Big Data and Cloud Computing

2020-2021

Assignment 2

In this assignment you will construct an inverted index by using the Apache Spark framework. We are interested in the term-level inverted file, where the inverted lists store document IDs and term frequencies. For more details you may refer to the InvertedIndex.ipynb notebook of Lecture 3, Section 2.

The input dataset is again the posts.csv file that stores blog posts from The Unofficial Apple Weblog (TUAW). This file is attached in this assignment and was also distributed with the notebooks of Lecture 3. The inverted index will be constructed by taking into consideration only the titles of the blog posts, similarly to what we did in the InvertedIndex.ipynb notebook.