PUBLIC PROJECT SUMMARY

The project's scope was a hierarchical clustering on dissimilarity matrices using social media data. The main objective was finding user-defined relationships between companies' Facebook pages and to confirm Datalyze's own page classification framework. The project's sponsor and data provider was Datalyze, a research and consulting company which specializes in social media and big data tools.

The research with various parameters and algorithms did not find a stable clustering structure to support Datalyze's framework but it still revealed interesting patterns on a limited set of observations. It showed that user activity does create an intricate set of linkages between companies' Facebook profiles which can be used for competitor analysis or targeted marketing campaigns.

The Company Behind the Data

Datalyze, an official spinoff of Eötvös Lóránd University, collects data from social media and other online platforms and sorts them into databases. It has more than 4 billion datapoints and covers more than 6 million Hungarian Facebook users through official Facebook APIs, which is augmented with data derived from users' visits of domestic websites.

This multi-source data is used to perform behavior analysis based on social media activity *and* web-browsing history. The result is a multitude of tools to map talk drivers and emotions which appear in online opinion streams and to identify and characterize various groups of online opinion leaders.

Datalyze's tools help clients gain insights both about their target audiences' emotional relation towards their products and services and about these audiences' view on and relationship to their competitors. The data is GDPR-compliant and anonymized but stored in a way that helps nano-level customer analysis and targeting. When analyzed properly with Datalyze's uniquely developed data science tools, it can also reveal non-trivial patterns which help make more robust business decisions.

The Project

The project had the very specific framework of clustering Hungarian companies' Facebook pages based on their users' activity across the Facebook universe. The basic building blocks are user activities: likes, dislikes and other emotional reactions to companies' Facebook contents. The project's aim was to build a hierarchical cluster structure to identify commonalities between various company pages based on their common users and cluster Facebook pages in a way which should mirror their human-defined thematic classifications.

The project's most important assumption was that users tend to follow themes, and their Facebook activity connects company pages of similar profiles. While an individual user may follow, like, or dislike companies in various product and service groups, a large number of common users between two companies means that these companies compete for the same or for similar clients. A large common user base means that these two companies are 'close' to each other while low or zero number of shared users imply little or no commonalities, or a far distance.

CEU eTD Collection

In data science terms this is a clustering solution based on a *distance matrix*, or in other terms *dissimilarity matrix*, like clustering cities in a hierarchical manner based on their geographical distances. Just as you can spot city clusters on a map, and identify them based on a distance matrix, it is also possible to create a hypothetical map of companies where distances between them are calculated based on the absolute or relative size of their common user bases.

Tools and Technology Used in the Project

Codes were written and run in Python, and the clustering algorithm which handle distance or dissimilarity matrix as input is located in the so-called SciPy package of Python. In terms of hardware, an Acer Swift SF314-15 laptop which runs 64-bit Windows 10 and has four Intel i5-7200U CPUs @2.50 GHz was used.

Steps to Build Clusters

The underlying data was accessed through a public Facebook API. Each datapoint is a user action, usually a 'like' on a Facebook page. The records, stored in csv files, have a very simple structure: page id, post id, time stamp, user id, and the user action (mostly a 'like'). Each file contains the records of user activity on a single Facebook page in a month. There were 83 727 csv files for 8 743 Facebook pages in 38 GB.

The first step was to create a data matrix to calculate the number of users any Facebook page shares with any other Facebook page. This is a symmetric matrix, where each row and each column are a page, and the value in every cell of the matrix is the number of users who were active on both pages. The matrix diagonals are the total number of users who liked, disliked or reacted in another way to the content on that page. This is eventually *a heatmap of connections*: the larger the number in a cell the stronger the two pages are linked by users.

Pages which share a large number of users serve a common user base, so they are in essence competitors. A hierarchical clustering can show for any page which other pages are its close and distant competitors. To measure the distance between any two pages the underlying data matrix of shared connections (the 'heatmap') had be transformed into a *distance or dissimilarity matrix*, using mathematical transformations of various kinds.

Once the distance matrix is defined and calculated several cluster building options can be implemented. These options define how the distance between two clusters of multiple elements is measured and have a significant impact on how observations are grouped into clusters at higher and higher levels of aggregation. At the end of the process, each observation is eventually merged into a single cluster which serves as a root for the system of sub-clusters.

Results

The clustering exercise did not only aim to draw a competitor landscape, but it should have cross-checked the thematic page classification framework by Datalyze. The company uses its internal definitions to classify pages into groups and subgroups based on the content of the pages. As many

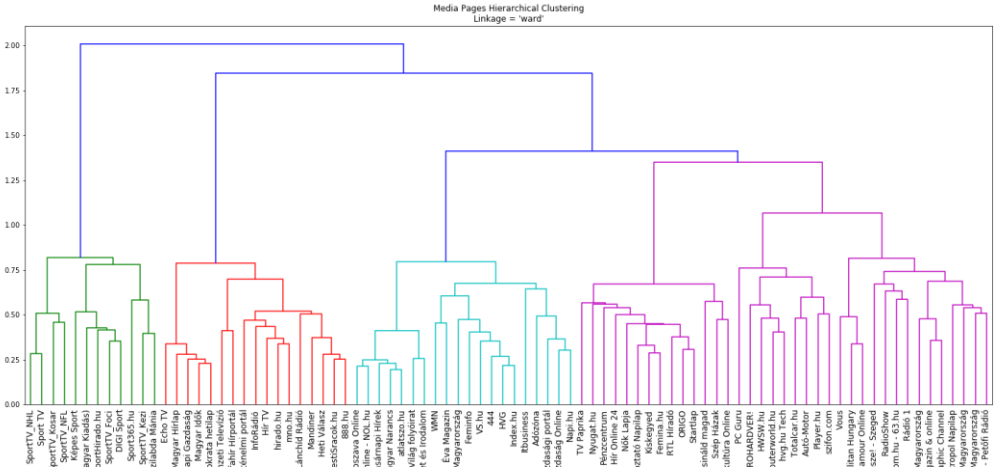
CEU eTD Collection

pages have multiple focus, this classification is far from straightforward and the data project was implemented either to confirm the framework or to reveal new patterns.

It turned out that user-defined linkages are way more complex than this relatively simple classification framework. Using the whole dataset, the clustering results were inconclusive, sometimes very sensitive to the clustering algorithm's parameters, and did not follow original, manually labeled categorization.

The procedure was then repeated on a much smaller, very targeted dataset: Facebook pages of media content providers. In this set only 86 pages were used, and the underlying target was finding any clear hierarchical cluster disregarding the manual category labels.

The results were much more consistent and meaningful than in the case of the total database. A very clear cluster hierarchy developed.



In this particular hierarchy six clusters can easily be identified:

1) Sports
2) Conservative and right wing-news and politics
3) Liberal and left-wing news and politics
4) Home + family + other
5) Tech + cars
6) Entertainment + tabloid-type contents

While content is important, other things, such as political orientation, can be a decisive factor in linkages. Uncovering this pattern of factors requires a deeper dive into the data with more emphasis on the diversity of the underlying population (Facebook and other social media pages). This will help a better understanding of the client's individual competition landscape and deliver a higher value-added service for Datalyze.

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection

CEU eTD Collection