Contents

Contents

1. Mining Massive Datasets
   - Introduction
   - Google Page Rank

Contents

1. Mining Massive Datasets
   - Introduction
   - Google Page Rank
   - Recommendation systems
2. Spark disgression

Contents

1. Mining Massive Datasets
   - Introduction
   - Google Page Rank
   - Recommendation systems
2. Spark disgression

Contents

1. Mining Massive Datasets
   - Introduction
   - Google Page Rank
   - Recommendation systems

2. Spark disgression

Outline

1. Mining Massive Datasets
   - Introduction
   - Google Page Rank
   - Recommendation systems
2. Spark disgression
   - Main ideas
   - Summary

Content

- For this last course we want to present other topics more related to big data issues :
- New type of algorithms on graph data, or e-business data
- Introduce Spark as the state-of-the art framework for ML at scale
- Slides from several MOOCs, advanced materials

▸ Mining Massive Datasets book site

▸ Spark MOOC from Berkeley (several course available)

▸ www.coursera.org/learn/recommender-systems (very detailed)

Outline

1. Mining Massive Datasets
   - Introduction
   - Google Page Rank
   - Recommendation systems
2. Spark disgression
   - Main ideas
   - Summary

- Invented and deployed by Google
- Algorithm on graph data
-

Outline

1. Mining Massive Datasets
   - Introduction
   - Google Page Rank
   - Recommendation systems
2. Spark disgression
   - Main ideas
   - Summary

- Invented in the 90's on e-business context and content website with huge catalog
- Unsupervised ML
- From simple computation , Matrix decomposition to complex, state-of-the-art framework like this year Amazon open source dsstne (destiny)
  https ://github.com/amznlabs/amazon-dsstne

Contents

Contents

Contents

1. Mining Massive Datasets
2. Spark disgression

Contents

1. Mining Massive Datasets
2. Spark disgression
   - Main ideas

Contents

1 Mining Massive Datasets

2 Spark disgression
- Main ideas
- Summary

Outline

1. Mining Massive Datasets
   - Introduction
   - Google Page Rank
   - Recommendation systems
2. Spark disgression
   - Main ideas
   - Summary

- Overcome the map-reduce V1 limitations : lot of reading/writing on disks
- In-memory computation
- Rewriting of Mahout into MLLib.
- figure four pillars of Spark
- Eco-system

Outline

1. Mining Massive Datasets
   - Introduction
   - Google Page Rank
   - Recommendation systems

2. Spark disgression
   - Main ideas
   - Summary

ML Cake

From the Yann Le Cun's lesson at College de France

Reinforcement Learning (cherry)

- The machine predicts a scalar reward given once in a while.
- **A few bits for some samples**

Supervised Learning (icing)

- The machine predicts a category or a few numbers for each input
- **10→10,000 bits per sample**

Unsupervised Learning (cake)

- The machine predicts any part of its input for any observed part.
- Predicts future frames in videos
- **Millions of bits per sample**