

Big Data Architectures For Machine Learning and Data Mining

Long talk topics

May 19th 2016

Web Technology and Information Systems Group

Michael Völske

`michael.voelske@uni-weimar.de`

Long Talks Overview

General Points

- ❑ Work on a more complex, big data mining or machine learning problem over the course of several weeks
- ❑ Study the relevant scientific literature and use it to guide your approach
- ❑ Group work, 2-3 people per topic
- ❑ Hence, there will be 3 groups in this seminar

Long Talks Overview

List of Topics

1. On-the-fly Indexing of Large Document Collections
2. Learning Text Summarization From Social Media Data
3. Analyzing a Large Citation Network
4. Visual Page Segmentation

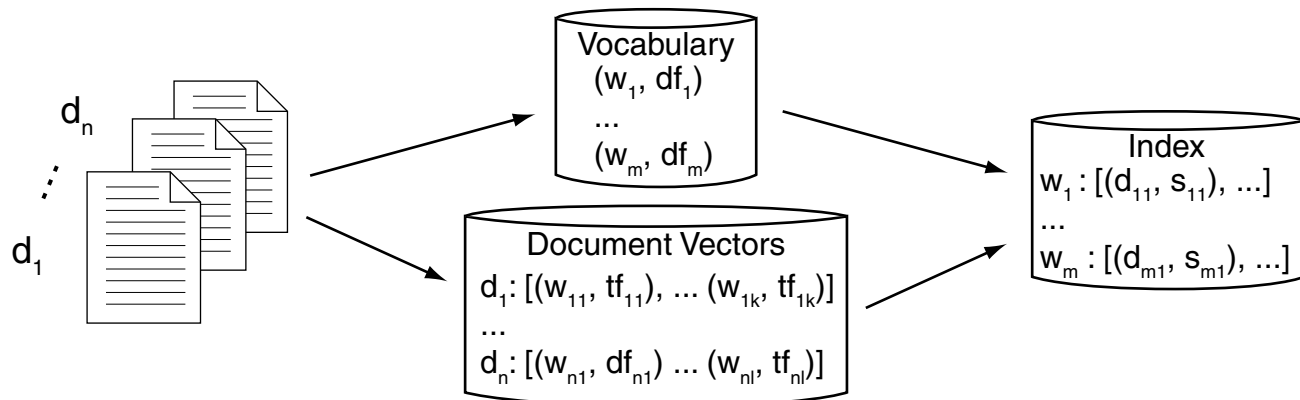
On-the-fly Indexing of Large Document Collections

The Task

- ❑ Take unstructured text dataset in HDFS
- ❑ Extract individual words, implement a term scoring function (e.g. TF-IDF)
- ❑ Create an in-memory inverted index on the cluster nodes
- ❑ Allow efficient querying with a basic UI (commandline or web)

The Data

- ❑ English Wikipedia Dump (~5 million pages)
- ❑ Must be pre-processed for this purpose (XML parsing, plaintext extraction)
- ❑ Optional: extend to larger datasets (e.g. ClueWeb)



Learning Text Summarization From Social-Media Data

The Task

- ❑ Long posts sometimes include a “Too Long; Didn’t Read” summary
- ❑ Can we use this to learn to better summarize text automatically?
- ❑ Extract training data from the Reddit corpus
- ❑ Review the state of the art in automatic summarization
- ❑ Implement a simple system of your own

The Data

- ❑ Reddit submissions (200 million) and comments (1.6 billion) 2006–2015
- ❑ Only a small sub-set contains “TL;DR” summaries
- ❑ The first step is to collect the relevant data



[–] **UserXYZ** 512 points 3 years ago*

A pizza place near me offers a free t-shirt if you eat an entire XXL pizza by yourself. My friend attempted the challenge and was about to quit with about a half of a slice left. This is when I started yelling at her to finish it for the t-shirt. She managed to finish the pie and won the shirt. 2 minutes later she sprinted outside and puked in front of the restaurant.

TL;DR Made my friend throw up for a free t-shirt.

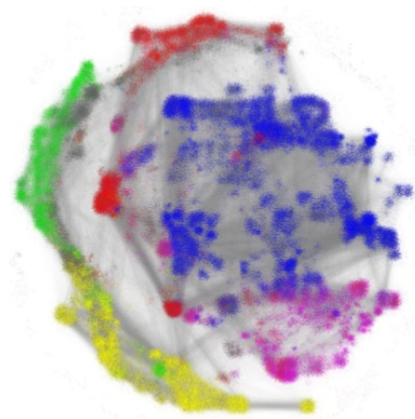
Analyzing a Large Citation Network

The Task

- ❑ Analyze, clean, and prepare a large graph dataset of scientific publications
- ❑ Learn the state of the art of large-scale graph processing
- ❑ Train a machine learning model to predict (for example) the number of citations that a publication will get
- ❑ Optional: investigate ways to visualize large graphs

The Data

- ❑ Microsoft Academic Graph
- ❑ > 100 million papers, 1 billion citations
- ❑ Data also includes information on authors, journals, conferences, and fields of study



Visual Page Segmentation

The Task

- ❑ You are given a page: segment it into content blocks
- ❑ Study already existing methods for segmenting printed documents (ICDAR workshop)
- ❑ Implement one such method
- ❑ Test it on images of web pages

The Data

- ❑ PRImA Layout Analysis Dataset (300 annotated magazine pages)
- ❑ Web pages (as many as you want)



Outlook

Organization

- ❑ Today: select topics and groups.
- ❑ Initial reading material and a data sample in the next few days.
- ❑ Access to betaweb cluster and full data soon; tutorial in 2 weeks.
- ❑ Work on the project until June 30th (Long talk presentations).
- ❑ Weekly meetings (Thu 17:00 or other appointment).
- ❑ Seminar paper due August 18th (two weeks after the exam period).

Long Talks

- ❑ About 30 minutes
- ❑ Present problem background and related work
- ❑ Dataset overview and stats
- ❑ Your approach and results
- ❑ Brief demo of your implementation