Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Projektseminar

www.ProminentPeople.info

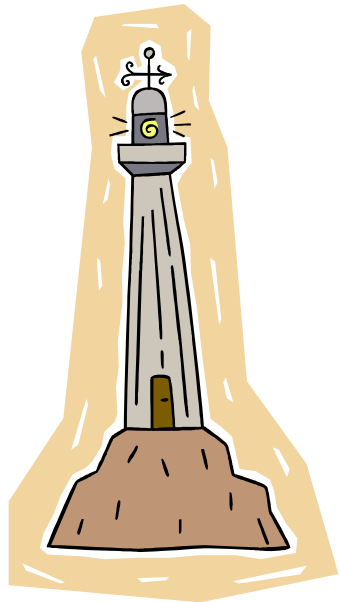
16. Oktober 2007

Jana Bauckmann

Alexander Albrecht

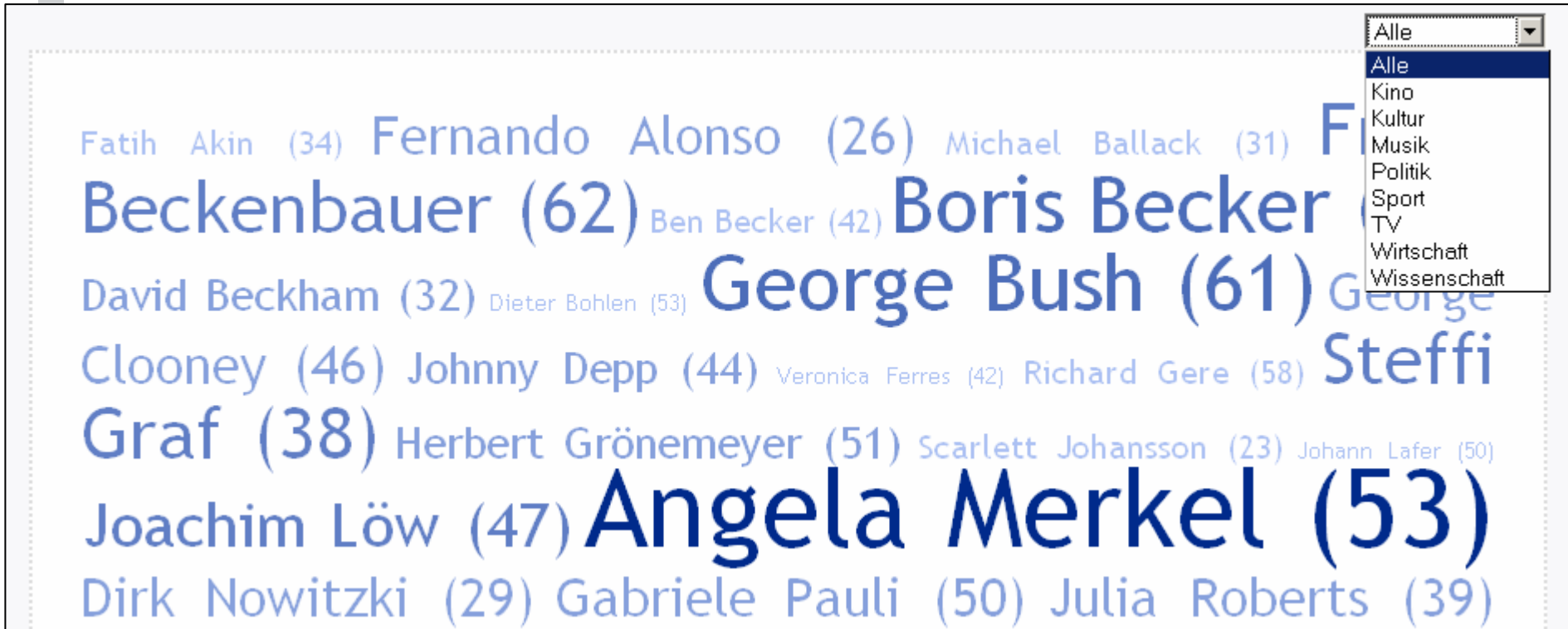
Überblick

- Ziele des Projektseminars
- Rahmenbedingungen
- Projektüberblick
  - Ausgangspunkt
  - Erste Schritte
  - www.ProminentPeople.info
- Projektplanung / Aufgabenteilung
- Vortragsthemen
- Ausblick / mögliche Erweiterungen
- Literatur
- Organisatorisches



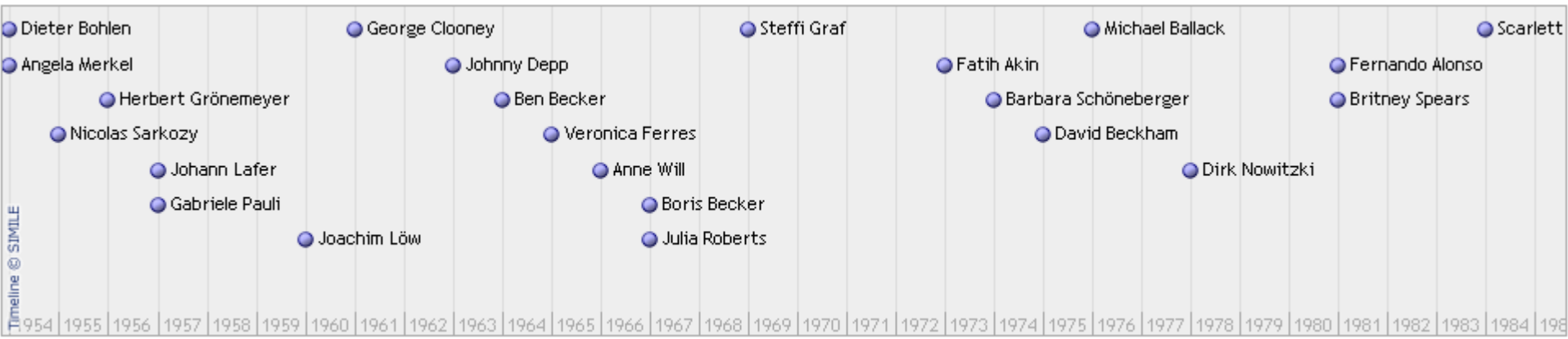
Ziele des Projektseminars

- Extraktion von Prominenten (Alter, Titel) aus News Texten – **N**amed **E**ntity **R**ecognition (NER)
- Visualisierung der Ergebnisse auf www.ProminentPeople.info als Name Cloud



Ziele des Projektseminars

- Datenanalyse, insbesondere für Altersangaben
  - Anfrage nach Alter (mit Widersprüchen)
  - Anfrage nach Geburtstag (auch aus statistischer Analyse)
  - Visualisierung über Zeitleiste, http://simile.mit.edu/timeline/



Rahmenbedingungen

- Regelmäßiger Termin am Dienstag, 11:00 - 12:30 Uhr, Raum A-1.1
- Bearbeitung der Aufgaben in Teams von zwei Studenten
- pro Termin
  - ein Vortrag (ca. 30 Minuten) – jeder ist mal dran
  - Für alle Gruppen:
    Vorstellung und Diskussion der bisherigen Ergebnisse
- 2 eingeladene Vorträge
  - Prof. Ulf Leser, Humboldt-Universität zu Berlin:
    Named Entity Recognition in der Bioinformatik
  - Alexander Löser, SAP Research:
    Unstructured Information Management
- Voraussetzungen: Java und SQL von Vorteil; mind. 3. Semester

Ausgangspunkt

- Datenbank mit Online-News, täglich aktualisiert

ID	NEWS	DATE	SOURCE	URL
1139	Charlize Theron zur attraktivsten Frau der Welt gekürt - Yahoo	Oct 10, 2007	Yahoo	http://de.news.yahoo.com/ap/20071010/ten-charlize-thero...
1140	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//E...	Oct 10, 2007	Yahoo	http://de.news.yahoo.com/ap/20071010/ten-armin-mueller-...
1141	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//E...	Oct 10, 2007	Yahoo	http://de.news.yahoo.com/ddp/20071010/ten-raucher-bezi...
1142	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//E...	Oct 10, 2007	Yahoo	http://de.news.yahoo.com/ap/20071010/ten-juanes-und-ca...
1143	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//E...	Oct 10, 2007	Yahoo	http://de.news.yahoo.com/ap/20071010/ten-fhrmann-findet...
1144	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//E...	Oct 10, 2007	Yahoo	http://de.news.yahoo.com/dpa2/20071010/ten-david-hass...
1145	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//E...	Oct 10, 2007	Yahoo	http://de.news.yahoo.com/ap/20071010/ten-unfall-clooney...
1146	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//E...	Oct 10, 2007	Yahoo	http://de.news.yahoo.com/ap/20071010/ten-moritz-bleibtre...
1147	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//E...	Oct 10, 2007	Yahoo	http://de.news.yahoo.com/dpa2/20071010/ten-kiefer-suthe...
1148	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Tr...	Oct 10, 2007	Focus	http://www.focus.de/panorama/welt/cleveland_aid_13548...
1149	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Tr...	Oct 10, 2007	Focus	http://www.focus.de/panorama/welt/kinderschaender_aid_...
1150	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Tr...	Oct 10, 2007	Focus	http://www.focus.de/panorama/welt/tier-tragoedie_aid_135...
1151	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Tr...	Oct 10, 2007	Focus	http://www.focus.de/panorama/boulevard/kollaps_aid_135...
1152	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Tr...	Oct 10, 2007	Focus	http://www.focus.de/panorama/welt/australien_aid_13538...
1153	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Tr...	Oct 10, 2007	Focus	http://www.focus.de/panorama/boulevard/new-york_aid_1...
1154	<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Tr...	Oct 10, 2007	Focus	http://www.focus.de/panorama/welt/tid-7611/justiz_aid_13...
1155	<!-- stern.de --><!DOCTYPE html PUBLIC "-//W3C//DTD X...	Oct 10, 2007	Stern	http://www.stern.de/lifestyle/leute/:David-Hasselhoff-Einer/...
1156	<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Tran...	Oct 10, 2007	Gala	http://www.gala.de/stars/news/index.html?id=9750?100ct...
1157	<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Tran...	Oct 10, 2007	Gala	http://www.gala.de/stars/news/index.html?id=9701?100ct...
1158	<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Tran...	Oct 10, 2007	Gala	http://www.gala.de/stars/news/index.html?id=9699?100ct...
1159	<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Tran...	Oct 10, 2007	Gala	http://www.gala.de/stars/news/index.html?id=9698?100ct...
1160	<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Tran...	Oct 10, 2007	Gala	http://www.gala.de/stars/news/index.html?id=9696?100ct...
1161	<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Tran...	Oct 10, 2007	Gala	http://www.gala.de/stars/news/index.html?id=9694?100ct...
1162	<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Tran...	Oct 10, 2007	Gala	http://www.gala.de/stars/news/index.html?id=9692?100ct...

Ausgangspunkt

- Datenbank mit Prominenten (Wikipedia)
  - über 92.000 Einträge
  - http://tools.wikimedia.de/
- Java-Sourcen für den Zugriff auf die Datenbank

Erste Schritte (NER)

- Parsen der News
- News in Wörter (Token) zerlegen
- Namen in den News finden
  - Listenbasiertes NER
- Prominente finden
- Altersangaben finden
  - NER mit regulären Ausdrücken

Scarlett Johansson mag Kalorien:
Films
sich
nicht
verde
BUNTE

Scarlett Johansson fliegt in den
Irak
Joha
Fuß
eins
BUNT

Scarlett geht durch eine schwierige Zeit: Hollywood-Schauspielerin Scarlett Johansson hat angeblich ihre New Yorker Nachbarn mit zu starkem Zigarettenrauch verärgert ...
BUNTE.T-Online.de Newsline, 28.11.06

```
Affäre, Amber, Apartment, Beziehung,
Freund, Gerüchten, Hartnett, Hollywood,
Johansson, Josh, Kettenrauchen, Krise,
Nachbarn, New, Sainsbury, Scarlett,
Schauspielerin, Yorker, Zeit,
Zigarettenkonsum, Zigarettenrauch
```

```
Amber, Hartnett, Johansson, Josh,
Sainsbury, Scarlett
```

```
Josh Hartnett (29)
Scarlett Johansson
```

Erste Schritte (NER) - Schwierigkeiten

- Parsen der News
  - Format HTML
  - ...
- News in Wörter (Token) zerlegen
  - Regeln zur Tokenisierung
  - ...
- Namen, Prominente, Alters-angaben finden
  - SQL
  - Reguläre Ausdrücke
  - ...

Scarlett Johansson mag Kalorien:
Films
sich
nicht
verde
BUNTE

Scarlett Johansson fliegt in den
Irak
Joha
Fuß
eins
BUNT

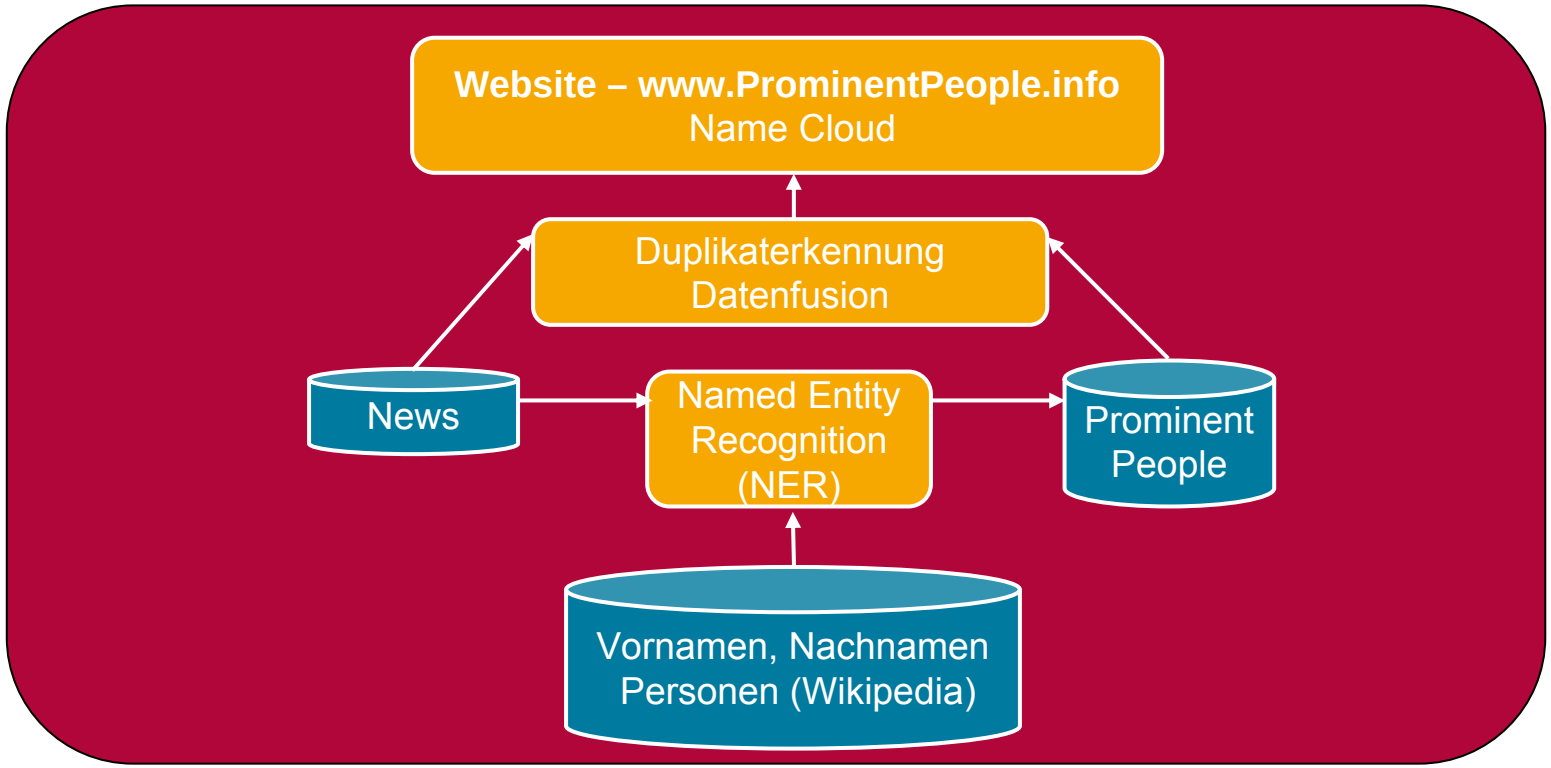
Scarlett geht durch eine schwierige Zeit: Hollywood-Schauspielerin Scarlett Johansson hat angeblich ihre New Yorker Nachbarn mit zu starkem Zigarettenrauch verärgert ...
BUNTE.T-Online.de Newsline, 28.11.06

```
Affäre, Amber, Apartment, Beziehung,
Freund, Gerüchten, Hartnett, Hollywood,
Johansson, Josh, Kettenrauchen, Krise,
Nachbarn, New, Sainsbury, Scarlett,
Schauspielerin, Yorker, Zeit,
Zigarettenkonsum, Zigarettenrauch
```

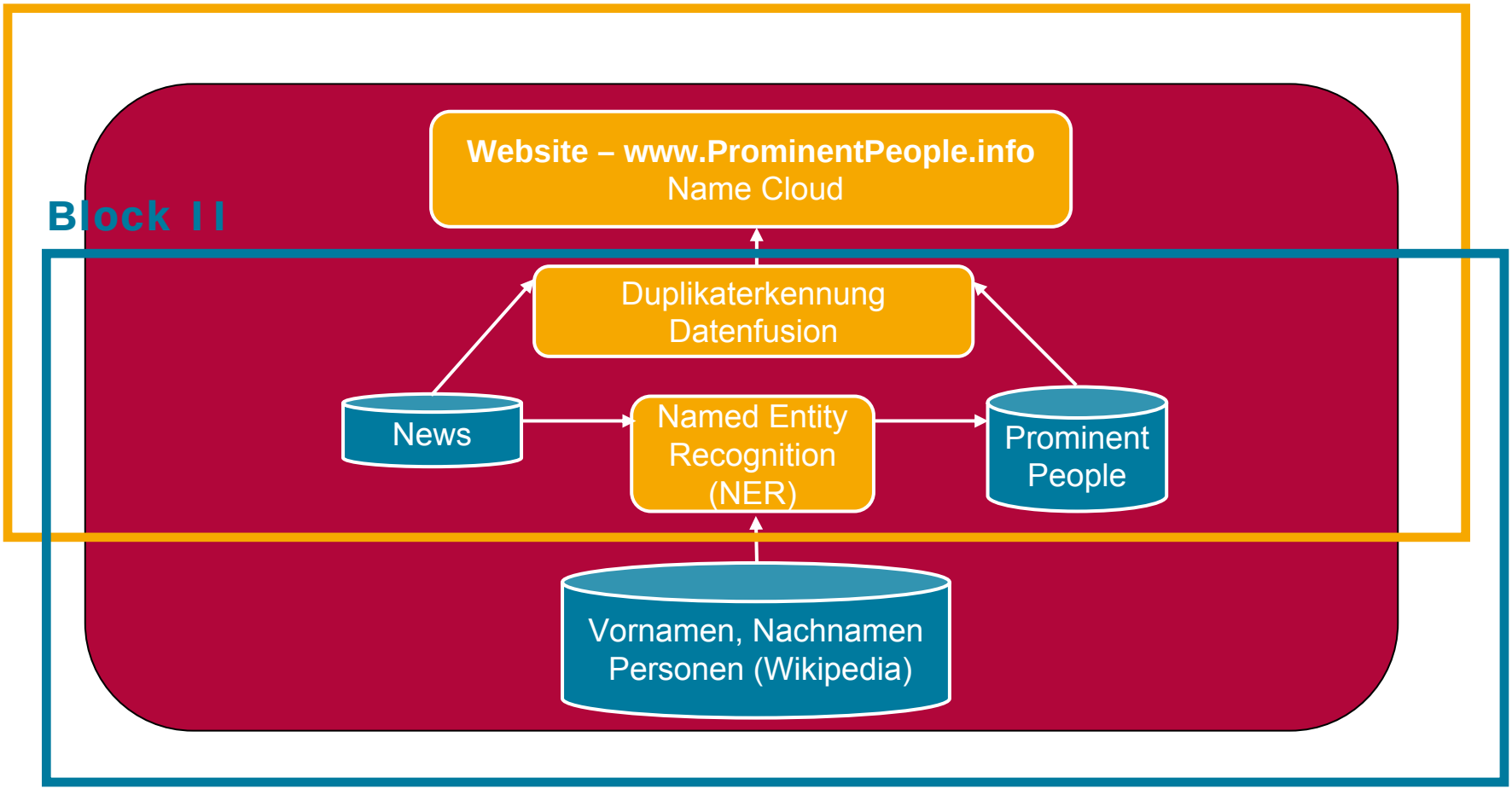
```
Amber, Hartnett, Johansson, Josh,
Sainsbury, Scarlett
```

```
Josh Hartnett (29)
Scarlett Johansson
```

Nächste Schritte
www.ProminentPeople.info



Projektplanung



Projektplanung - Aufgabenverteilung

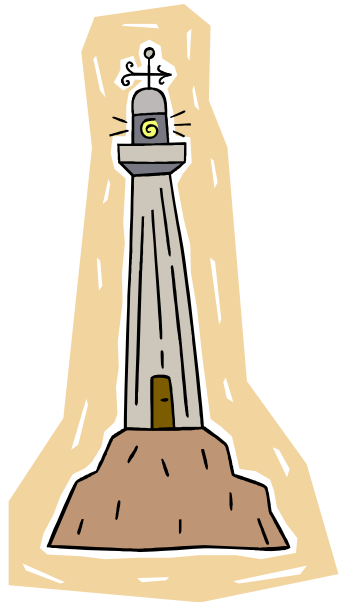
- Teams von 2 Studenten bearbeiten Block I oder Block II
- Block I & Block II
  - Listenbasierte NER-Techniken auswählen und implementieren
  - Duplikaterkennung & Datenfusion der gefundenen Personendaten
  - Zusatzinformationen wie das Alter, den Titel (Diplom/Doktor/Professor) finden
- Block I
  - Darstellung als Name Cloud, …
- Block II
  - Vornamen, Nachnamen, Personen, Geburtsdatum, Personen-Kategorien, … aus Wikipedia extrahieren und in die Datenbank schreiben

Vortragsthemen

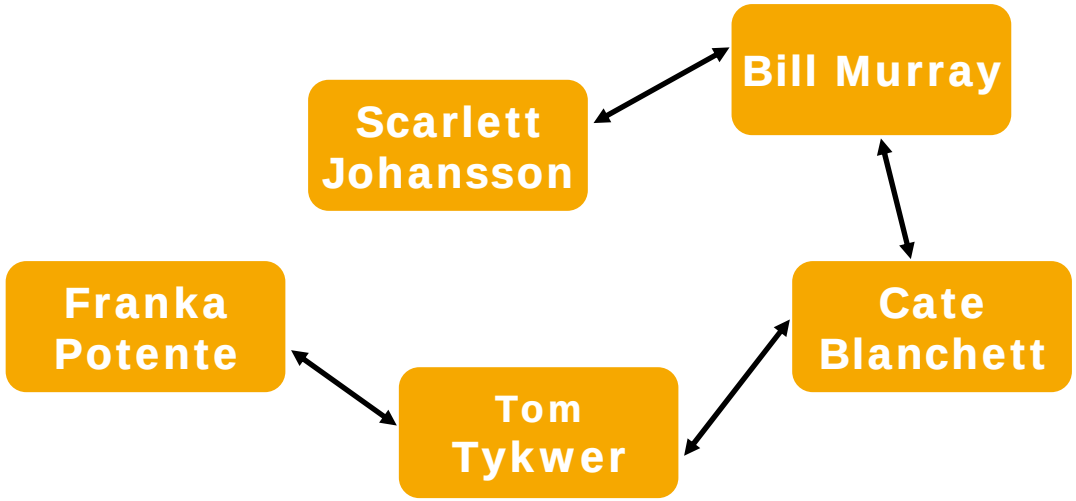
- SQL / JAVA / Web
  - Jana & Alex am 23.Oktober 2007
- SQL für DB2
  - 30. Oktober
- Datenbankentwurf & Datenbankzugriff mit JDBC
  - 06. November
- Reguläre Ausdrücke mit Java
  - 13. November
- Crawling the Web
  - 20. November
- Listenbasierte NER-Verfahren
  - 27. November

Ausblick

- Entdecken von Personenbeziehungen
- Geocodierte News-Feeds
- ProminentPeople-Webservice



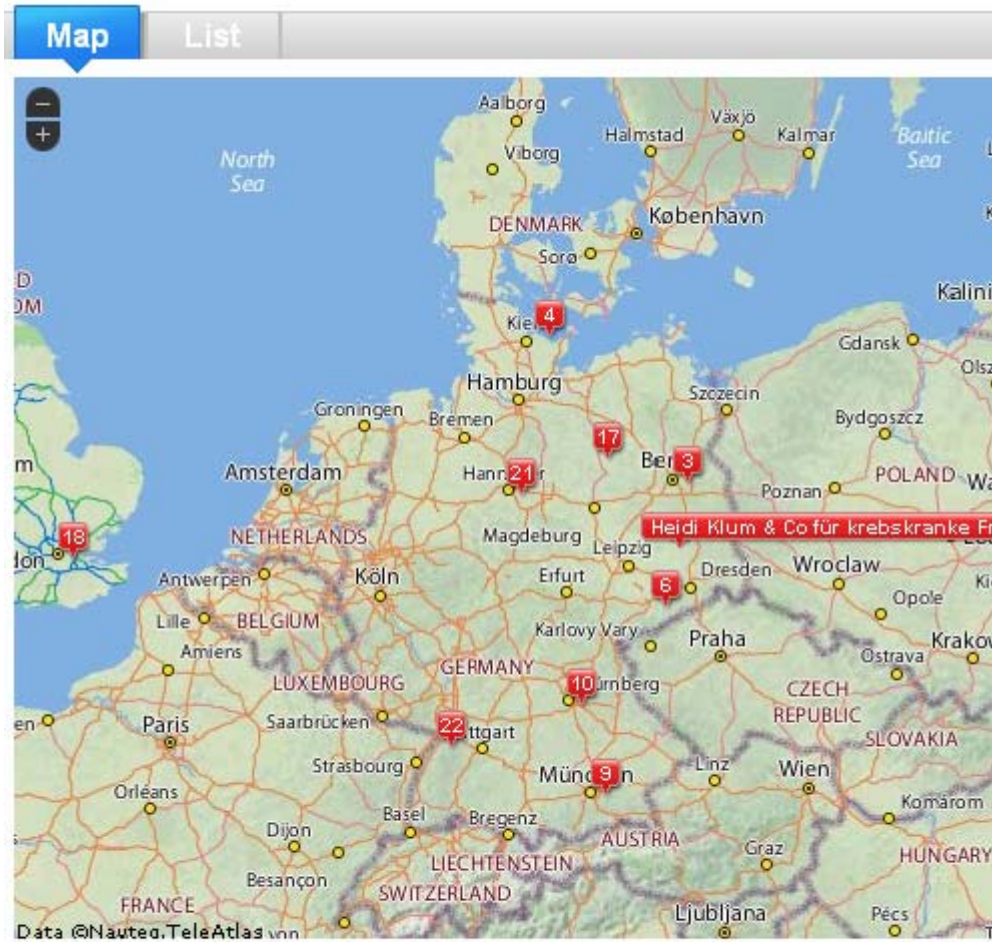
Entdecken von Personenbeziehungen



- Mögliches Vorgehen
  - Verbindung zwischen zwei Prominenten finden (*kennt*-Relation)
  - Transitive Hülle der *kennt*-Relation ermitteln
  - Transitive Hülle als Datenbank-Sicht speichern
  - SQL-Anfragen auf der transitiven Hülle formulieren

Geocodierte News-Feeds

- Orte mit hoher Prominenten-dichte finden
- Webservice GeoNames verwenden
  - RSS-Einträgen Geo-Koordinaten zuweisen
  - RSS to GeoRSS Conversion
  - www.GeoNames.org



ProminentPeople-Webservice

- Datenbank-basierter Webservice
  - RSS to ProminentRSS Conversion
  - Prominente in RSS-Einträgen finden und Name mit Zusatzinformationen (Geburtsdatum,...) in das RSS-Feed einbinden.
  - Zugriff auf neue News-Quellen/RSS-Feeds durch Webservice-Aufruf
  - Verknüpfung mit Yahoo Pipes/Flickr

Literatur

- Mining the Web: Discovering Knowledge from Hypertext Data, Soumen Chakrabarti, ISBN: 1558607544
- Reguläre Ausdrücke, Jeffrey E. F. Friedl, ISBN: 3897213494
- Text Mining: Predictive Methods for Analyzing Unstructured Information, Sholom M. Weiss, ISBN: 0387954333

Organisatorisches

- Leistungserfassung
  - Implementierung in Java zum gewählten Block + ca. 5 Seiten Dokumentation
  - 30-minütiger Vortrag zum gewählten Thema + Diskussion
  - Live-Demo
- Benotung für Projekterfolg, Vortrag und Dokumentation (3 ECTS Credit Points)

- Zum Auswahlverfahren:
  - Rückmeldung per mail an uns (beide!) bis 17.10., 18:00 Uhr,
  - mit Angabe des gewünschten Blockes
  - dann: blockweises Losverfahren
  - Nachricht per mail bis 18.10.