From Wikipedia to Thousands of Wikis

Scaling up DBpedia

Heiko Paulheim, Sven Hertling
Ingredients to the Team Project

- A hot topic
- A lot of data
- A challenging task
- A bunch of motivated, talented people (i.e.: you)
Ingredient 1: A Hot Topic
Ingredient 1: A Hot Topic

IBM big data platform

IBM is unique in having developed an enterprise-class big data platform that allows you to address the full spectrum of big data business challenges.

Understanding Big Data

Google BigQuery
Ingredient: A Hot Topic

• A lot of intelligent applications need access to massive knowledge
  – but they can't simply read the web or look up on Wikipedia as humans
  – so we need machine-accessible knowledge

• Examples for machine-accessible knowledge bases
  – DBpedia, Freebase, YAGO, Google Knowledge Graph, ...
Ingredient 2: A Lot of Data
Ingredient 2: A Lot of Data

• DBpedia 2016
  – The most recent release
  – Data extracted from Wikipedia in 100+ languages
  – More structured extraction backed by ontology in 28 languages
  – Also includes Wikidata

• English DBpedia
  – 6M entities (persons, organizations, places, ...)
  – 9.5B facts about those entities
Ingredient 2: A Lot of Data

- Wikipedia (English)
  - 5.3 million pages
  - Good coverage of head entities

- Wikifarms (e.g., Wikia)
  - 360,000 Wikis (!)
  - 43 million pages
  - Good coverage of tail entities
Ingredient 3: A Challenging Task

- Scaling up from on Wiki to thousands of Wikis
  - Distributed processing
  - Knowledge integration

- Specific tasks
  - Running extraction code on Wikifarm
  - Data profiling (i.e., understanding the data)
  - Data cleansing
  - Interlinking to DBpedia
Ingredient 4: You!

- ...a bunch of motivated, talented students
Ingredient 4: You!

• What do you have to expect
  – a challenging task
  – finding solutions on your own
  – working with unknown technologies
    (e.g., DBpedia Extraction Framework, large computing machines, …)
  – programming (yes, actual programming!)
  – working from a Linux console

• What can you take home
  – experience in Big Data (one of the hot topics in computer science!)
  – active contribution to one of the most famous Linked Open Data sets
  – working on cutting edge research topics
Technical Fact Sheet

• Participants: four students
  – Duration: one semester

• Language: German or English

• Requirements:
  – Programming skills
  – Knowledge in data mining, web mining, and/or data integration

• Suitable for both for Wifo and MMDS students
Questions?