Verschlagwortung digitaler Texte

Verschlagwortung

- Zuordnung von Schlagwörtern zu einem Dokument (Text) zur Erschließung der darin enthaltenen Sachverhalte
- Manuelle Verschlagwortung
  - Schlagwörter meist aus einem kontrollierten Vokabular
- Computergestützte Verschlagwortung
  - Schlagwörter werden maschinell vorgeschlagen, manuell ausgewählt
- Automatische Verschlagwortung
  - statistisch durch Ermittlung von Worthäufigkeiten

Volltextindexierung

- Erfassung sämtlicher Wörter eines Textes
- Stoppwörter werden nicht beachtet

hohe Anzahl an Stichwörtern

bei der Suche keine Kenntnis über das Ordnungssystem erforderlich

Suche über Volltextindex => aufwendig

Termgewichtung

- einfaches Verfahren zur Termgewichtung:

- Verhältnis zwischen
  Häufigkeit eines Begriffs in einem Text
  und
  Anzahl der Dokumente, in denen der Begriff vorkommt

- Gewichtung eines Begriffs ist hoch, wenn es wenige Texte im Korpus gibt, in denen der Begriff enthalten ist und der Begriff im zu indexierenden Text häufig vorkommt

Termgewichtung

- Beispiel: Korpus mit 200 Texten
- "auf":
  - Häufigkeit im Text = 9, kommt in allen (200) Texten des Korpus vor:
  - 9/200 = 0.045
- "Staatssekretär":
  - Häufigkeit im Text = 9, kommt in 5 Texten vor
  - 9/5 = 1.8

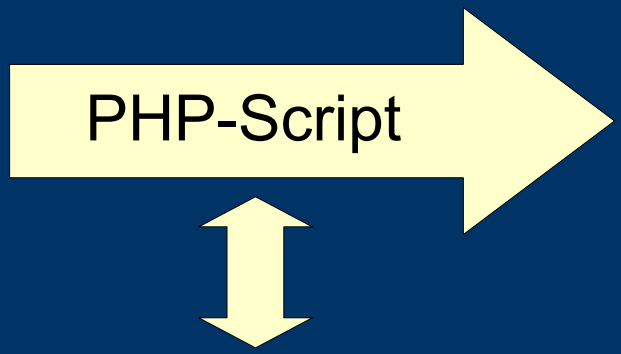
Korpus

- 200 Artikel der taz die tageszeitung
- XML-Dateien (je Artikel eine Datei)
- Stuttgart-Tübingen Tagset (STTS)

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE corpus SYSTEM "d:\XML-Soft\corpus.dtd">
[...]
<clause complete="-">
    <token lemma="d" wclass="ART">
        Die
    </token>
    <token lemma="institutionell" wclass="ADJA">
        institutionelle
    </token>
[...]
```

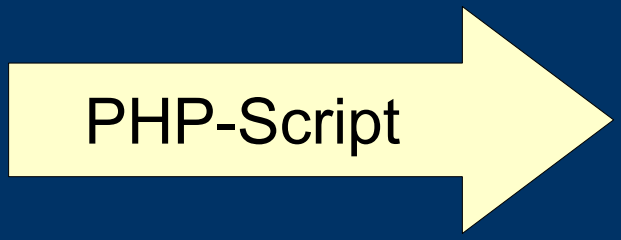
Verarbeitung

1. XML
<token>



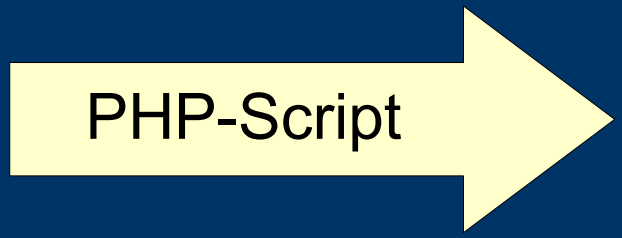
Datenbank

CSV

2. 

Berechnungen in DB

3. Datenbank
Schlagwörter



XML
sonstige

erweiterte XML-Daten (Beispiel)
(Gewichtung als Attribut)

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE corpus SYSTEM "d:\XML-Soft\corpus.dtd">
[...]
<clause complete="-">
    <token lemma="d" wclass="ART" q="0.8450">
        Die
    </token>
    <token lemma="institutionell" wclass="ADJA" q="0.5000">
        institutionelle
    </token>
    <token lemma="Kompetenzschwäche" wclass="NN" q="3.0000">
        Kompetenzschwäche
    </token>
[...]
```

Beispieltext

Tendenz zur Lästigkeit
Die institutionelle Kompetenzschwäche Michael Naumanns
und wie er sie nutzen kann. Was der Kulturbeauftragte darf und was nicht.
Staatstragende Überlegungen von Elke Gurlit

Niemand wird bestreiten, daß Gerhard Schröder mit der Etablierung des Bundeskulturbeauftragten ein Coup gelungen ist. Nicht nur die staatliche Kulturpolitik, sondern auch das Räsonieren über Kultur hat in den letzten Monaten einen enormen Bedeutungszuwachs erfahren. Die tägliche Naumann-Meldung gehört zum unverzichtbaren Repertoire des Feuilletons. Man gewinnt fast den Eindruck, Michael Naumann handele als Beauftragter unterbeschäftigter Kulturredaktionen. Zum besseren Verständnis der Stellung des Kulturbeauftragten lohnt ein Blick auf das Beauftragtenwesen, das sich in der Bundesrepublik flächendeckend ausgebreitet hat. Wir kennen beispielsweise die Datenschutzbeauftragten , die betrieblichen Immissionsschutzbeauftragten und die Gleichstellungsbeauftragten in der Verwaltung. Ungeachtet aller Unterschiede im Detail lassen sich gemeinsame Grundstrukturen ausmachen: Die Beauftragten vertreten Interessen, die im normalen Gang der Verwaltungs- oder Unternehmensgeschäfte zuwenig Beachtung finden. [...]

gewichtete Terme

(freq = Häufigkeit im Text, Texte = Anzahl der Texte, in denen Lemma vorkommt, q = Quotient)

Lemma	freq	Texte	q
Bundeskulturbeauftragter	5	1	5.0000
Kulturbeauftragte	14	3	4.6667
Kompetenzschwäche	3	1	3.0000
Naumann	11	5	2.2000
Kulturhoheit	2	1	2.0000
Bundesbeauftragte	2	1	2.0000
parlamentarisch	12	6	2.0000
Lästigkeit	2	1	2.0000
Staatssekretär	9	5	1.8000
Kulturpolitik	5	3	1.6667
Beauftragte	5	4	1.2500

[...]

gewichtete Terme

(freq = Häufigkeit im Text, Texte = Anzahl der Texte, in denen Lemma vorkommt, q = Quotient)

Lemma	freq	Texte	q
[...]			
groß	1	175	0.0057
erst	1	179	0.0056
man	1	183	0.0055
ander	1	194	0.0052
alle	1	191	0.0052
oder	1	191	0.0052
nach	1	197	0.0051
so	1	195	0.0051
wie	1	199	0.0050
daß	1	200	0.0050

Beispieltext *(Lemmata mit q>1 in rot)*

Tendenz zur Lästigkeit
Die institutionelle Kompetenzschwäche Michael Naumanns
und wie er sie nutzen kann. Was der Kulturbeauftragte darf und was nicht.
Staatstragende Überlegungen von Elke Gurlit

Niemand wird bestreiten, daß Gerhard Schröder mit der Etablierung des Bundeskulturbeauftragten ein Coup gelungen ist. Nicht nur die staatliche Kulturpolitik, sondern auch das Räsonieren über Kultur hat in den letzten Monaten einen enormen Bedeutungszuwachs erfahren. Die tägliche Naumann-Meldung gehört zum unverzichtbaren Repertoire des Feuilletons. Man gewinnt fast den Eindruck, Michael Naumann handele als Beauftragter unterbeschäftigter Kulturredaktionen. Zum besseren Verständnis der Stellung des Kulturbeauftragten lohnt ein Blick auf das Beauftragtenwesen, das sich in der Bundesrepublik flächendeckend ausgebreitet hat. Wir kennen beispielsweise die Datenschutzbeauftragten, die betrieblichen Immissionsschutzbeauftragten und die Gleichstellungsbeauftragten in der Verwaltung. Ungeachtet aller Unterschiede im Detail lassen sich gemeinsame Grundstrukturen ausmachen: Die Beauftragten vertreten Interessen, die im normalen Gang der Verwaltungs- oder Unternehmensgeschäfte zuwenig Beachtung finden. [...]

Schlagwort oder nicht?

- 'Auswahl' der Schlagwörter anhand Gewichtung
- mögliche Kriterien:
  - nach Rang (z.B. die ersten vier Ränge)
  - fester Grenzwert (z.B. q>1)
  - Vergleich

    z.B. q > relative Häufigkeit (fairer Vergleich? fraglich!)

    Beispiel "Staatssekretär":

    q = 1.8 > 0.075

    (Häufigkeit im Korpus / Anzahl der Texte im Korpus)

Gewichtungsmethode tf-idf

- tf-idf (term frequency - inverse document frequency)
- *term frequency* ist das Verhältnis
  - Häufigkeit eines Terms im Text zu
  - Anzahl der Terme im Text
- *inverse document frequency* ist das Verhältnis
  - Gesamtzahl der Texte im Korpus zu
  - Anzahl der Texte, in denen der Term vorkommt

Berechnung tf-idf

- tfidf = tf * log(idf)
- Beispiel "Staatssekretär":
  - der Text hat N = 1427 Wörter
  - "Staatssekretär" kommt n = 9 mal vor
  - Anzahl der Texte im Korpus T = 200
  - Anzahl der Texte, in denen "Staatssekretär" vorkommt d = 5

tfidf = 9 / 1427 * log(200 / 5) = 0.0232

Gewichtung mit tf-idf

Lemma	tfidf
Kulturbeauftragte	0.0412
parlamentarisch	0.0295
Naumann	0.0284
Staatssekretär	0.0233
Bundeskulturbeauftragter	0.0186
Kulturpolitik	0.0147
staatlich	0.0144
Beauftragte	0.0137
Kompetenzschwäche	0.0111
kulturell	0.0100
institutionell	0.0098
[...]	