Capabilities and Costs of Running NLP Pipelines on Big Data Resources in Service-Oriented Architectures

Soheila Sahami
Leipzig University
sahami@ informatik.uni-leipzig.de

Thomas Eckart
Leipzig University
teckart@ informatik.uni-leipzig.de

Tag
Datum
Zeit
Raum

Natural Language Processing (NLP) is tending to analyze and to rely on the massive information which can be extracted from increasing input data(Boyd, D). Also the usage of tool chains in Service Oriented Architectures (SOAs) is an ongoing topic which gains more and more interest by NLP researchers and developers.

The implicit assumption of lack of reliable information about technical costs of executing processes -the amount of resources that are required to complete a NLP task in an acceptable time span- causes that most of the available Web-based NLP tool chains are hardly able to process "Big Data". In this contribution, we describe how both topics can be merged to combine a user-friendly SOA processing pipeline in NLP and Big Data technology. The final aim is to support flexible pipeline configuration with the goal to provide realistic run time estimations.

As the first step a variety of typical NLP tasks, including sentence segmentation, pattern-based text cleaning, tokenizing and language identification were implemented using Apache Spark (Karau, H. et al.) in a distributed and scalable environment, Hadoop. Using this approach, we could decrease the run times in comparison with a non-distributed implementation with comparable hardware configuration. In the second step (currently in progress) this back-end will be connected with the NLP processing environment WebLicht (Hinrichs, E.W. et al.) to allow the users to access these new services. Based on the real-world data the existing run time profiles will be continuously improved to allow realistic estimations of run times based on a variety of possible resource configurations. This will provide the end users a more transparent view on SOA environments, a more realistic view on actual costs of used services and helps them to make convenient decisions considering research endeavors and temporal constraints.

References: • Boyd, D. et al.(2012):Critical Questions for Big Data. *Information, Communication & Society* 15, 662–679. • Hinrichs, E.W. et al.(2010): WebLicht:Web-Based LRT Services for German. *ACL 2010 System Demonstrations*, 25–29. • Karau, H. et al.(2014):*Learning Spark:Lightning Fast Big Data Analytics*. O'Reilly Media.