Table of Content

1. Introduction

Pokec is the most popular on-line social network in Slovakia. The popularity of the network has not changed even after the coming of Facebook. Pokec has been provided for more than 10 years and connects more than 1.6 million people. Datasets contain anonymized data of the whole network. Profile data contains gender, age, hobbies, interest, education etc. Profile data are in Slovak language. Friendships in Pokec are oriented.

2. Problems

- We can't extract frequent patterns from eye-balling the data
- We were not able to tell what different types of data are there, what type of data can be mined and what can't be mined.
- What patterns can be received from the data?
- What algorithms can be used for data mining?
- What form of data should be, for frequency pattern mining?

3. Solutions

1. Feature selection

Data study: We studied what type of data it is, what patterns can be deduced from the data. We saw that it is data from social media, so trends, regional information can be deduced by it.

So we selected 3 features; 1) Regional, 2) Level of Education, 3) Age.

1. Regional

Slovakia's data has a total of 8 admin regions attached with districts and cities. We separated regions from cities in a sentence using NLP. Therefore, we were able to get regional data.

2. Level of Education

We eliminated the extra string and categorized the level of education into 4 categories.1)zakladne:basic, 2) stredoskolske:high school, 3) vysokoskolske:university, 4) ucnovske:apprentice

3. Age.

For age, we made bins of 20s, starting from age 0 to age 100, we made bins of 0-20, 21-40 onwards.

2. Algorithm Selection

1. A Priori

We used A Priori algorithm and we got Association rules out of it, using the Association rules, we were able to get Corelation analysis.

2. FP Growth

We used the FP Growth algorithm for Frequent Patterns.

3. PCY

For constraint base patterns, we used the PCY algorithm.

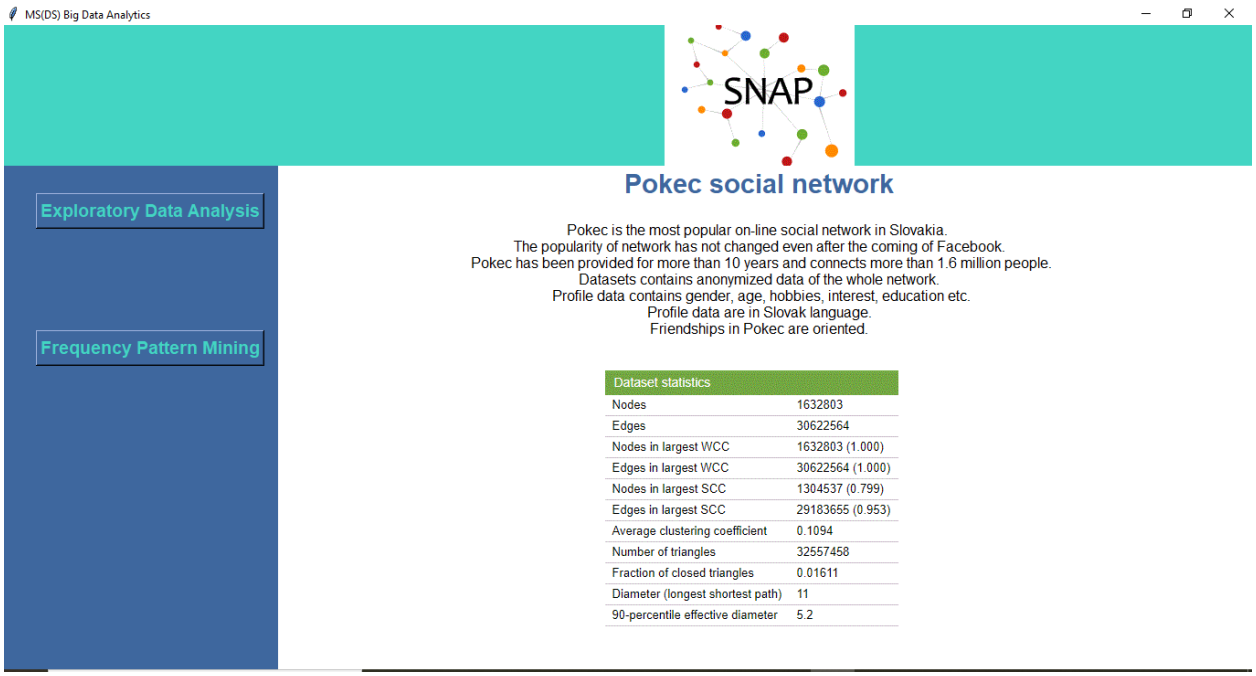
3. Algorithms Review

4. Visualization

5. Screenshots of GUI.

The front window of the GUI of Slovakia's social media looks like this. The right corner is the introduction about the data. And the left pane has buttons that lead to visualizations.

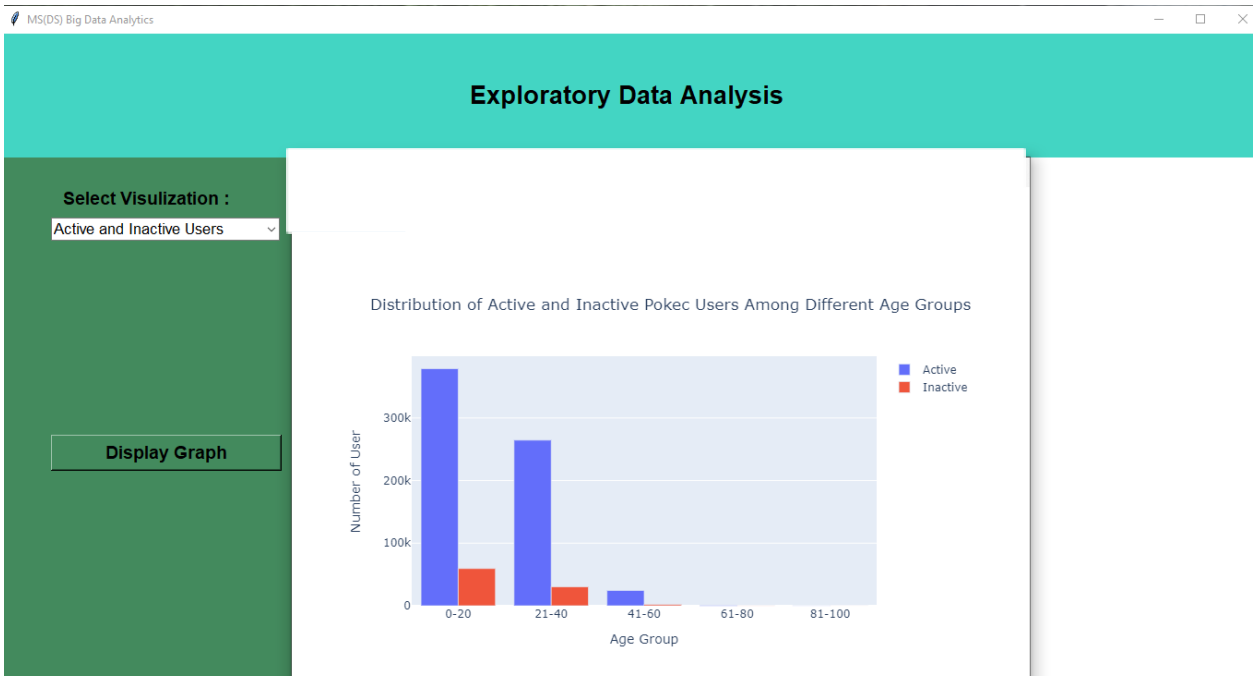
Exploratory Data Analysis is one kind of visualization and Frequent Pattern Mining is the other.



Below is the window of Exploratory Data Analysis, with 3 types of visualizations, mentioned in a dropdown menu;

1. Active and Inactive Users,
2. Spoken Languages Distribution
3. Top 5 Languages Spoken

By clicking on any of the options, their graph displays on the blank plane.

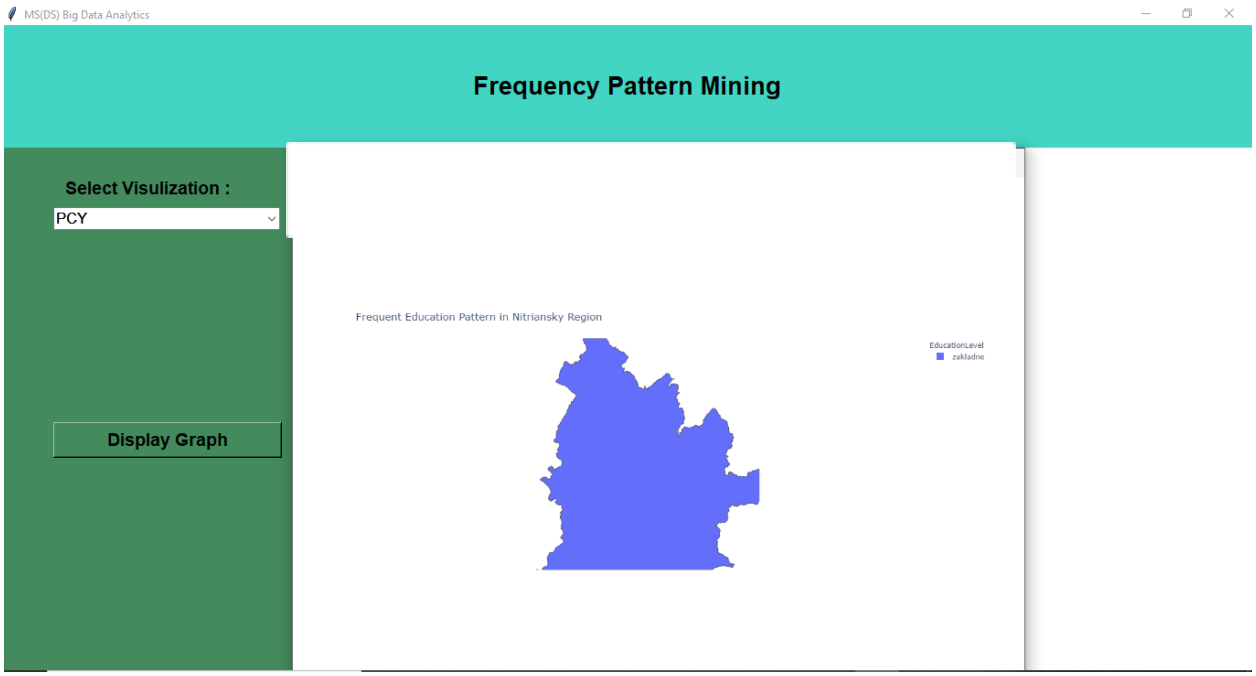


Below is the window of Frequency Pattern Mining, with 3 types of visualizations, mentioned in a dropdown menu;

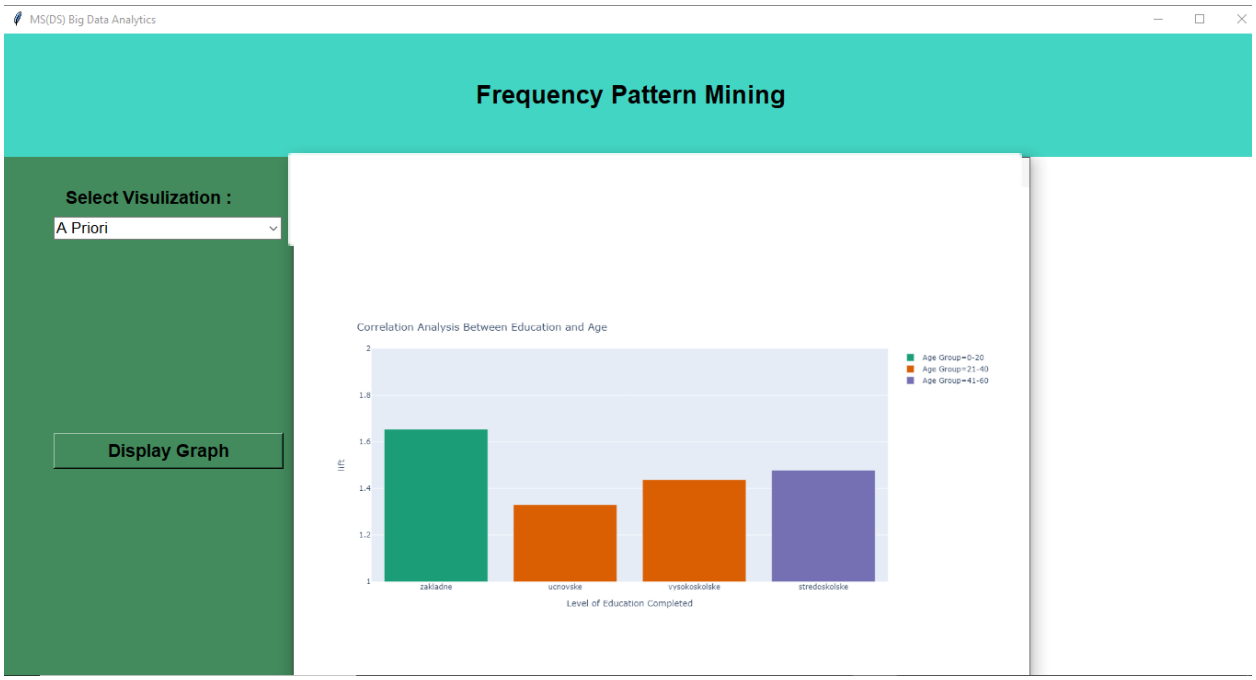
1. A Priori
2. PCY
3. FP Growth.

By clicking on any of the options, their graph displays on the blank plane.

Visualization for PCY.



Visualization of A Priori



MS(DS) Big Data Analytics

Frequency Pattern Mining

Select Visulization :

F P Growth

Display Graph