Algorithmen und Analyse auf bibliographischen Daten

peterr und Lusy

7. November 2011

Eigenschaften des Datensatzes

- enthält ca. 700 000 Einträge
- mit 19 verschiedenen Themengebieten(setSpecs) z.B. cs, math
- nur der Themenbereich Physik wird in Themengruppen aufgeteilt
- 11 Einträge haben keine Metadaten und damit keinen Titel, Autor usw.

Aufbau der Datensatz

Header

```
<identifier>oai:arXiv.org:0704.0001</identifier>
<datestamp>2008-11-26</datestamp>
<setSpec>physics:hep-ph</setSpec>
```

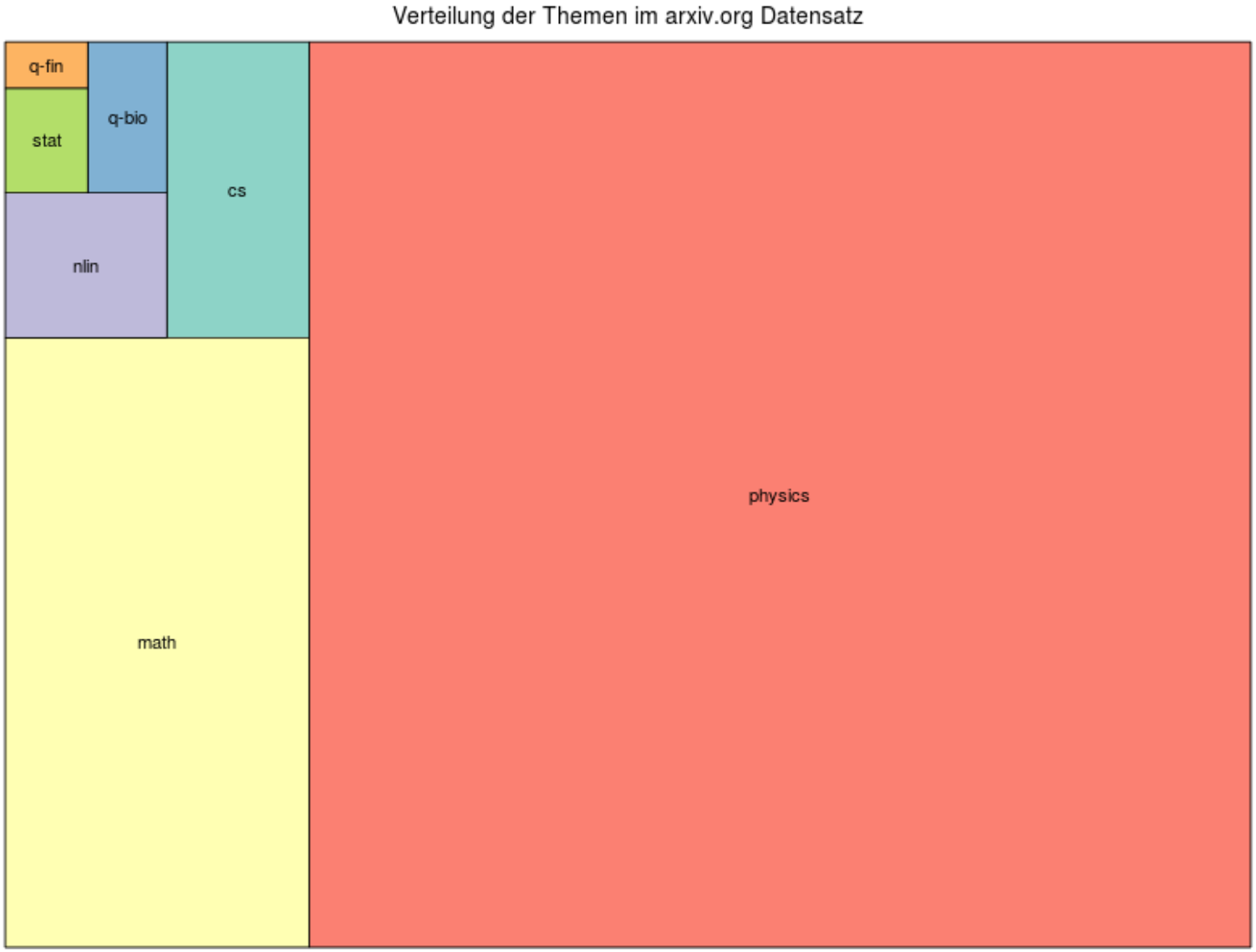
Metadaten

```
<dc:title>Titel des Papers</dc:title>
<dc:creator>Author 1</dc:creator>
<dc:creator>Author 2</dc:creator>
<dc:subject>..</dc:subject>
<dc:description>Beschreibung</dc:description>
<dc:description>Comment</dc:description>
<dc:date>2007-04-02</dc:date>
<dc:date>2007-07-24</dc:date>
<dc:type>text</dc:type>
<dc:identifier>http://arxiv.org/abs/0704.0001</
    dc:identifier>
<dc:identifier>Phys.Rev.D76:013009,2007</dc:identifier>
```

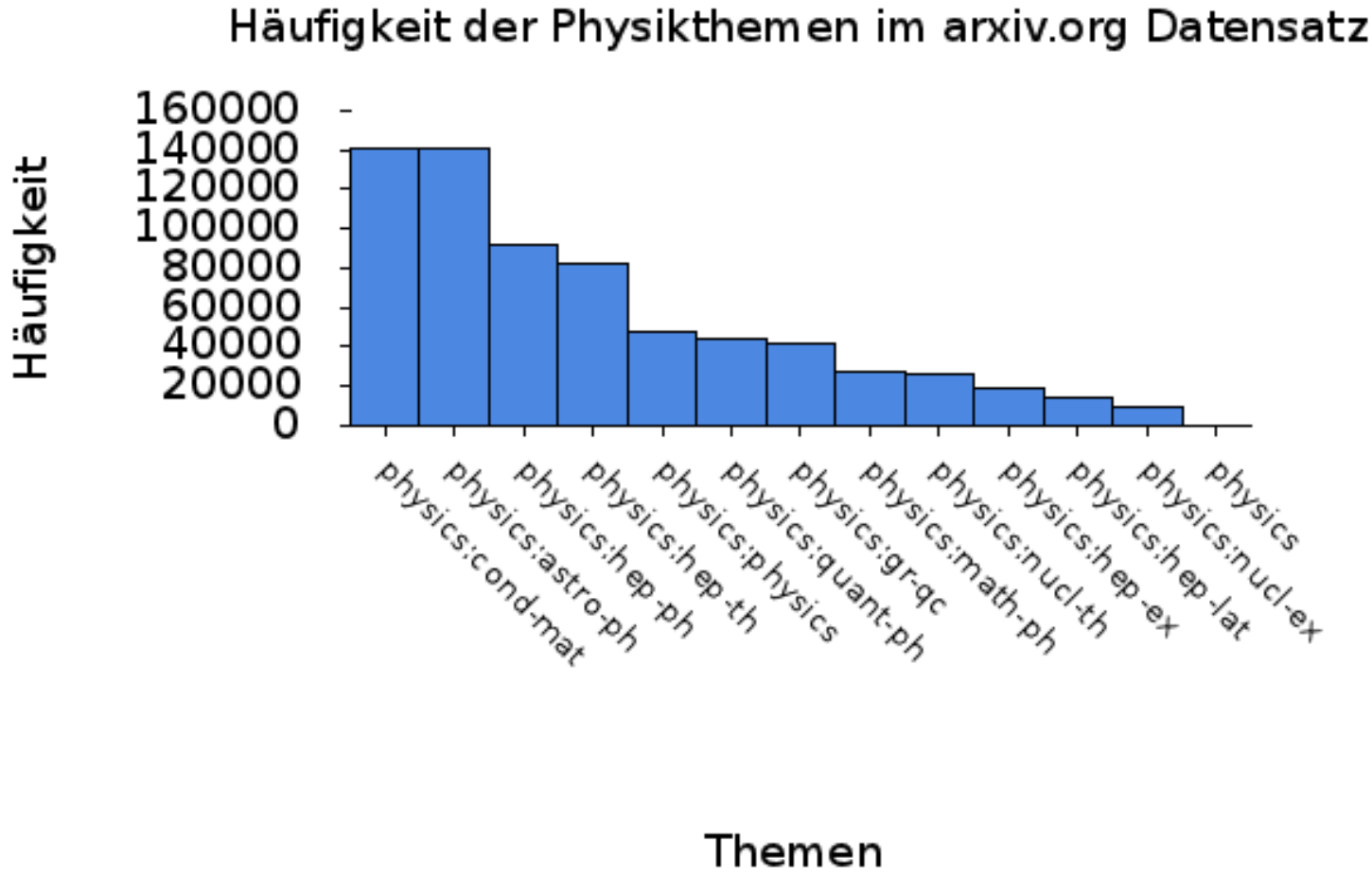
Parsen der Daten

- Parser in Python geschrieben
- kompletter Datensatz in den Speicher
- nur der Themenbereich Physik wird in Themengruppen aufgeteilt
- 11 Einträge haben keine Metadaten und damit keinen Titel, Autor usw.

Verteilung der Themen



Aufschlüsselung von physics



Häufigkeit von Themen pro Publikation

Assoziationsregeln auf setSpec

Probleme

Weitere Analysen