*R*eport: Text Processing Fundamentals using MapReduce

Adam Domaslawski (12241331) Roman Furzykov (12326229) Elias Hirsch (12341086)

SS 2024

1 Introduction

This report aims to summarize the effects of the group work on the application of MapReduce techniques on large text corpora, Amazon Review Dataset 2014 containing 142.8 million reviews. To deal with such an extensive dataset, the Hadoop infrastructure is provided by TU Wien.

2 Problem Overview

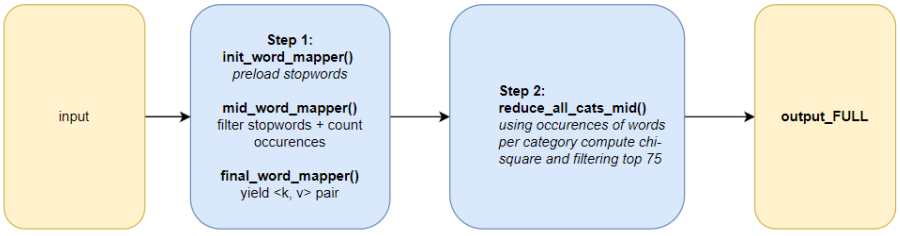
The problem presented to us, in essence, is a very trivial task in itself, computation of the chi$\hat{2}$ metric for words in texts within different categories. The information received may be used for further analysis or classification tasks. In our particular case, we used the Pearson's chi-squared test:

$$X^2_{t,c} = \frac{N(AD - BC)^2}{(A + B)(A + C)(B + D)(C + D)}$$

The task required a certain degree of efficiency, so the query itself should be optimal. The optimal execution time for the previous semester was around 20 min, so we used this time as a goal.

3 Methodology and Approach

The used MapReduce scheme was organized in the following two steps. The first step consists of four substeps. In first substep stopwords were preloaded. In second substep review is parsed, filtered with stopwords, category of review is parsed and review text as well. Categories were stored in dictioary as keys with dictionaries containing word counts as values. Then in third substep ¡key, value¿ pair is yielded where k is a category name and value words count dictionary. In the second step, a reducer was applied in which, using word count dictionaries for each category, the chi-sqaure statistic was computed.



4 Results and Conclusions

The obtained time of calculation is **7min 15 seconds**. This is the best so far result our team was able to achieve.