Enhanced clustering

Gianluca Bortoli
DISI - University of Trento
Student id: 179816
gianluca.bortoli@studenti.unitn.it

Martin Brugnara
DISI - University of Trento
Student id: 182904
martin.brugnara@unitn.it

ABSTRACT

Keywords

Big data, Data mining, URL categorization, topic extraction, geolocalized URLs, Latent Dirichlet Allocation

1. INTRODUCTION

2. RELATED WORK

3. PROBLEM DEFINITION

4. SOLUTION

5. CONCLUSIONS AND FUTURE WORK

6. REFERENCES

[1] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, and M. Zaharia. Spark sql: Relational data processing in spark. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 1383–1394, New York, NY, USA, 2015. ACM.

[2] D. Beazley. Understanding the python gil. 2010.

[3] A. B. Bondi. Characteristics of scalability and their impact on performance. In *Proceedings of the 2Nd International Workshop on Software and Performance*, WOSP '00, pages 195–203, New York, NY, USA, 2000. ACM.

[4] D. M. B. et al. Latent dirichlet allocation. *Journal of Machine Learning Research*, (3):993–1022, 1 2003.

[5] S. Gopalani and R. Arora. Article: Comparing apache spark and map reduce with performance analysis using k-means. *International Journal of Computer Applications*, 113(1):8–11, March 2015. Full text available.

[6] H. L. Lei Gu. Memory or time: Performance evaluation for iterative operation on hadoop and spark. *High Performance Computing and Communications*, 11 2013.

[7] A. Rajaraman and J. D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2011.