Construction and Analysis of a Known-Item Question Corpus for the ClueWeb09

Daniel Wägner

Bauhaus-Universität Weimar
daniel.waegner@uni-weimar.de

Bachelorverteidigung
17. Juli 2013

Gliederung

Motivation

Known Items

Was sind Known Items?

- Known Items sind Objekte, auf die ein Nutzer zuvor bereits zugegriffen hat und von denen er einen gewissen Grad an Kenntnis bzw. Informationen besitzt.
- Aufgabenstellung einer Known-Item-Suche ist das Wiederfinden oder Wiederzugänglichmachen eines solchen Known Items.

Known Items

Beispiele für Known-Item-Typen

- Dokumente
- e-Mails
- Websites
- Filme
- Musikstücke
- Bücher, ...

Known Items

Beispiele für Known-Item-Typen

- Dokumente, e-Mails } Personal Information Management (PIM)
- Websites — Web Page Re-finding
- Filme, Musikstücke } Multimedia Retrieval
- Bücher, ...

Known Items

Beispiele für Known-Item-Typen

- Dokumente
- e-Mails
- Websites
- Filme
- Musikstücke
- Bücher, ...

} Webis-KIQ-13

Known Items

Repräsentation von Known Items

- Dokumente
- e-Mails
- Websites · URL
- Filme · ISAN, (IMDb ID)
- Musikstücke · ISRC, ISWC, ISMN
- Bücher, ... · ISBN, ...

Known Items

Repräsentation von Known Items

- Dokumente
- e-Mails
- Websites — URL
- Filme
- Musikstücke
- Bücher, ... } Wikipedia-Eintrag

Known Items

Repräsentation von Known Items

- Dokumente
- e-Mails
- Websites → URL
- Filme, Musikstücke, Bücher, ... → Wikipedia-Eintrag

(URL, Wikipedia-Eintrag) → ClueWeb09

ClueWeb09

Informationen zum ClueWeb09 Dataset

- Statischer Webcrawl, enthält ca. eine Milliarde Webseiten in zehn Sprachen.
- Davon ca. 500 Millionen englische Webseiten, inklusive einem nahezu vollständigem Dump der englischen Wikipedia.
- Erstellungszeitraum Januar und Februar 2009.
- 5 TB komprimiert, 25 TB unkomprimiert.
- Zum Zugriff auf den ClueWeb09-Korpus wird in dieser Arbeit die an der BU Weimar entwickelte ChatNoir-Suchengine verwendet.
- Seit Anfang 2013 wird mit dem ClueWeb12 ein aktualisierter Datensatz herausgegeben. Für die vorliegende Arbeit stand dieser jedoch nicht rechtzeitig zur Verfügung.

Known Items

Quellen für Known-Item-Korpora

- Dokumente, e-Mails } Personal Information Management (PIM)
- Websites — Web Page Re-finding
- Filme, Musikstücke } Multimedia Retrieval
- Bücher, ...

Known Items

Quellen für Known-Item-Korpora

- Dokumente, e-Mails } Persönlicher Arbeitsplatz, Pseudo-Desktops
- Websites — Search-Logs, Query-Logs, Web-Directories
- Filme, Musikstücke } Audio-/Video-Korpora
- Bücher, ...

Known Items

Quellen für Known-Item-Korpora

- Dokumente, e-Mails — Blanc-Brude & Scapin, IUI '07; Kim & Croft, CIKM '09
- Websites, Filme — Beitzel et al., SIGIR '03; Hauff, Hagen, Beyer & Stein, IiX '12
- Musikstücke
- Bücher, ...

Known Items

Quellen für Known-Item-Korpora

- Dokumente
- e-Mails
- Websites
- Filme
- Musikstücke
- Bücher, ...

} Webis-KIQ-13 $\longrightarrow$ YAHOO! Answers

Yahoo! Answers

YAHOO! Answers

Informationen zur Plattform

- In Betrieb seit 13.12.2005, laut eigenen Angaben 250 Mio. Nutzer.
- Deckt Themen in 26 Haupt- mit über 300 Unterkategorien ab.
- Nutzer können Fragen stellen, Antworten vorschlagen und für die beste Antwort auf eine Frage abstimmen.
- Operiert unter einem Punktesystem, welches zur Beteiligung anregen und hervorragende Antworten auszeichnen soll.
- Stellt eine öffentliche API zum Zugriff auf Fragen und Antworten bereit.

Yahoo! Answers

Home > All Categories > Entertainment & Music > Movies > Resolved Question

Harry

Resolved Question Show me another »

Looking for a film title I can't remember.?

Hello, the film was about a man who was a sniper, then a think Morgan freeman offers him a job to kill a person, the man packs up his things, and goes to the mission when he arrives Morgan shoots him and when he tries to escape he flips a police officer, sorry that's all I remember, will give 5 stars to the person with he correct title.

11 months ago Report Abuse

Additional Details
It's not the first three answers sorry.
11 months ago

Not wanted either it nevermind.
11 months ago

Jennifer

Best Answer - Chosen by Asker

[redacted]

Report Abuse

[redacted]

1 Interesting! Email Comment (0) Save

This question about "Looking for a film t... " was originally asked on Yahoo! Answers United Kingdom

Other Answers (6) **Show:** All Answers

Yahoo! Answers

Home > All Categories > Entertainment & Music > Movies > Resolved Question

Harry

Resolved Question Show me another »

Looking for a film title I can't remember.?

Hello, the film was about a man who was a sniper, then a think Morgan freeman offers him a job to kill a person, the man packs up his things, and goes to the mission when he arrives Morgan shoots him and when he tries to escape he flips a police officer, sorry that's all I remember, will give 5 stars to the person with he correct title.

11 months ago Report Abuse

Additional Details
It's not the first three answers sorry.
11 months ago

Not wanted either it nevermind.
11 months ago

Jennifer

Best Answer - Chosen by Asker

The only sniper movie I can think of is Shooter, but that's Danny Glover
Source(s):
http://www.imdb.com/title/tt0822854/

11 months ago Report Abuse

Asker's Rating: ★★★★★
Thanks this was the film I watched it ages ago forgot the cast

1 Interesting! Email Comment (0) Save

This question about "Looking for a film t... " was originally asked on Yahoo! Answers United Kingdom

Other Answers (6) **Show:** All Answers

Korpusaufbau

Korpusaufbau

```
(remember) AND (title) AND (movie)
(forgot) AND (name) AND (film)
(forgot) AND (title) AND (song)
(forgot) AND (url) AND (website OR (web site))
(remember OR forgot) AND (name OR title) AND (book)
```

Anfragen an die Yahoo! Answers API

- Jeweils neun Anfragen für die Themen Filme, Musik und Websites (Hauptschwerpunkte des Korpus) sowie je eine Anfrage für zehn Nebenschwerpunkte (u.a. Bücher, TV-Serien, Comics)
- Nur beantwortete Fragen wurden berücksichtigt.
- Von der Yahoo! Answers API wurden insgesamt 24,759 verschiedene Fragen zurückgegeben.

Korpusaufbau

Auswahl der Fragen für den Webis-KIQ-13-Korpus

- Zunächst wurden jene Fragen entfernt, deren beste Antwort nicht vom Fragesteller ausgewählt wurde.
- Für die verbleibenden Fragen wurde einzeln überprüft, ob es sich um ein Known-Item-Problem handelt und die Antwort das gesuchte Objekt enthält. Andere Fragetypen wurden verworfen.
- Im nächsten Schritt wurden die Known Items für den Korpus ausgewählt, deren URL im ClueWeb09 enthalten ist.
- Sofern die Frage False Memories enthielt, wurden diese markiert und eine Korrekturnotiz angebracht.
- Zuletzt wurden die Known-Item-Fragen mit den Annotationen in den Webis-KIQ-13-Korpus zusammengeführt
- Aufwand: ca. 200 h für 8825 Fragen, bzw. 80 s pro Frage

Korpusaufbau

Auswahl der Fragen für den Webis-KIQ-13-Korpus

- Zunächst wurden jene Fragen entfernt, deren beste Antwort nicht vom Fragesteller ausgewählt wurde.
- Für die verbleibenden Fragen wurde einzeln überprüft, ob es sich um ein Known-Item-Problem handelt und die Antwort das gesuchte Objekt enthält. Andere Fragetypen wurden verworfen.
- Im nächsten Schritt wurden die Known Items für den Korpus ausgewählt, deren URL im ClueWeb09 enthalten ist.
- Sofern die Frage False Memories enthielt, wurden diese markiert und eine Korrekturnotiz angebracht.
- Zuletzt wurden die Known-Item-Fragen mit den Annotationen in den Webis-KIQ-13-Korpus zusammengeführt
- Aufwand: ca. 200 h für 8825 Fragen, bzw. 80 s pro Frage

Korpusaufbau

Auswahl der Fragen für den Webis-KIQ-13-Korpus

- Zunächst wurden jene Fragen entfernt, deren beste Antwort nicht vom Fragesteller ausgewählt wurde.
- Für die verbleibenden Fragen wurde einzeln überprüft, ob es sich um ein Known-Item-Problem handelt und die Antwort das gesuchte Objekt enthält. Andere Fragetypen wurden verworfen.
- Im nächsten Schritt wurden die Known Items für den Korpus ausgewählt, deren URL im ClueWeb09 enthalten ist.
- Sofern die Frage False Memories enthielt, wurden diese markiert und eine Korrekturnotiz angebracht.
- Zuletzt wurden die Known-Item-Fragen mit den Annotationen in den Webis-KIQ-13-Korpus zusammengeführt
- Aufwand: ca. 200 h für 8825 Fragen, bzw. 80 s pro Frage

Korpusaufbau

Auswahl der Fragen für den Webis-KIQ-13-Korpus

- Zunächst wurden jene Fragen entfernt, deren beste Antwort nicht vom Fragesteller ausgewählt wurde.
- Für die verbleibenden Fragen wurde einzeln überprüft, ob es sich um ein Known-Item-Problem handelt und die Antwort das gesuchte Objekt enthält. Andere Fragetypen wurden verworfen.
- Im nächsten Schritt wurden die Known Items für den Korpus ausgewählt, deren URL im ClueWeb09 enthalten ist.
- Sofern die Frage False Memories enthielt, wurden diese markiert und eine Korrekturnotiz angebracht.
- Zuletzt wurden die Known-Item-Fragen mit den Annotationen in den Webis-KIQ-13-Korpus zusammengeführt
- Aufwand: ca. 200 h für 8825 Fragen, bzw. 80 s pro Frage

Korpusaufbau

Auswahl der Fragen für den Webis-KIQ-13-Korpus

- Zunächst wurden jene Fragen entfernt, deren beste Antwort nicht vom Fragesteller ausgewählt wurde.
- Für die verbleibenden Fragen wurde einzeln überprüft, ob es sich um ein Known-Item-Problem handelt und die Antwort das gesuchte Objekt enthält. Andere Fragetypen wurden verworfen.
- Im nächsten Schritt wurden die Known Items für den Korpus ausgewählt, deren URL im ClueWeb09 enthalten ist.
- Sofern die Frage False Memories enthielt, wurden diese markiert und eine Korrekturnotiz angebracht.
- Zuletzt wurden die Known-Item-Fragen mit den Annotationen in den Webis-KIQ-13-Korpus zusammengeführt
- Aufwand: ca. 200 h für 8825 Fragen, bzw. 80 s pro Frage

Korpusaufbau

Auswahl der Fragen für den Webis-KIQ-13-Korpus

- Zunächst wurden jene Fragen entfernt, deren beste Antwort nicht vom Fragesteller ausgewählt wurde.
- Für die verbleibenden Fragen wurde einzeln überprüft, ob es sich um ein Known-Item-Problem handelt und die Antwort das gesuchte Objekt enthält. Andere Fragetypen wurden verworfen.
- Im nächsten Schritt wurden die Known Items für den Korpus ausgewählt, deren URL im ClueWeb09 enthalten ist.
- Sofern die Frage False Memories enthielt, wurden diese markiert und eine Korrekturnotiz angebracht.
- Zuletzt wurden die Known-Item-Fragen mit den Annotationen in den Webis-KIQ-13-Korpus zusammengeführt
- Aufwand: ca. 200 h für 8825 Fragen, bzw. 80 s pro Frage

Korpusaufbau

	Filme	Musik	Websites	Gesamt
Erhaltene Fragen	5896	6481	5343	24759
Chosen by Voters	-3718	-4112	-3637	-15934
Chosen by Asker	2178	2369	1706	8825

Korpusaufbau

	Filme	Musik	Websites	Gesamt
Erhaltene Fragen	5896	6481	5343	24759
Chosen by Voters	-3718	-4112	-3637	-15934
Chosen by Asker	2178	2369	1706	8825
Sonstige Fragen	-768	-1451	-1624	-5419
Known-Item-Fragen	1410	918	82	3406

Korpusaufbau

	Filme	Musik	Websites	Gesamt
Erhaltene Fragen	5896	6481	5343	24759
Chosen by Voters	-3718	-4112	-3637	-15934
Chosen by Asker	2178	2369	1706	8825
Sonstige Fragen	-768	-1451	-1624	-5419
Known-Item-Fragen	1410	918	82	3406
Nicht im ClueWeb09	-250	-219	-20	-651
Webis-KIQ-13	1160	699	62	2755

Korpusaufbau

	Filme	Musik	Websites	Gesamt
Erhaltene Fragen	5896	6481	5343	24759
Chosen by Voters	-3718	-4112	-3637	-15934
Chosen by Asker	2178	2369	1706	8825
Sonstige Fragen	-768	-1451	-1624	-5419
Known-Item-Fragen	1410	918	82	3406
Nicht im ClueWeb09	-250	-219	-20	-651
Webis-KIQ-13	1160	699	62	2755
False Memories	81	74	4	240

Annotation von False Memories

Known Item	False Memory	Korrektur
Shooter (film)	[...] Morgan freeman offers him a job to kill a person [...]	wrong actor: Danny Glover, not Morgan Freeman
Tokio Hotel	What's the english emo rock band [...] They are american [...]	origin: German band, not English or American
An American Tail	[...] a Disney cartoon about a little mouse [...]	company: Amblin Entertainment, not Disney
theforgottenlair.net	[...] it went somethin like the underground lair [...]	URL: "forgotten", not "underground"

Korpusanalyse

Text- und Readability-Maße

Methodik

- Untersucht wurden die Länge der Fragen und Antworten in Zeichen, Silben, Wörtern und Sätzen.
- Auf Basis dieser Text-Maße wurden die folgenden Lesbarkeits-Indizes berechnet: ARI, Flesch-Kincaid Grade Level, Gunning fog index und SMOG.
- Zu Vergleichszwecken wurden zusätzlich zu den im Webis-KIQ-13 enthaltenen Einträgen die 10,000 aktuellsten Fragen in den Kategorien Film und Musik von der API angefordert.

Text- und Readability-Maße

Auswertung

- Known-Item-Fragen sind im Mittel länger als andere Fragen.
- Im selben Themengebiet unterscheiden sich Fragen zu Known Items für gleiche Readability-Indizes nur unwesentlich von sonstigen Fragen.
- Dagegen unterscheiden sich die Einschätzungen der Lesbarkeit des gleichen Textes z.T. stark unter den verwendeten Readability-Formeln.
- Klassische Readability-Indizes haben nur eine eingeschränkte Aussagekraft für Q&A-Plattformen wie Yahoo! Answers.

Text- und Readability-Maße

Auswertung

- Known-Item-Fragen sind im Mittel länger als andere Fragen.
- Im selben Themengebiet unterscheiden sich Fragen zu Known Items für gleiche Readability-Indizes nur unwesentlich von sonstigen Fragen.
- Dagegen unterscheiden sich die Einschätzungen der Lesbarkeit des gleichen Textes z.T. stark unter den verwendeten Readability-Formeln.
- Klassische Readability-Indizes haben nur eine eingeschränkte Aussagekraft für Q&A-Plattformen wie Yahoo! Answers.

Text- und Readability-Maße

Auswertung

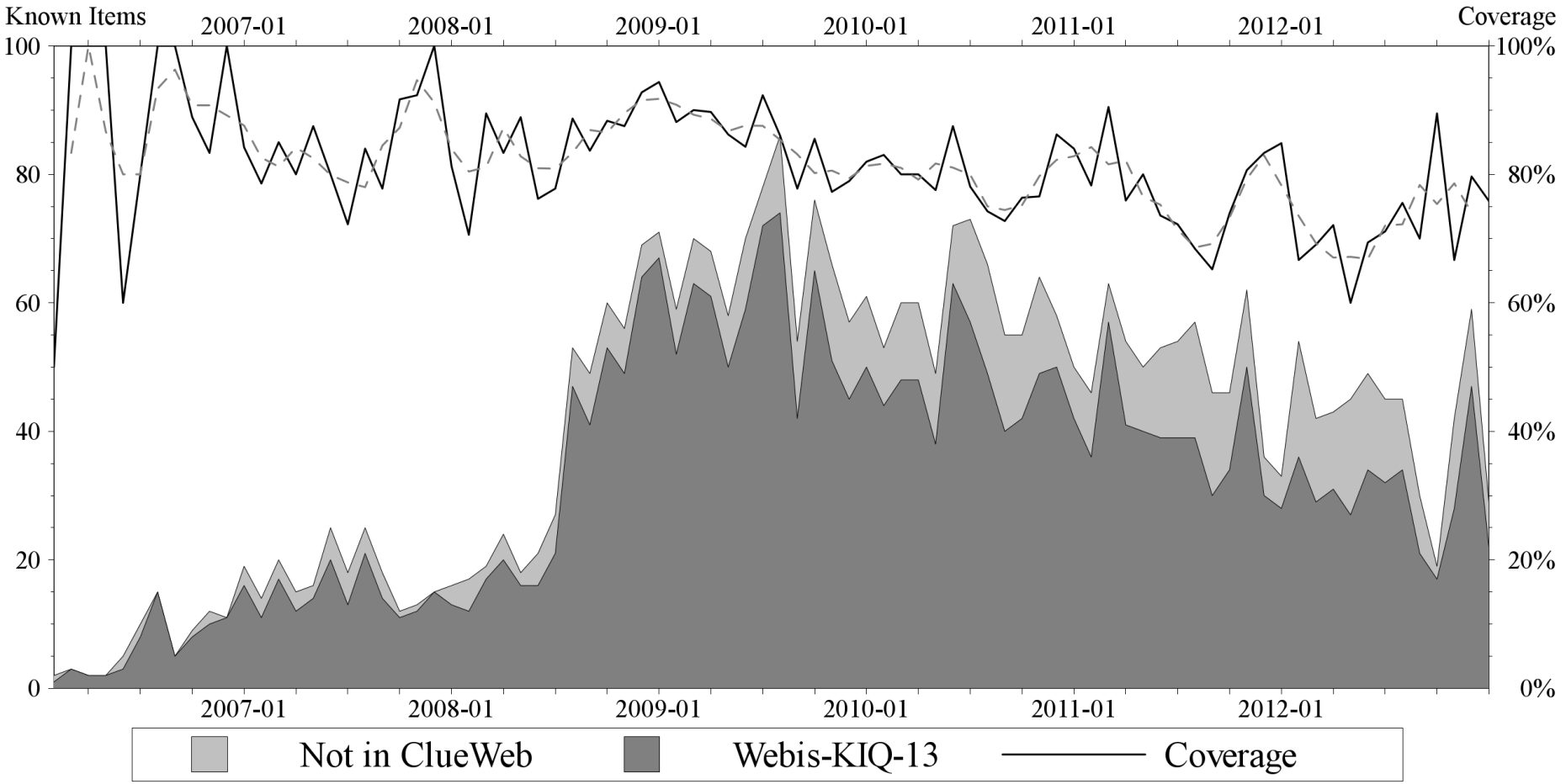
- Known-Item-Fragen sind im Mittel länger als andere Fragen.
- Im selben Themengebiet unterscheiden sich Fragen zu Known Items für gleiche Readability-Indizes nur unwesentlich von sonstigen Fragen.
- Dagegen unterscheiden sich die Einschätzungen der Lesbarkeit des gleichen Textes z.T. stark unter den verwendeten Readability-Formeln.
- Klassische Readability-Indizes haben nur eine eingeschränkte Aussagekraft für Q&A-Plattformen wie Yahoo! Answers.

Text- und Readability-Maße

Auswertung

- Known-Item-Fragen sind im Mittel länger als andere Fragen.
- Im selben Themengebiet unterscheiden sich Fragen zu Known Items für gleiche Readability-Indizes nur unwesentlich von sonstigen Fragen.
- Dagegen unterscheiden sich die Einschätzungen der Lesbarkeit des gleichen Textes z.T. stark unter den verwendeten Readability-Formeln.
- Klassische Readability-Indizes haben nur eine eingeschränkte Aussagekraft für Q&A-Plattformen wie Yahoo! Answers.

ClueWeb09-Abdeckung (Verlauf)



ClueWeb09-Abdeckung (nach Jahren)

	2006	2007	2008	2009	2010	2011	2012
Webis-KIQ-13	68	176	369	701	578	477	364
Nicht im ClueWeb09	8	15	60	112	148	140	142
Gesamt	76	191	429	813	726	617	506
Abdeckung	90%	92%	86%	86%	80%	77%	72%

ClueWeb09-Abdeckung (nach Domain)

	Wikipedia	IMDb	Andere	Keine URL	Alle
Webis-KIQ-13	2618	3	134	–	2755
Nicht im ClueWeb09	405	66	94	86	651
Gesamt	3023	69	228	86	3406
Abdeckung	87%	4%	59%	–	81%

ClueWeb09-Abdeckung

Auswertung

- Erwartung: höhere Abdeckungsraten mit dem ClueWeb12 realisierbar.
- Mit einem aktuellen Wikipedia-Dump (Anfang 2013, ca. 9 GB komprimiert) lässt sich bereits eine höhere Abdeckung an Known Items als mit dem ClueWeb09 (ca. 5 TB komprimiert) erzielen.
- Aber: Andere Websites werden nicht abgedeckt.
  Dumps werden nur für ein Jahr vorrätig gehalten.

ClueWeb09-Abdeckung

Auswertung

- Erwartung: höhere Abdeckungsraten mit dem ClueWeb12 realisierbar.
- Mit einem aktuellen Wikipedia-Dump (Anfang 2013, ca. 9 GB komprimiert) lässt sich bereits eine höhere Abdeckung an Known Items als mit dem ClueWeb09 (ca. 5 TB komprimiert) erzielen.
- Aber: Andere Websites werden nicht abgedeckt.
  Dumps werden nur für ein Jahr vorrätig gehalten.

ClueWeb09-Abdeckung

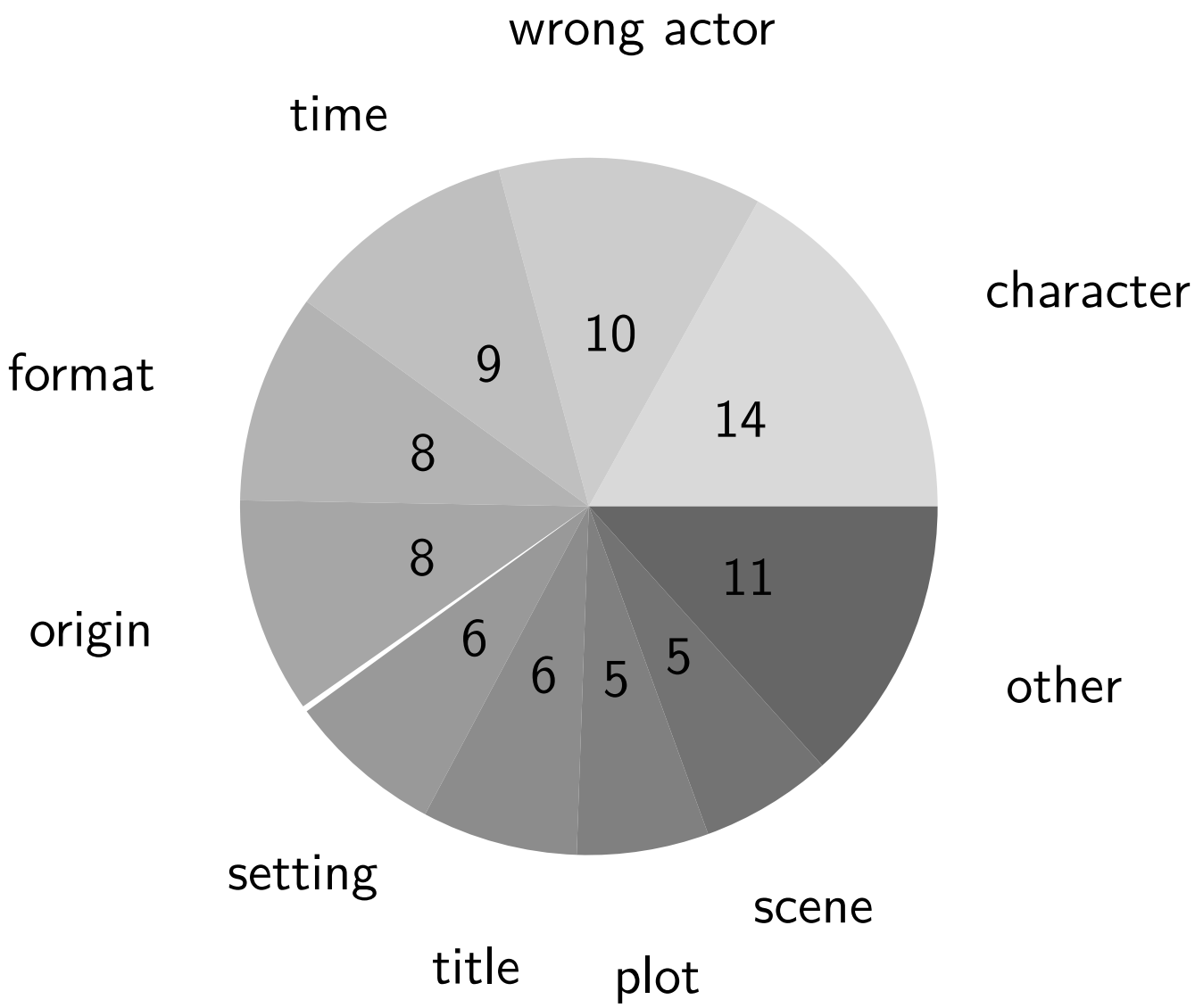
Auswertung

- Erwartung: höhere Abdeckungsraten mit dem ClueWeb12 realisierbar.
- Mit einem aktuellen Wikipedia-Dump (Anfang 2013, ca. 9 GB komprimiert) lässt sich bereits eine höhere Abdeckung an Known Items als mit dem ClueWeb09 (ca. 5 TB komprimiert) erzielen.
- Aber: Andere Websites werden nicht abgedeckt.
  Dumps werden nur für ein Jahr vorrätig gehalten.

False Memories

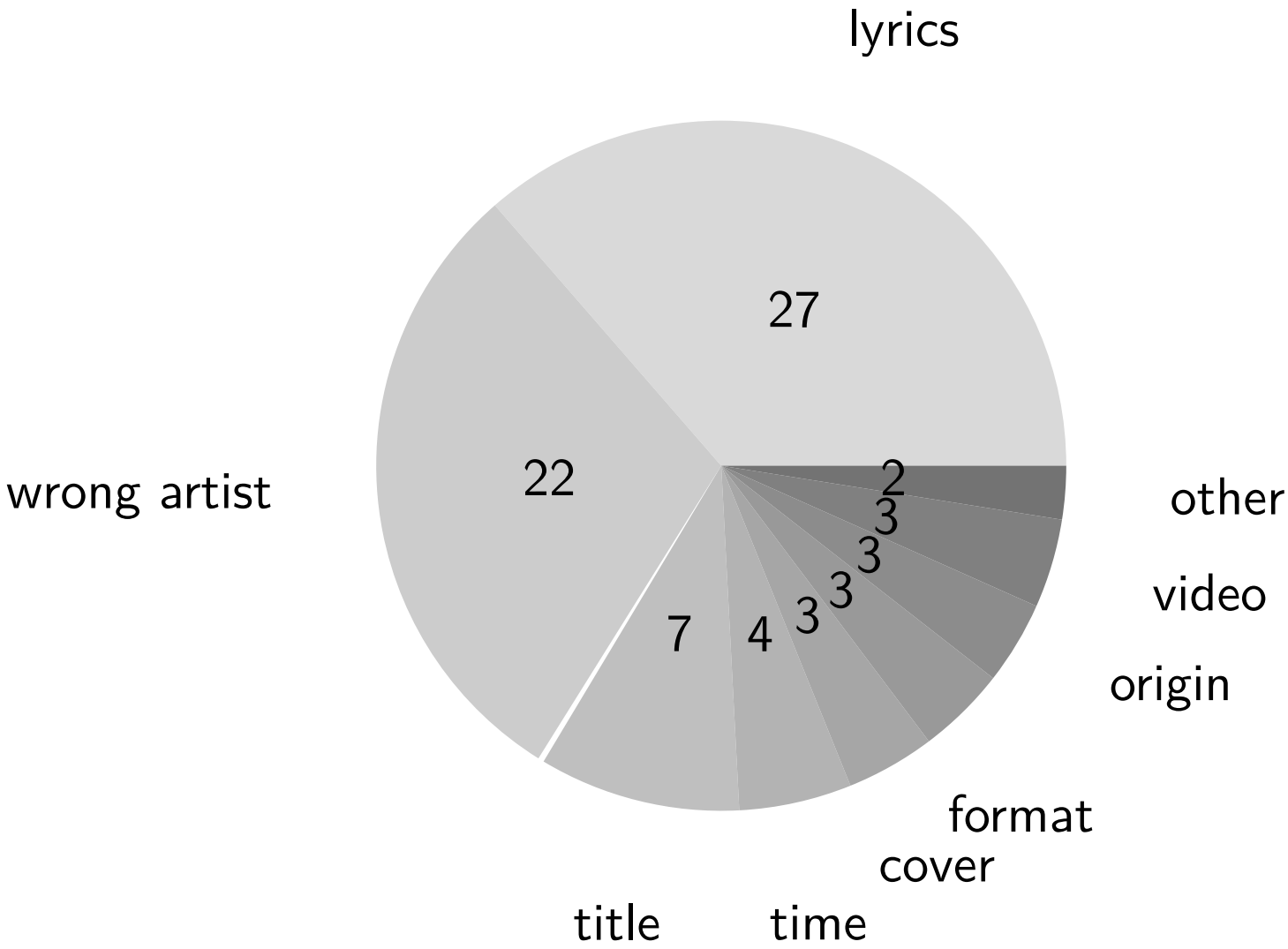
Kategorie	Falsche Erinnerungen bezüglich...	#
character	der Eigenschaften eines Charakters oder mehrerer Charaktere	34
lyrics	dem Text eines Lieds oder Gedichts	29
title	dem Titel des Known Items	27
wrong artist	der Zuordnung eines Künstlers zu einem Musikstück	22
format	dem Format, in dem etwas hergestellt oder veröffentlicht wurde	21
time	der Zeit der Herstellung oder Veröffentlichung	18
origin	dem geographischen Hintergrund eines Items bzw. einer Person	15
wrong actor	der Zuordnung eines Schauspielers zu einem Film	11
plot	dem Plot eines belletristischen Werks	9
setting	dem Setting eines belletristischen Werks	9
company	einem mit dem Known Item verknüpften Unternehmen	6
scene	einer bestimmten Szene in einem Film	5
prop	einer Requisite in einem Film oder Theaterstück	5
mix-up	Verwechslung mehrerer Known Items und ihrer Eigenschaften	5
URL	der URL einer Website	4

False Memories (Filme)



81 Fragen

False Memories (Musik)



74 Fragen

Häufig nachgefragte Known Items

Platz	Known Item	Anzahl
1	Once Upon a Forest	10
2	Internet Archive/Wayback Machine	6
2	Open Water 2: Adrift	6
2	The Pagemaster	6
5	Explorers (film)	4
5	Hammer House of Mystery and Suspense	4
5	Little Nemo: Adventures in Slumberland	4
5	Mystery Science Theater 3000	4
5	Spirited Away	4
5	Tamara (2005 film)	4
5	Tenchi Muyo!	4
5	The Phantom Tollbooth (film)	4
5	The Road To El Dorado	4
5	The Uninvited (2009 film)	4
5	YuYu Hakusho	4

Zusammenfassung

Zusammenfassung

Webis Known-Item Question Corpus 2013 (Webis-KIQ-13)

- 2755 Known-Item-Fragen, gemappt auf Einträge im ClueWeb09
- 240 annotierte False Memories

Zusammenfassung

Webis Known-Item Question Corpus 2013 (Webis-KIQ-13)

- 2755 Known-Item-Fragen, gemappt auf Einträge im ClueWeb09
- 240 annotierte False Memories

Ausblick

- Manuelle Anfrage-Generierung aus Known-Item-Fragen
- Mapping von Known Items auf das ClueWeb12 oder direkt auf einen Wikipedia-Dump
- Nutzung als Trainingsdaten für die automatische Klassifizierung von Yahoo! Answers-Fragen
- Schnittmengen mit dem Gebiet des Multimedia Retrieval
- Auswirkungen von False Memories in Suchanfragen