Homework2 Demo

Mahout vs Spark ML

Mahout	Spark ML
Mature(before version 0.9)	Rapidly Growing
Moving from HadoopMapReduce to Spark (after version 0.10)	Using Spark

DataFrame-based API vs RDD-based API

DataFrame-based	RDD-based
User-friendly Pipeline Structure	
Become primary ML API (after 2.0)	Enter maintenance mode (expected to be removed after Spark3.0)

“standard” procedure

- Data Cleaning
- Data Vectorization
- Fitting data into model
- Result

Q1

- ALS: Alternating Least Squares
- ALS on Implicit Feedback
- User-Based Collaborative Filtering (Not supported in spark ML)
- Item-Based Collaborative Filtering (Not supported in spark ML)

Q2

- Text Vectorization
- Feature Extraction(Word Count,TF-IDF)
- Data Cleaning

Q3

Useful python package

- nltk: natural language tool kit
- Sklearn: Machine Learning tools
- Panda: database manager