Ukrainian Catholic University

Faculty of Applied Sciences

Data Science Master Program

Predicting population density using bus routes data

Linear Algebra final project report

Volodymyr Lut

June 2019

Contents

Abstract

When trying to predict where migrants would settle down, one do not have a big intuition. However, busines could be helpful in such cases. In this project, PageRank algorithm would be used to determine the most valuable nodes in the network of Polish cities, created by UA-PL international bus routes. This approach would be used to determine the areas where Ukrainians live in Poland at most.

1 Introduction

The paper should be treated as a way to apply existing business/market knowledge in non-trivial tasks. In contains information about the PageRank algorithm, it's mecahnics, comparison with existing results based on other business data.
This paper also explains key concepts of PageRank agorithm from the perspective of Linear Algebra. All source codes are provided in referenced GithHub repository.

2 Problem, Data, and Intuition

2.1 Emigrants population density estimation

Once a person crosses a border of destination country it becomes hard to estimate, where this person would settle. However, this data is important to provide governmental services for those people. During the Presidetial Election in April 2019 thousands of Ukrainians in Poland and other countries were not able to vote, becuse embassies were not able to handle such a demand.
This situation becomes even more tricky in Poland, where, seasonally, only 1.7 mil of Ukrainians are legally working.
However, Poland is a neighbouring country to Ukraine and bus lines are probably the most major transportation way in this migrations. There are 482 regular, daily bus lines, heading from Ukraine to Poland cities. Data about them would be used in this research.

2.2 Data

In this project data from the Ukrainian Open Data Portal was used.

The dataset contains information about international bus routes and bus stops in Ukraine. Copies of those datasets (actual on Monday, June 10 2019) are contained under the raw_data directory of this repository.

Each bus line contains information about it's stops. Together, they created a network, connecting 6446 bus stops in Ukraine and Poland.

During the preprocessing, addresses of those bus stops were gecoded using Google Maps API, meaning, that data was enriched with information about country, region, city (or disctrict) each stop belogs to, along with their lattitude and longtitude.

More Data Preprocessing insights could be found in the DataPreprocessign Jupyter Notebook in the root of this repository.

2.3 Intuition

Since every stop is connected to another stop via the chedule of the bus line, they create a network of connections between cities. Scaling out to city (or district) we can construct a network where some areas would be connected to some other areas.

Bus lineas are operated by business, so they should follow general market agenda: there would be more bus lines to the cities which have bigger demand of such transportation. In general, intuition behind this project is following: **there would be more bus connections between Ukrainian and Polish cities where more Ukrainians live in or travel to.**

This is where PageRank comes. This alghoritm is used by Google to provide rough estimation of how good the website is, accounting the quantity and quality of links pointing to it. In general, it could be used over our network to determine the most important areas bus lines to Poland are heading to.

3 PageRank

PageRank is named after Google's co-founder Larry Page. It uses the link structure as an indicator of an individual page's value. It interprets a link from page A to page B as a vote, by page A, for page B. This idea is extremely simple, beacause the importance of each page is the average of the importance of every page that points to it.

3.1 PageRank from Linear Algebra perspective

Since the importance of everybody page (here and later - node) relies on the importance of other nodes, some iterating process should be performed over the network.

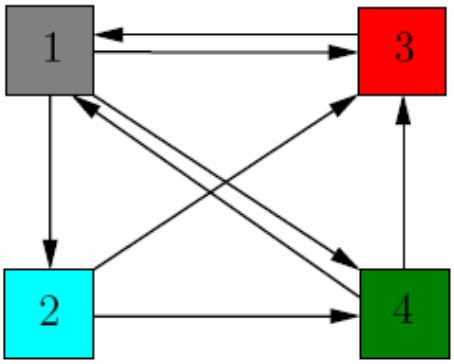
Assigning each node the same "importance" (hereunder I would use the word importance to ease reading; Generally saying, I mean the evolution of weight of other nodes pointing to current node under this term) and calculating the number outgoing links (in the nature of this project - road between two stops), evenly sharing to each neighbouring node the inportance of current node. At each step, new importance would be calculated, eventually converging to a situation, where importance of each node would be very close to the importance of same node at the previous step.

In other words, this is a Markov chain, that could be represented as an eigenvalue problem

$$v = Av_{previous}$$

Where A is a network, described in a matrix form, v is an importance vector of current iteration, $v_{previous}$ is an importance vector of previous step (which would later evolve to an eigenvector of A).

For example, for a network of 4 nodes of following structure



The connection matrix (network) would be described as set of columns (one column for every node), where each elment represents a link to other node (if exists) with a weight of $\frac{1}{n}$ where n denotes the number of outgoing links from this node.

$$A = \begin{pmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{pmatrix}$$

Iterative process for this network would be following

$$\mathbf{v} = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}, \quad \mathbf{Av} = \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix}, \quad \mathbf{A^2 v} = \mathbf{A(Av)} = \mathbf{A} \begin{pmatrix} 0.37 \\ 0.08 \\ 0.33 \\ 0.20 \end{pmatrix} = \begin{pmatrix} 0.43 \\ 0.12 \\ 0.27 \\ 0.16 \end{pmatrix}$$

$$\mathbf{A^3 v} = \begin{pmatrix} 0.35 \\ 0.14 \\ 0.29 \\ 0.20 \end{pmatrix}, \quad \mathbf{A^4 v} = \begin{pmatrix} 0.39 \\ 0.11 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad \mathbf{A^5 v} = \begin{pmatrix} 0.39 \\ 0.13 \\ 0.28 \\ 0.19 \end{pmatrix}$$

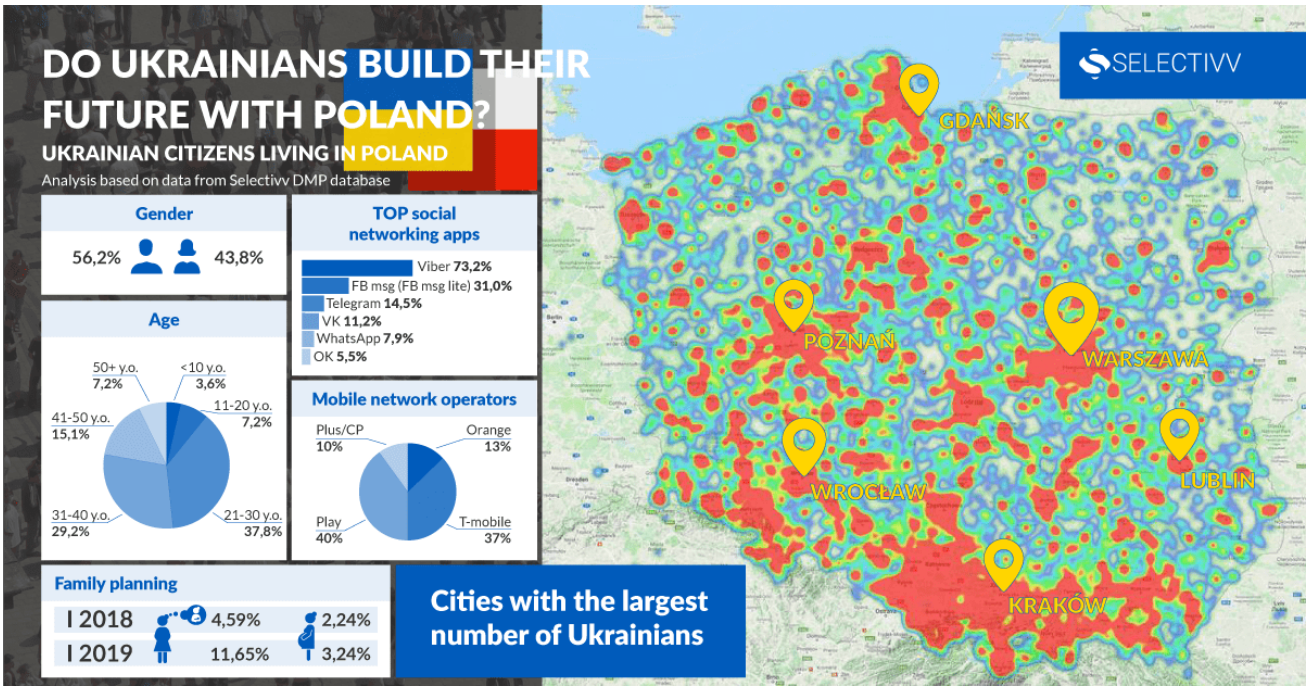
$$\mathbf{A^6 v} = \begin{pmatrix} 0.38 \\ 0.13 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad \mathbf{A^7 v} = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}, \quad \mathbf{A^8 v} = \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}$$

Implementation of the iterative process (power method) applied to the network of Polish cities could be found in Jupyter Notebook in the repository provided.

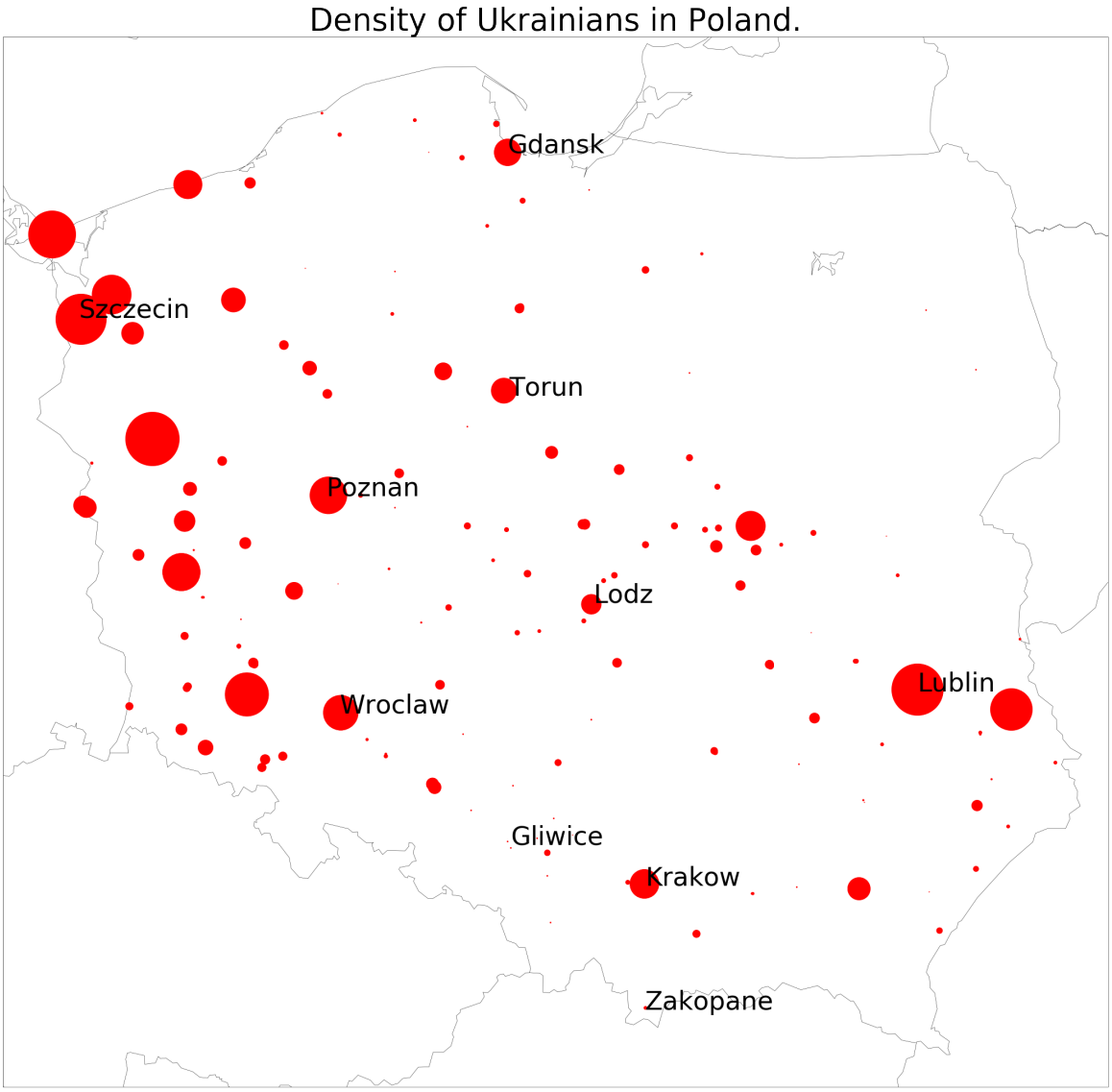
4 Results and Conclusions

4.1 Results evaluation

Polish company Selectivv was using cellphone network operators data to detemine which phones where previously using a SIM card of a Ukrainian operator and where do their owners live. This is the most precise approach that could be used to understand the geographical trend of seasonal migration.



To compare with this data, the chart was created, where each red node's size describes the importance of each city / region in the overall network of Polish cities. Each node location is aggregated mean of geographical coordinates of stops, groupped by city or region. The network contains of 201 big cities or regions in Poland, Ukrainian bus lines are heading to.



Overall the estimation of the density of population performed good on detection of wide Warszaw aglomeration, Krakow, Wroclaw, Poznan and Gdansk clusters, but, according to the nature of data is not precise and biased in the direction of German border, becuse this is where nost of bus lines are heading too.

4.2 Conclusions

PageRank algorithm is an effective and elegant tool for the detection of important nodes in datasets. However, trying to apply it to the real-world data, like bus lines in this exact project could be tricky, because PageRank is too tolerant to the data it works with.

However, additional data preparation could help. Filtering out stops in airports and border-crossing points on the Polish and German border could help in estimation of

density population, because a lt of transferring passengers would be cut off.
The algorithm showed, that there is less Ukrainians in the Podlaskie Wojewodztwo - there is no bus lines heading there, and much smaller population of Ukrainians living there, according to Selectivv's research.
This algorithm should not be used in such tasks, where multiple variables could take effect and bias the output; however, it could drop some useful insights at the very first stages of similar researches.

References

[1] Slectivv *Do Ukrainians build their future with Poland? – the latest Selectivv study.* https://selectivv.com/en/czy-ukraincy-wiaza-swoja-przyszlosc-z-naszym-krajem-najnowsze-badanie-selectivv/

[2] Raluca Tanase, Remus Radu *PageRank Algorithm - The Mathematics of Google Search.*
`https://pi.math.cornell.edu/ mec/Winter2009/RalucaRemus/Lecture3/lecture3.html`

[3] Wei-Chien-Benny Chin, Tzai-Hung Wen *Geographically Modified PageRank Algorithms: Identifying the Spatial Concentration of Human Movement in a Geospatial Network*

`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4593571/`

[4] Supportive code
`https://github.com/volodymyrlut/international-bus-routes-analysis`