Popular Erasmus destinations

Alessandro Lotta, Youssef Ben Khalifa, Yong Xiangji

November 8, 2022

Contents

List of Algorithms

1 Motivation

The goal of this project is to analyze and identify the most popular Erasmus destinations at the university level. This particular analysis can be useful for students who are planning to go abroad for their Erasmus experience, but also for the universities that want to find out how they compare to the other universities around the world in terms of popularity.

1.1 Dataset description

The datasets we used for the project were found on the *Official European Data* portal, where for each record, were specified:

- Departing country;

- Destination country;

- Type of Erasmus (Junior student, Senior student, Staff, Trainees ecc.);

- Reference Period;

The datasets were provided as separate .csv files, one for each year from starting from 2014 to 2019, for a total of 6 files and roughly 160.000 records. Due to the large amount of data, we decided to focus on the most recent years, from 2017 to 2019, and consider only Senior university students. All the data we decide to use will be properly pre-processed and sampled, in order to obtain a more manageable dataset.

1.2 Graph Model

Throughout the project we may use different graph models, depending on the analysis we want to perform. The main graph model will be modelled using the various universities as nodes, and the edges are weighted by the number of students that went to that university. The resulting graph will be a directed graph, in which each edge is defined by the starting node, that represents the university from which the student departed, and the ending node, that represents the university where the student went. On each node will then be specified various attributes, like the name of the university, the country it is located in, the number of students that went there.
Another graph may be modelled using the countries as nodes instead of the universities: this particular graph can be used to measure the amount of ingoing and outgoing students from a country, and can be used to identify the most popular countries for Erasmus students.

2 Methods Adopted

Our goal is to find the most important nodes in the network, to do so we are going to employ techniques and metrics used in social network analysis. In particular we are going to use the

PageRank algorithm and HITS link analysis algorithm.

Page Rank is used to rank each node in the network by giving a score of importance, in which the most important nodes are the ones that have a high number of incoming edges.

HITS algorithm is used to identify the most important hub and authority institutions in the Erasmus program. Hubs are institutions with a high quantity of people moving out of there, authorities instead are institutions in which there is a big quantity of incoming people.

At the end we are going to compare the results obtained using both the algorithms.

If they don't work we will use approximation algorithms to find the most important nodes in the network. One idea is to adopt sampling on the existing graphs before applying the mentioned algorithms.

3   Intended Experiments

The implementations for the PageRank is available at this link, it is based in the python NetworkX library and it provides also ways of applying the approximate PageRank algorithm by adjusting specified values. The implementation for HITS algorithm is also part of the NetworkX library, it can be found here