera, I., Ghahramani, Z.: Automatic discovery of the statistical types of variables in a dataset. In: Int'l Conf. on Machine Learning. (2017) 3521-3529
- [8] Wick, M.L., Rohanimanesh, K., Schultz, K., McCallum, A.: A unified approach for schema matching, coreference and canonicalization. In: Proc. 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. KDD '08 (2008) 722-730

LEHIGH UNIVERSITY