Evaluation of Ad-Hoc Information Retrieval Systems on Web Crawls with Redundant Documents

Jan Philipp Bittner

October 19, 2020

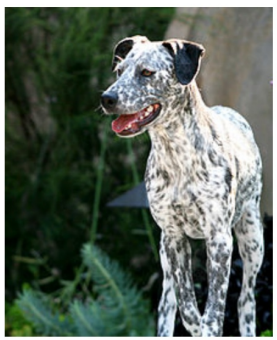
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid

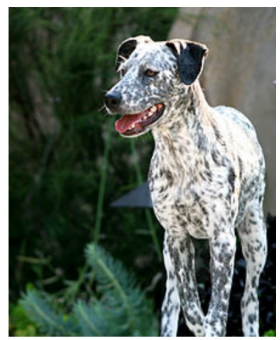


A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

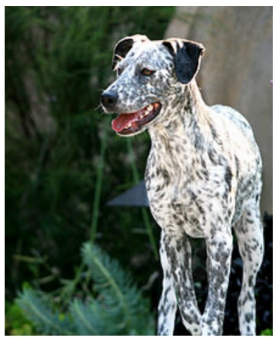
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid

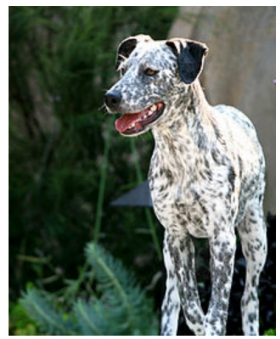


A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

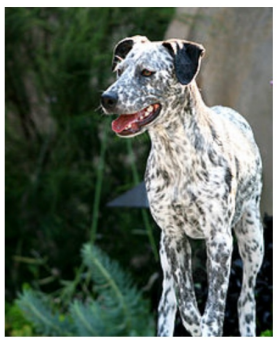
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid

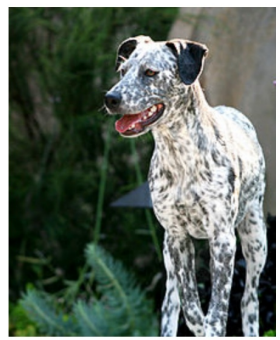


A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

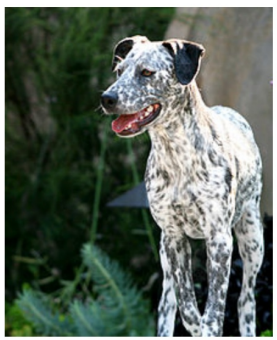
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



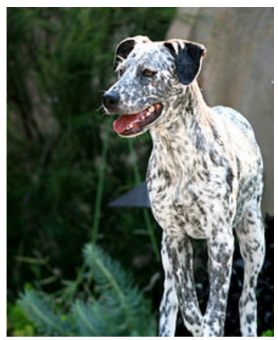
A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

✓

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

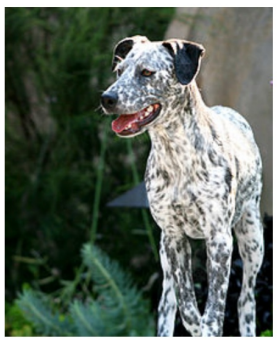
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



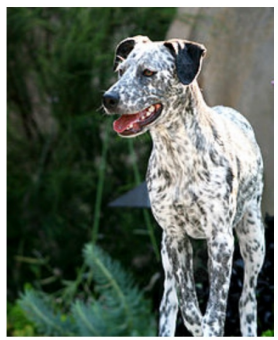
A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

✓

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

✓

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

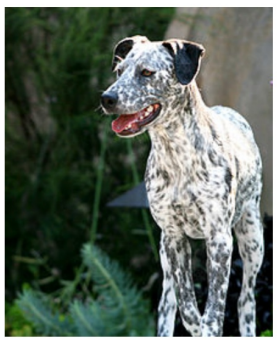
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



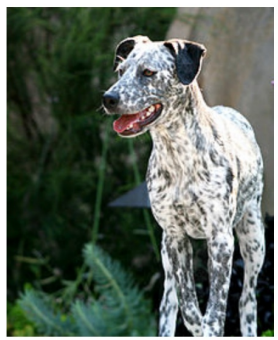
A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

✓

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

✓

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

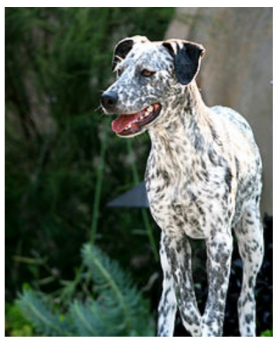
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- ❑ Benchmark nach Cranfield-Paradigma: Dokumente unabhängig voneinander bewerten
- ❑ Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



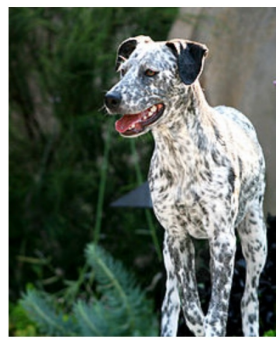
A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

✓

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

✓

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

✗

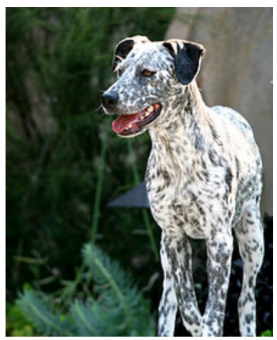
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid

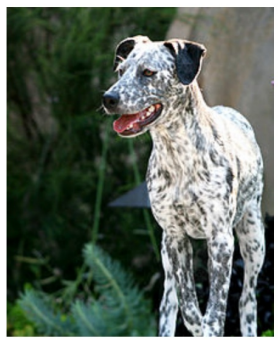


A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

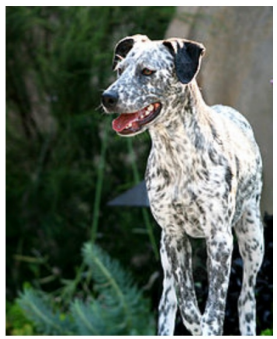
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid

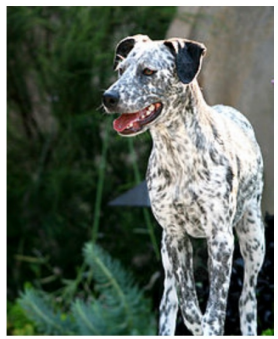


A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

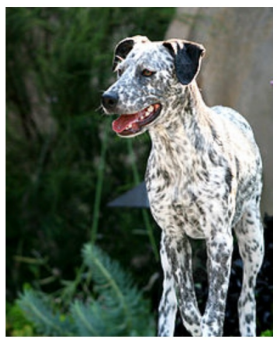
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



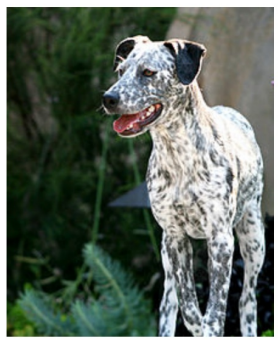
A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

✓

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

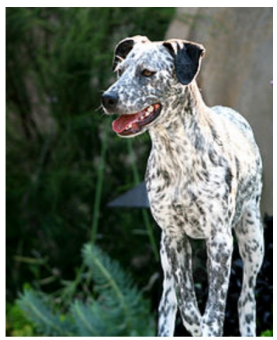
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- ❑ Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- ❑ Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



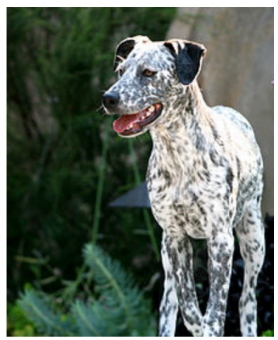
A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

✓

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

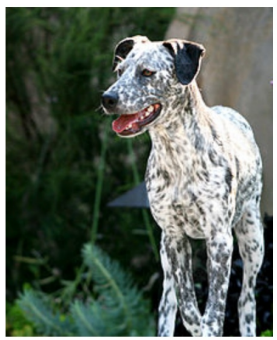
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



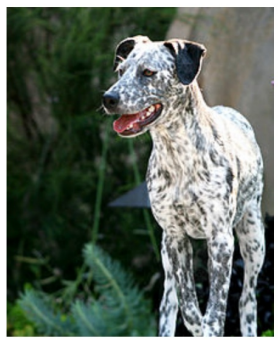
A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

✓

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

✗

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

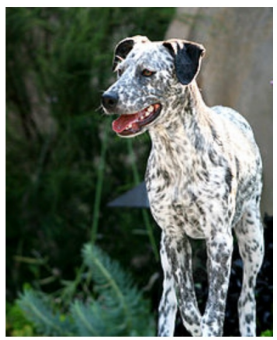
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



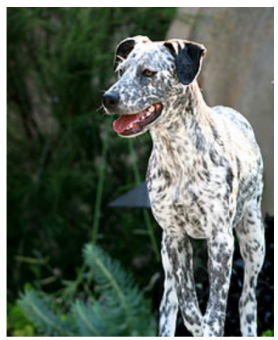
A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

✓

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

✗

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

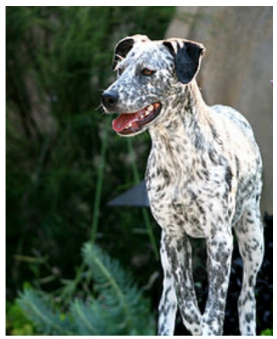
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

- Nutzersicht nach Novelty-Principle: Duplikate sind irrelevant
- Query: "designer dog breeds"

WIKIPEDIA

Dog hybrid



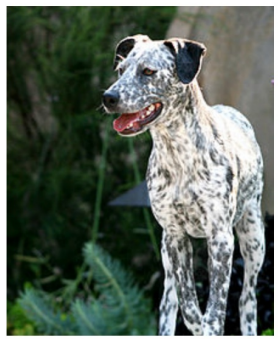
A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred dog breeds, traditionally called a ***crossbred dog***. Designer dogs are

✓

WIKIPEDIA

Dog hybrid



A pointer - dalmatian cross.

A **dog hybrid** (also called a designer dog) is a new term for an individual dog with ancestry in two different purebred do

Redirected from "Shihpom"

crossbred dog. Designer dogs are

✗

WIKIPEDIA

Mathematics

Mathematics (from Greek: μάθημα, *máthēma*, 'knowledge, study, learning') includes the study of such topics as quantity (number theory),[1] structure (algebra),[2] space (geometry),[1] and change (mathematical analysis).[3][4][5] It has no generally accepted definition.[6][7]

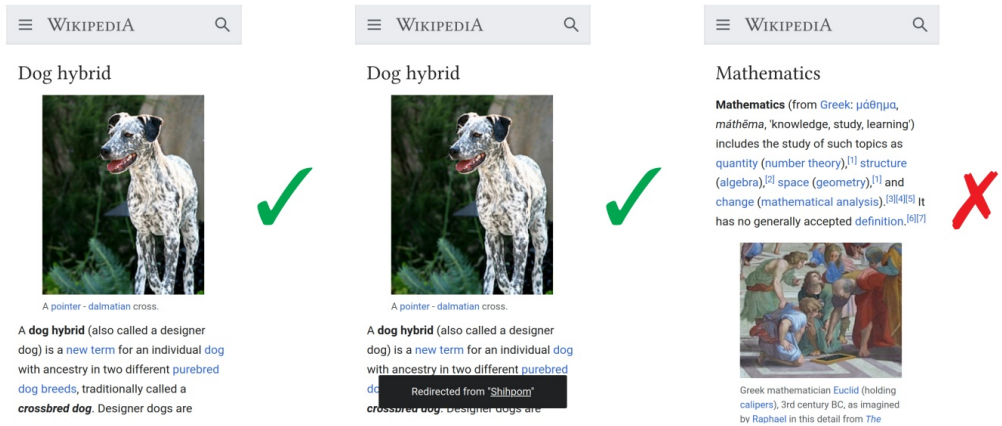
Greek mathematician Euclid (holding calipers), 3rd century BC, as imagined by Raphael in this detail from *The*

✗

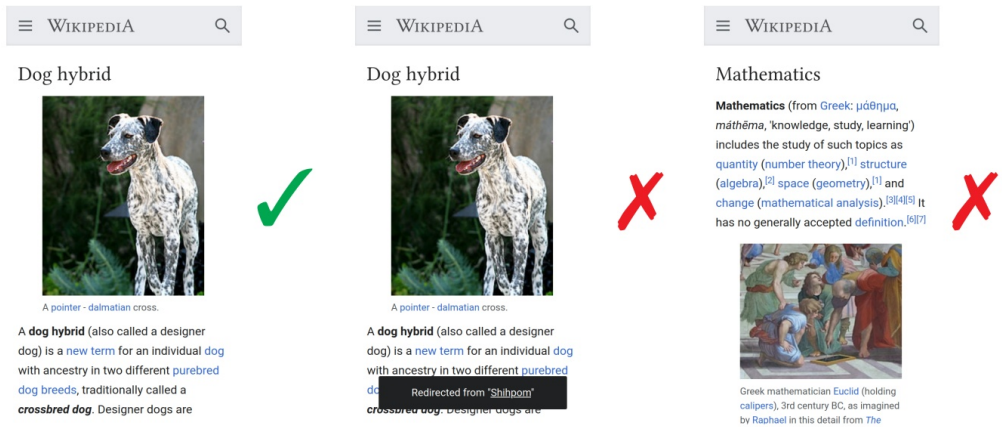
Einfluss von Near-Duplicates auf IR-Evaluationen

Motivation

Bewertung (Cranfield):



Realität (Novelty):



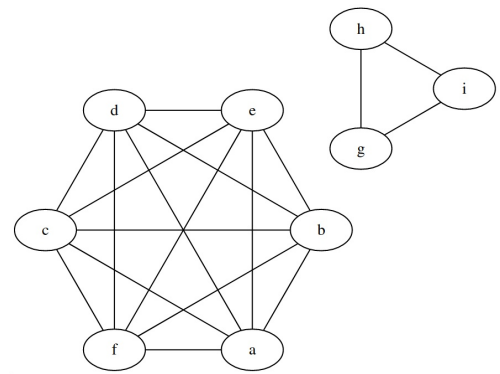
Beeinflussen Duplikate die Auswertung?
Wie viele Duplikate gibt es in den Corpora?

Einfluss von Near-Duplicates auf IR-Evaluationen

Pilotstudie - Retrieval-Equivalence

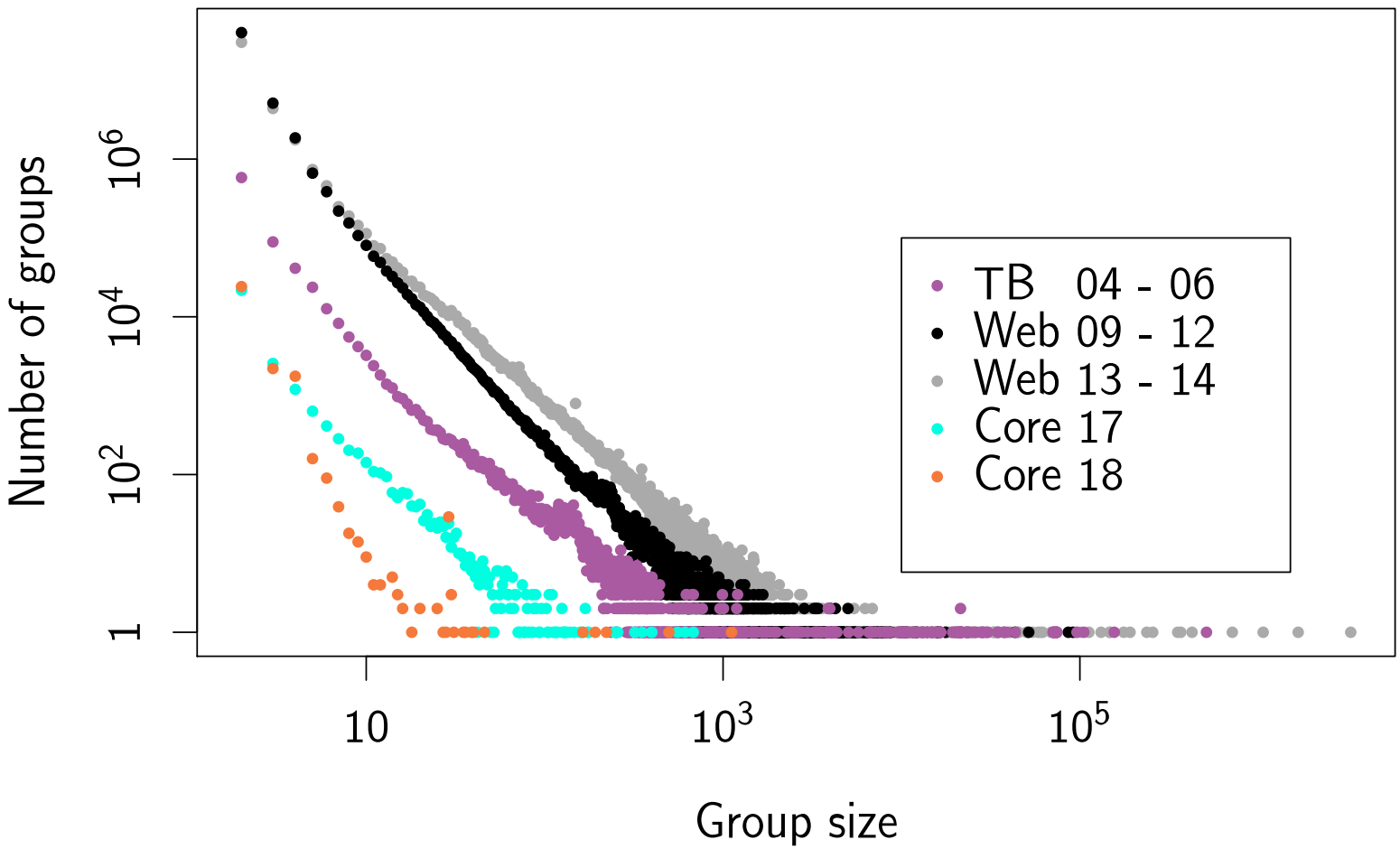
- Duplikate "kostengünstig" erkennen mittels Hashing
  - Äquivalenz bei Indexierung feststellbar (Retrieval-Equivalence)
- Pipeline: Text extrahieren, filtern, stemmen, hashen
- Gruppieren der Dokumente nach Hash-Wert
- Ergebnis: Gruppen äquivalenter Dokumente

Shared-Task	**Docs**	**Duplicates**
Terabyte 2004 - 2006	$2.5 * 10^7$	23.39%
Web 2009 - 2012	$1.0 * 10^9$	7.74%
Web 2013 - 2014	$7.3 * 10^8$	14.71%
Core 2017	$1.8 * 10^6$	2.11%
Core 2018	$5.9 * 10^5$	12.52%



Einfluss von Near-Duplicates auf IR-Evaluationen

Pilotstudie - Retrieval-Equivalence

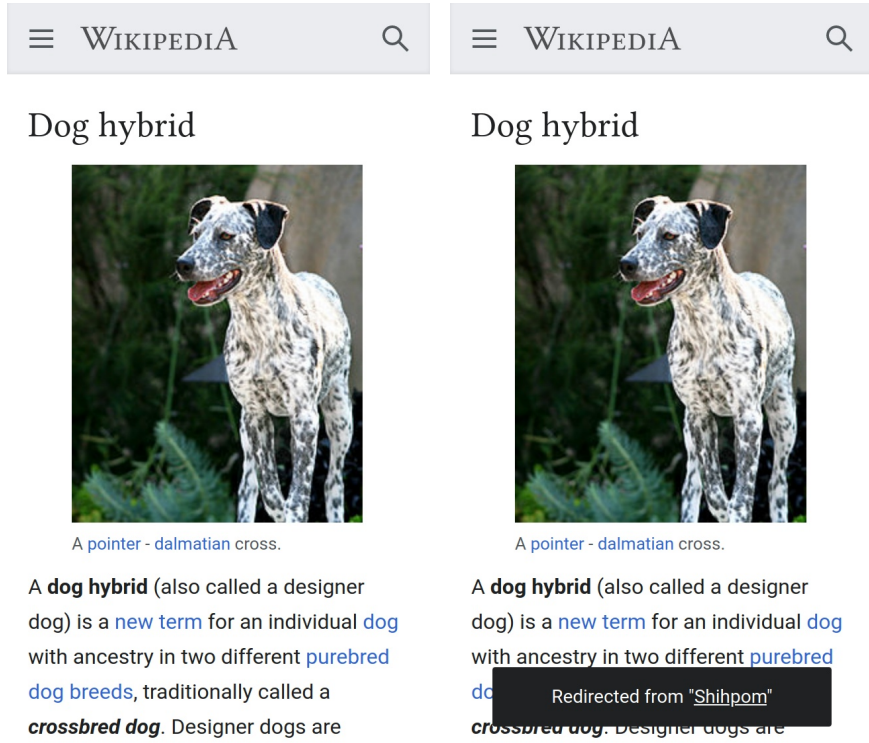


Einfluss von Near-Duplicates auf IR-Evaluationen

Pilotstudie - Retrieval-Equivalence

- Reicht Retrieval-Equivalence?
  - Nein, Near-Duplicates werden zum Teil nicht erkannt

- Dokumente mit unterschiedlichem Hash

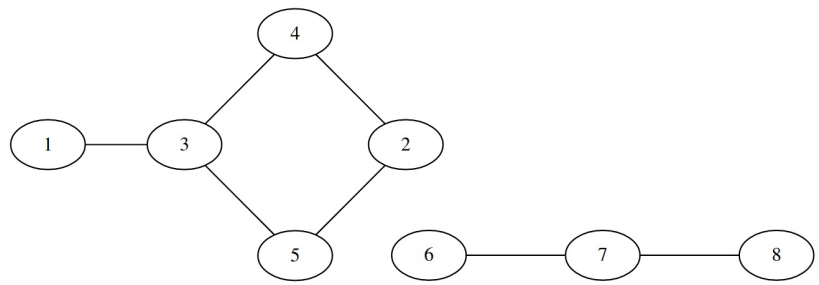


Einfluss von Near-Duplicates auf IR-Evaluationen

Content-Equivalence

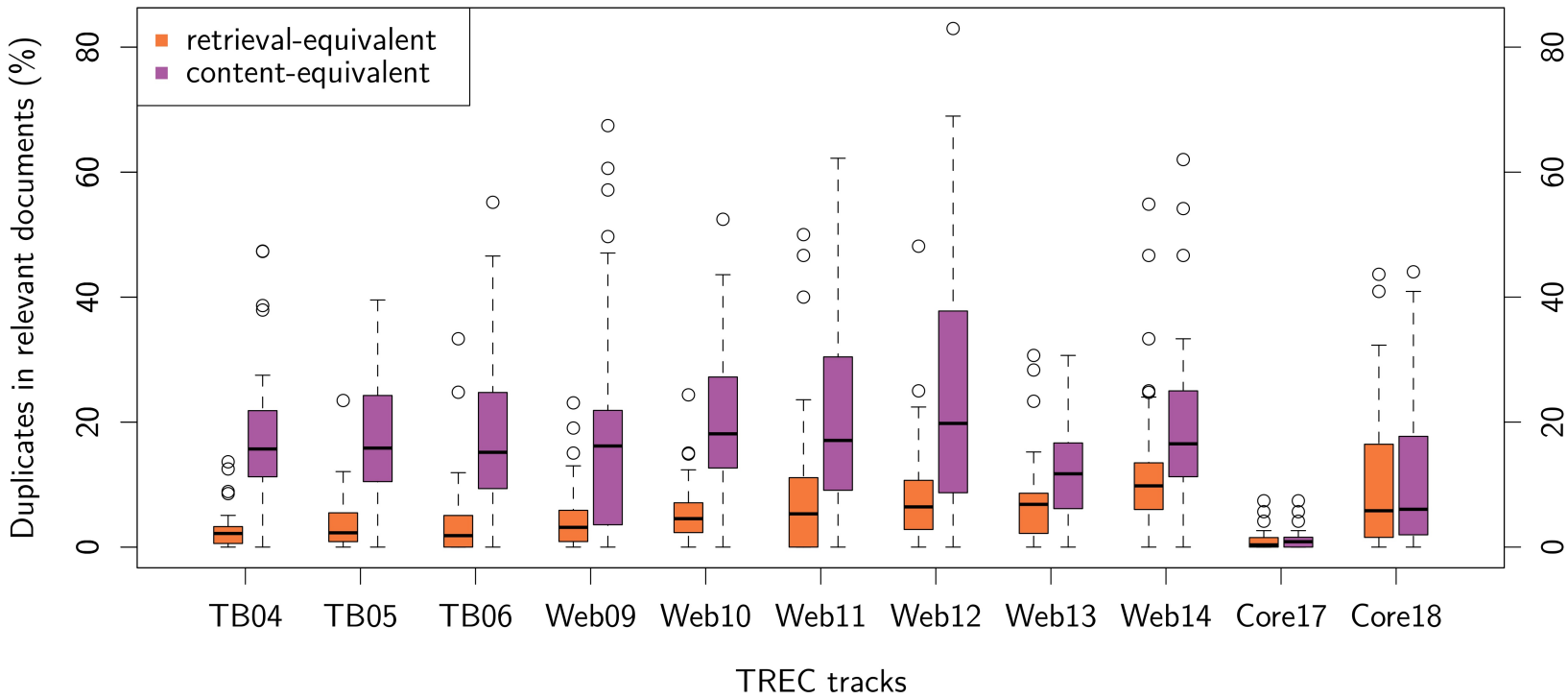
- Zwei Dokument sind Content-Equivalent, wenn sie hinreichende Überschneidungen haben
- Messung durch Bernstein und Zobel's S3-Score
  - Normalisierte Anzahl der gemeinsamen n-Gramme
  - Threshold trennt Duplikate und Unikate
  - Berechnung ganzer Corpora zu teuer $\rightarrow$ bewertete Dokumente
- Pipeline: Text extrahieren, filter, stemmen und **in n-Gramme zerlegen**
- Anzahl gemeinsamer n-Gramme zweier Dokumente zählen

- Kette von äquivalenten Dokumenten möglich
  - Beispiel: Versionsverläufe



Einfluss von Near-Duplicates auf IR-Evaluationen

Content-Equivalence



Einfluss von Near-Duplicates auf IR-Evaluationen

Impact

- Duplikat-sensitive Auswertung
  - Bewertung nach Novelty-Priciple
  - Vorher-Nacher-Vergleich

- nDCG: Gütemaß mit Werten zwischen 0 und 1, je höher desto besser
- Submissions schneiden schlechter ab
  - Duplikate werden angezeigt

Track		avg_{nDCG}	ΔnDCG
Terabyte	2004	0.425	-5.1%
	2005	0.586	-3.8%
	2006	0.654	-4.4%
Web	2009	0.323	-8.9%
	2010	0.302	-14.1%
	2011	0.341	-9.0%
	2012	0.295	-17.3%
	2013	0.324	-4.6%
	2014	0.380	-7.9%
Core	2017	0.560	-0.3%
	2018	0.541	-4.3%

Einfluss von Near-Duplicates auf IR-Evaluationen

Impact

- Duplikat-sensitive Auswertung
  - Bewertung nach Novelty-Priciple
  - Vorher-Nacher-Vergleich

- τ: Rangkorrelation zwischen konventioneller Auswertung und Duplikat-sensitiver Auswertung
- Sehr unterschiedlich starke Ausprägungen

	Track	τ	τ@5
Terabyte	2004	0.96	0.80
	2005	0.95	0.20
	2006	0.94	1.00
Web	2009	0.89	0.80
	2010	0.49	0.42
	2011	0.85	0.40
	2012	0.72	0.61
	2013	0.86	0.80
	2014	0.87	1.00
Core	2017	0.99	1.00
	2018	0.92	1.00

Einfluss von Near-Duplicates auf IR-Evaluationen

Impact

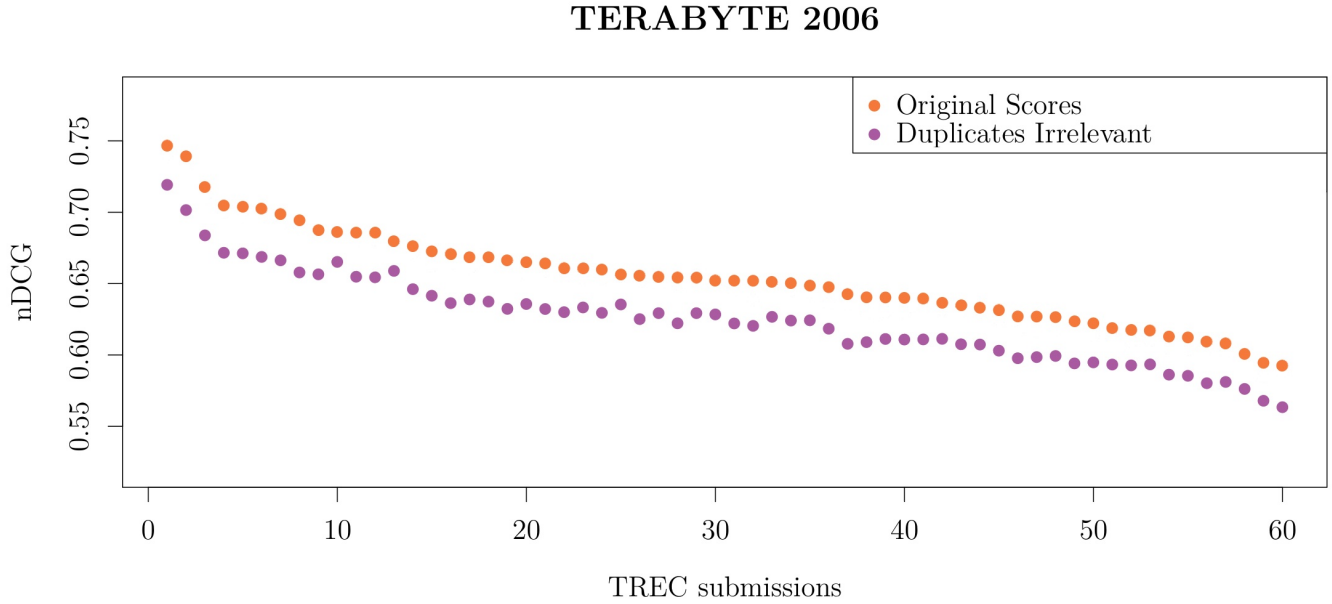
- Duplikat-sensitive Auswertung
  - Bewertung nach Novelty-Principle
  - Vorher-Nacher-Vergleich

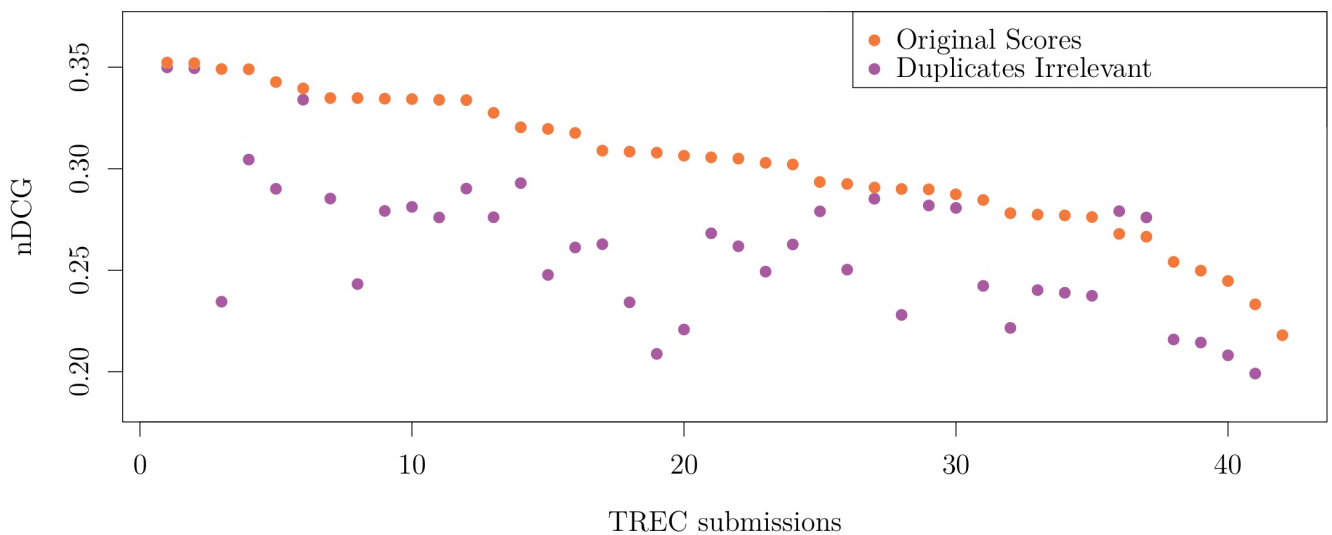
- i: Ideales System, setzt Novelty-Principle um
- max_i: Ideales System, das die meisten Plätze verliert
- med_i: Ideales System, das im Median bezüglich der Platzverluste liegt

Track		med_I	max_I
Terabyte	2004	-9.0	-19
	2005	-15.5	-27
	2006	-29.5	-53
Web	2009	-8.5	-24
	2010	-19.5	-39
	2011	-8.0	-13
	2012	-9.0	-16
	2013	-4.0	-8
	2014	-4.0	-11
Core	2017	-1.0	-9
	2018	-11.0	-26

Einfluss von Near-Duplicates auf IR-Evaluationen

Impact - Terabyte 2006 vs. Web 2010





Einfluss von Near-Duplicates auf IR-Evaluationen

Was können Organisatoren von Shared-Tasks machen?

- Duplikat-freie Copora
  - \+ Prozesse bleiben unverändert
  - – Erstellungskosten steigen mit Größe
- Duplikat-sensitive Auswertung
  - \+ Vollständige Auswertung nach Novelty-Principle
  - – Auswertung erfordert Duplikatsgruppen und zusätzliches Auswertungsprogramm
- riskante Topics entfernen
  - \+ Kompatibel mit allen Auswertungen
  - – Auswahlkriterien fragwürdig
  - – verbirgt Systeme, die Duplikate schlecht behandeln

Einfluss von Near-Duplicates auf IR-Evaluationen

Take Home Message

- Auch neuere Web-Corpora enthalten Duplikate
- Rankings werden durch Duplikate verzerrt
- Je weniger Duplikate, desto weniger Verzerrung
- Kein Universalkonzept zur Einflussminderung
  - beste Option: Duplikat-sensitive Auswertung

Einfluss von Near-Duplicates auf IR-Evaluationen

Take Home Message

- Auch neuere Web-Corpora enthalten Duplikate
- Rankings werden durch Duplikate verzerrt
- Je weniger Duplikate, desto weniger Verzerrung
- Kein Universalkonzept zur Einflussminderung
  - beste Option: Duplikat-sensitive Auswertung

Vielen Dank für Ihre Aufmerksamkeit! Nun zu ihren Fragen.

Kontakt: jan.bittner@student.uni-halle.de