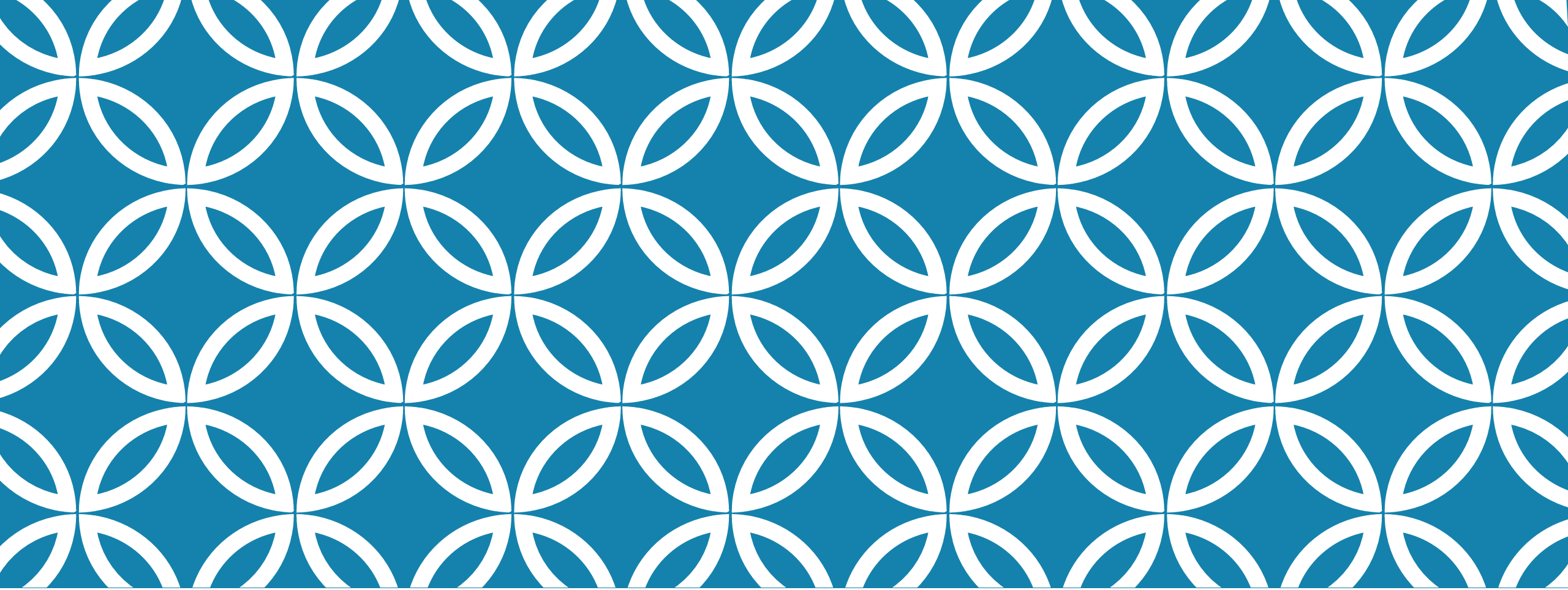


A DISTRICT RECOMMENDER

IBM Data Science Capstone Project

Georgios Lekkas

June 4, 2020

THE PROBLEM

A person moving to a new city for work or studies, must choose the areas where to look for accomodation

- “If I have no time or budget for on-site exploration, which is the right place for me?”
- City guides are for tourists
- Expat or student forums have little information and carry subjective opinions

The new “District Recommender” service comes to fill that gap

- Supports decisions with objective data
- Takes personal preferences into account

DATA SOURCES

1. Data on the districts of each city
   Web scraping from Wikipedia
2. Geolocation of each district
   Online information provider: ArcGIS by ESRI
3. Venues in each district
   Via the Foursquare API; categorized by type
4. Taxonomy of venues
   Via the Foursquare API
5. User input
   City of destination
   Preferred districts in other cities & their ratings

Initial testing areas

Three European cities

- London, England
- Glasgow, Scotland
- Copenhagen, Denmark

Example of user input

Destination: Copenhagen, Denmark

Preferences:

District code	District_name	City	Rating
E9	Hackney, Homerton	London	3.5
N1	Barnsbury, Canonbury, Islington	London	4.0
W11	Holland Park, Notting Hill	London	5.0

METHODOLOGY

The problem maps directly to a class of problems and solutions studied by Data Science, known as Recommender Systems

Two types of recommender systems:

- **Content-based** recommenders: the system tries to figure out the user's favourite aspects of an item, and then recommends items that present those aspects.
- **Collaborative filtering** recommenders: the system find other users that have similar preferences and opinions to the target user, and then recommends items that those others have liked

We decide to build a **content-based recommender**. We must

- Find out the user's favorite aspects from the districts and ratings given
- Use user preferences to recommend districts in the destination city

THE PROCESS

The acquired data is processed to produce a recommendation in five steps:

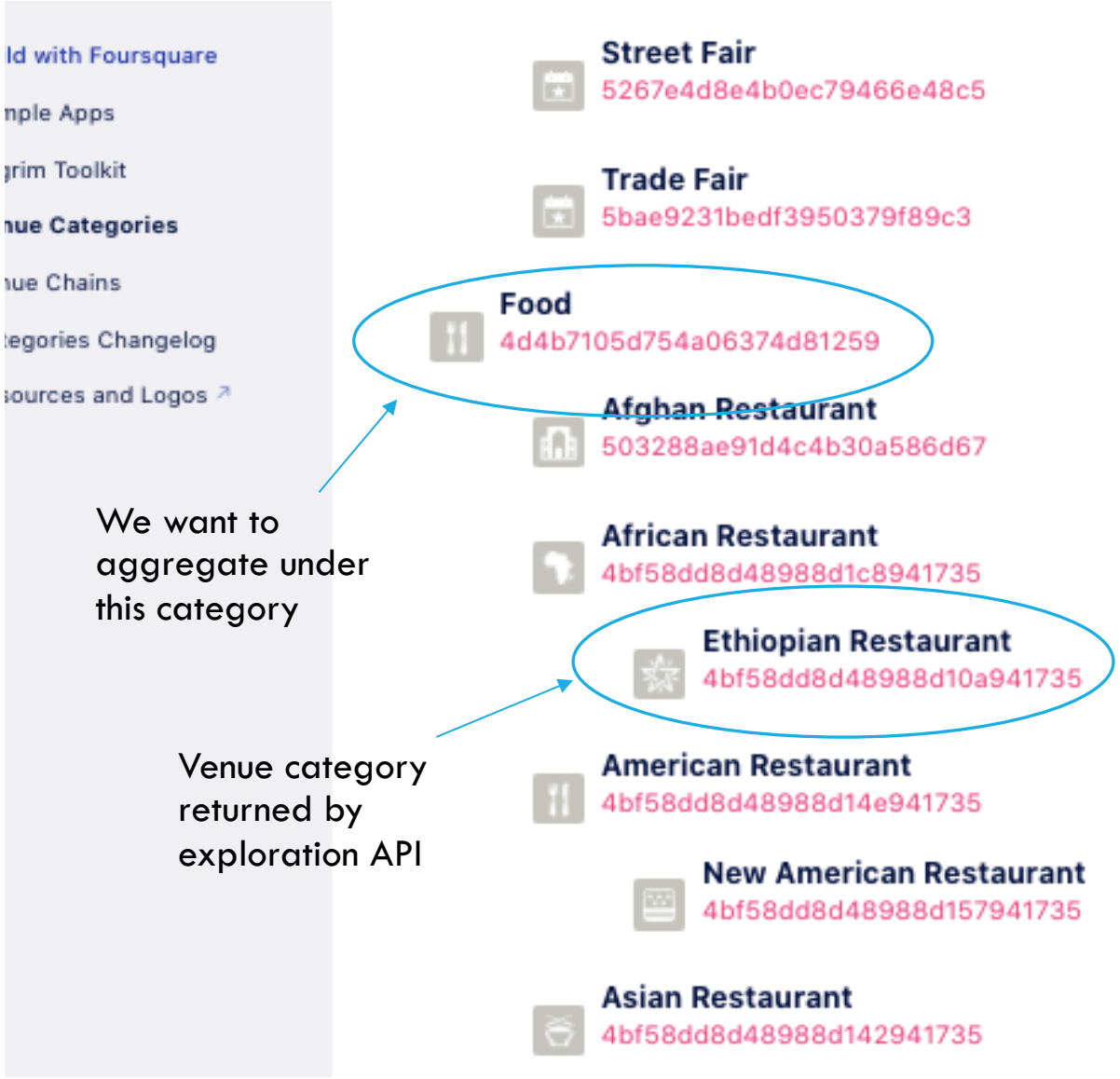
1. **Transform** the dataset of districts and venue categories
2. **Select** the districts that correspond to the user's preferences
3. Use preferences to **calculate a user profile vector** of weights assigned to Venue categories
4. Obtain a numeric **preference score for each unknown district**
5. **Present** the top scoring alternatives, in text form and on a map

1A/ DATA TRANSFORMATION: CATEGORY AGGREGATION

The Foursquare API returns detailed (leaf-level) categories

Designate specific categories in the category hierarchy, that can serve as **aggregation points** for all categories below it

We experiment with High-Level Categories



1B/ DATA TRANFORMATION: SCALING

- The complete list of venues for each district was one-hot encoded, based on the high-level category
- The aggregated results by district using mean(), was unevenly scaled across categories

District_code	Arts & Entertainment	Food	Nightlife Spot	...	Shop & Service	Travel & Transport
1000-1499	0.261364	0.443182	0.102273	...	0.102273	0.056818
1500-1799	0.000000	0.538462	0.076923	...	0.076923	0.153846
1800-1999	0.261364	0.443182	0.102273	...	0.102273	0.056818

Scaling

- Data was scaled using a QuantileTransformer with a uniform distribution output
- The probability density function of each feature is mapped to a uniform distribution

District_code	Arts & Entertainment	Food	Nightlife Spot	...	Shop & Service	Travel & Transport
1000-1499	0.916667	0.494186	0.451550	...	0.362403	0.527132
1500-1799	0.000000	0.742248	0.284884	...	0.207364	0.860465
1800-1999	0.916667	0.494186	0.451550	...	0.362403	0.527132

2/ DATA SELECTION

The subset of cities' encoded data that belongs to the districts known and rated by the user

District code	Arts & Entertainment	…	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport
E9	0.521318	…	0.612403	0.618217	0.866279	0.000000	0.476744	0.393411
N1	0.000000	…	0.689922	0.926357	0.542636	0.992248	0.220930	0.271318
W11	0.432171	…	0.701550	0.655039	0.449612	0.000000	0.759690	0.306202

3/ CALCULATE USER PROFILE

Express user preferences on venue categories as numbers

User profile = (User Ratings) x (Venue Category Data of user-preferred districts)

Venue High Level Category	Weight
Arts & Entertainment	3.985465
College & University	0.000000
Food	8.410853
Nightlife Spot	9.144380
Outdoors & Recreation	7.450581
Professional & Other Places	3.968992
Shop & Service	6.350775
Travel & Transport	3.993217

Nightlife and Outdoors & Recreation now weigh nearly as much as, or more than Food. These preference weights seem in line with the character of London areas chosen by the user

4/ CALCULATE SCORES OF UNKNOWN DISTRICTS

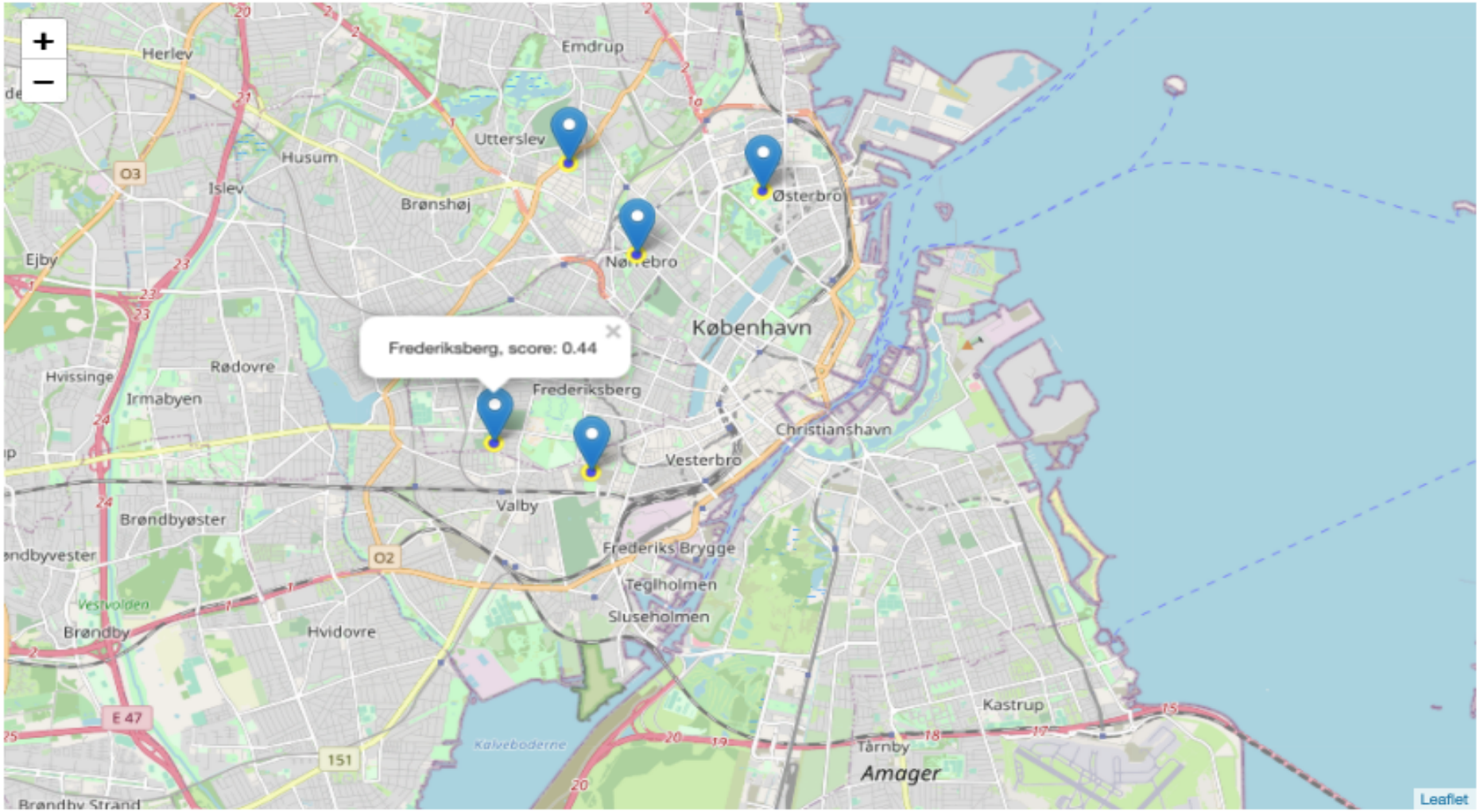
To arrive at a single number of 'likeability' for each district, we'll calculate a **weighted average**

The formula is

District score = (Venue Category Data) x (User profile vector)

District_code	District_name	Score
2200	Copenhagen N	0.610662
2400	Copenhagen NV	0.546174
1500-1799	Copenhagen V	0.451454
2000	Frederiksberg	0.443338
2100	Copenhagen Ø	0.432967

5/ PRESENT BEST RECOMMENDATIONS



CONCLUSIONS AND FUTURE DIRECTIONS

The Recommender produces results in line with expectations, it could be envisaged as support to human decision-making, although not tested and refined enough yet

- The trendy London neighborhoods preferred by the user are similar to the proposed districts of Nørrebro, Vesterbro, Østerbro and Frederiksberg. Many venues for Food, Art, Nightlife, Shops and Recreational activities

A lot of space for improvement, some possible ways:

- Additional data sources, for example data on property prices
- Improve the representation of the district profiles.
  - How to weigh one vs. dozens of venues of the same type?
  - Search of improved district representations supported by clustering and visualization of districts
- Improve scoring calculations to increase effect of low ratings
- Improve accuracy of district profiling by aggregating on different venue categories
  - Investigate intermediate levels between the top-most and the lowest ones
- Reduce dimension of districts, using other criteria for geographical decomposition