ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Computer Human Interaction in Learning and Instruction - CHILI Lab

Massive Open Online Courses reach on Twitter, Stack Overflow and Github.

Bachelor project done by:
Inès Bahej

Under the direction of:
Prof. Pierre Dillenbourg

Supervised by:
Łukasz Kidziński

Section: Communication Systems

Academic year: 2015/2016

Introduction

Massive online open courses (MOOCs) represent the future of teaching. First of all, they are accessible to anyone with an electronic device and internet. Hence it is a good way to provide education for all . Besides, they are an innovative way for large scale teaching, which makes the MOOCs part of the globalization.

The aim of this project is to analyze the reach of the MOOCs today, in order to evaluate their popularity in different parts of the world. The aim is also to see to what extend MOOCs' users take advantage of them: maybe by creating new projects or by contributing to some existant ones. Finally, it would be interesting to see if MOOCs are as efficient and credible as the traditional way of teaching, by analyzing the sentiment of MOOCs users on social media.

In order to achieve these goals, we will analyze the data from three websites:

- the social media Twitter. By extracting tweets with specific hashtags and keywords, we will be able to analyze people's opinion on MOOCs and see where in the world people tweet the most about them thanks to geolocalization.
- Stack Overflow, a question and answer site for programmers. We analyze some of the questions posted by MOOCs' users to see how often they encounter problems with particular assignments.
- Github, a web-based git repository hosting service. Mining some of the repositories will tell us if some projects have been concretized thanks to MOOCs. It will also be helpful to track repositories that publish some assignments solutions, which could let some students to plagiarize.

Table of contents

Part I: Mooc reach

1- Mooc Reach on twitter

Thanks to the Twitter REST API, it is easy to data mine tweets that contain keywords we want. Nevertheless, there are some limitations set by Twitter: we are not allowed to retrieve more than 100 tweets per query. A solution to avoid this problem would be to retrieve tweets by dates of creation and store them. However, Twitter API has another limitation: it is not possible to retrieve tweets that are older than a week. Therefore, we will unfortunately only analyze tweets that are a week old. A way to retrieve more tweets about the same subject is to find tweets with keywords that lead to the same ideas. For instance, instead of searching only « Mooc », we can also search for its plural « Moocs » in order to have more tweets to analyze, which would lead to more credible results.

To organize our data, I created a web application using Django. The goals is to:

- visualize where in the world people tweet the most about MOOCs
- check whether people have a good or bad opinion/attitude regarding the MOOCs
- hashtags that are often associated with MOOCs.

We will also do the same researches for the Scala MOOC given by Prof. Odersky, in order to see how popular is this EPFL course. Finally, the platform can serve as a tool for research on any other topic.

1.1- Map

To visualize spacial distribution of the reach, we represent tweets on a Map. We extracted time zones from tweets containing the required keyword in order to represent on a map where people tweet the most about it. The tool that has been used is Mapbox. Because of privacy concerns, it is not possible to retrieve timezone from all the tweets. However, the results are quite representative as we can see in the Figure 1.1.

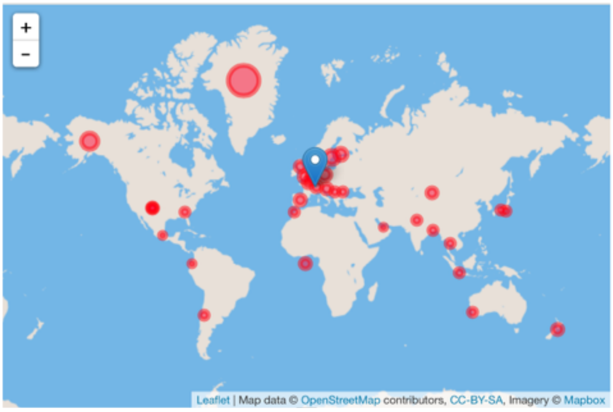
+
-
Leaflet | Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © Mapbox

Figure 1.1: Map representing locations of tweets with the keyword «moocs »

1.2- Pie chart

The goal here is to check whether people have a positive or negative opinion about MOOCs. Using IBM tool Alchemy, it is easy to analyze people's sentiment toward MOOCs. A simple alternative is to gather all the tweets that contain our keyword and the smiley « :) » or « :(» for positive and negative attitude. The results are represented in a pie chart as in Figure 1.2.

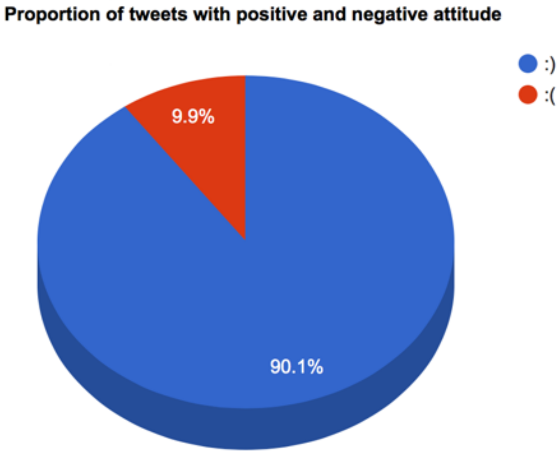

Figure 1.2: Pie chart representing the proportion of tweets with positive and negative attitude

1.3- Hashtags

Finally, it is interesting to see what keywords are most often related to the MOOCs. Therefore a word cloud is a quite good tool. For instance, the Figure 1.3 represent a word cloud generated on 11/16/2015. We clearly see that the main words related to the moods are « Big Data », « Data science », « Mathematics » and « Machine Learning ». Hence, these domains might be the subject of the most popular MOOCs.

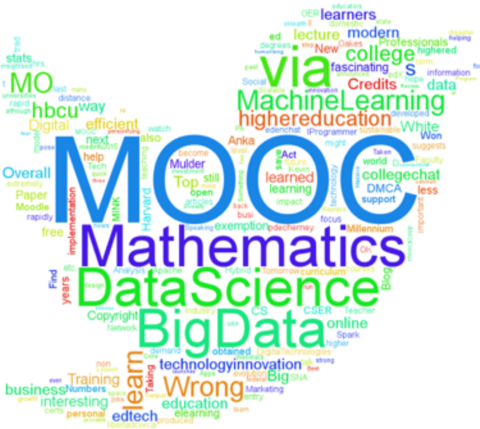

Figure 1.3: Word cloud of the keyword « MOOC »

2- Mooc reach on Stack Overflow

Data mining Stack Overflow is a great way to investigate popularity of a MOOC. For instance, we can clearly see in Figure 1.4 that from 2008, there is a linear growth of the number of questions containing the keyword « Scala». We also represent the number of repositories per week, to see if for particular assignments, people have difficulties. On the MOOC-reach website, we represent the data from September, to analyze the questions about this year's session.

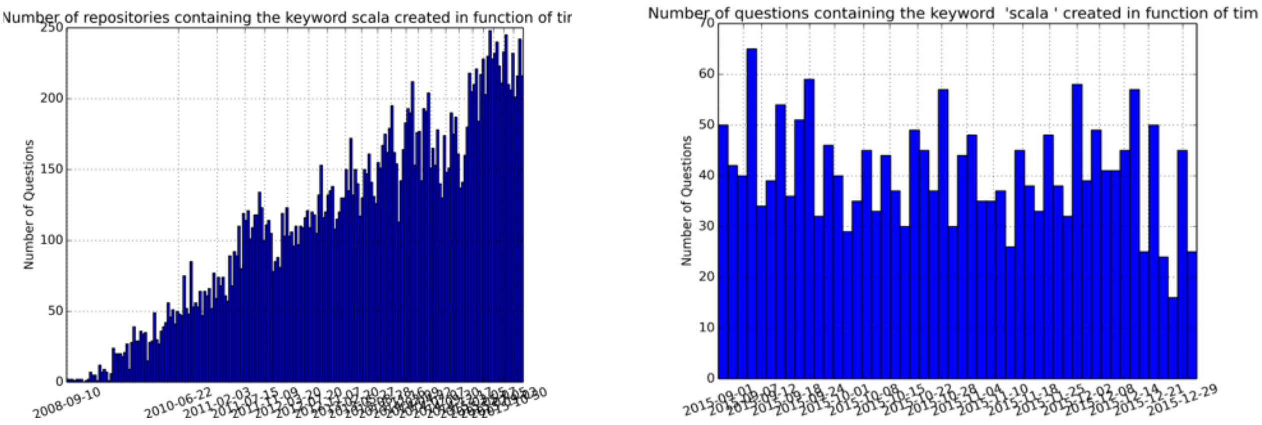

Figure 1.4: Number of questions concerning Scala from 2008 (left) and from the beginning of the semester (right)

Nevertheless, we don't have any concrete results when it comes to data mine Stack Overflow with the keyword « MOOCs ». Indeed, we can only retrieve questions that contain the keyword in the title. Unfortunately, there are very few questions with this keyword in the title which makes the data mining of Stack Overflow not very useful for this case.

3- Mooc reach on Github

It is interesting to see with which languages people code the most for a particular MOOC. To obtain these results, I extracted all the repositories that contain the required keyword in their title. From each of these repositories, I extracted the main programming language. Again, because of some API limitations, only 800 repositories can be analyzed per query. However, the results are quite representative. We finally obtain the following results for the keyword « MOOCs »

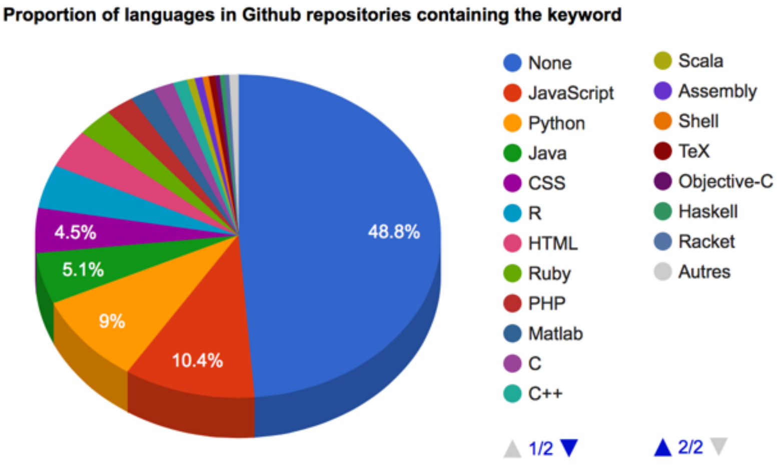

Figure 1.5:Repartition of the languages of the repositories containing the keyword « MOOCs »

We clearly see that people began to create repositories related to MOOCs from 2013, whereas platforms like Coursera were created in 2012. That translates a real expansion of the MOOCs.

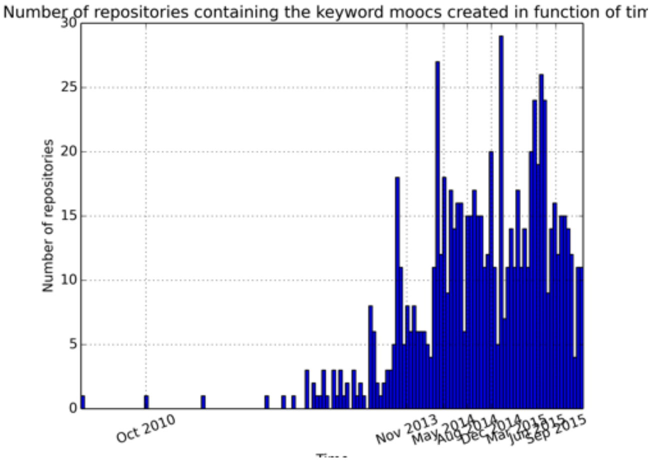

Figure 1.6: Number of repositories containing the keyword « MOOCs » in function of the time

Part II: The real impact of MOOCs

1- Scala MOOC efficiency

Nearly 70000 students attended the Scala MOOC. Therefore, it would be interesting to see how many of them used their acquired skills to create concrete projects. In this case, we will focus on Github repositories. By matching emails from Scala MOOC database and emails from GitHub, it is easy to find repositories that have the most stars, which are the most popular.

We obtained the results on figure 2.1. Nearly 120 emails matched. However, there is no real correlation between the grades of the students and the number of stars of their respectives repositories. In general, most of the people that have repositories with a high number of stargazers had indeed good grades in the Scala MOOC.

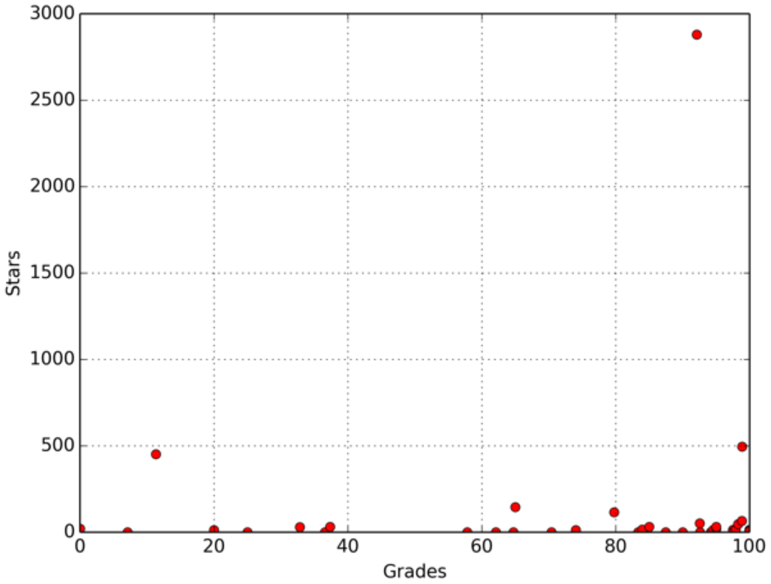

Figure 2.1: Number of stars of repositories containing the keyword « MOOCs » in function of the repositories owners' grades at the scala mooc.

We decide to separate GitHub users in two groups. The first group corresponds to all users that have more than a twenty stars for their repositories. In contrary, the second one corresponds to users that have less than twenty stars. By calculating the average grade of these two groups, we can see that there are no big differences.

To be sure that there is no real correlation between the number of stars of a repository and the grades of its owner, we do the Student's t-test. Finally, we don't have significant results to prove that there is a correlation.

Conclusion

MOOCs are without doubts becoming more and more popular. Data mining Twitter, Stack Overflow and Github confirmed our expectations. Regarding the Scala MOOC, analyzing the evolution of the number of repositories (Github) and the number of questions (Stack Overflow) in function of time is an efficient way to have a good weekly feed-back of students reaction to this MOOC. Finally, by matching emails of Github repositories and emails of the Scala MOOC database, we could see if this MOOC has been useful for some students and helped some of them to create meaningful projects.

Resource page

Bibliography

- *Mining the Social Web*, Matthew A. Russell
- *21 Recipes for Mining Twitter*, Matthew A. Russell

Online resources

- Twitter REST API: https://dev.twitter.com/rest/public
- GitHub API: https://developer.github.com/v3/
- Stack Exchange API: https://api.stackexchange.com/docs
- Django tutorial: http://tutorial.djangogirls.org/fr/index.html
- HTML, CSS and Javascript tutorial: http://www.w3schools.com/default.asp
- Mapbox: https://www.mapbox.com
- Google Developers: https://developers.google.com/chart/interactive/docs/gallery/piechart
- Word cloud package: https://github.com/amueller/word_cloud
- Stack Exchange package: https://github.com/lucjon/Py-StackExchange

Installed packages:

Wordcloud
Py-stackexchange
Twitter
Github
Requests
Json
Statistics