

UTOPIA: Automatic Pivot Table Assistant

Whanhee Cho, Anna Fariha




Messy spreadsheet data causes vain pivot tables

Movie Title	Genre	Gross
The Shawshank Redemption	prison drama	20 M
The Dark Knight	superhero action, crime, epic drama	500 M
The Matrix	action, epic sci-fi	200 M

Traditional pivot table tools struggle with messy data

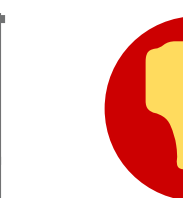
- Messy spreadsheet data is common [1]
- Data parsing requires additional expertise
- Resolving synonymous variants requires manual effort
- Users don't prefer switching environments

Which movie genre achieves the most gross? Let's find out this through a Microsoft Excel pivot table with Genre and Gross.



Genre	Sum of Gross
prison drama	20 M
epic drama, superhero action, crime	500 M
action, epic sci-fi	200 M

Multiple Values




Genre	Sum of Gross
prison drama	20 M
epic drama	500 M
superhero action	500 M
...	...

Synonymous Variants

Who wants to know the total gross of epic drama, superhero action, and crime genre movies?
Multiple values in the Genre cause this problem.

I split the data with Microsoft Power Query. But, some values share **the same semantic meaning**.



Genre	Sum of Gross
action	700 M
crime	500 M
drama	520 M
sci-fi	200 M

I finally got the desired result, but I had to find and replace variants about **240 times**. It took me **3 hours**!

UTOPIA provides ideal pivot tables without data preprocessing

Upload Data

UTOPIA

Data Overview

Title	Year	Genre	Gross	Director	Casts
The Shawshank Redemption	1994	prison drama	28 M	Frank Darabont	Tim Robbins, Mo ...
The Dark Knight	2008	superhero action, crime, epic drama	535 M	Christopher Nolan	Christian Bale, H ...
The Matrix	1997	action, cyberpunk sci-fi	171 M	Lana Wachowski	Lilly Wachowski, ...

Select Attributes

- ☐ Title
- ☒ Year
- ☒ Genre
- ☒ Gross
- ☐ Director
- ☐ Casts
- ☐ Certificate
- ☐ Runtime

Pivot Table Fields

Column: Year

Row: Genre

Value and Aggregate: Gross

SUM AVG MIN MAX CNT

Pivot Table

Sum of Gross	Year		
Genre	2010	2011	2012
crime	300 M	138 M	41 M
action	558 M	164 M	1416 M
action	8 M	4 M	304 M
superhero action	251 M	0 M	1071 M
team action	293 M	146 M	0 M
epic action	6 M	14 M	41 M
drama	621 M	886 M	673 M
drama	231 M	764 M	393 M
epic drama	82 M	1 M	4 M
prison drama			

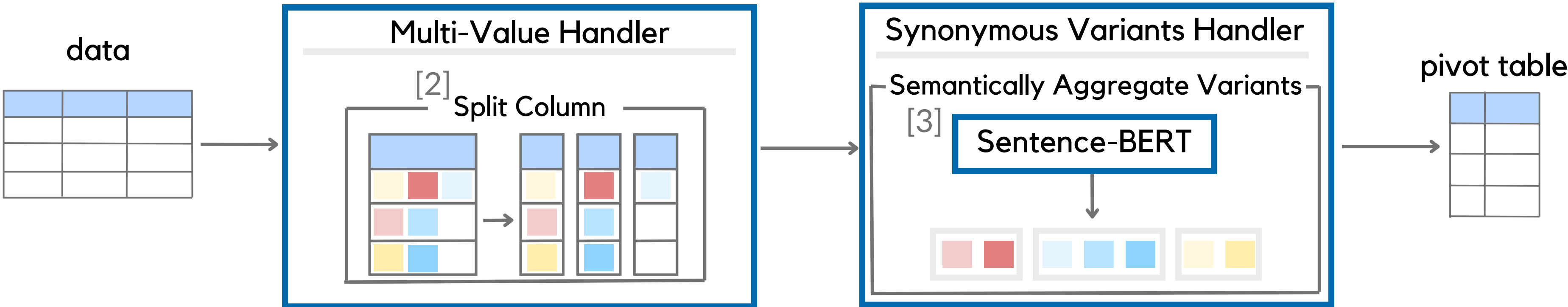
Synonymous variants are grouped

UTOPIA **semantically aggregates** synonymous variants. Expand or collapse rows to view groups.

Each cell contains a single value

UTOPIA **parses** multi-valued attributes even in the **absence of delimiters**.

UTOPIA uses SOTA parsing technology and embeddings



[1] P Li, Y He, C Yan, Y Wang, and S Chaudhuri. 2023. Auto-Tables: Synthesizing Multi-Step Transformations to Relationalize Tables without Using Examples. PVLDB

[2] M Raza and S Gulwani. 2017. Automated Data Extraction Using Predictive Program Synthesis. AAAI

[3] N Reimers and I Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP