# Phase 1: Data Selection and Data Profiling

# Agenda

# Entity Matching – General

# Data Selection & Language Detection

## DATA SELECTION

PRODUCT

- Biggest entity type
- Cluster IDs given in another corpus

LOCALBUSINESS, HOTEL & RESTAURANT

- LB 3rd biggest entity type
- Phone number and geo location as identifiers

## APPROACH

Step 1

Cleaned the tables with TLD-based approach

- Use **.com, *.net, *.org, *.uk*

Step 2

Used fastText Language Detection on every row

## STATISTICS *

PRODUCT

| | *Before* | *After* |
|---|---|---|
| Top100 | 100 | 75 |
| Min3 | $\sim 1.66\ M$ | $\sim 435\ K$ |

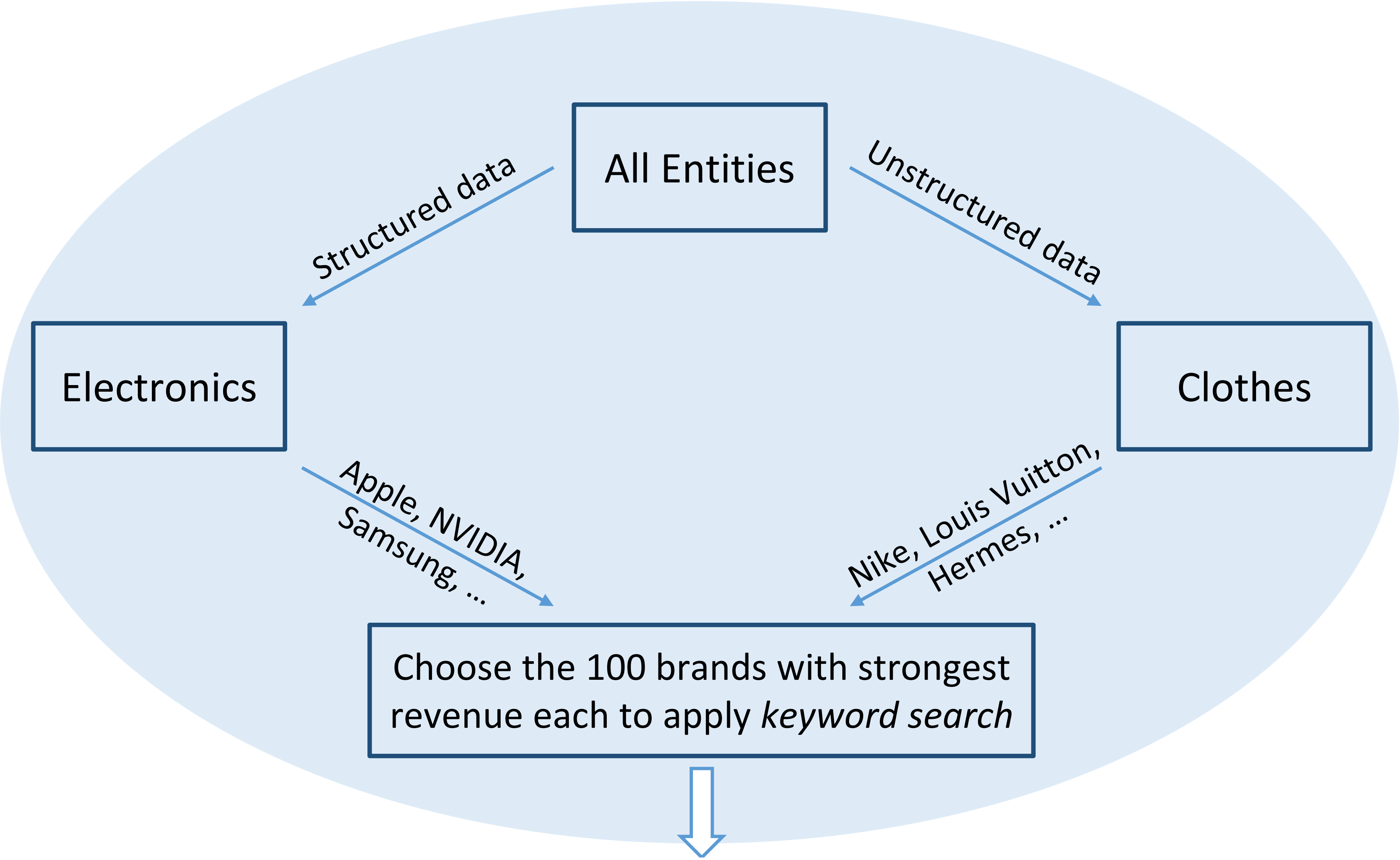
LOCALBUSINESS, HOTEL & RESTAURANT

| | *Before* | *After* |
|---|---|---|
| LB Top100 | 100 | 53 |
| LB Min3 | $\sim 50.5\ K$ | $\sim 11.7\ K$ |
| Hotel Top100 | 100 | 52 |
| Hotel Min3 | $\sim 13\ K$ | $\sim 1\ K$ |
| Restaurant Top100 | 100 | 64 |
| Restaurant Min3 | $\sim 6.4\ K$ | $\sim 1\ K$ |

* These statistics are referring to Step 1; statistics for Step 2 can be found in the Appendix (Appendix 1)

# Entity Matching – Product

# General Approach

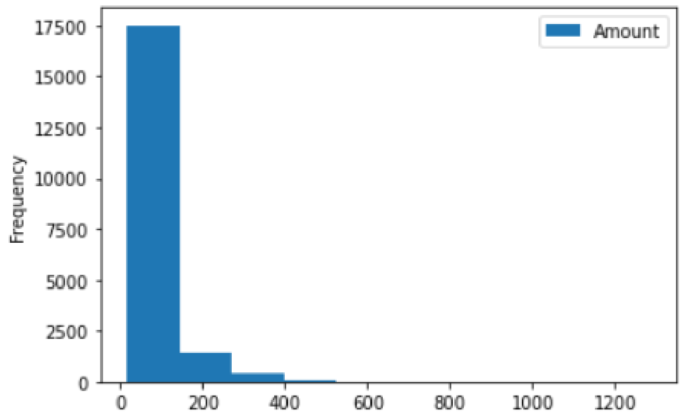



Either use „brand“ column or first 3 words in „name“ column

# Distribution of Tables per Product Cluster

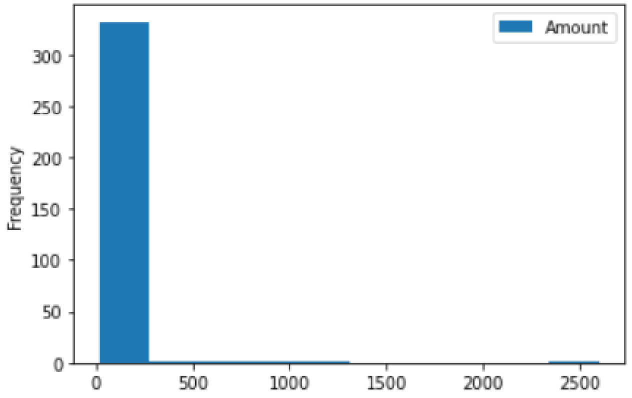
More than 15 tables per cluster

All



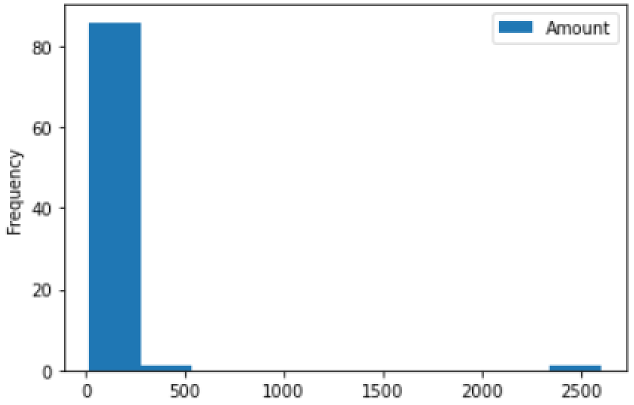

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Amount** | 19367.0 | 59.0 | 69.0 | 16.0 | 22.0 | 33.0 | 63.0 | 1285.0 |

Electronics




| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Amount** | 416.0 | 48.0 | 148.0 | 16.0 | 18.0 | 25.0 | 50.0 | 2600.0 |

Clothes

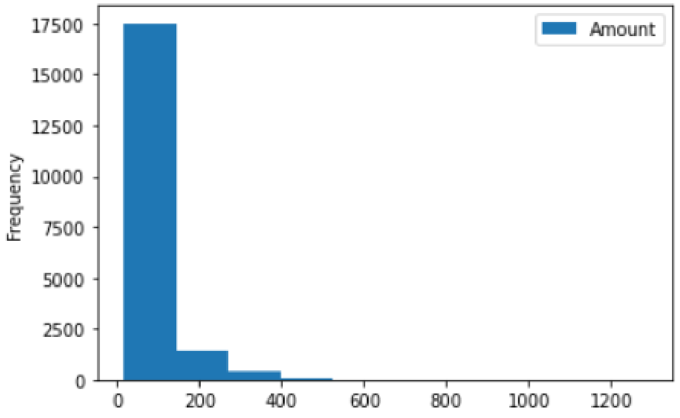



| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Amount** | 88.0 | 62.0 | 276.0 | 16.0 | 17.0 | 18.0 | 28.0 | 2600.0 |

# Distribution of Tables per Product Cluster

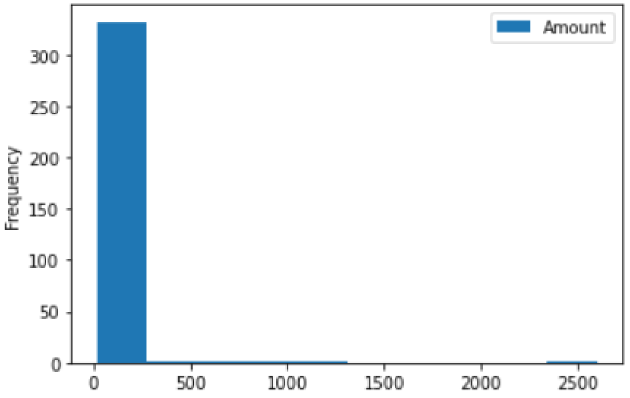
More than 15 tables per cluster

## All



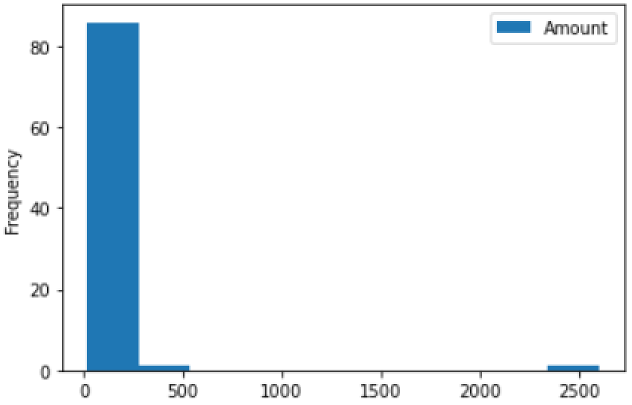

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Amount** | 19367.0 | 59.0 | 69.0 | 16.0 | 22.0 | 33.0 | 63.0 | 1285.0 |

## Electronics




| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Amount** | 416.0 | 48.0 | 148.0 | 16.0 | 18.0 | 25.0 | 50.0 | 2600.0 |

## Clothes




| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Amount** | 88.0 | 62.0 | 276.0 | 16.0 | 17.0 | 18.0 | 28.0 | 2600.0 |

More than 10 tables per cluster:

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Amount** | 829.0 | 31.0 | 107.0 | 11.0 | 12.0 | 16.0 | 25.0 | 2600.0 |

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Amount** | 268.0 | 29.0 | 159.0 | 11.0 | 12.0 | 13.0 | 17.0 | 2600.0 |

# Suitable Tables for Entities

APPROACH

- Combine information of cluster_ids with brands
- Get information about top clusters that have brands associated
- Get baseline entities by looking into each cluster
- Compute top 15 nearest entities for one baseline by Doc2Vec and discard duplicates

| cluster_id | name |
|---|---|
| 16617 | nanobeam-ac-gen2 |
| 16617 | ubiquiti nanobeam ac gen 2 5ghz 19dbi radio and antenna |
| 16617 | ubiquiti cpe 5 ghz nanobeam ac, gen2 |
| 16617 | ubiquiti networks nbe-5ac-gen2 5ghz nanobeam ac gen2 19dbi row |
| 16617 | ubiquiti access point 5 ghz nanobeam ac, gen2 |

| | |
|---|---|
| 18640 | magnet - women belong in all places where decisions are being made. - ruth bader ginsburg |
| 18640 | sony zeiss 55mm f/1.8 prime e mount lens |
| 18640 | lente sony sonnar t* fe 55mm f/1.8 za |
| 18640 | i'm not bossy, i'm the boss - crew socks |
| 18640 | blue q blue q damessokken 'i'm not bossy, i'm the boss' |
| 18640 | i'm not bossy. i'm the boss. socks |
| 18640 | sony sony alpha sonnar t fe 55mm f1.8 za - e-mount (full frame - for a7/a7r) |

# Possible Entities

| | | | |
|---|---|---|---|
| 652 | 99153 | asus geforce gt 1030 phoenix overclocked single fan 2gb gddr5 pcie 30 graphics card | 29 |
| 45841 | 2751278 | asus nvidia geforce gt710 2gb gddr5 graphics card | 28 |
| 40392 | 285492 | geforce gt 1030 2048mb gddr5 pciexpress graphics card 90yv0at0m0na00 | 14 |
| 25477 | 1021594 | 882277 asus geforce gtx 1650 dual oc 4gb gddr5 graphics card dual gtx1650 o4g | 11 |

| | | | |
|---|---|---|---|
| 1081 | 1524820 | sony a7 iii full frame mirrorless interchangeable lens camera optical with 3 inch lcd black ilce7m3 b | 95 |
| 26842 | 3078421 | alpha a6100 mirrorless camera with 16 50mm lens black | 14 |
| 4022 | 1961922 | sony hvl f45rm compact radio controlled gn 45 camera flash with 1 display black | 25 |
| 51060 | 2473758 | sony alpha a7s iii mirrorless digital camera body_p_7624 | 9 |

| | | |
|---|---|---|
| 939889 | dolce gabbana pour homme eau de toilette 125 ml 519 | 11 |
| 944406 | dolce gabbana pour homme eau de toilette | 18 |
| 1225336 | 14439 dolce gabbana pour homme eau de toilette vaporizador 3423473020776 | 7 |
| 1006344 | 339 dolce  gabbana femme eau de parfum spray 50 ml | 10 |

| | | |
|---|---|---|
| 2732926 | mens shoes nike air force 1 07 black 315122001 55827.aspx | 14 |
| 58083280 | mens shoes nike air force 1 07 white 315122111 21544.aspx | 24 |
| 1424369 | mens shoes nike air force 1 high 07 lv8 wb flax 882096200 165302.aspx | 8 |
| 58591075 | mens shoes nike sportswear air force 1 mid 07 triple black 315123001 74438.aspx | 11 |
| 58592784 | mens shoes nike sportswear air force 1 mid 07 white 315123111 68768.aspx | 19 |

# Entity Matching – LocalBusiness

# Problem and Matching Strategy

PHONE NUMBERS

- Branches share headquarter phone number
- *Same* numbers are *differently* formatted

GEO-LOCATION

- Locations might refer to headquarter
- *Different* entities share the *same* building

SOLUTION

- Use the combination of both identifiers
  - Find matching entities across phone numbers
  - Apply geo-matching on only this subset
- *Assumption:* entries with the same phone number *AND* the same geo-location are matches and will form a cluster

# Preprocessing Phonenumbers

- Use ***phonenumbers*** library
  - Goal: normalize phone numbers to *E.164 Format*

Introduces new complexity: country codes

- Library only accepts *ISO-alpha 2* country codes like
  - "DE" , "US" ,  "IN"
- Solution: ***pycountry*** library and manual cleansing
  - Saving majority but not all of our data

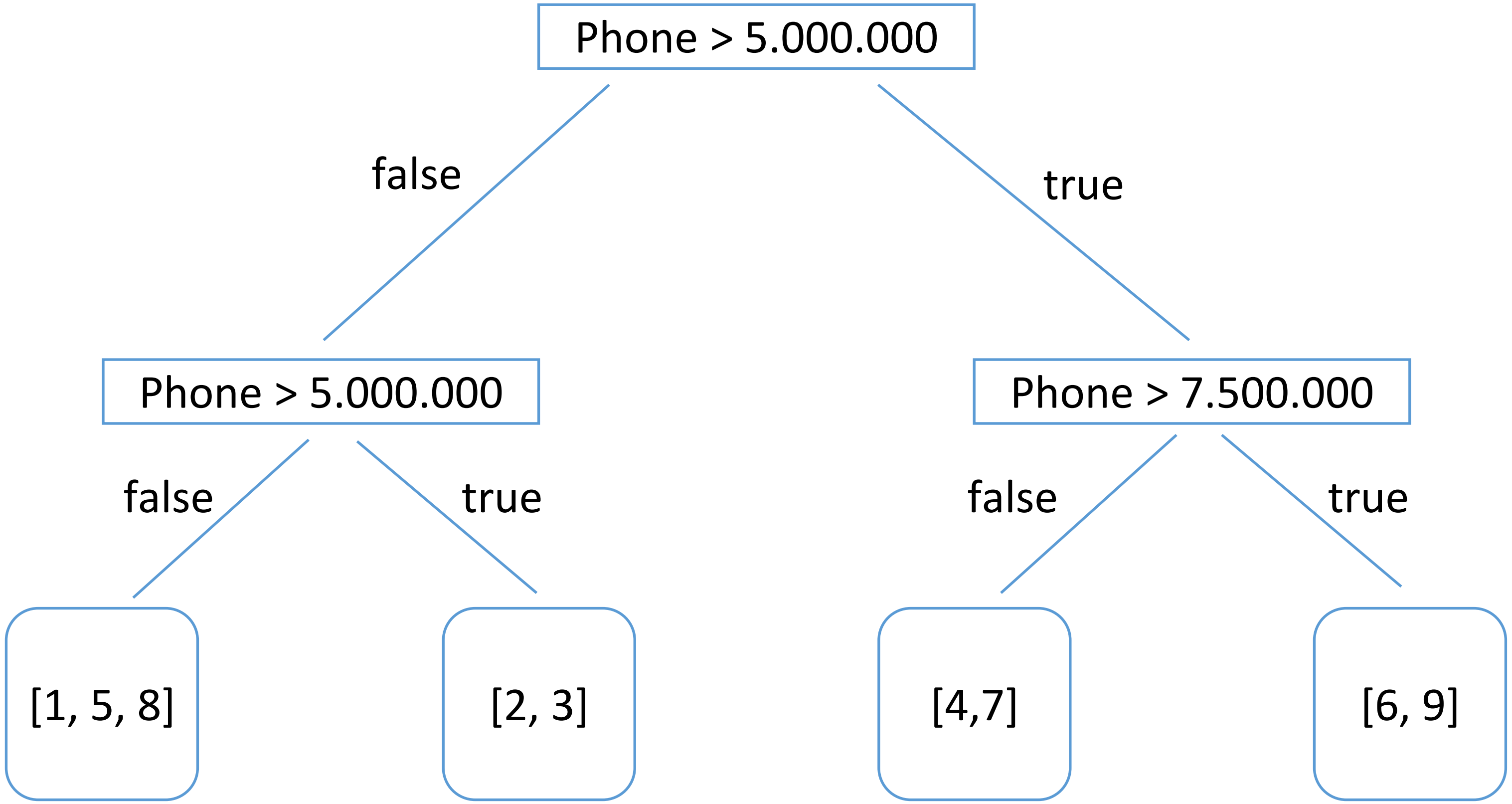
- Apply ***phonenumbers*** library to dataset
  - Parse *number* and *country code* into *phone-object*
  - Only keep *valid phonenumbers*
  - Convert every phone number to E.164 format

"0044 2083661177"
"020 8366 1177"
"+442083661177"

"USA"
"United States"
"德國"
"DEU"
"آلمان"
"de"
"DEUTSCHLAND"

| phone_object | E.164 format |
|---|---|
| Country Code: 1 National Number: 3034443535 | +13034443535 |

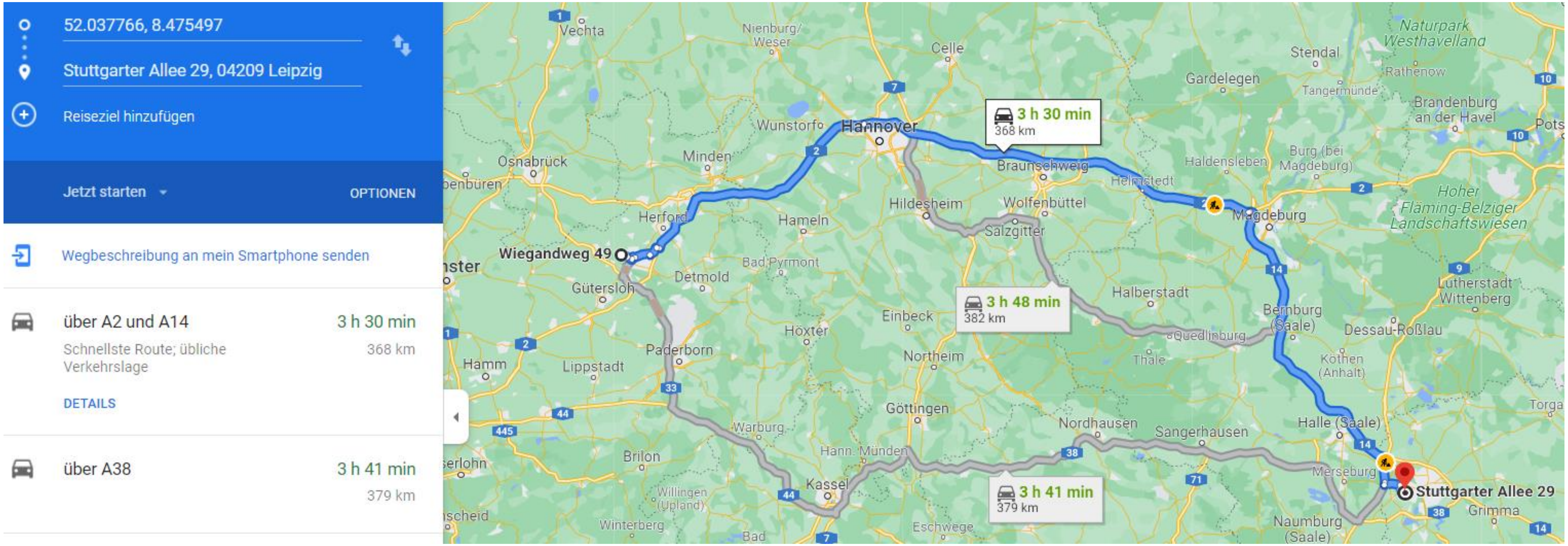
# Matching Phonenumbers



# Scales of Latitude and Longitude

Difference between: Lat: 52.037766, Lon: 8.475497

Lat: 51.322514, Lon:12.287662



# Preprocessing Geo-Locations

- Preprocessing:
  - Split Geo-Location in two columns
  - Remove entries that cannot be converterted to float
  - Definition Latitude: Between –90 and 90
  - Definition Longitude: Between –180 and 80

- Using GeoPy
  - Calculate distance between two tupels

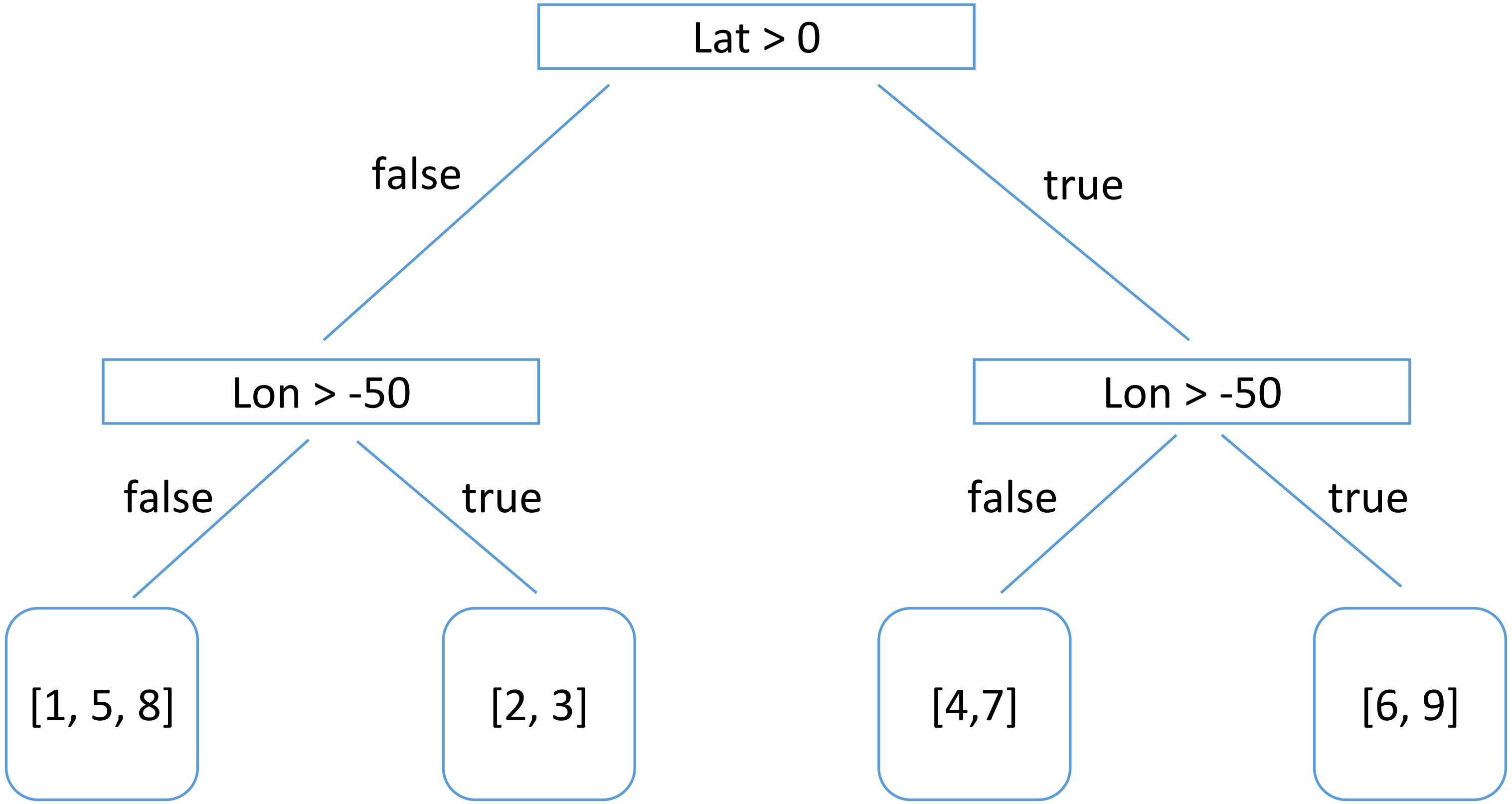
```
{lat:  52.037766, lon: 8.475497}

{lat:'51.322514', lon:
12.287662}
```

| latitude | longitude |
|---|---|
| 50.001 | 8.270224 |

| MatchingGeoPoints | Difference |
|---|---|
| [21907, 32] | [0.004145070470485397, -1] |

# Matching Geo-Locations



# Entity Matching – Example Result

| MatchingGeoPoints | Difference |
|---|---|
| [21907, 32] | [0.004145070470485397, -1] |

```
data.loc[data['indexValue'] == 32][['name', 'address', 'page_url', 'E.164 format', lat, lon]]
```

| | name | address | page_url | E.164 format | latitude | longitude |
|---|---|---|---|---|---|---|
| **32** | Tyson's Tacos | {'addresslocality': 'Austin', 'postalcode': '7... | https://www.cookingchanneltv.com/restaurant-gu... | +15124513326 | 30.30964 | -97.715239 |

```
data.loc[data['indexValue'] == 21907][['name', 'address', 'page_url', 'E.164 format', lat, lon]]
```

| | name | address | page_url | E.164 format | latitude | longitude |
|---|---|---|---|---|---|---|
| **21907** | Tyson's Tacos | {'addresslocality': 'Austin', 'postalcode': '7... | https://theinfatuation.com/austin/reviews/tyso... | +15124513326 | 30.309656 | -97.7152 |

# Entity Matching – Final Datasets

I. Considered entities and tables
- Local Business
- Restaurant
- Hotel

II. Concatenate top100 + minimum3 over each entity
- Control for potential overlap
- Only keep entries in line with matching strategy

III. Apply combined matching strategy
- Get final matching files

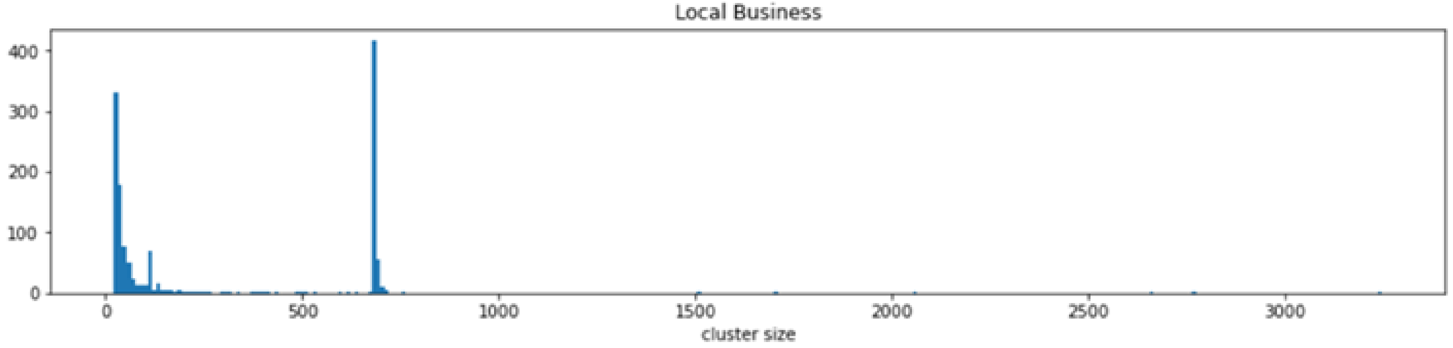
STATISTICS

| Entity | Before | After |
|---|---|---|
| Local Business | ~ 3.6 Mio | ~ 470.000 |
| Restaurant | ~ 408.000 | ~ 130.000 |
| Hotel | ~ 525.000 | ~ 68.000 |

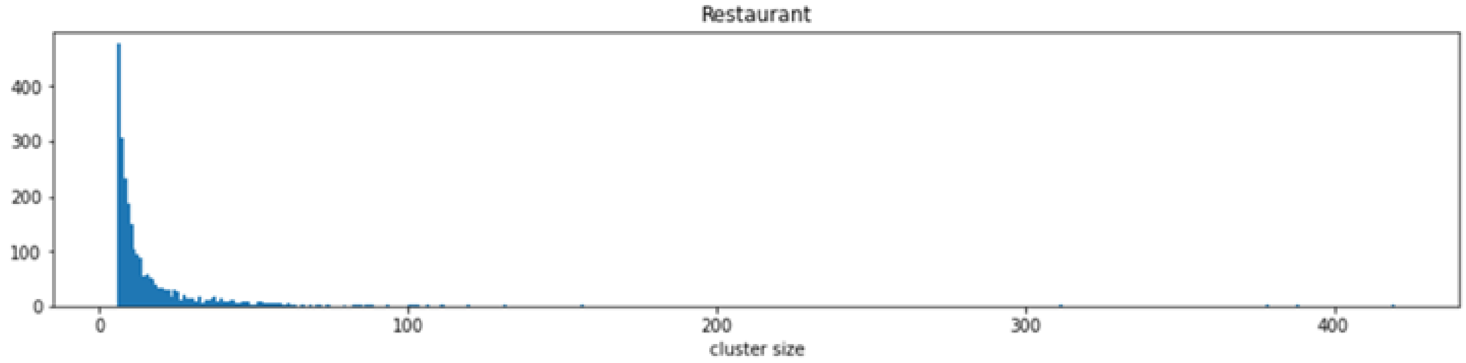
| Entity | Length Matching File | Remaining Tables |
|---|---|---|
| Local Business | ~ 108.000 | 1.123 |
| Restaurant | ~ 42.000 | 459 |
| Hotel | ~ 28.000 | 189 |

# Entity Matching – Final Datasets

- Histogram of Local Business (*20 < cluster size < max*)



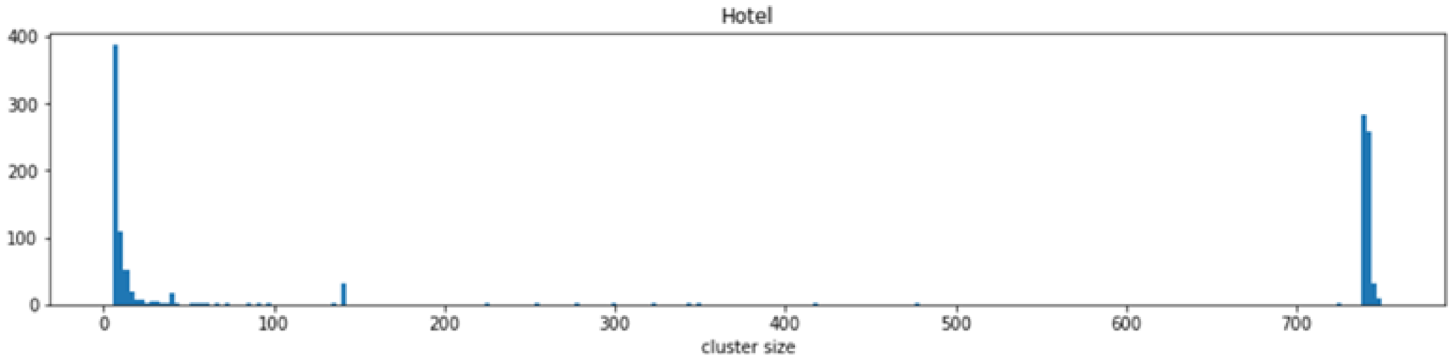

- Histogram of Restaurant (*5 < cluster size < max*)



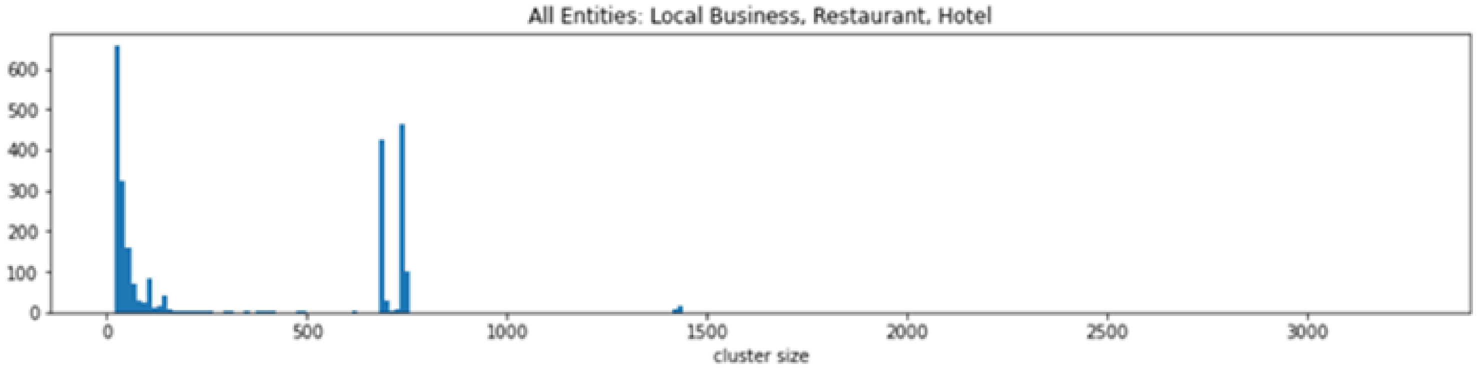

| | | Local Business | Restaurant |
|---|---|---|---|
| | **count** | 107567 | 42236 |
| **clusters statistics** | **# clusters** | 25283 | 10919 |
| | **mean** | 19.76 | 5.91 |
| | **std** | 129.69 | 22.47 |
| | **min** | 2 | 2 |
| | **0.25** | 2 | 2 |
| | **0.5** | 2 | 3 |
| | **0.75** | 4 | 5 |
| | **max** | 12704 | 2032 |
| | **second largest** | 3245 | 420 |

# Entity Matching – Final Datasets

- Histogram of Hotel ( *5 < cluster size < second largest*)




- Histogram of all Entity Types ( *20* < cluster size < *second largest)*




Overlapping does exist!

| | | Local Business | Restaurant | Hotel | All |
|---|---|---|---|---|---|
| | count | 107567 | 42236 | 28879 | 178682 |
| clusters statistics | # clusters | 25283 | 10919 | 6993 | 42606 |
| | mean | 19.76 | 5.91 | 67.53 | 24.33 |
| | std | 129.69 | 22.47 | 215.55 | 136.65 |
| | min | 2 | 2 | 2 | 2 |
| | 0.25 | 2 | 2 | 2 | 2 |
| | 0.5 | 2 | 3 | 3 | 2 |
| | 0.75 | 4 | 5 | 4 | 5 |
| | max | 12704 | 2032 | 5183 | 12704 |
| | second largest | 3245 | 420 | 2477 | 5183 |

Next to Do:

- For some clusters, elements mainly are from a single table

⇨ track cluster elements to original tables and build histograms on matches across tables

# Schema Matching

IV

# Task at a glance (Phase 1)

*Data Understanding*

Become familiar with the structure of data

- Tables
- Statistics files
- Columns
- Technical characteristics

*Data Preparation*

Data processing and profiling

- Statistical gathering of the chosen tables after language detection to remove non-English tables

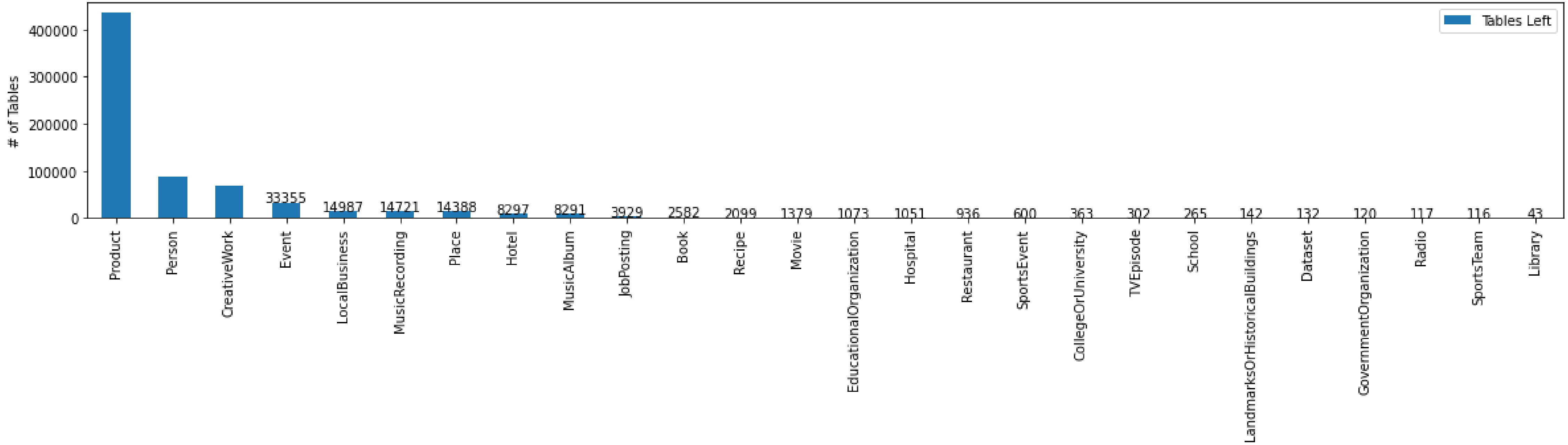
*Selection of Columns/Tables*

Find 200 column type labels

- Analyzed statistics & occurrence of columns for each class
- Identified 200 most common columns with at least 100 respective tables
- Identified at least 3 similar columns within the 200 column type labels

# Distribution of Tables across Classes

Distribution of remaining tables after English language detection



# Selection of Tables



**Data Used:**

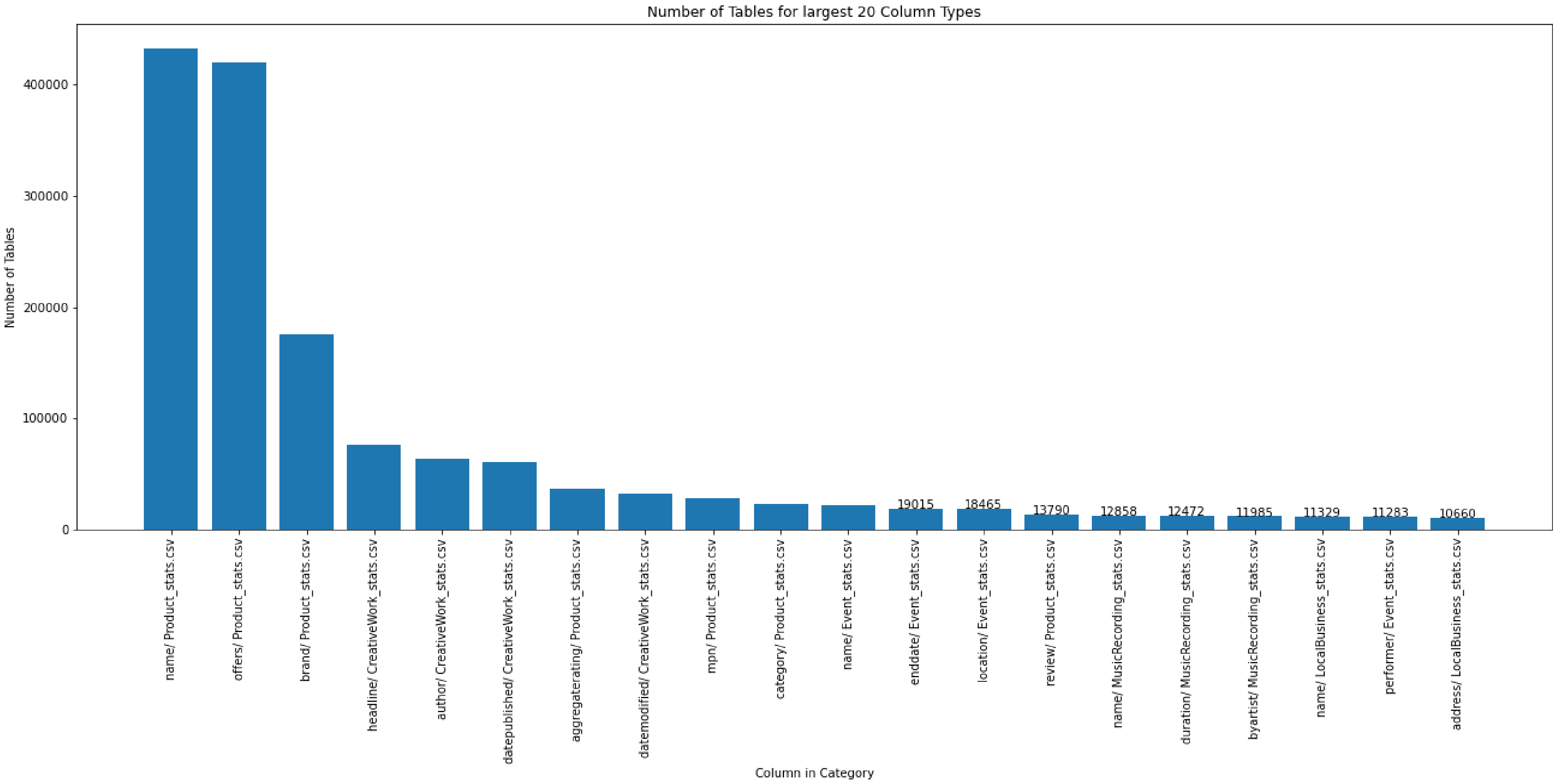
- 19 classes used
- filtered on English tables only
- Top100 & Minimum3 files

**Process:**

- Counter of columns of each class – how many times is a column (entity) in a class?
- Division by size of class – how well is the column (entity) represented in the class?
- Removal of not interesting columns (entities) such as url, images, row_id etc.
- Filtering most common columns by absolute count of tables -> get top 210
- Find overall categories manually to cluster columns (eg. person_name, rating etc.)
- Categorization of similar column types

# Statistics of selected tables

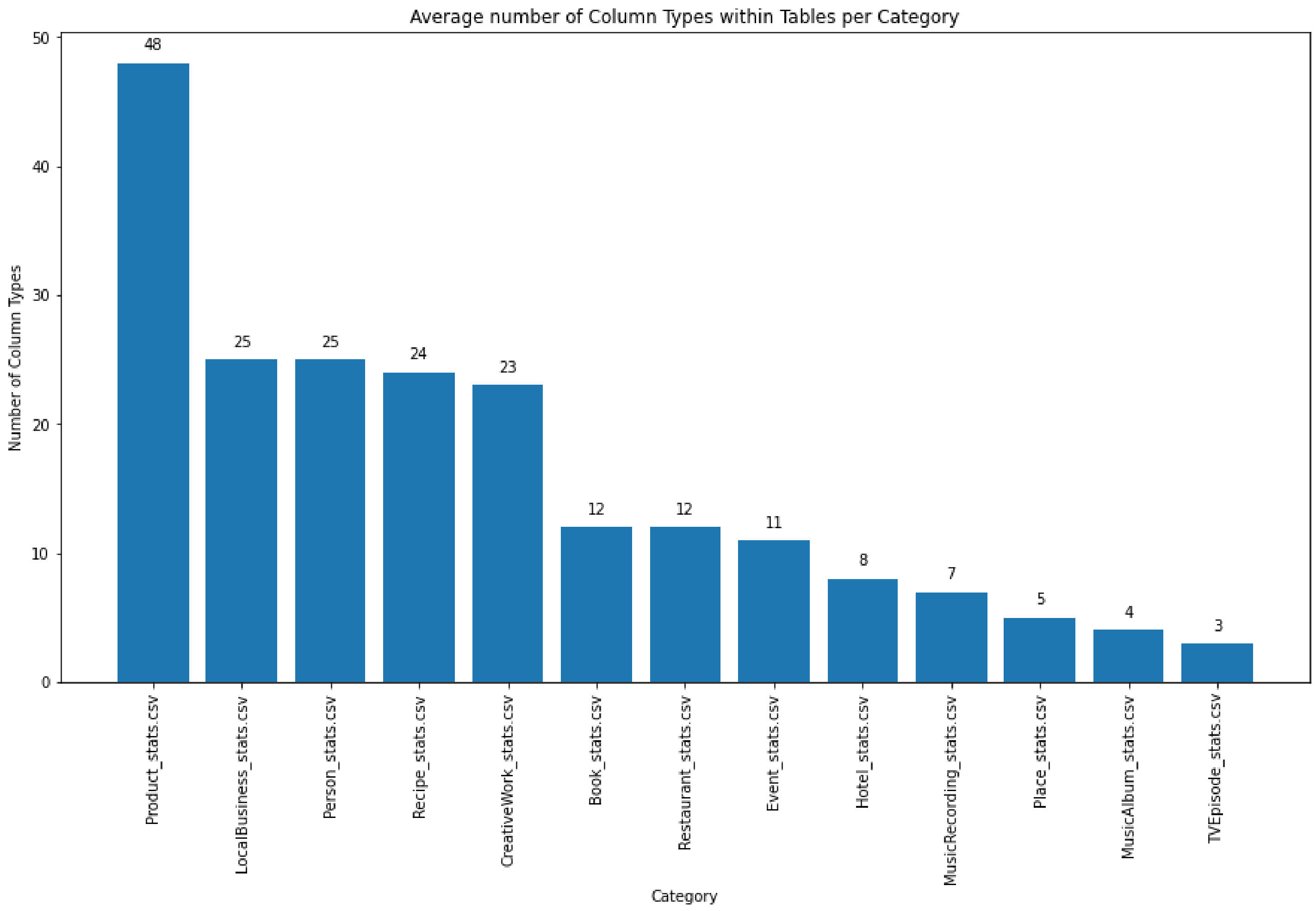
**Tables per entity ***




* Full overview in repository

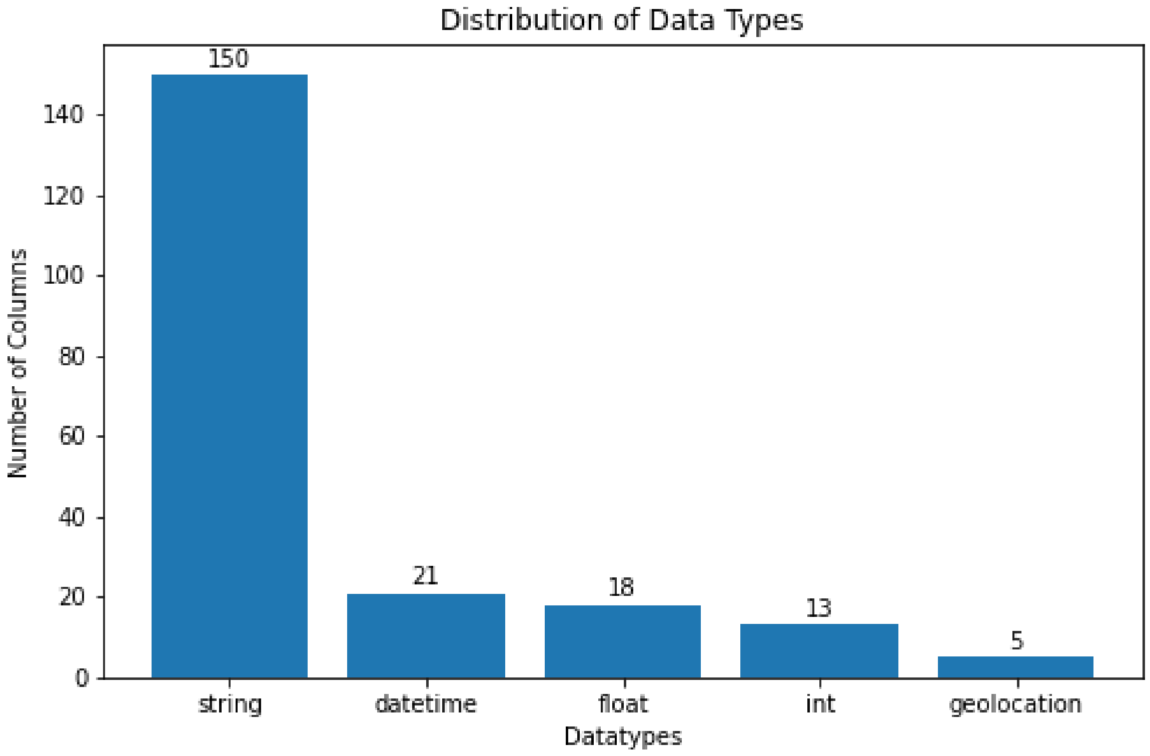
# Statistics of selected tables

**Number of entities used from each class**



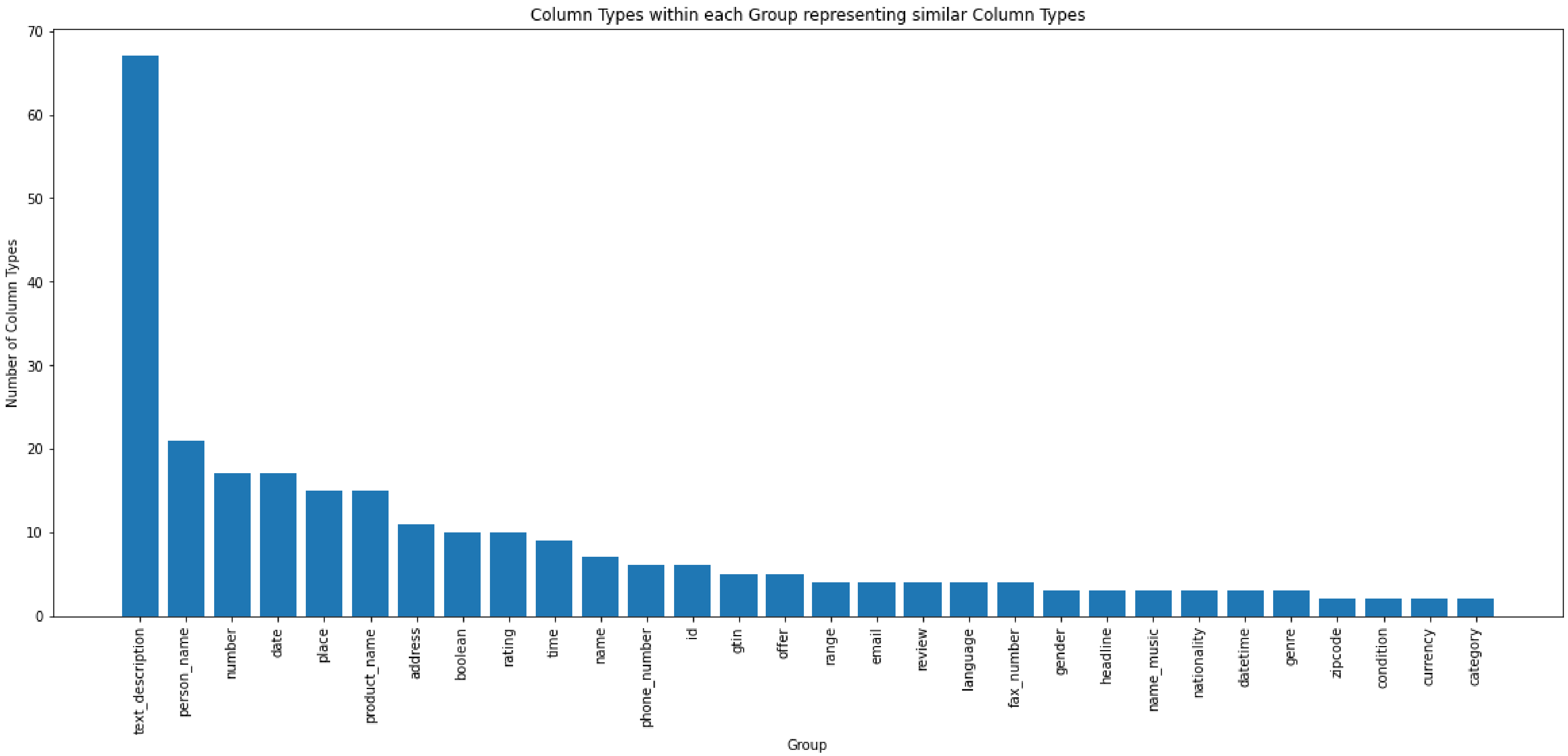
# Statistics of selected tables

**Data types of chosen tables**



# Statistics of selected tables

## Categorized Columns representing similar Column Types



# Deep dive: Selection of Groups (min 3 columns per group)

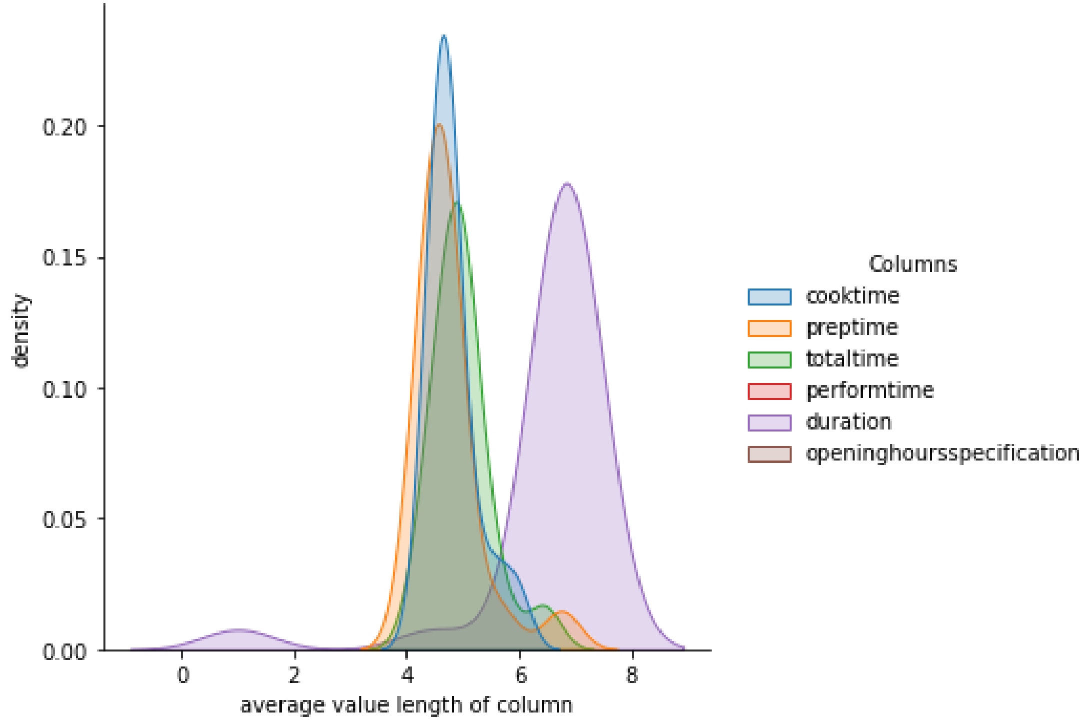
**Length similarity:**

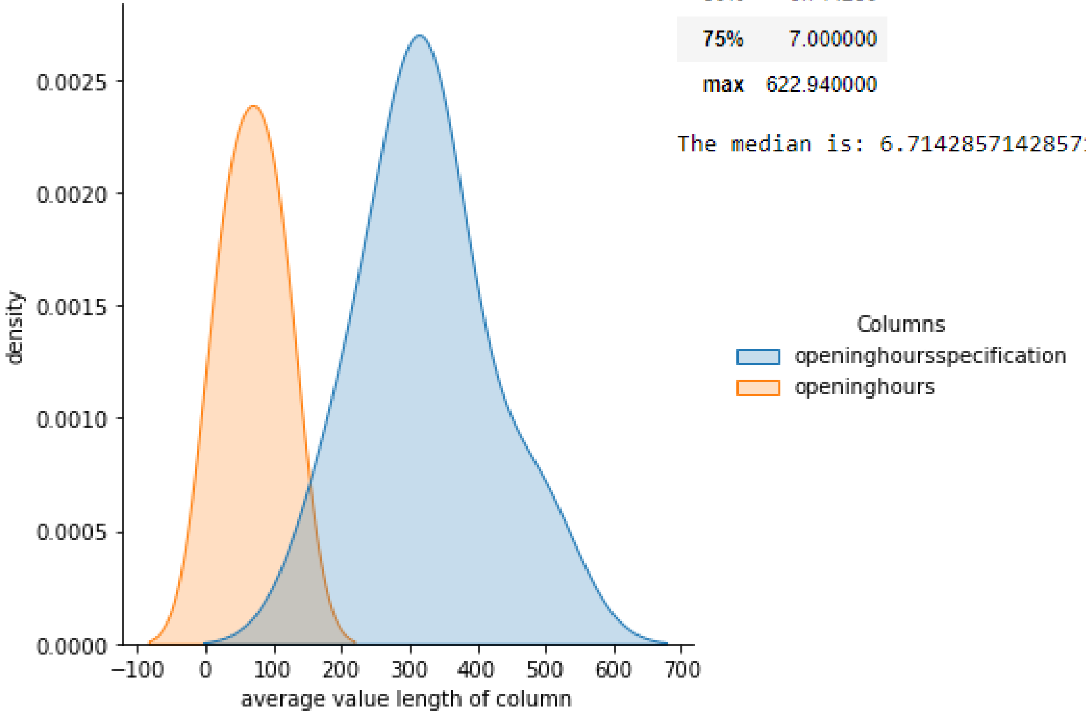
- Pick an overall category (such as “time” which includes 8 columns)
- Get a few tables with columns from the selected category
- Compute the mean length of the values in one column of one table
- Compare the mean length of the column and class with the other columns and classes

**Cosine similarity:**

- Pick an overall category (such as “rating” which includes 7 columns)
- Get a few columns from every column that is of interest
- Preprocess data to prepare for embeddings (cleaning etc.)
- Sentence embedding with pretrained fasttext model
- 1-1 comparison of cosine similarities (matrix)

# Length similarity measure with one category (Example “Time”)




| | |
|---|---|
| **count** | 749.000000 |
| **mean** | 40.870512 |
| **std** | 101.189523 |
| **min** | 1.000000 |
| **25%** | 4.797980 |
| **50%** | 6.714286 |
| **75%** | 7.000000 |
| **max** | 622.940000 |

```
The median is: 6.714285714285714
```

# Cosine Similarity within a Group (Example ”Rating”)

| | product_ratingvalue | product_aggregaterating | product_bestrating | product_worstrating | recipe_aggregaterating | localbusiness_aggregaterating | creativework_aggregaterating | hotel_aggregaterating | restaurant_aggregaterating | book_aggregaterating |
|---|---|---|---|---|---|---|---|---|---|---|
| product_ratingvalue | 1 | | | | | | | | | |
| product_aggregaterating | 0.358 | 1 | | | | | | | | |
| product_bestrating | 0.918 | 0.309 | 1 | | | | | | | |
| product_worstrating | 0.49 | 0.029 | 0.539 | 1 | | | | | | |
| recipe_aggregaterating | 0.314 | 0.665 | 0.329 | 0.1 | 1 | | | | | |
| localbusiness_aggregaterating | 0.329 | 0.692 | 0.341 | 0.1 | 0.998 | 1 | | | | |
| creativework_aggregaterating | 0.379 | 0.846 | 0.365 | 0.087 | 0.948 | 0.96 | 1 | | | |
| hotel_aggregaterating | 0.33 | 0.723 | 0.339 | 0.096 | 0.996 | 0.998 | 0.969 | 1 | | |
| restaurant_aggregaterating | 0.316 | 0.674 | 0.33 | 0.099 | 0.999 | 0.998 | 0.951 | 0.997 | 1 | |
| book_aggregaterating | 0.37 | 0.873 | 0.359 | 0.082 | 0.943 | 0.955 | 0.992 | 0.967 | 0.967 | 1 |

| < 0.4 | >0.8 |
|---|---|

# Next steps

Finalize groups

Train/Test sets + downsampling

Experiments

# Evaluation

V

# Evaluation

UNIVERSITÄT MANNHEIM

## TEAMWORK

☺

- High motivation
- Various backgrounds
- Communication within the team
- Communication with Ralph

🙄

- Communication with other subgroups

## TECHNICAL

😎

- Programming tasks
- Usage of server

😵

- Parallel Computing
- Server connection

# Appendix 1: Preprocessing with Language Detection – Visualization after Step 2