Extracting tourists movements' information from Twitter data sets:
What are the most visited places in the Netherlands?

Mickaël Nkunku
Enschede, Netherlands

Thibault Defeyter
Enschede, Netherlands

Jose Ailton Filho
Enschede, Netherlands

Fernando Augusto Machado
Enschede, Netherlands

I. Introduction

The era of the Internet of things we are currently living in is still improving and everyone owning a means to access Internet can express onself. Data published in any way on the network is stored somewhere and can be identified. That identification procedure usually includes retrieving more information than one can see when posting something. Those pieces of information can be anything: IP address, geo-localization, name of the web browser and name of the operating system used are few instances. To put it in a nutshell, from each data set can be extracted a lot of information which anybody who is provided an access to can do analytics on.

The current context enables to use previously unavailable information and create services as to improve users' experience when connected to the network. Information about tourists' movements, for instance, could allow companies and public institutions to do the necessary so that travelers and tourists have a better experience, and even prevent them from doing unnecessary expenses. However, useful information cannot be accessed in a direct way: it must be extracted. Having at disposal an extracting information means which could process huge data sets would be indeed a useful tool in order to analyze such data.

For this research topic, we have chosen to output a table with relevant attributes. Our table will be filled by analyzing huge data sets of tweets which are also used by the *twiqs.nl* website. From each tweet can be extracted useful pieces of information, in particular, geolocalization details. Some papers propose their method to find users' hometowns like the Principal Components Analysis mentionned in [1]. As for the technology to be used to approach this research topic, Pig Latin, over the Hadoop cluster, is the most suited. User defined functions will also be implemented in order to increase its efficiency in our favor.

II. Materials and methods

The tweets from *twiqs.nl* are provided in a compressed JSON format - *GZIP* -, thus requiring to have a proper class so that information can directly be read from the compressed format. The *JsonLoader* class created by Twitter - *Elephant Bird* project - could be a method to deal with the issue. However, this implies using Java as programming language to define user defined functions. Preferring untyped programming language, we decided to use Python instead. Using the latter, we defined the Mapper and Reducer classes so that the MapReduce operation can be executed over the Hadoop cluster on the huge data sets of tweets. For each tweet, many pieces of information can be retrieved using its attributes: for our research question, we decided to focus on the names of the destinations. As for what can be considered as a trip destination for a given movement, if the *city* attribute value is different from the location from which the tweet was written, then it is a trip destination. Then, counting the number of occurrences for each city using a map enables to output the total number of visitors for any town.

III. Results and discussion

As final result, we produced a table giving, for each dutch city, the number of tweeters who went there on a given day. This table can be used to determine, for instance, the most visited place in the Netherlands during the Christmas period. Graphical representations are also much easily possible thanks to our final result. Running out ot time, we did not have time to implement one. In order to state the legitimacy of our data, it could have been great to compare to an official source.

IV. Conclusion

Our table enables to answer the research question at different levels, and using graphical representations is the best way for users to determine when and where to go to places not crowded. Adding legitimacy to our data by comparing it to official sources would give more righteousness to our computations.

References

[1] *A Multi-Indicator Approach for Geolocalization of Tweets*, Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mhlhuser.

[2] *Representation and Communication: Challenges in Interpreting Large Social Media Datasets*, Mattias Rost, Louise Barkhuus, Henriette Cramer, Barry Brown.

[3] *Photo2Trip: Generating Travel Routes from Geo-Tagged Photos for Trip Planning*, Xin Lu, Changhu Wang, Jiang-Ming Yang, Yanwei Pang, Lei Zhang.

[4] *Knowledge Discovery from Geo-Located Tweets for Supporting Advanced Big Data Analytics: A Real-Life Experience*, Alfredo Cuzzocrea, Giuseppe Psaila, and Maurizio Toccu.

[5] *Predicting the Future With Social Media*, Sitaram Asur, Bernado A. Huberman.

[6] *Role of social media in online travel information search*, Zheng Xiang, Ulrike Gretzel.

[7] *Text and Structural Data Mining of Influenza Mentions in Web and Social Media*, Courtney D. Corley, Diane J. Cook, Armin R. Mikler and Karan P. Singh.

[8] *Predicting Tie Strength With Social Media*, Eric Gilbert and Karrie Karahalios.