Maschinelle Sprachverarbeitung

Übung

Aufgabe 1: Analyse eines PubMed-Korpus

Mario Sänger

mario.saenger@informatik.hu-berlin.de

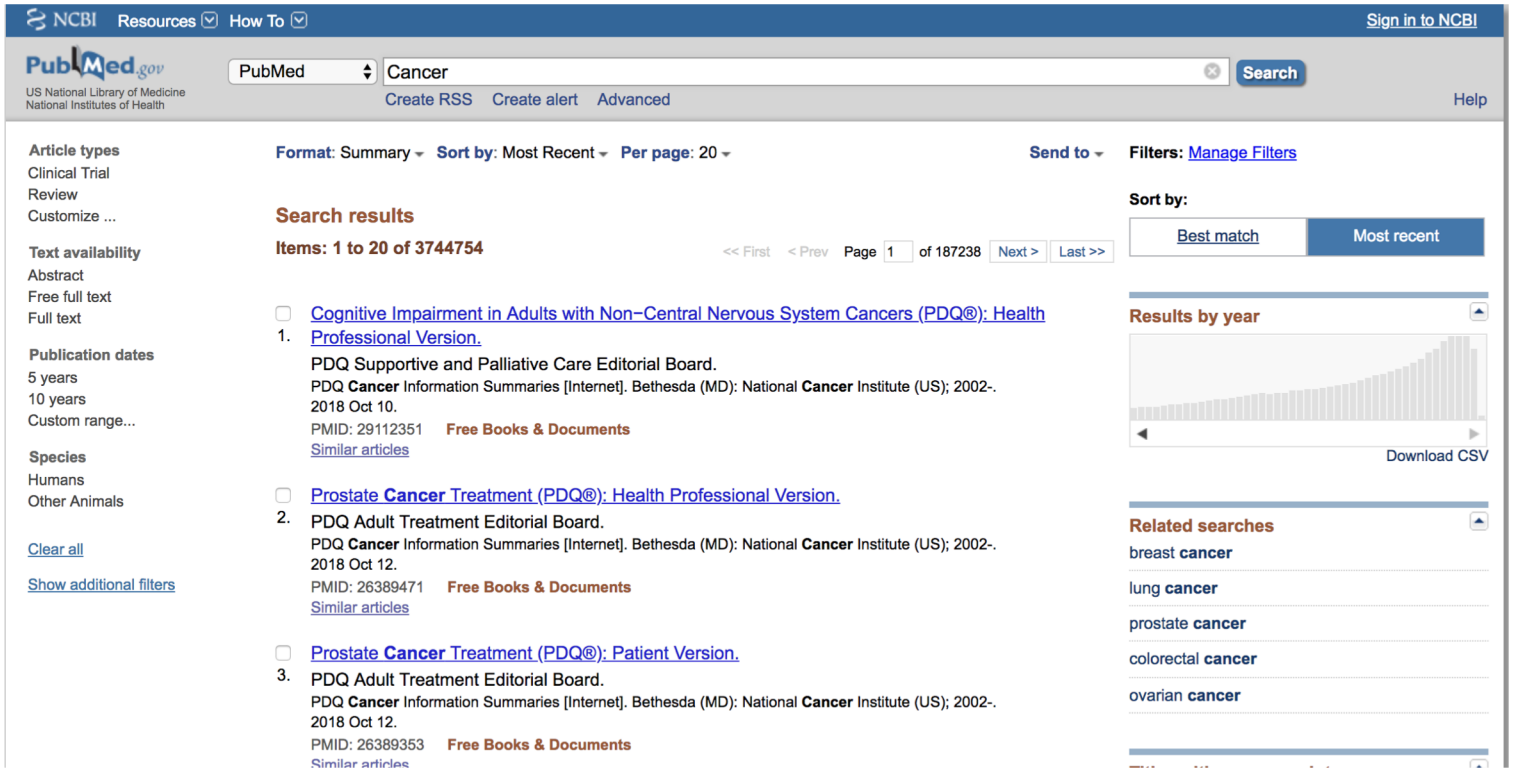
PubMed

- Suchmaschine für biomedizinische Artikel
  - Entwicklung und Pflege durch National Center for Biotechnology Information (NCBI)
  - Zugriff auf verschiedene (Meta-) Datenbanken
  - Dokumentation von medizinischen Artikeln in Fachzeitschriften

- PubMed umfasst über 28 Millionen Referenzen
  - Indexierung von über 5600 Fachzeitschriften
  - Wächst jährlich um ca. 500k Dokumente
  - Zugriff auf 13 Millionen Abstracts und 3,8 Millionen freie Volltexte

PubMed

- Zugriff via Webbrowser oder Programmierschnittstelle
  - https://www.ncbi.nlm.nih.gov/pubmed/



PubMed

- Indexierte Informationen:
  - Zuordnung einer eindeutigen ID zu jedem Artikel (PubMedID)
  - Angaben zum Journal und Veröffentlichungszeitpunkt
  - Titel des Artikels und ggf. Abstrakt oder Volltext
  - Verschlagwortung mittels Medical Subject Headings (MeSH) https://www.nlm.nih.gov/mesh/
  - Autorenname und Affiliationen
  - Weitere Metadaten (z.B. Sprache, Verarbeitungsangaben)

- Datenbereitstellung in einem XML-Format
  - https://www.nlm.nih.gov/bsd/licensee/data_elements_doc.html

Aufgabenstellung

- Entwicklung eines Programms zum Analysieren eines PubMed-Korpus
  - https://box.hu-berlin.de/f/19f49c6cf7604bd8bb65/?dl=1
- Implementierung eines eigenen XML-Parsers
  - SAX / DOM / Eigenentwicklung „from scratch“
  - Benutzung von Third-Party-Libraries zum Parsen von XML-Dateien (e.g. Jackson, Saxon) ist nicht erlaubt
  - Keine Validierung der XML-Struktur (mittels DTD) notwendig
- Analyse der Wortverteilung und Ermittlung verschiedener Kennzahlen

Kennzahlen

- Anzahl an Dokumenten
- Anzahl von Wörtern in ArticleTitle und AbstractText

```
<PubmedArticle>
    <MedlineCitation Status="MEDLINE" Owner="NLM">
        <PMID Version="1">16753163</PMID>
        <ArticleTitle>Identification of [...]</ArticleTitle>
        <Abstract>
            <AbstractText>Guanylyl cyclase C (GC-C) is a single
               transmembrane receptor for [...]
            </AbstractText>
        </Abstract>
    </MedlineCitation>
</PubmedArticle>
```

Kennzahlen

- Anzahl an Dokumenten
- Anzahl von Wörtern in ArticleTitle und AbstractText
  - Whitespaces (\s+) bilden Trennzeichen für Wörter
  - Alle Wörter in lowercase konvertieren
  - Anzahl an Wörtern insgesamt + Anzahl verschiedener Wörter („distinct“)
  - Keine separate Kennzahl für AbstractTitle und AbstractText!

- Die 50 häufigsten Wörter, deren Vorkommenshäufigkeit und (kumulativer) prozentualer Anteil am Gesamtkorpus

Kennzahlen

- Achtung: Mehrere AbstractText-Elemente mit zusätzlichen Attributen möglich

```
<Abstract>
    <AbstractText Label="OBJECTIVE" NlmCategory="OBJECTIVE">
      To test the hypothesis that the cumulative endometriosis [...]
    <AbstractText>
    <AbstractText Label="DESIGN" NlmCategory="METHODS">
       Retrospective cohort study including infertility [...]
    </AbstractText>
</Abstract>
```

Kennzahlen

- Anzahl an Artikeln je Journal + Anzahl an Artikeln je Jahr

```
<Journal>
    <ISSN IssnType="Print">0016-6480</ISSN>
    <JournalIssue CitedMedium="Print">
        <Volume>149</Volume>
        <Issue>1</Issue>
        <PubDate>
            <Year>2006</Year>
            <Month>Oct</Month>
        </PubDate>
    </JournalIssue>
    <Title>General and comparative endocrinology</Title>
    <ISOAbbreviation>Gen. Comp. Endocrinol.</ISOAbbreviation>
</Journal>
```

Kennzahlen

- Anzahl an Artikeln je MeSH-Heading
  - Kombination des Description-Eintrags mit Qualifier-Angabe
  - Verbindungszeichen „//“

```
<MeshHeadingList>
    <MeshHeading>
        <DescriptorName UI="D010802">Phylogeny</DescriptorName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName UI="D011506">Proteins</DescriptorName>
        <QualifierName UI="Q000378">metabolism</QualifierName>
    </MeshHeading>
</MeshHeadingList>
```

=> „Phylogeny“ und „Proteins//metabolism“

Kennzahlen

- Anzahl an Artikeln je MeSH-Heading
  - Kombination des Description-Eintrags mit Qualifer-Angabe
  - Verbindungszeichen „//“

```
<MeshHeadingList>
    <MeshHeading>
        <DescriptorName UI="D011248">
           Pregnancy Complications
        </DescriptorName>
        <QualifierName UI="Q000209">etiology</QualifierName>
        <QualifierName UI="Q000601">surgery</QualifierName>
    </MeshHeading>
</MeshHeadingList>
```

=> „Pregnancy Complications//etiology“ und „Pregnancy Complications//surgery“

Abgabedetails

- Aufgabe muss mit Java oder einer Java-VM kompatiblen Sprache gelöst werden (z.B. Scala, Kotlin, ..)
- Programm lässt sich wie folgt starten:

```
java -jar uebung1-gruppeX.jar ordner/
```

- Programm liest und analysiert alle XML-Dateien (*.xml) im übergebenen Ordner
- Ausgabe der Ergebnisse in die Standardausgabe (stdout)

Programmausgabe (Beispiel)

1. Artikelanzahl, Wortanzahl (insgesamt/distinct) und Top-50 Wörter (inkl. kumulativen Anteil)

```
Artikel: 60000

Wörter: 10477600 (491354 distinct)
the : 531489 (5,07 %)
of : 456729 (4,36 %)
and : 376277 (3,59 %)
[...]
cell : 13112 (0,13 %)
Top 50 : 3515229 (33,55 %)
```

Programmausgabe (Beispiel)

2. Anzahl an Jahresangaben (insgesamt/distinct) und Top-10 Jahre (inkl. kumulativen Anteil)

```
Jahresangaben: 57739 (53 distinct)
2017 : 18428 (31,92 %)
2014 : 16525 (28,62 %)
2015 : 11262 (19,51 %)
2016 : 1434 (2,48 %)
[...]
1996 : 397 (0,69 %)
1992 : 389 (0,67 %)
1991 : 369 (0,64 %)
Top 10 : 50047 (86,68 %)
```

Programmausgabe (Beispiel)

3. Anzahl an Journalangaben (insgesamt/distinct) und Top-10 Journale (inkl. kumulativen Anteil)

```
Journale: 60000 (4701 distinct)
Oecologia : 7295 (12,16 %)
PloS one : 1287 (2,15 %)
Scientific reports : 673 (1,12 %)
[...]
Methods in molecular biology (Clifton, N.J.) : 348 (0,58 %)
Kardiologiia : 337 (0,56 %)
Top 10 : 12429 (20,72 %)
```

Programmausgabe (Beispiel)

4. Anzahl an MeSH-Angaben (insgesamt/distinct) und Top-10 MeSHs (inkl. kumulativen Anteil)

```
MeSH: 426314 (69839 distinct)
Humans : 22406 (5,26 %)
Female : 11887 (2,79 %)
Male : 11516 (2,70 %)
Animals : 6861 (1,61 %)
[...]
Adolescent : 2544 (0,60 %)
Treatment Outcome : 2032 (0,48 %)
Top 10 : 77006 (18,06 %)
```

Abgabe

- Upload eines ZIP-Archivs *uebung1-gruppeX.zip*
  - Ausführbares JAR-Archiv und Quellcode des Programms
  - Feedback zur benötigten Zeit

- Testet das Programm vor der Abgabe auf gruenau!
  - Verwendet (auch) kleinere Teil- oder Testkorpora

- Abgabe bis spätestens 04.11.2018, 23:59 Uhr über:
  - https://hu.berlin/ue_masprach1819_ass1

Wettbewerb

- Implementiert eine schnelle (und korrekte!) Lösung
- Wir messen die Laufzeit des Programms mit "time"
  - Parallelisierung macht sicherlich Sinn
- Die drei besten Teams bekommen 3/2/1 Punkte!

RoadMap nächste Wochen

- Gruppenbildung bis 22.10.2018
  - Per Email an mario.saenger@informatik.hu-berlin.de
- Optionale Übung am 25.10.2018
  - Rückfragen zur ersten Aufgabe + sonstige Fragen
- **Achtung: Übung am 01.11.2018 entfällt!**
  - Rückfragen per Mail aber jederzeit möglich
- Abgabe 1. Übungsaufgabe bis 04.11.2018, 23:59 Uhr