

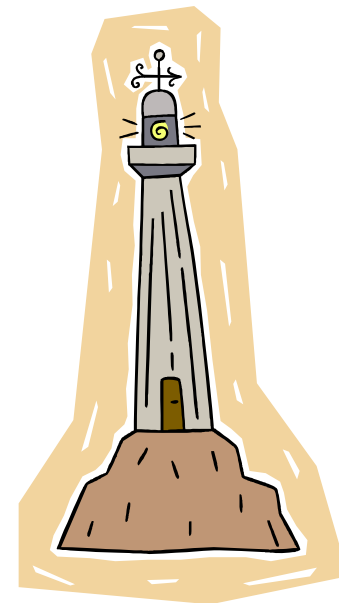
Projektseminar
www.ProminentPeople.info

16. Oktober 2007
Jana Bauckmann
Alexander Albrecht

Überblick

2

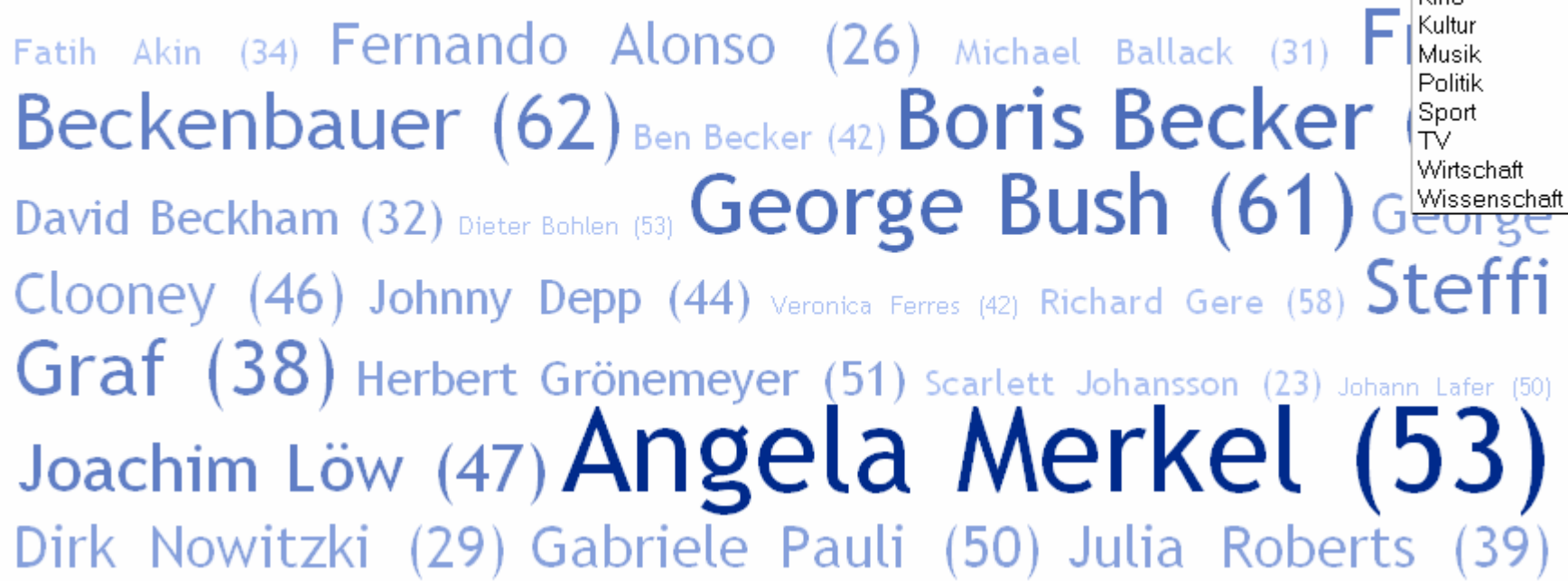
- Ziele des Projektseminars
- Rahmenbedingungen
- Projektüberblick
 - Ausgangspunkt
 - Erste Schritte
 - www.ProminentPeople.info
- Projektplanung / Aufgabenteilung
- Vortragsthemen
- Ausblick / mögliche Erweiterungen
- Literatur
- Organisatorisches



Ziele des Projektseminars

3

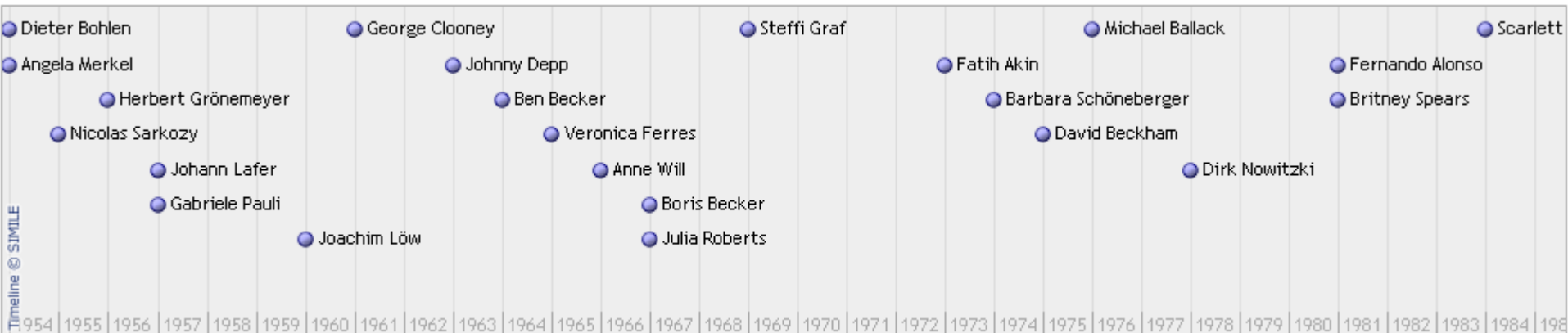
- Extraktion von Prominenten (Alter, Titel) aus News Texten – **Named Entity Recognition (NER)**
- Visualisierung der Ergebnisse auf www.ProminentPeople.info als Name Cloud



Ziele des Projektseminars

4

- 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

5

- 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

6

- 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-juan-es-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

7

- Datenbank mit Prominenten (Wikipedia)
 - über 92.000 Einträge
 - <http://tools.wikimedia.de/>

- Java-Sourcen für den Zugriff auf die Datenbank



The screenshot shows the Wikipedia article for Scarlett Johansson. At the top, there is a navigation bar with tabs for 'Artikel', 'Diskussion', and 'Seite bearbeiten'. Below this, a warning message states: 'Aus technischen Gründen kommt es momentan zu Anzeigefehlern bei einigen Bildern.' The main heading is 'Scarlett Johansson'. The article text begins with: 'Scarlett Johansson (* 22. November 1984 in New York City, New York) ist eine US-amerikanische Schauspielerin.' To the right of the text is a photograph of Scarlett Johansson. Below the main text, there is a 'Inhaltsverzeichnis' (Table of Contents) with sections for '1 Biografie', '2 Filmografie', '3 Quellen', and '4 Weblinks'. Further down, there is a 'Biografie' section with a 'Familie' subsection. The 'Familie' section starts with: 'Johanssons Vater Karsten ist ein dänischer Architekt, ihre Mutter Melanie entstammt einer jüdisch-amerikanischen Familie polnischer Herkunft. Sie wurde am 22. November 1984 in New York als Scarlett Marie Johansson geboren. Da ihre Eltern geschieden waren, lebte sie einige Zeit bei ihrem Vater in New York und einige Zeit bei ihrer Mutter in Los Angeles. Als Kind besuchte Scarlett die Schauspielschule Lee Strasberg in New York. Sie machte 2002 ihren Abschluss an der Professional Children's School in Manhattan. Ihre Schauspielkarriere begann 1992 in Los Angeles in dem Off-Broadway-Theaterstück *Sophistry*, in dem sie mit Ethan Hawke auf der Bühne stand.'

Erste Schritte (NER)

8

- 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
Fußs
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 Newslines, 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

9

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

Scarlett Johansson mag Kalorien:

Films
sich
nicht
verde
BUNTE

Scarlett Johansson fliegt in den

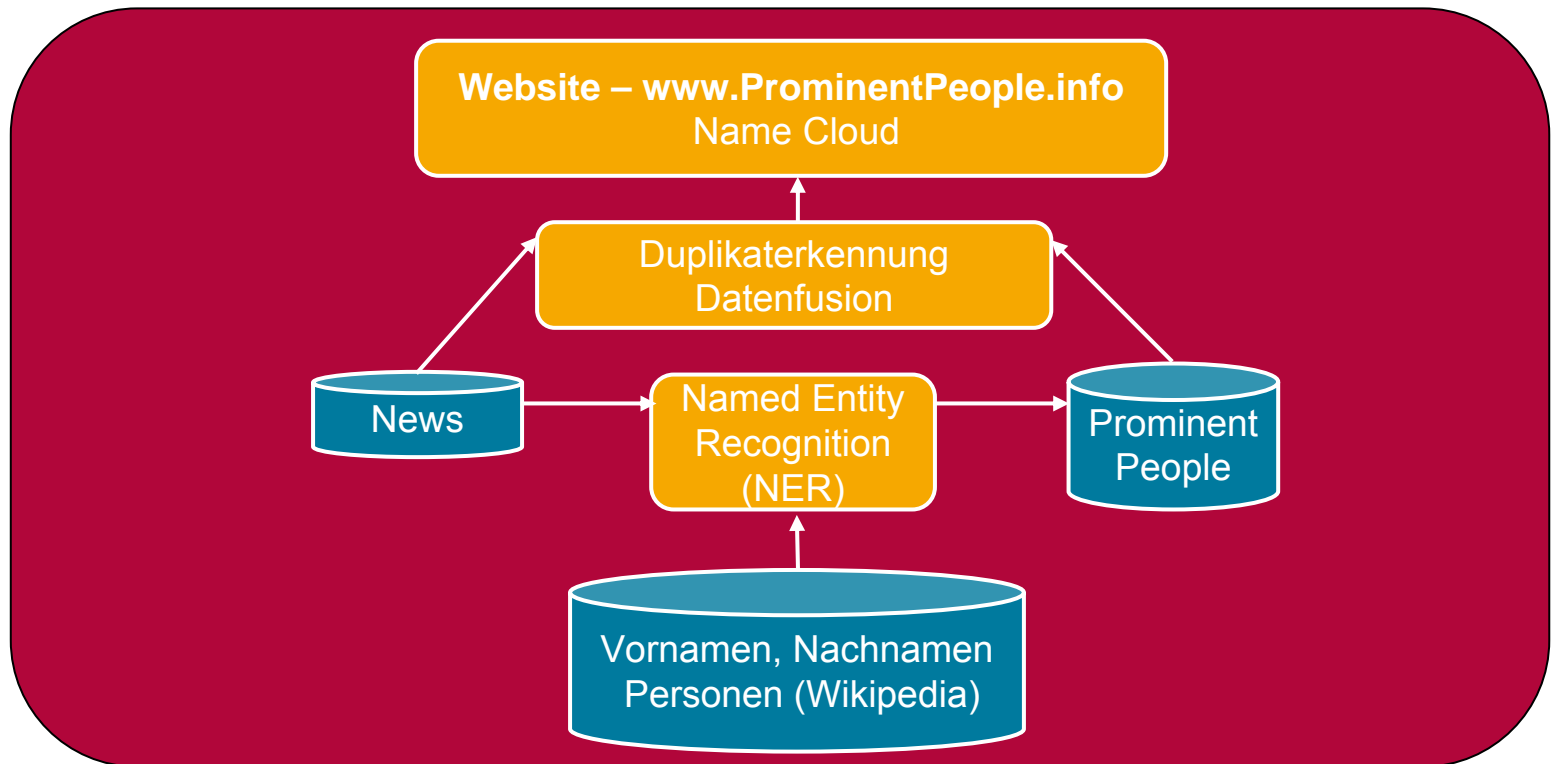
Irak
Joha
Fußs
eins
BUNTE

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 Newslines, 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

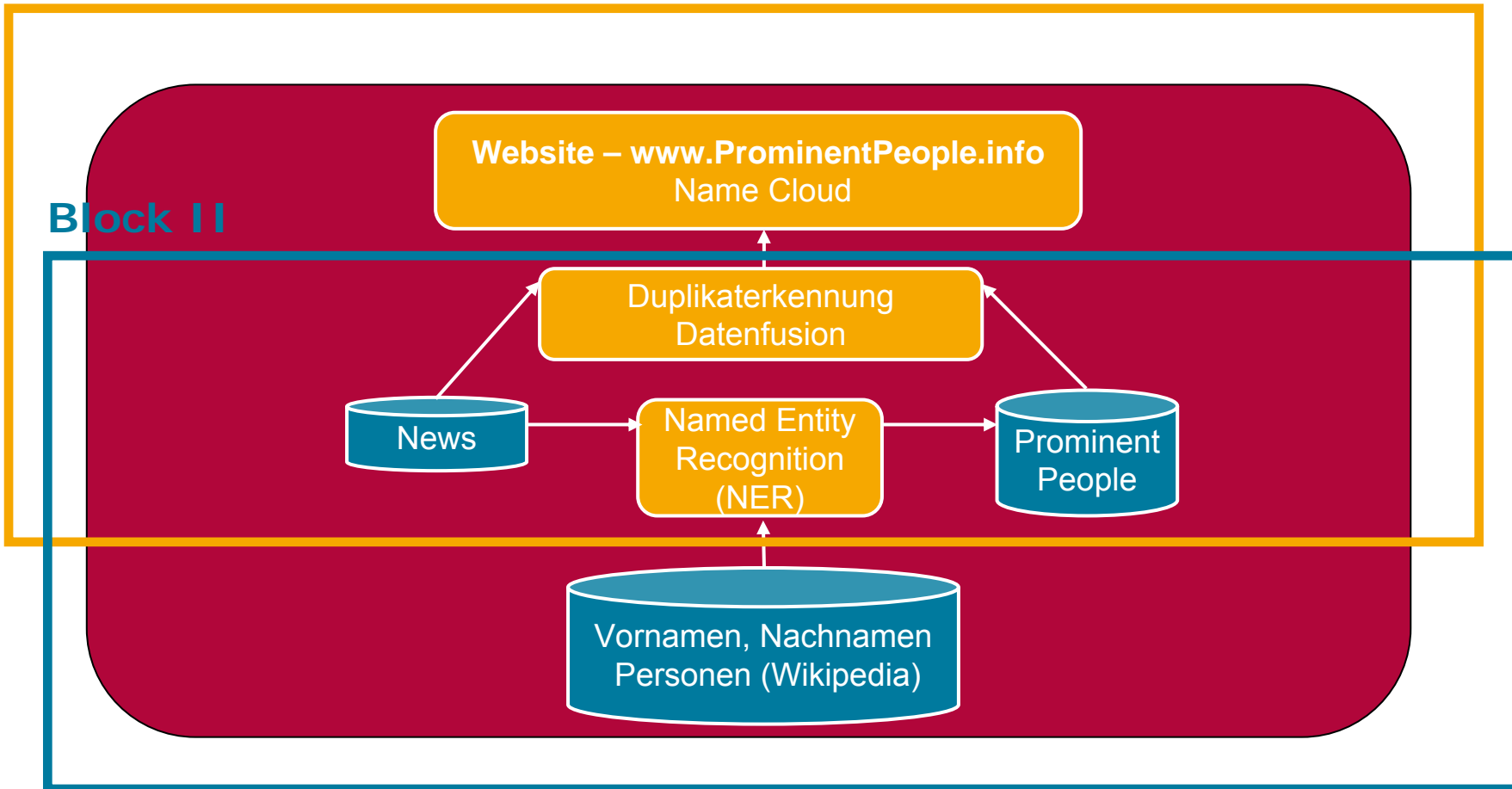
Josh Hartnett (29)
Scarlett Johansson



Projektplanung

11

Block I



Projektplanung - Aufgabenverteilung

12

- 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

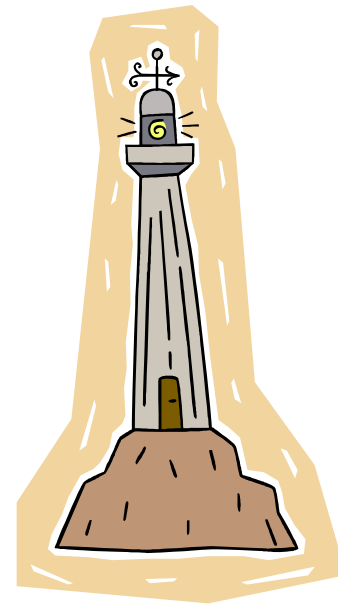
13

- 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

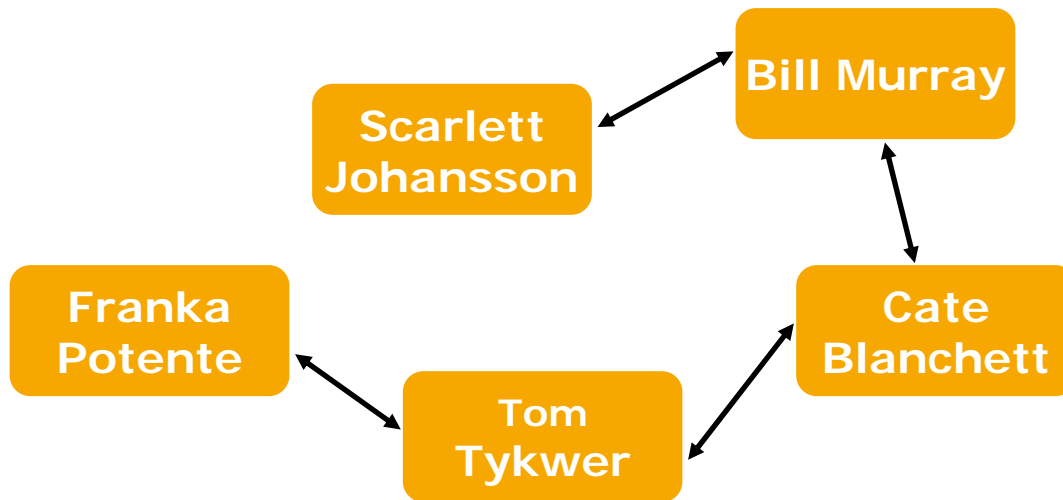
Ausblick

14

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



Entdecken von Personenbeziehungen



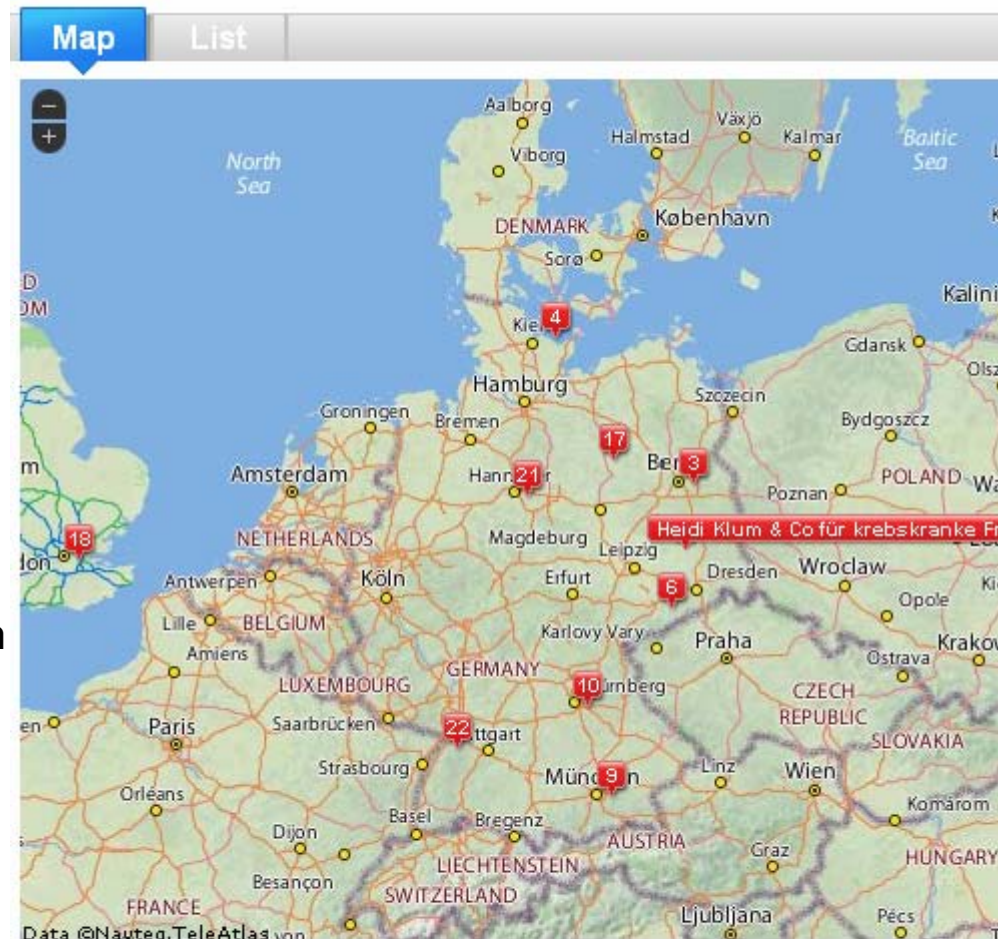
■ Mögliches Vorgehen

- 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

16

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



3. Heidi Klum & Co für krebserkrankte Frauen in Berlin

Die eine gab sich unnahbar, die andere zeigte die Herzlichkeit eines deutschen "Frol...
Hollywood-Schauspielerin Kim Basinger und Topmodel H...

ProminentPeople-Webservice

17

- 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

- 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

19

- 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.