Visualizing CDLI Accounting Corpora

Google Summer of Code 2020

Logan Born | Maxim Ionov

Overview

Information extraction from Sumerian accounting tablets

Makes data **more approachable** and enables **quantitative analysis**

Motivation

CDLI data contains >100k Sumerian administrative tablets

Numeric data, good for statistical analyses

...but recorded in non-numeric format: e.g. “250 sheep” written as “4(gesz2) 1(u) udu”

Challenges:

- Opaque to non-experts
- Potentially complex and slow to interpret (subtraction, very small fractions)
- Large-scale quantitative analysis is impossible

Numeral Conversion

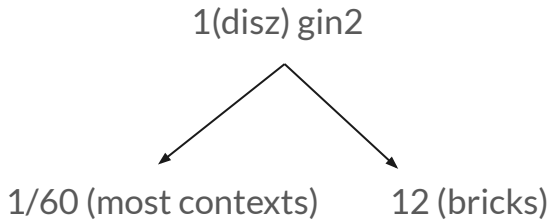
Dictionary-based conversion of Sumerian numerals to Arabic notation.

Adjust value based on surrounding context:

1(disz)	vs.	**1(disz)** gin2	vs.	**3(disz)**	**1/3(disz)**	**9(disz)**	**2/3(disz)** gin2
1		1/60		3	1/3	9/60	⅔/60

Numeral Conversion

Multiple number systems with some shared signs:



Conversion returns multiple readings in case of ambiguity.

Also supports subtraction with *la2* and fractions with *igi-...-gal2*.

Commodity Identification

Often straightforward:

1(gesz2) 5(disz) **udu** = 65 **sheep**

Harder cases:

2(ban2@c) **dumu-nita** = 20 L **male children**

Commodity Identification

Often straightforward:

1(gesz2) 5(disz) **udu** = 65 **sheep**

Harder cases:

[sze-bi] 2(ban2@c) **dumu-nita** = 20 L **barley** for the **male children**

Commodity Identification

Often straightforward:

1(gesz2) 5(disz) **udu** = 65 **sheep**

Harder cases:

[sze-bi] 2(ban2@c) **dumu-nita** = 20 L **barley** for the **male children**

4(asz@c) la2 1(barig@c) **szum2 sikil gal-gal** gur saggal = 1140 L **pure large garlic**

Commodity Identification

Rule-based:

- no training data for a machine learning model
- limited linguistic resources

Commodity Identification

Word Order: adjective follows noun; a word immediately after a numeral is usually a counted object

mu us2-sa si-mu-ur4-ru{ki} lu-lu-bu-um a-ra2 **1(u) la2 1(disz)-kam-asz ba-hul**

“The year after Simurru and Lulubu were **destroyed** for the **ninth** time”

Part-of-Speech Tags: to identify nouns; projected from English translations scraped off ePSD

Commodity Identification

Determinatives: identify likely commodities containing {ku6}, {gi}, {gisz}, etc.

Wordnet hypernyms: classify synsets as commodity-like or not: “food”, “clothing”, “metal” vs. “person”, “geographic_region”

Jointly classify all words based on a feature vector derived from these rules.

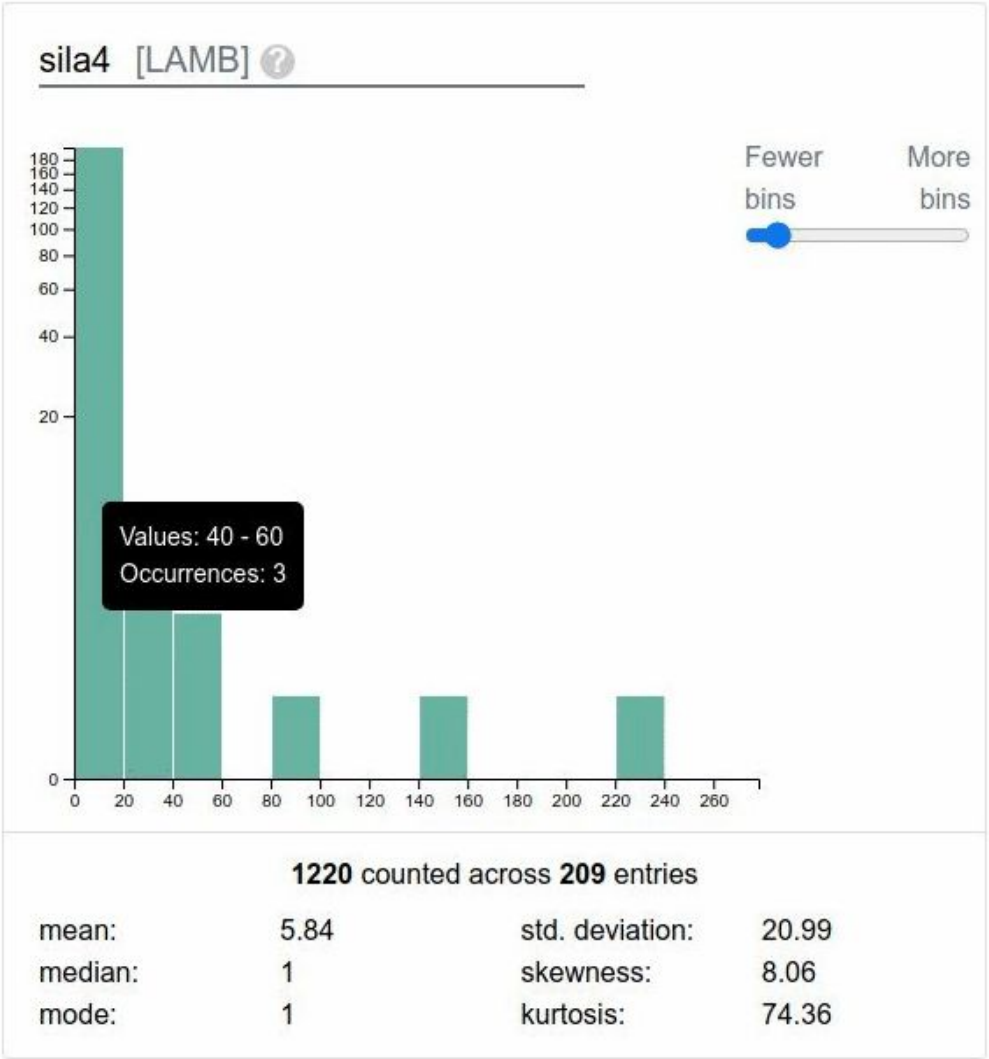
Visualizations

Demonstrate possible uses for the information extracted in previous sections.

Included in CDLI framework.

Histogram and Summary Statistics

- Overview of item's distribution
- Identify unimodal vs. multimodal distributions

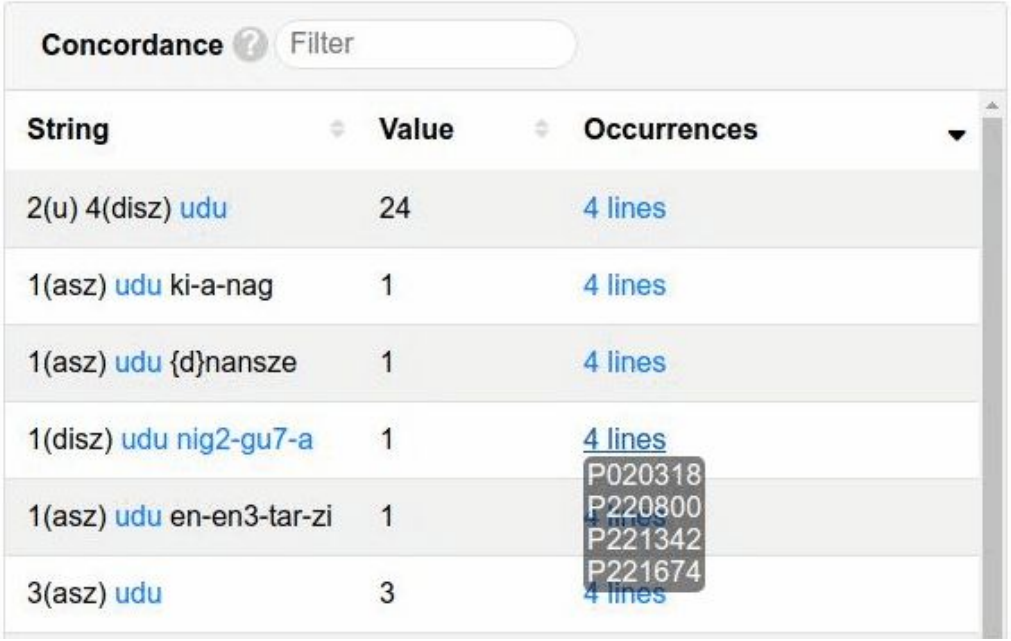


Similar Items

- Do other items have similar distributions?

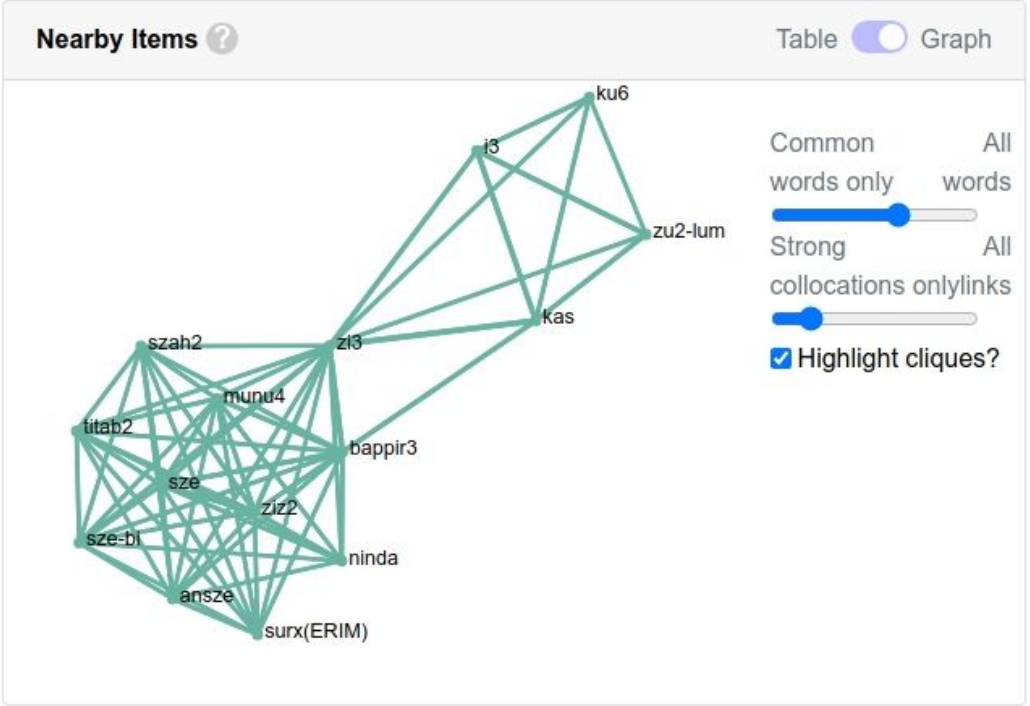
Concordance

- Item in context
- Most common values
- High- and low-value contexts



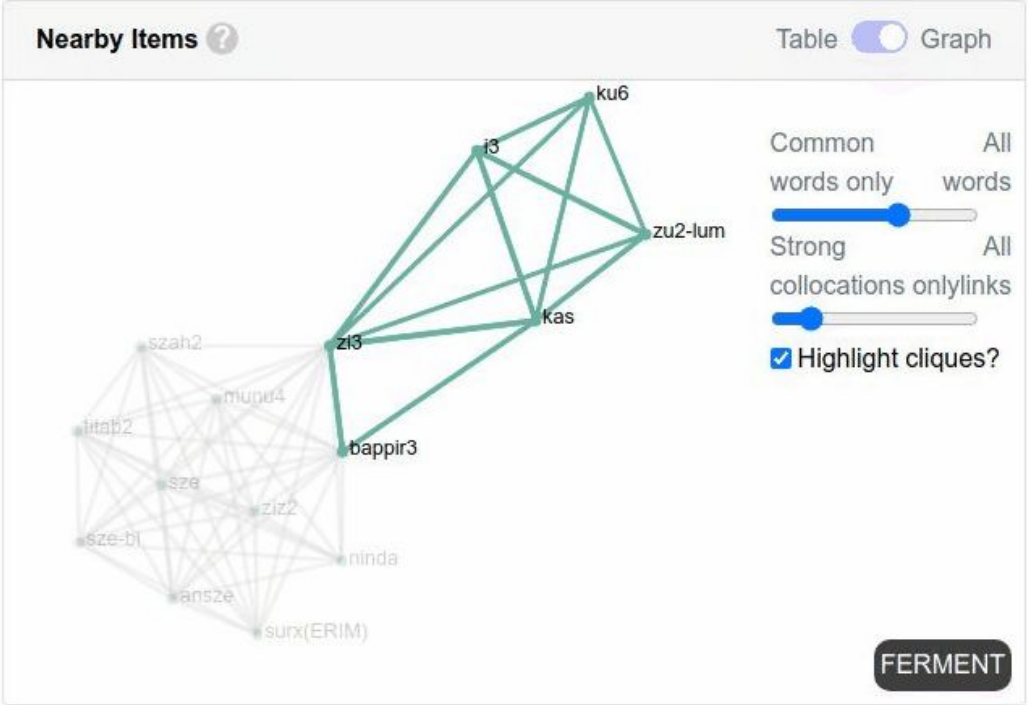
Nearby Items

- Identify administrative subgenres



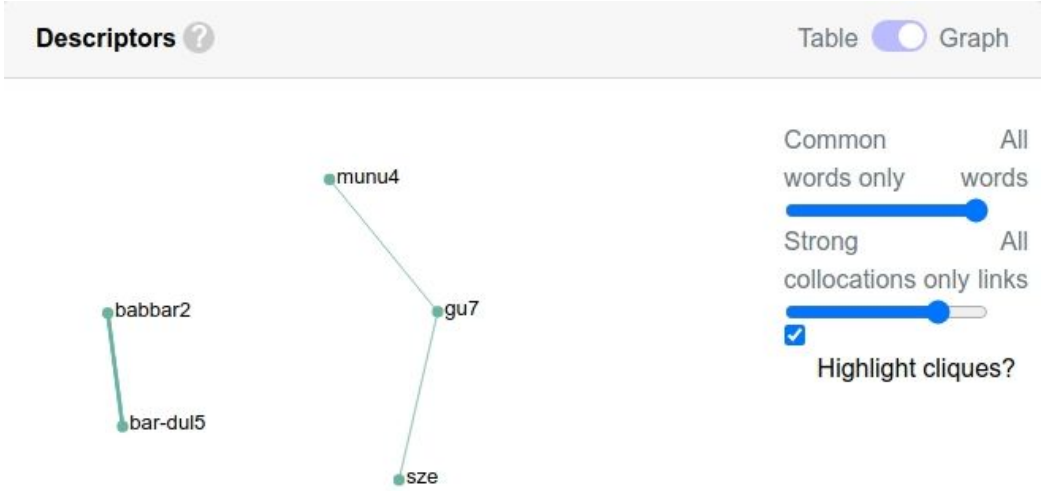
Nearby Items

- Identify administrative subgenres



Descriptors

- Identify use cases for a counted object



Future Work

Obtain more accurate dictionary for better commodity identification.

Use labeled data from current system to train a more versatile model.

Conclusion

Convert Sumerian numerals to Arabic notation

Extract information about counted objects

Visualize extracted information to highlight potential uses

Thank You!

With special thanks to the CDLI and my mentor Max Ionov.

Code is available at
https://github.com/cdli-gh/cdli-accounting-viz

Questions? loborn@sfu.ca