

# MI1a Information Retrieval

Worum geht es hier?

Sommersemester 2013

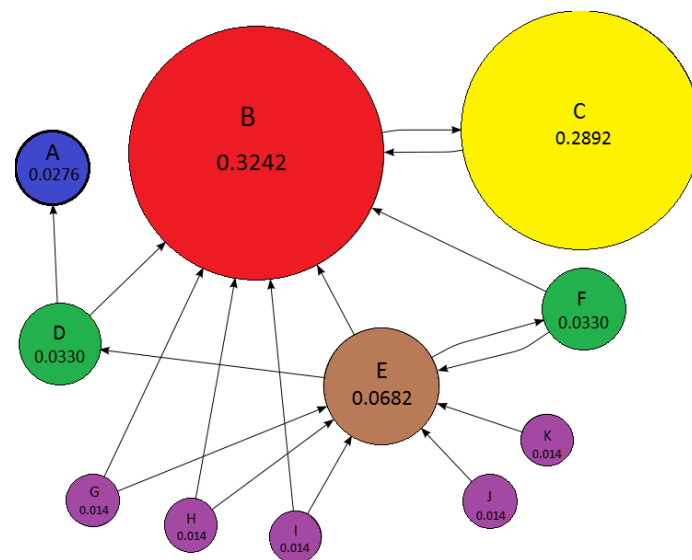
# Organisatorisches

- Der Kurs wird in Form eines Projektseminars durchgeführt
- AP durch Mitarbeit im Kurs und schriftliche Ausarbeitung
- Jeden Mittwoch 14:30 – 18:00 Uhr (evtl. auch kürzer)



# Aufgabe

- Erstellt als Team eine Suchmaschine auf Basis von Tweets
- Als Datenbasis werden Tweets mit Links heruntergeladen
- Diese Links verweisen auf „unsere Dokumente“
- Hashtags und Text in Tweets können analog zu Ankertexten im klassischen Web-Retrieval genutzt werden
- Zudem werden die Websites für das Retrieval genutzt
- Vorteil von Twitterdaten: Aktualität und Relevanz (?)





**Jens Terliesner**

View my profile page

201  
TWEETS

133  
FOLLOWING

67  
FOLLOWERS

Compose new Tweet...

Who to follow · Refresh · View all



**The British Library** @britishli...

Followed by Tamara Heck and oth...

Follow



**Martin Fenner** @mfenner

Followed by Heinz Pampel and ot...

Follow

Browse categories · Find friends

New York Trends · Change

#ProudToBeAFanOf

#unpopularopinionnight

#WhyBeInARelationshipIf

#janoskiansworldtour

#Louisville

Cuba

Masters

Peyton Siva

Official Video

Now Playing

© 2013 Twitter About Help Terms Privacy  
Blog Status Apps Resources Jobs  
Advertisers Businesses Media Developers

## Tweets



**JASIST** @JASIST

31s

RDA: Resource description & access—a survey of the current state of the art [ow.ly/jyc0R](http://ow.ly/jyc0R)

Expand



**Life at Google** @googlejobs

1m

In this new video, an Interaction Designer explains how he helped change how users interact w/Google Search on iOS [goo.gl/cMjFe](http://goo.gl/cMjFe)

View media



**Jay Rosen** @jayrosen

8m

Q & A with @carr2k [bit.ly/17oSWAK](http://bit.ly/17oSWAK) Plug his comments on the shifting relationship between journalists and the American vernacular.

Retweeted by Bora Zivkovic

Expand



**Bora Zivkovic** @BoraZ

1m

Lauren Friedman @fidera Q&A @open-Notebooks @trachselviv on her @newyorker story on science of sex abuse [bit.ly/10QgLxj](http://bit.ly/10QgLxj)

Expand



**Studierendenservice** @HHU\_Aktuell

10m

Studieren ohne Abschluss? Das wichtigste auf einen Blick [tinyurl.com/c4tdo5w](http://tinyurl.com/c4tdo5w)

Expand



**Mark Dang-Anh** @mdanganh

18m

#toread "An Investigation of Influentials and the Role of Sentiment in Pol. Comm. on Twitter during Election Periods" [tandfonline.com/doi/abs/10.108...](http://tandfonline.com/doi/abs/10.108...)

Expand



**Bora Zivkovic** @BoraZ

20m

Tutorial 23: How to avoid giving your work to a "predatory open access publisher" [svpow.com/2013/04/09/tur...](http://svpow.com/2013/04/09/tur...)

Expand



**Lambert Heller** @Lambo

20m

Schöne Video Lecture (75 Minuten) über #OpenScience PT @dunkelmunkel Open Science im Web [wp.me/pzK5sz-2U...](http://wp.me/pzK5sz-2U...)

View media



**Barbara Leisner** @LeisnerBarbara

44m

Mein Blog "(Nord)DeutscheStunde" ist umgezogen: Alte Inhalte in neuem Gewand [wp.me/p3cJAq-7U...](http://wp.me/p3cJAq-7U...)

Retweeted by Doerte

View summary

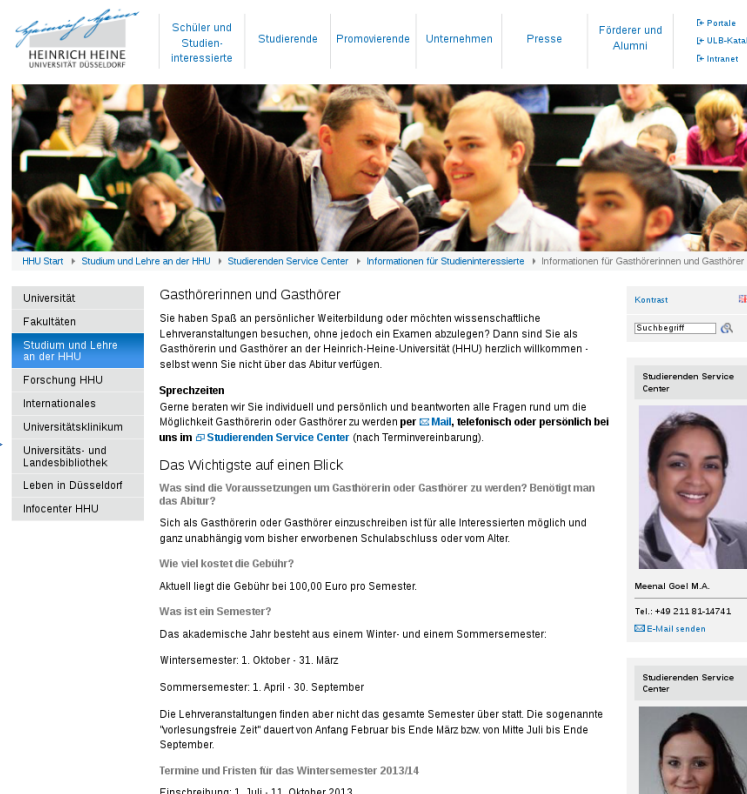


**Europeana** @EuropeanaEU

23m

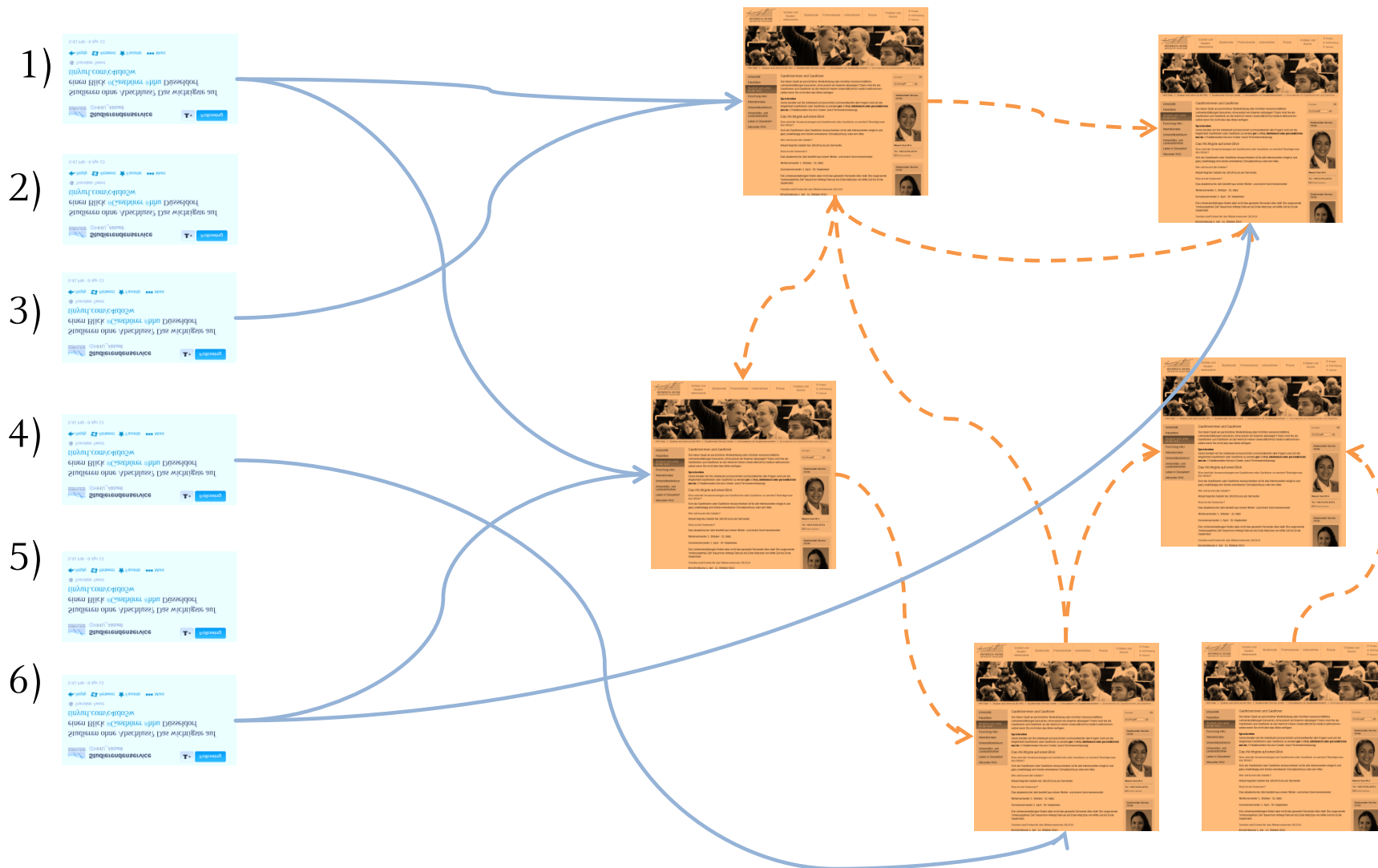
Check out our new theme on the @EurFashion Tumblr. From Gauguin to Missoni, explore connections between art & fashion [europeanafashion.tumblr.com](http://europeanafashion.tumblr.com)

# Welche Informationen haben wir zu einem Dokument?



1. Hashtags aus Tweets
2. Worte in Tweets
3. Website mit Title

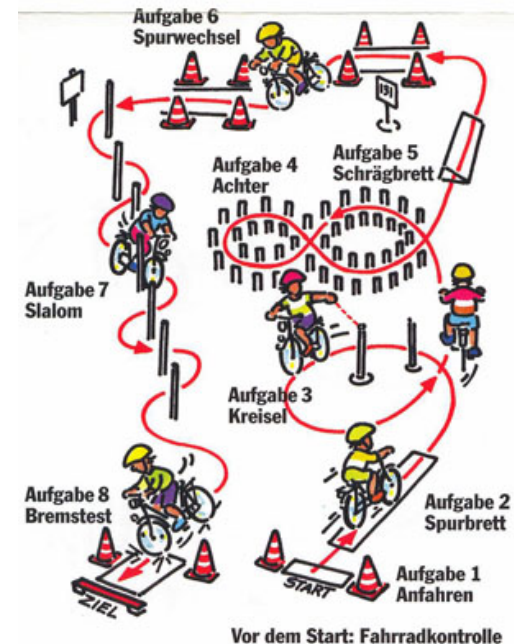
# Was sind Links in Tweets



# Was muss gemacht werden?

1. Datenbankmodell erstellen
2. Tweets über API herunterladen und in Datenbank speichern
3. Kurzlinks auflösen und zusammen mit langer Variante in DB speichern
4. Website Parsen und Daten in Datenbank speichern
5. Stemming
6. Textstatistik
7. Suchfunktion
8. Verschiedene Rankingmethoden berücksichtigen
  - a) Klassische Rankingmethoden ( $WDF * IDF$ )
  - b) Tweetspezifisches Ranking
9. Frontend erstellen

**Ausblick:** Evaluation und Retrievaltest



# Mögliche Rankingfaktoren

- Rankingfaktoren aus Tweets
  1. Hashtags
  2. Text in Tweets
  3. Aktualität
- Rankingfaktoren aus Websites
  1. Textstatistik

Die Rankingfaktoren können kombiniert werden






# Gibt es so etwas schon?

[Jobtweet.de](http://jobtweet.de)



172442 Jobs Nur FÜR Dich Nur bei




**jobtweet.de**

suchmaschine 




(z.B. Entwickler Hamburg)

Die Suche nach "suchmaschine" ergab 51 Resultat(e)  
Resultate 1 bis 10

PHP PROGRAMMIERER SEO /  
Suchmaschinenmarketing (m/w): Anbieter: nicht  
oeffentlich ... <http://t.co/ltvkkW7QU8> #jobs  
#seo  
2013-04-10 08:07:53  
zum jobtweet

  Gefällt mir  Zeige deinen Freunden, dass dir das gefällt.

Teamleiter (m/w) SEO  
(Suchmaschinenoptimierung) in Berlin  
<http://t.co/NrxLv0Q4x4> #seo #jobs  
2013-04-10 07:32:51  
zum jobtweet

  Gefällt mir  Zeige deinen Freunden, dass dir das gefällt.

SEO/ Suchmaschinenoptimierer für Linkaufbau  
(Online-Marketing-Manager/in): Anbieter: Br...  
<http://t.co/Voy3gjrVSc> #jobs #seo  
2013-04-09 16:31:42

# Gibt es so etwas schon?

Tame: <http://tame.tazaldoo.com/>

The screenshot shows the Tame (Twitter Analytics) interface. At the top, there's a search bar with 'düsseldorf' entered. Below the search bar, it says '1000 Tweets der letzten 18 Stunden für globale Suche nach düsseldorf analysiert. Mehr laden...'. A progress bar indicates '18 Stunden'. The interface is divided into three columns: Themen, Nutzer, and Inhalte.

| Themen           | Nutzer                | Inhalte                                 |
|------------------|-----------------------|---|
| 1 #düsseldorf 87 | 1 @youtube 9          | 1 Düsseldorf: Anwohner schmücken sc     |
| 2 #germany 14    | 2 @sksmidiaberlin 6   | 2 Abitur-Schummelei mit iPhones: 1. S   |
| 3 #job 13        | 3 @melbad_mel 6       | 3 Nuten in Düsseldorf Ladies Eierlecke  |
| 4 #f95 11        | 4 @news_muenchen 5    | 4 Düsseldorf: Eine Million Euro für neu |
| 5 #deutsch 11    | 5 @thelawdoofficial 4 | 5 'Staalproducenten maken jaren prijs   |
| 6 #berlin 11     | 6 @fannyheather 4     | 6 The Adidas Originals Germany Fußball  |
| 7 #stuttgart 11  | 7 @einbruch 3         | 7 The Adidas Originals Germany 1974     |
| 8 #jobs 10       | 8 @goebelmasse 3      | 8 Ehrenamtspreis: SPD zeichnet Schla    |
| 9 #adidas 10     | 9 @cridderwall 3      | 9 Nacht der Museen 2013 - Düsseldorf    |
| 10 #jailbreak 9  | 10 @elasticsearch 3   | 10 Der Flughafen Düsseldorf hat einen r |

At the bottom, there's a footer with links: Start, How-to, FAQ, Über uns, Presse, Impressum, Datenschutz. It also mentions 'Investieren Sie in uns' with a logo and 'www.companisto.de (noch 38 Tage)'. Social media links for Twitter, Facebook, and Google+ are provided. The tazaldoo logo is also present.

# Programmiersprachen und Kenntnisse

- Python oder PHP (kann flexibel besprochen werden)
- Grundlegende Kenntnisse von SQL-basierten Datenbanken werden vorausgesetzt
- Die fertige Suchmaschine soll auf einem eigenen Server laufen (wird von uns bereitgestellt)

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>

<?php

    /*
        archive.inc.php          (include for newsletter archive)
    */

    // fit if necessary:
    define( "PATH_TO_SCRIPT", "../" );

    if (empty($base_url["jax_nl"])) $base_url["jax_nl"] = ".";

    // Global variables (have to be fit)
    require_once ( dirname(__FILE__) . "/" . PATH_TO_SCRIPT . "settings/newsletter.settings.inc.php" );

    // (If exists) include localization file for the favoured language
    if ( file_exists( dirname(__FILE__) . "/" . PATH_TO_SCRIPT . "languages/" . $language . ".inc.php" ) )
    {
        define( "LOC_LANG", $language );
    }
    else
    {
        define( "LOC_LANG", $default_language );
    }

    require_once ( dirname(__FILE__) . "/" . PATH_TO_SCRIPT . "languages/" . LOC_LANG . ".inc.php" );

?>

<head>
    <title><?php echo "$newsletter_title - Archive"; ?></title>
    <link rel="stylesheet" href="<?php echo "$base_url["jax_nl"])/css/$css_file";?>"
    <meta http-equiv="Content-Type" content="text/html; charset=<?php echo $charset ?>"
</head>
```

# Projektarbeit und Projektgruppen

## Achtung:

Das Gelingen dieses Projektseminars liegt in euren Händen.  
Es muss programmiert werden! Dabei gibt es Unterstützung!

Wer vorzeitig aussteigt verschafft seinem Team Mehrarbeit.

# Und los geht's...

## Wie können Expertenteams gefunden werden?

- Wer kann besonders gut programmieren?
- Wer kann eine Gruppe leiten?
- Wer hat Erfahrung mit Datenbanken?
- ...

## Kommunikation

- E-Mail
- Repositories
- Wiki
- ...

