Seminar Map/Reduce Algorithms on Hadoop

Topics

Alex, Christoph

Organisatorisches

- Prioritisierte Liste mit allen vorgestellten Themen bis heute 23:59 an Alexander.Albrecht@hpi.uni-potsdam.de
- Vergabe der Themen und Festlegen der Teams bis morgen, 21.04., 23:59 per E-Mail
- Nächstes Treffen am 27.04. als individuelle Teambesprechung im Raum A 1-7, ca. 30 Minuten pro Team
- Für alle Teams: Bis nächste Woche Hadoop lokal aufsetzen, Wordcount lokal und auf dem Cluster zum Laufen bringen

- Bewertung: Abschlusspräsentation, Beteiligung, Ausarbeitung (5 Seiten), Anwesenheit

Overview

- [1] (Similarity) Join --- Alex
- [2] Pic Latin, Cascading --- Alex
- [3] (K-Means) Clustering of DBPedia Subjects --- Christoph
- [4] Phrase Subsumption Computation --- Christoph
  - find occurrences of n-grams in texts
  - compute subsumptions
- [5] compute frequent item sets / association rules --- Alex
- [6] compute tf/idf --- Christoph

Similarity Join

Table R

Name	CITY
Christoph	Berlin
Prof. Felix Naumann	*Potsdam*
Alex	Berlin
…	…

Table S

Name	CITY
Felix Naumann	*Potsdam-Babelsberg*
Alex	Berlin
Christoph	Berlin
…	…

- Similarity-Join
  - $\{(r, s) \mid r \in R, s \in S, sim(r, s) \geq t\}$
- Aufgabe
  - Start: Test-Szenario entwerfen, z.B. Finden redundanter Websites, und auf Hadoop kopieren
  - Implementierung des Similarity Joins auf Hadoop

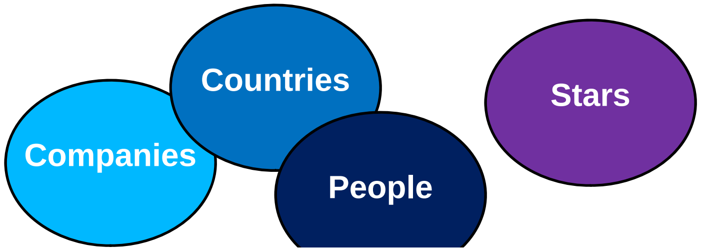
Evaluation
Pic Latin, Cascading, …

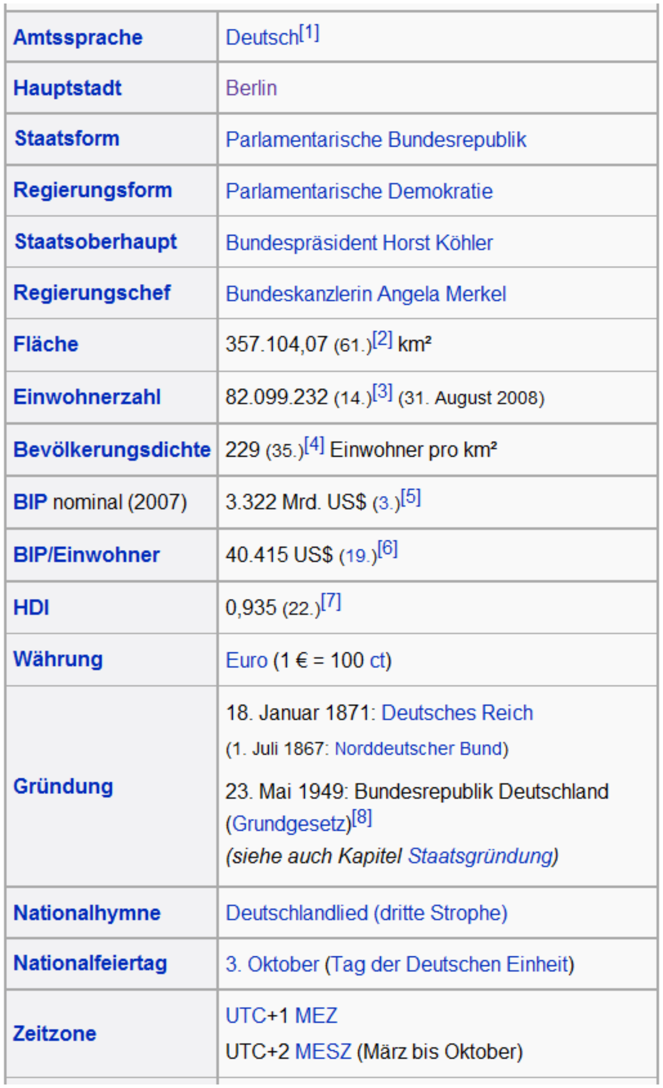
- Verarbeiten großer Datenmengen mit einer Skriptsprache statt Java
- Einfache Operatoren, z.B. `JOIN, FILTER, FOREACH, GROUP, PROJECTIONS`
- Weniger komplex als native Hadoop Java Code
- Beispiel WordCount

  -
    ```
    myinput = load '/user/hadoop01/demo.txt'
    words = FOREACH myinput GENERATE FLATTEN(TOKENIZE(*));
    grouped = GROUP words BY $0;
    counts = FOREACH grouped GENERATE group, COUNT(words);
    store counts into '/user/hadoop01/output' using PigStorage();
    ```
- Aufgabe
  - Start: Einrichten von PIG auf dem Hadoop Cluster & erste Evaluation, z.B. WordCount gegen nativen Hadoop Code testen
  - Umfassende Evaluation (Experimente) und Vergleich mit anderen Lösungen, z.B. Cascading

Clustering of DBPedia Subjects

- Christoph
- (Subject, Predicate, Value)
  - `[Deutschland, Amtssprache, Deutsch]`
  - `[Deutschland, Fläche, 357104]`
  - `[Deutschland, HDI, 0,935] …`
  - `[Engalnd, Fläche, 130395] …`
  - `[IBM, Umsatz, 103,6] …`
  - Cluster subjects based on predicte presence
  - Start: check out DBPedia Infoboxes, review k-means





Amtssprache	Deutsch[1]
Hauptstadt	Berlin
Staatsform	Parlamentarische Bundesrepublik
Regierungsform	Parlamentarische Demokratie
Staatsoberhaupt	Bundespräsident Horst Köhler
Regierungschef	Bundeskanzlerin Angela Merkel
Fläche	357.104,07 (61.)[2] km²
Einwohnerzahl	82.099.232 (14.)[3] (31. August 2008)
Bevölkerungsdichte	229 (35.)[4] Einwohner pro km²
BIP nominal (2007)	3.322 Mrd. US$ (3.)[5]
BIP/Einwohner	40.415 US$ (19.)[6]
HDI	0,935 (22.)[7]
Währung	Euro (1 € = 100 ct)
Gründung	18. Januar 1871: Deutsches Reich (1. Juli 1867: Norddeutscher Bund) 23. Mai 1949: Bundesrepublik Deutschland (Grundgesetz)[8] *(siehe auch Kapitel Staatsgründung)*
Nationalhymne	Deutschlandlied (dritte Strophe)
Nationalfeiertag	3. Oktober (Tag der Deutschen Einheit)
Zeitzone	UTC+1 MEZ UTC+2 MESZ (März bis Oktober)

Subsumption Computation

- Christoph
- 327.000 texts from PhenomicDB
- 10.000 phrases

1. locate phrases in the texts
2. Find out probability of occurrence of x when y is present

A genomewide search finds major susceptibility loci for gallbladder disease on chromosome 1 in Mexican Americans.

…

A high level of megakaryocyte colony stimulating activity, comparable to the levels present in sera from adults with aplastic anemia, was detected in the serum from the TAR infant.

…

Congenital anomalies also included facial capillary hemangiomata, intracranial vascular malformation, sensorineural hearing loss, and scoliosis.

...

- P(x|y) high ➔ x subsumes y / x is a superconcept of y
- Helps building term hierarchies
- Start: review “Deriving concept hierarchies from text”, Sanderson, 1999

Association Rule Computation

- Alex
- 327.000 texts from PhenomicDB
- 10.000 phrases

1. locate phrases in the texts
2. Determine support and confidence for all rules X ⇒ Y where |X|=|Y|=1

A genomewide search finds major susceptibility loci for gallbladder disease on chromosome 1 in Mexican Americans.

…

A high level of megakaryocyte colony stimulating activity, comparable to the levels present in sera from adults with aplastic anemia, was detected in the serum from the TAR infant.

…

Congenital anomalies also included facial capillary hemangiomata, intracranial vascular malformation, sensorineural hearing loss, and scoliosis.

…

- support and/or confidence X ⇒ Y high ➔ x is a superconcept of y
- Helps building term hierarchies
- Start: review “Fast Algorithms for Mining Association Rules”, Agrawal and Srikank, 1993

tf/idf Computation

- Christoph

$$tf.idf_d(t) = tf_d(t) * \log \frac{1}{df(t)}$$

- Measures the importance of a term with regard to the corpus
- For each term compute tf/idf for every document
- Helps identifying terms specific for a corpus

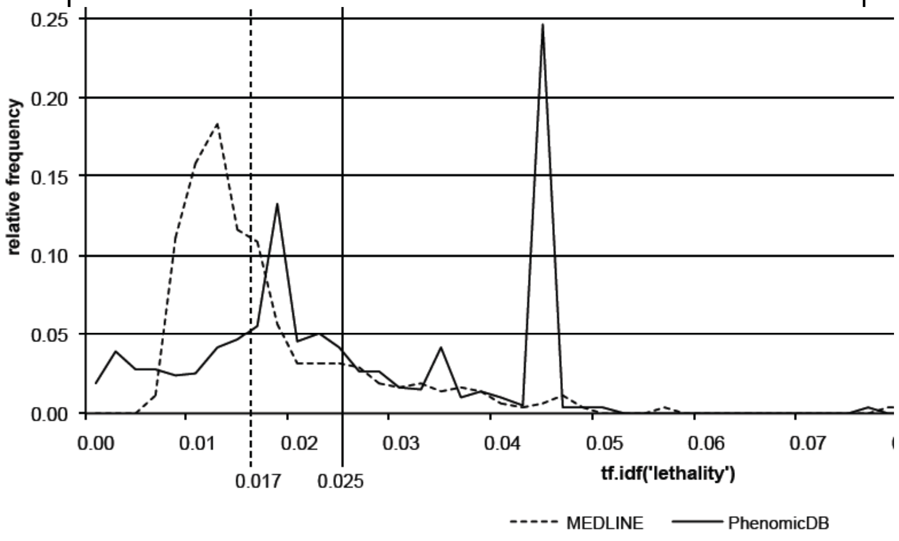
A genomewide search finds major susceptibility loci for gallbladder disease on chromosome 1 in Mexican Americans.

…

A high level of megakaryocyte colony stimulating activity, comparable to the levels present in sera from adults with aplastic anemia, was detected in the serum from the TAR infant.

…

Congenital anomalies also included facial capillary hemangiomata, intracranial vascular malformation, sensorineural hearing loss, and scoliosis.



End

Questions ?

- [1] (Similarity) Join --- Alex
- [2] Pic Latin, Cascading --- Alex
- [3] (K-Means) Clustering of DBPedia Subjects --- Christoph
- [4] Phrase Subsumption Computation --- Christoph
- [5] compute frequent item sets / association rules --- Alex
- [6] compute tf/idf --- Christoph