Introduction to Clustering Toastmasters Clubs with Python

By Ricky Soo

Introduction

Toastmasters is a non-profit organization that trains members in communication and leadership skills. Members organize regular meetings in "clubs" where they practise public speaking, presentation and impromptu speaking skills. Members follow an education program called Pathways, featuring experiential learning, self-paced learning, peer evaluation and mentoring.

Founded in California in 1924, Toastmasters has spread its wing to 143 countries. Now there are more than 358,000 members in more than 16,800 clubs worldwide. To manage the many clubs around, the clubs are divided hierarchically into regions, districts, divisions, and areas.

Typically, each club consists of around 20 members. 4 to 6 nearby clubs are organized into an "area". 3 to 6 areas constitute a "division". 6 to 12 divisions form a "district". About 10 districts make up a "region". There are now 14 regions worldwide.

The regions are numbered from Region 1 to Region 14. Districts are also numbered such as District 102 where I am active in. Divisions are named with alphabetical letters such as Division A and Division B. Areas are identified with a number after the Division name, such as Area A1 and Area A2. The information of all regions and districts can be found at https://www.toastmasters.org/~/media/35503AED4D20498FBDA2AA75559FF2E0.ashx

The fiscal year of Toastmasters starts in July and ends in June the next year. Each year, officers are elected or appointed to serve as District Officers to manage the clubs and develop their leadership skills. At the end of each fiscal year, each district has the option to "realign" clubs to group them in a way that helps to manage, market and strategize for the clubs, areas, divisions in the district.

The Project

There are Toastmasters clubs of various sizes and conditions. There are bigger clubs with more than 50 members, and there are clubs with only a few members. There are restricted clubs such as corporate clubs whose membership is open only to the employees of a sponsoring company, or university clubs which are open for students of a university, and there are community clubs where anyone 18 years old and above can join as a member. There are clubs which produce good results in membership growth and members' achievements, and there are clubs having challenges to recruit members or hold regular meetings. There are clubs that are close to one another geographically, and there are clubs which are located far from the others.

The purpose of this project is to use machine learning algorithms with the Python programming language to group similar clubs into clusters in order to better understand them. The insights gained could help in formulating strategies to grow and support the clubs, promoting the clubs to the general public and assisting in the yearly realignment exercise.

The specific areas of benefits include:

- **Management and support** – District Officers might need different strategies to support clubs of different sizes and conditions in order to help them to be effective clubs serving their members.
- **Marketing** – Clubs need to formulate marketing strategies to recruit new members. Understanding the neighbourhood where a club is located can help gain insights on the potential market out there for new members.
- **Realignment** – The yearly realignment exercise typically does not seek to group clubs of similar nature together in order to be fair to give District Officers a variety of experiences in leading them. But clustering the clubs could help identify the similar clubs and avoid aligning them together. However, the realignment does seek to group clubs that are near to one another for easier logistics.

The scope of this clustering project includes 90 Toastmasters clubs in District 102 located in the state of Selangor in Malaysia. This is the district where I served as a District Officer twice before and so is familiar to me. These clubs currently belong to Division B, C, D, E and H in District 102. A summary of all clubs in District 102 for the year 2019-2020 can be found at http://dashboards.toastmasters.org/2019-2020/Club.aspx?id=102

The Data

The data for this project comes mainly from 3 sources – Toastmasters web site, Foursquare data and domain knowledge.

Toastmasters web site – The Toastmasters web site publishes a public dashboard showing the performance reports of all clubs in all areas, all divisions, districts and regions. The web site also contains a club page for each club showing its name, location and contact information.

Foursquare data – With the location information, Foursquare API is used to understand the neighbourhood where the clubs are located. Some neighbourhoods are popular with places with many people checking in. The popularity can give an indication on how active people in the vicinity of the Toastmasters clubs, as such the potential members who might be able to visit the clubs and join as members.

Domain knowledge – As a two-time District Officer myself for the year 2017-18 and 2019-2020, I have gained sufficient knowledge into the working of Toastmasters clubs. This helps me to identify the important club features that go into explaining the nature of the clubs, the similarities and dissimilarities, and the strategies that the clubs might need to be effective.

In this project, a dataset of 90 clubs is drawn up completed with data below:

- **Club name** – The name of a club used to identify a club. Not used in modelling.
- **Club number** – The club identification number of a club assigned by Toastmasters. Not used in modelling.

- **Address** – The local address of the club which gives input to Python library to obtain the longitude and latitude of the club location.
- **Location** – The longitude and latitude of a club location to be used in clustering nearby clubs together.
- **Membership** – The number of members in the club. This gives an indication of the club membership strength.
- **Net growth** – The net increase or decrease in the membership number. This gives an indication of the club marketing effectiveness.
- **Goals achieved** – The performance of each club is measured by the number of goals achieved in the Distinguished Club Program. The program consists of 10 goals and hence, this data can range from 0 to 10. This gives an indication of club quality of a club in serving its members.
- **Community club** – A community club opens its membership to the public to join. But a restricted club is only for the employees or students of a sponsoring company or university. A community club might have bigger market potential, but a restricted club might better align the club to the purpose of the sponsoring company or university. This data takes the value 0 or 1.
- **Online attendance** – A club might allow meeting attendance by online means. A club with online attendance might attract members regardless of geographical boundaries. This data takes the value 0 or 1.
- **Neighbourhood activities** – The level of activities in the neighbourhood where a club is located. This is measured by the total number of check-ins in the Foursquare places within 1km radius of the club location. A more active neighbourhood might indicate bigger market opportunities for the club.

All data is taken from the year-end result of the 2019-2020 fiscal year on July 13th, 2020. The club data is taken from the publicly available data on Toastmasters web site.

This project is completed as the capstone project in fulfilment of the Applied Data Science Capstone course on Coursera at https://www.coursera.org/learn/applied-data-science-capstone