Department of Statistics 2019: Mapping Inequalities Online Using Data

Candidate Number: 10140

Contents

List of Tables

List of Figures

1 Introduction

Aims:

1. Literature review on natural language processing for GH/SE data and information retrieval
2. Exploratory and descriptive analysis of small dataset
3. Explore small dataset and employ ML methods to get predict quality of question/commit
4.

2 Brief Literature Review

2.1 A Summary of the Research

Brinton *et al.* (2014):

Studies behaviour in courses offered by MOOC provider during summer of 2013. State that social learning is a key element of scalable education on MOOC and transpires through online discussion forums, they want to understand forum activities. Two NB features: First is that there is a high decline rate - discussion begin with a lot of energy and then depletes over duration of course. Second is that discussion are *high-volume* and *noisy* (information overload), i.e. 30% or more of courses produced new discussion threads at rates making reading by students and teachers infeasible. Also, much discussion is off-topic.

Brinton *et al.* (2014) explore reasons for decline of activity on MOOC forums and find effective ways of classifying threads to rank their relevance. They use linear regression models to analyze forum activity and observe that, for example, teachers getting involved is correlated with increase in discussion volume, but does not affect depletion.

They propose a unified generative model for discussin threads, allowing them to choose efficient thread classifiers as well as design an effective algorithm to rank relevance.

They want to address information overload (which actually falls into field of information retrieval) by forming a simple model and thus improving the online learning experience. Contrary to IR, they want to highlight the unique characteristics of MOOC dynamics when compared to Yahoo!, Q&A and StackExchange or social media sites.

Their methodology for addressing information overload:

- First few days see a lot of small-talk in forums which need to be classified and filtered out.
- Small talk then fades away, thus need to rank relevance of new threads over time.

Therefore need effective classifier for discussion-thread and algorithm for ranking relevance.

“We propose a unified generative model for thread discussions that simultaneously guides (i) the choice of classifiers, (ii) the design of algorithms for extracting important topics in each forum, and (iii) the design of a relevance ranking algorithm based on the resulting topic extraction algorithm.”

“We crawled the forum content from Coursera’s server at a rate of 1 to 3 pages per second using Python and the Selenium library. Finally, we used Beautifulsoup to parse the html into text files. In total, our data set consists of approximately 830K posts (Section 3 presents more details).”

“Through our analysis, we presented a large-scale statistical analysis of a MOOC platform (Coursera), in which we made a number of interesting observations; for instance, that active participation of the teaching staff is associated with an increase in discussion volume but does not reduce the participation decline rate. We also presented two proof-of-concept algorithms for keyword extraction and relevance-ranking of discussion threads, each of which was demonstrated to be effective, through human evaluation when necessary.”

Stadtfeld *et al.* (2019):

“The findings underline the importance of understanding social network dynamics in educational settings. They call for the creation of university environments promoting the development of positive relationships in pursuit of academic success.”

(**???**):

“describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics.”

(**???**):

“This paper explores a simple and efficient baseline for text classification.”

(**???**):

(**???**):

“If a researcher seeks to see trends of programming language use, type of tools built, number and size of contributions and so on, the publicly available data can give solid information about the descriptive characteristics of the GitHub environment."

"We recommend that researchers interested in performing stud- ies using GitHub data first assess its fit and then target the data that can really provide information towards answering their research questions."

"Perhaps the biggest threat to validity to any study that uses GitHub data indiscriminately is the bias towards per- sonal use. While many repositories are being actively devel- oped on GitHub, most of them are simply personal, inactive repositories. Therefore, one of the most important ques- tions to consider when using GitHub data is what type of repository one's study needs and to then sample suitable repositories accordingly."

"While we believe there to be a need for research on the identification and automatic classification of GitHub projects according to their purpose, we suggest a rule of thumb. In our own experience, the best way to identify active software development projects is to consider projects that, during a recent time period, had a good balance of number of com- mits and pull requests, and have a number of committers and authors larger than 2."

"Based on our work, we believe a simple way to determine whether a repository actively works with another might be to identify if commits have flown from one to the other in both directions, but this strategy requires further validation."

(**???**):

"Our study shows that active GitHub committers ask fewer questions and provide more answers than others. Moreover, we observe that active StackOverflow askers distribute their work in a less uniform way than developers that do not ask questions. Finally, we show that despite the interruptions incurred, the StackOverflow activity rate correlates with the code changing activity in GitHub."

3 Datasets I Have Found

4 Empirical Methodology

4.1 Empirical Model

This is an equation:

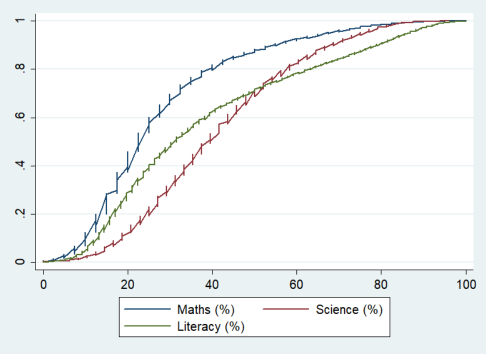
$$A_{it} = f(T_i^{(t)}, S_i^{(t)}, P_i^{(t)}, B_i^{(t)}, I_i), \tag{4.1}$$

Table 1: **Learner achievement (%)**

Subject	Mean	Q1	Median	Q3
Maths	27	15	23	35

Source: Own calculations in Stata using 2004 Grade 6 Intermediate Phase Systemic Evaluation.

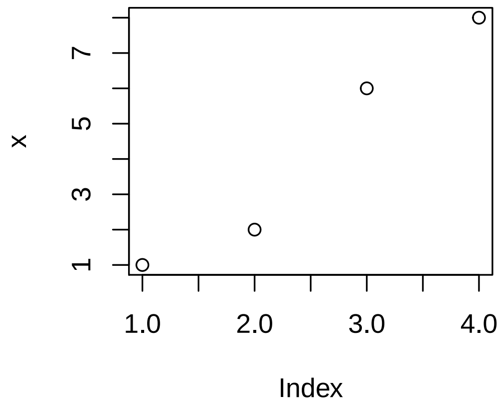
Figure 1: **Cumulative graph for subject scores**



Source: Own calculations in Stata using 2004 Grade 6 Intermediate Phase Systemic Evaluation.

Again, you can reference the figure in-text: figure 1 is a figure and it displays etc. etc. etc.

This is an organic figure generated with an R chunk that is executed when the document is knitted (the image is also saved in the folder `final-article-template_files`):



Regarding R chunks: If you want a chunk's code to be printed, include set `echo = TRUE`. `message = FALSE` stops R printing package loading details and setting `warning = FALSE` should suppress most warnings.

5 Recommendations for Further Research

6 Concluding Remarks

7 References

Brinton, C. G. *et al.* (2014) ‘Learning about social learning in MOOCs: From statistical analysis to generative model’, *IEEE Transactions on Learning Technologies.* IEEE, 7(4), pp. 346–359. doi: 10.1109/TLT.2014.2337900.

Stadtfeld, C. *et al.* (2019) ‘Integration in emerging social networks explains academic failure and success’, 116(3), pp. 792–797. doi: 10.1073/pnas.1811388115.

References