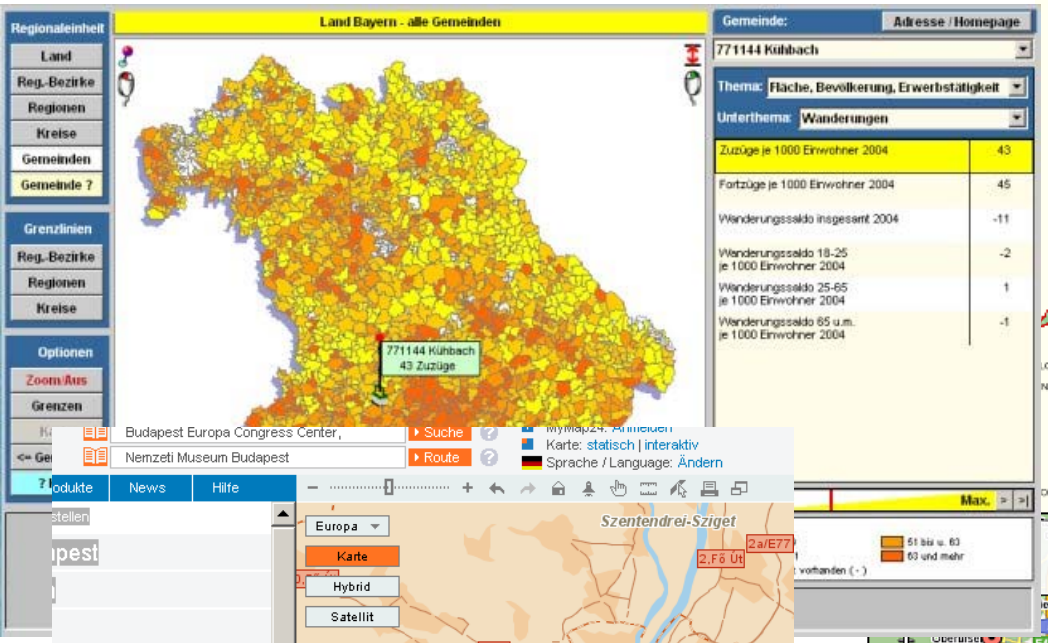
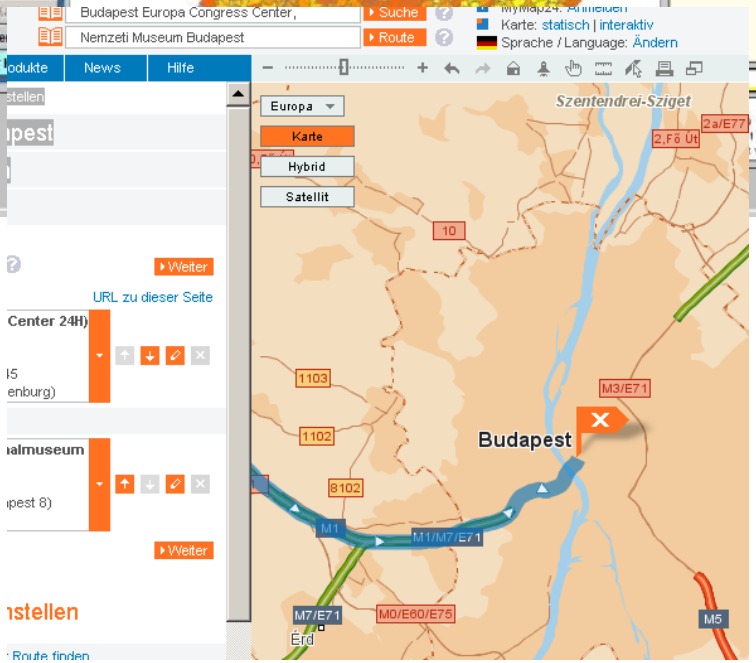
Thomas Mandl
Informationswissenschaft
Universität Hildesheim
mandl@uni-hildesheim.de

Fachgruppentreffen
Fokussierte Suche
Bamberg 8.5.2008

Die Klassifikation von geographischen Anfragen in einem Query-Log

Geographic Information Systems





Geographisches IR

- Erweiterung Information Retrieval
  - Anfragen mit geographischen Referenzen
- Cross-Language Evaluation Forum (CLEF)
  - Evaluierungsinitiative für sprachübergreifendes Retrieval
  - http://www.clef-campaign.org/
- Track GeoCLEF seit 2004
  - Evaluierung von geographischem Information Retrieval

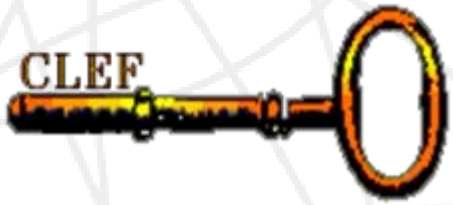
Initial Aim of GeoCLEF

- Aim: to evaluate retrieval of multilingual documents with an emphasis on geographic search (GIR)
  - "*find me news stories about riots near Dublin*"

(Fred Gey @ CLEF Workshop 2005)

Ziele von CLEF

- Infrastruktur für Forschung und Entwicklung cross- und multi-lingualer Information Retrieval Systeme
  - Multilingual Information Retrieval Systeme evaluieren
  - Testsuites entwickeln:
    Dokumente, Anfragen, Relevanzurteile, stat. Auswertungsverfahren
- CLEF 2008
  - Registration is still possible
  - E.g. Robust Word Sense Disambiguation Task
    - http://ixa2.si.ehu.es/clirwsd/



Participation

CLEF Year	2005	2006	2007
Nr. of Participants	11	17	13 New: Query Classification Task 6 participants
Nr. of submitted Experiments	117	149	108

Query Classification Task

- Goal: find geo queries in a log of real queries
- New in 2007

- Organized by Xing Xie (Microsoft Research Asia, Beijing, China)

Why Query Classification?

- User interface adaptation
  - Result set design should be different for navigational and informational queries (Cutrell & Guan 2007)
  - Geo queries may require map as result (Google)
- Retrieval Performance
  - „More work needs to be done on customizing methods for each topic“ (Harman 2005)
  - Named entities in query could be used for system parameter selection (Mandl & Womser-Hacker 2006)

Data

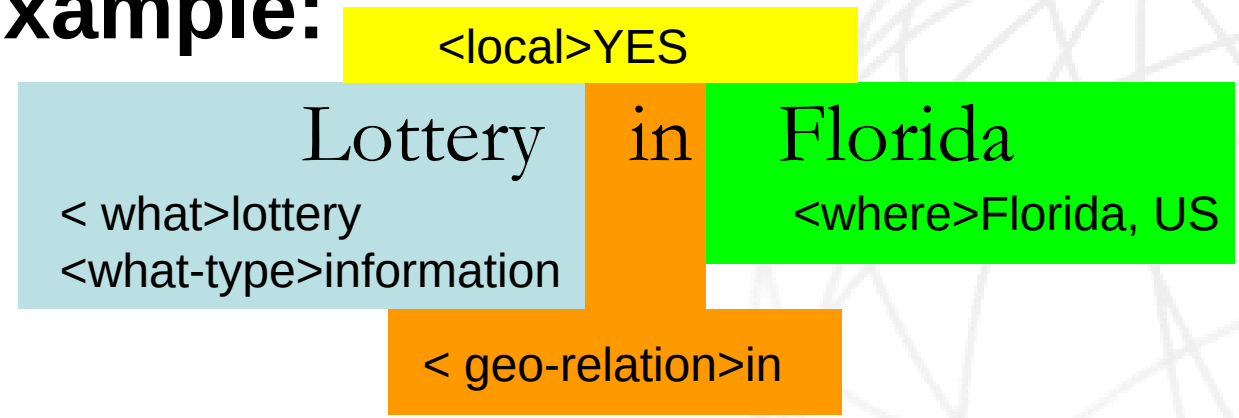
- Query log from the MSN search engine
  - in English
  - 800.000 queries
  - collected August 2006
  - 500 queries were labelled and used for evaluation
    - 100 queries for training
    - 400 for testing

Beaver County times
BEDDING
bedroom furniture
Beer
Behr Paint
BEIJING
BELFAST TELEGRAPH
belinda
Belize real estate
Belkin
BELL SOUTH
BELLA VISTA
BELLAGIO HOTEL
bellagio las vegas
BELLE
belleville
Belleville News Democrat
bellevue
bend it like beckham
bendigo bank
Benfica

Task

- Find queries with a geographic scope
  - Extract where component
  - Extract geo-relation-type
  - Extract what component
  - Classify what type {information, yellow page, map}

Example:



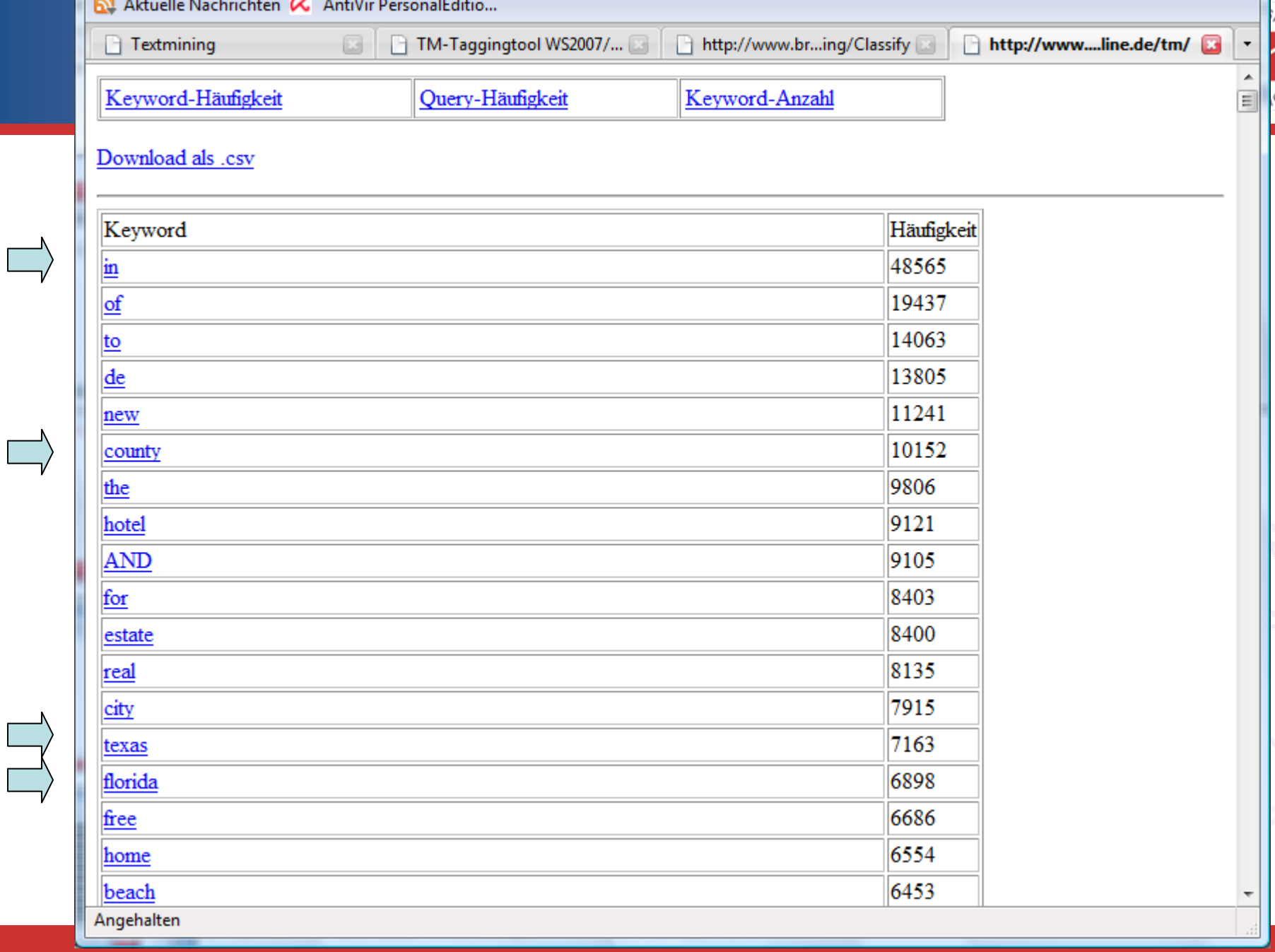
Geo-relation-types

- 27 classes
- Examples:
  - In
  - On
  - Near
  - Along
  - Distance
  - North_of
  - North_west_of
  - North_to
  - …

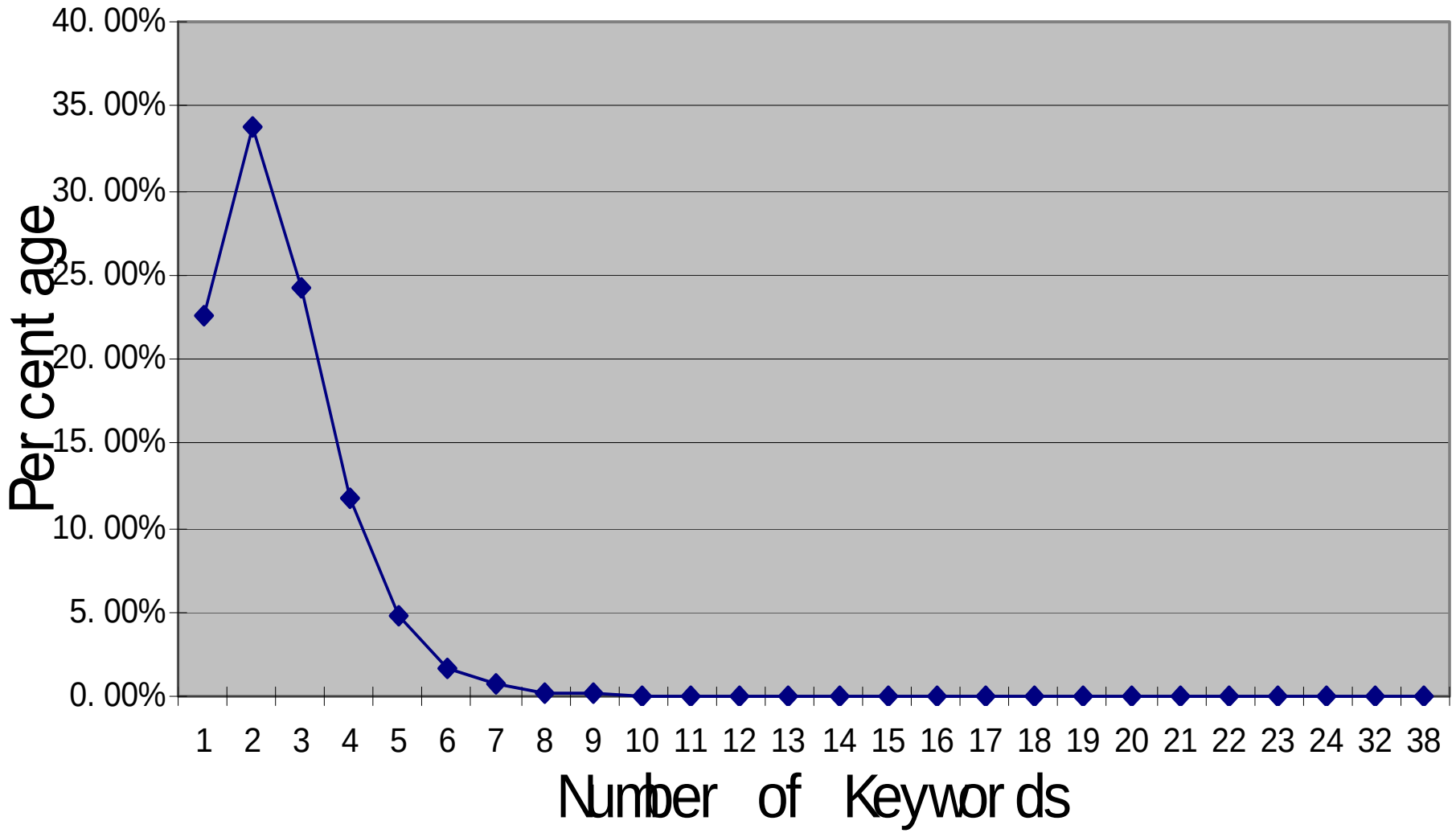
Statistik

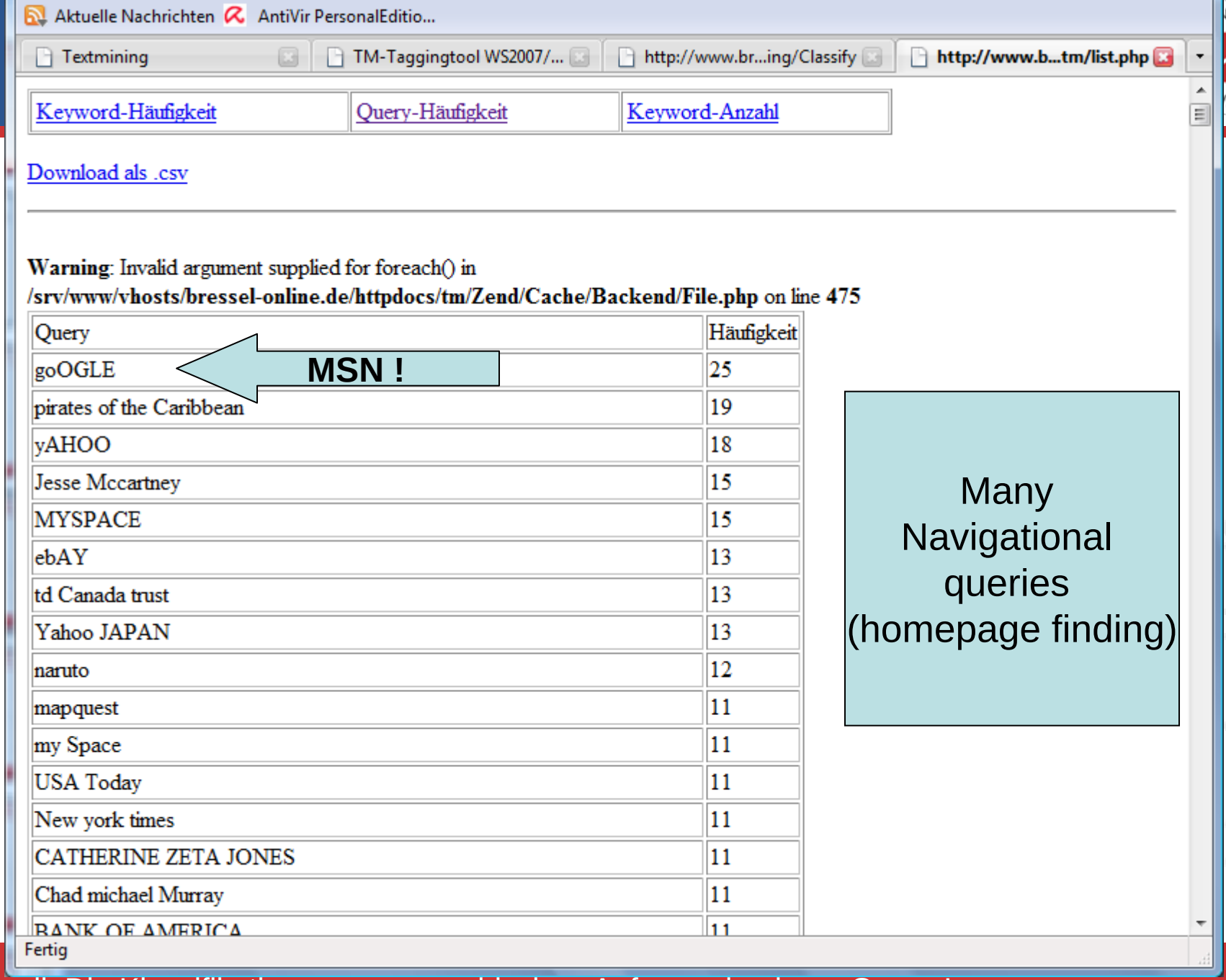
- 70.000 Queries with Geo Relations
- Davon 50.000 mit Geo Eigennamen

Li et al. 2007


Aktuelle Nachrichten
AntiVir PersonalEditio...
Textmining
TM-Taggingtool WS2007/...
http://www.br...ing/Classify
http://www....line.de/tm/
Keyword-Häufigkeit
Query-Häufigkeit
Keyword-Anzahl
Download als .csv
Keyword
Häufigkeit
in 48565
of 19437
to 14063
de 13805
new 11241
county 10152
the 9806
hotel 9121
AND 9105
for 8403
estate 8400
real 8135
city 7915
texas 7163
florida 6898
free 6686
home 6554
beach 6453
Angehalten

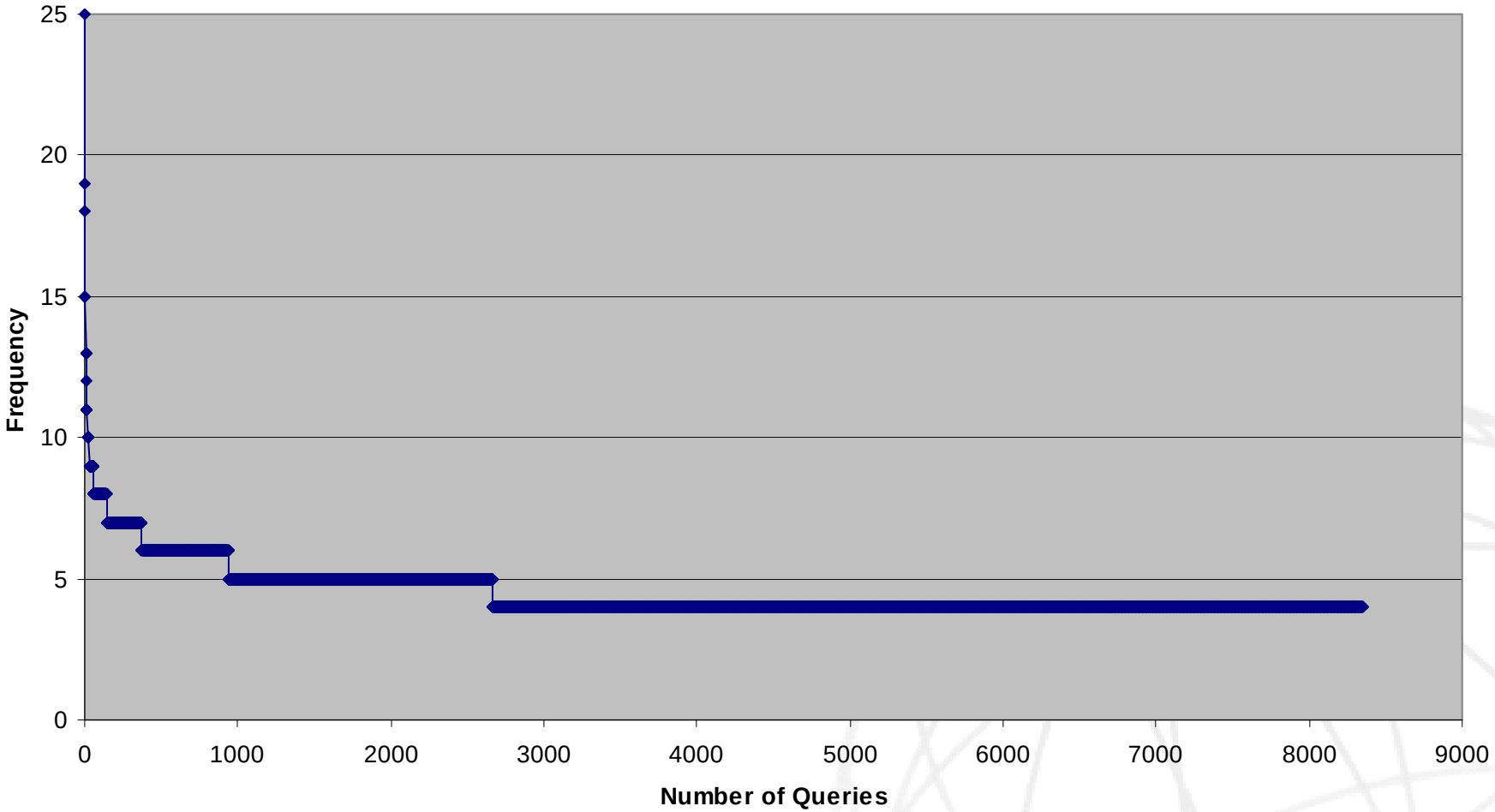
Length of Queries



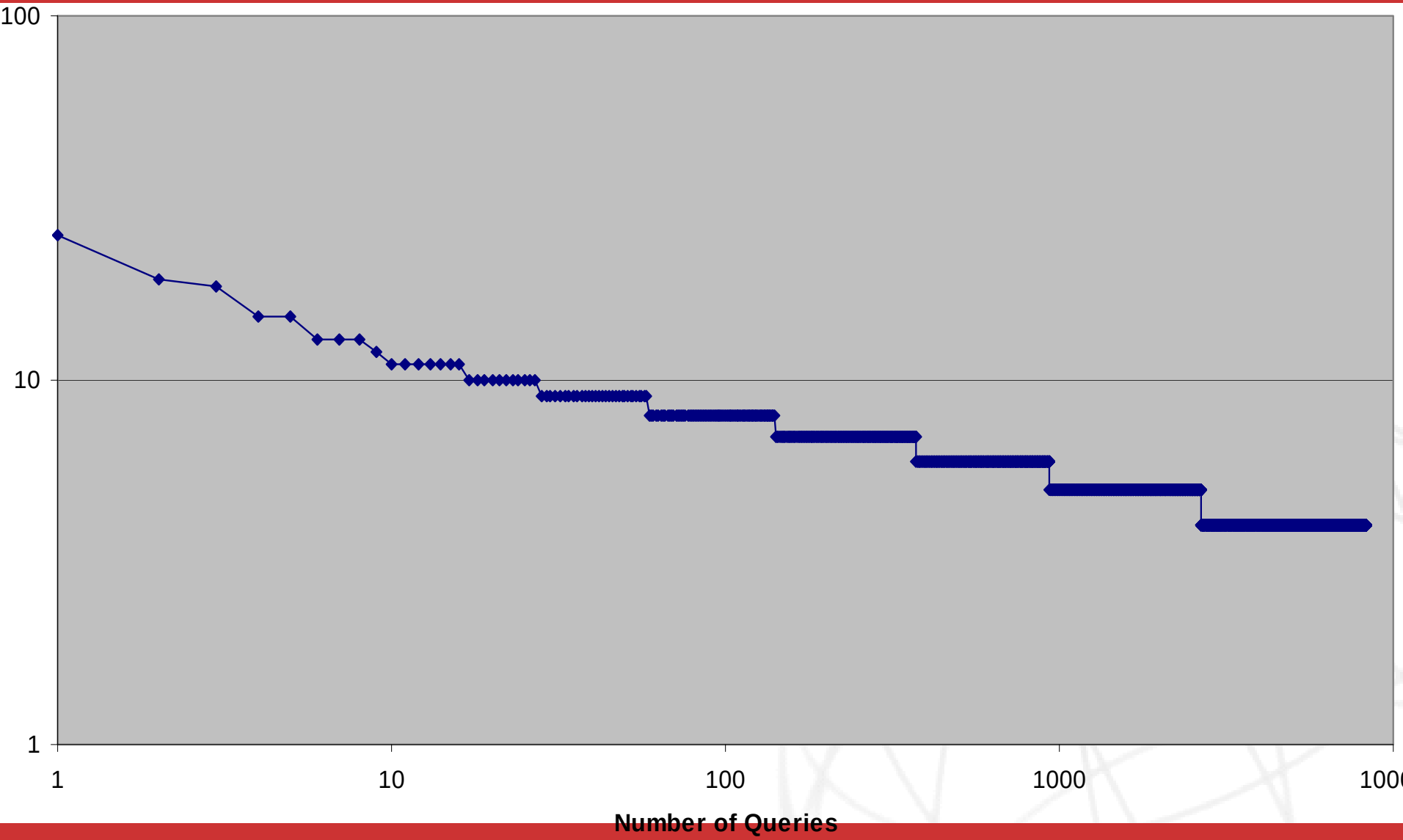


Many Navigational queries (homepage finding)

Most frequent Queries



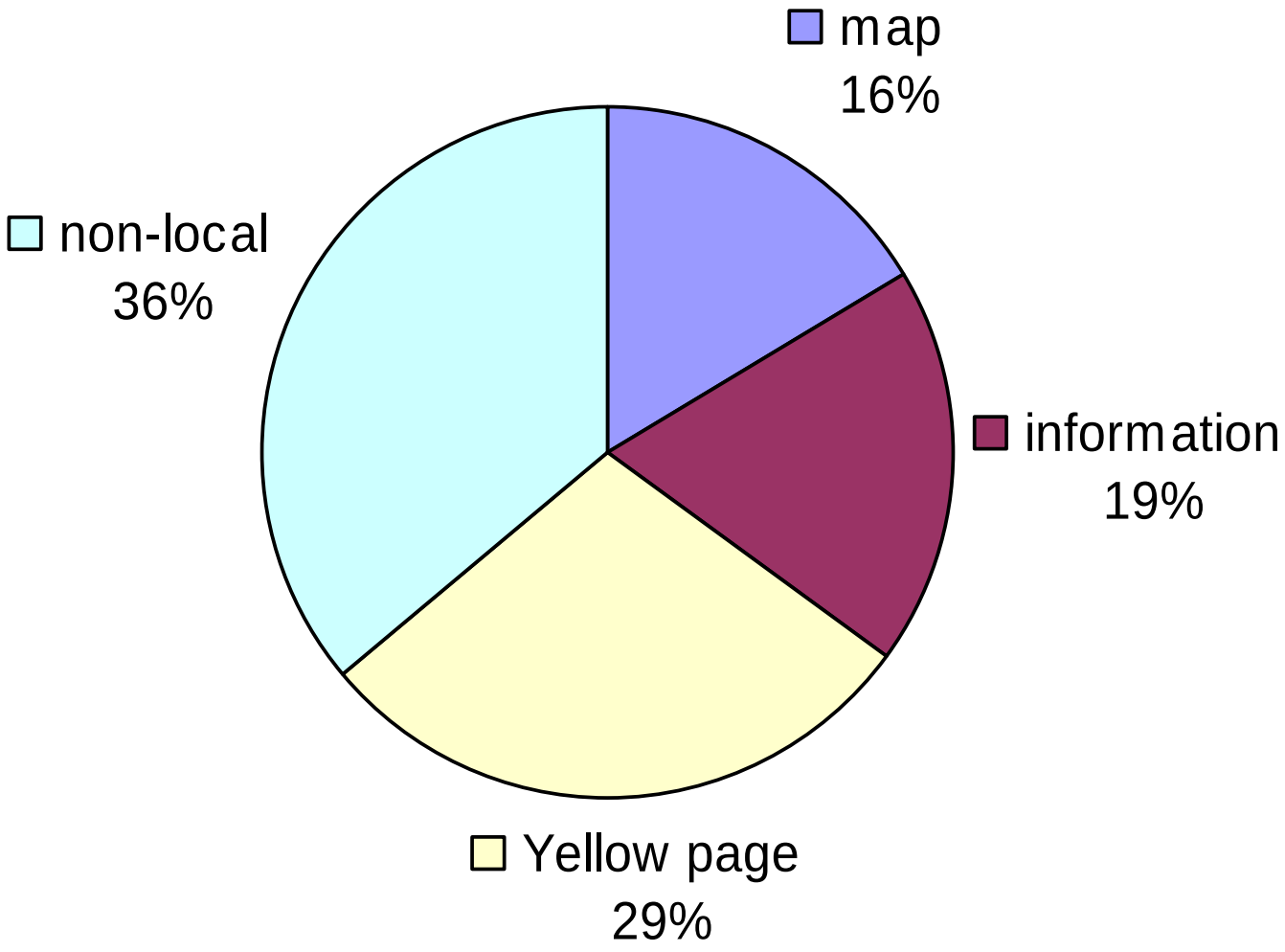
Most frequent Queries



Evaluation Set

1. Choose 800 queries randomly from the query set.
2. Remove the typos and the ambiguous queries from the 800 ones manually.
3. Select the queries with special geo-relations from the remainder queries in the query set manually and add them to the evaluation set.
4. Select 500 queries for the final evaluation set.

Query Types in Evaluation Set



Comparison to other Studies

- 20% of all Web queries have a geographic context

  (Sanderson & Kohler 2004)

- Analysis of 1000 Altavista Queries
  - Informational 48%
  - Navigational 20%
  - Transaction oriented 30%

  (*Broder 2002)*

Evaluation Metrics

- Three assessors
  - individually assessed all system answers
  - reached an agreement
- Fully Correct classified query instances
- Recall, precision and combined F1-Score

Problems

- Ambiguity
  - Place names: *Li et al. 2007*
    - Washington
  - Place names / other named entities
    - Machida (city and actor in Japan)
  - Between classes: *Lana-Serano et al. 2007*
    - Atlanta medical (informational or yellow page?)
    - Airport (geo or not?)
    - calabria chat (geo or not?)

Results

Team	Precision	Recall	F1
Ask	**0.625**	0.258	0.365
Csusm	0.201	0.197	0.199
Linguit	0.112	0.038	0.057
Miracle	0.428	**0.566**	**0.488**
Talp	0.222	0.249	0.235
Xldb	0.096	0.08	0.088

Approaches of Participants

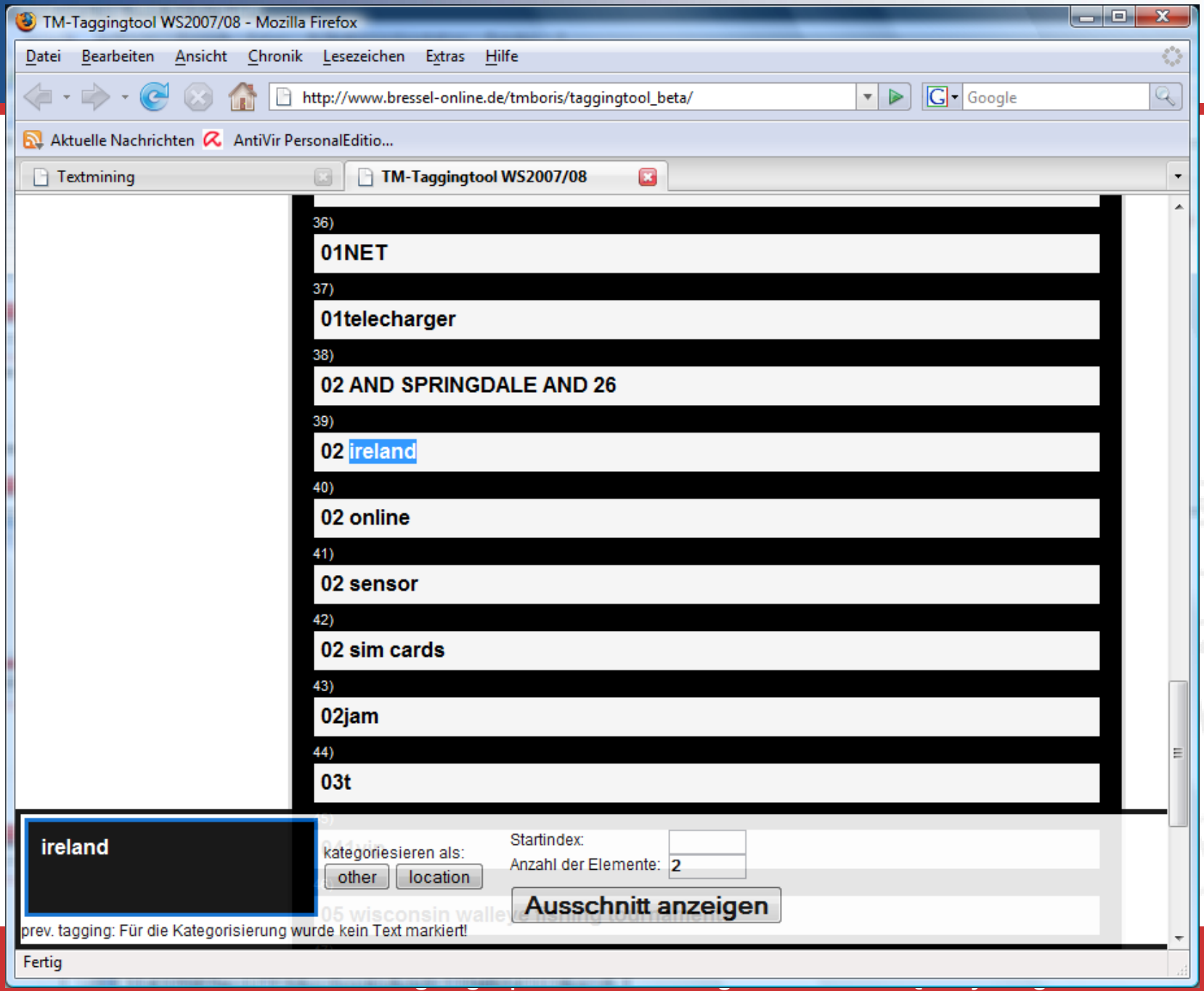
- Gazeteers for location identification
  - Large databases of geo names
- Pre-defined Rules

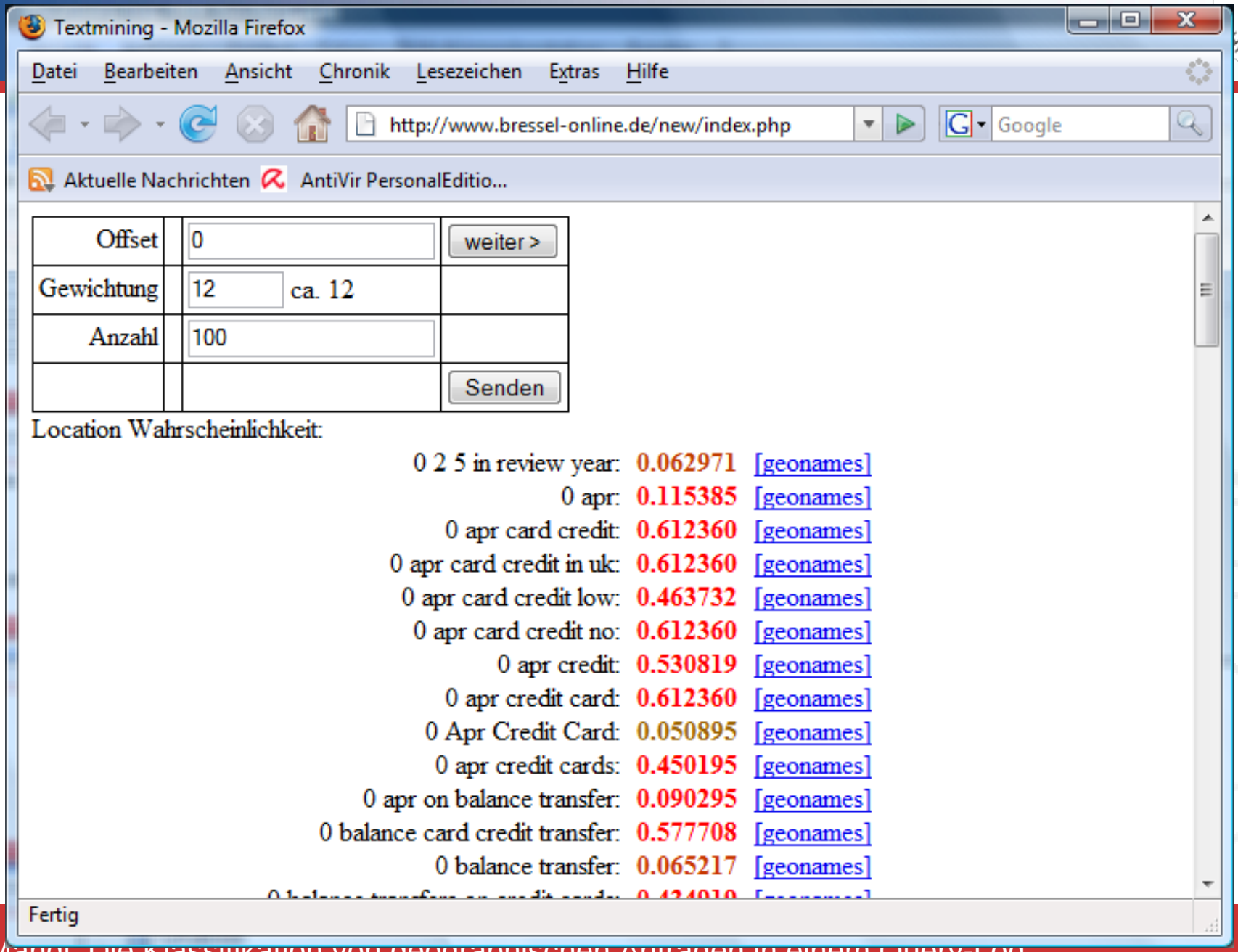
- Issues
  - Low Perfomance
  - Few training classes for many geo-types

Approaches

- Small Project in Hildesheim

- Classification
  - Statistical classifier based on keyword frequencies
  - Geo-Name database
  - Tagging Tool to gather more trainings examples


TM-Taggingtool WS2007/08 - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.bressel-online.de/tmboris/taggingtool_beta/
Google
Aktuelle Nachrichten
AntiVir PersonalEditio...
Textmining
TM-Taggingtool WS2007/08
36)
01NET
37)
01telecharger
38)
02 AND SPRINGDALE AND 26
39)
02 ireland
40)
02 online
41)
02 sensor
42)
02 sim cards
43)
02jam
44)
03t
ireland
kategoriesieren als:
other
location
Startindex:
Anzahl der Elemente: 2
Ausschnitt anzeigen
prev. tagging: Für die Kategorisierung wurde kein Text markiert!
Fertig


Textmining - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.bressel-online.de/new/index.php
Google
Aktuelle Nachrichten
AntiVir PersonalEditio...
Offset
0
weiter >
Gewichtung
12
ca. 12
Anzahl
100
Senden
Location Wahrscheinlichkeit:
0 2 5 in review year: 0.062971 [geonames]
0 apr: 0.115385 [geonames]
0 apr card credit: 0.612360 [geonames]
0 apr card credit in uk: 0.612360 [geonames]
0 apr card credit low: 0.463732 [geonames]
0 apr card credit no: 0.612360 [geonames]
0 apr credit: 0.530819 [geonames]
0 apr credit card: 0.612360 [geonames]
0 Apr Credit Card: 0.050895 [geonames]
0 apr credit cards: 0.450195 [geonames]
0 apr on balance transfer: 0.090295 [geonames]
0 balance card credit transfer: 0.577708 [geonames]
0 balance transfer: 0.065217 [geonames]
Fertig

Negative Example

- „Credit Card“ -> classified as Geo Query
- Positive Training examples
  - „Credit River“
  - „Port Credit“
  - „Card Gulch“
  - „Card Bay“

Vielen Dank für Ihre

Aufmerksamkeit