Ciencia de Dados em Larga Escala, 23/24

Inês Dutra and Zafeiris Kokkinogenis
DCC-FCUP
ines@dcc.fc.up.pt (room: 1.31)

Feb 15th, 2024

Practical Class # 1: Apache Beam

- Why apache beam?
  - allows for building machine learning pipelines
  - it hides optimization and implementation details

Practical Class # 1: Apache Beam

- Objectives of this class
  - understand the apache beam constructions: Pipeline, PCollection, Ptransform, ParDo and DoFn
  - understand the syntax and orchestration of these components
  - application of this knowledge to build a simple pipeline

Practical Class # 1: Apache Beam

- Tasks
  - Go to the Apache Beam Programming Guide and read about the components and how and where to use them in a program
  - Move to this notebook for a light introduction on how to create pipelines
  - Create your own pipeline that:
    - Takes a large text file as input (you can find some reasonably large text files here)
    - Parses each line into words
    - Performs a frequency count on the tokenized words
  - Compare the performance of this program with:
    - an ordinary sequential python implementation to solve the same problem
  - run the programs in:
    - your own machine