

Digital Libraries WS 2018/2019

Übungsblatt 1

Aaron Winziers - 1176638; Michael Wolz - 1195270

18. November 2018

Aufgabe 1

a)

- Das Schiff „Titanic“
- Den Film „Titanic“
- Den damit verbundenen Unfall mit dem Eisberg

b)

	Google	Yahoo
Schiff	2	1
Film	2	7
Unfall	2	1
Sonstige	3	1

Eine Zuordnung ist nicht immer eindeutig möglich, da in der Sprache oftmals eine Bezeichnung für mehrere Bedeutungen verwendet wird.

c)

Der Suchbegriff könnte erweitert werden, so dass das Informationsbedürfnis besser abgedeckt wird. So könnte der Suchbegriff dann z.B. „Titanic Schiff“, „Titanic Unfall“ oder „Titanic Film“ lauten.

Aufgabe 2

a)

Welche Software gibt es, um eine digitale Bibliothek zu erstellen.

Suchbegriff: „digital library software“

Category:Digital library software - Wikipedia
https://en.wikipedia.org/wiki/Category:Digital_library_software ▼ Diese Seite übersetzen
Pages in category "Digital library software". The following 23 pages are in this category, out of 23 total. This list may not reflect recent changes (learn more).

Greenstone Digital Library Software: Welcome
www.greenstone.org/ ▼ Diese Seite übersetzen
Greenstone is a suite of software for building and distributing digital library collections. It provides a way of organizing information and publishing it on the web or ...

What is the most important software in Digital Libraries? - ResearchGate
https://www.researchgate.net/.../What_is_the_most_important_sof... ▼ Diese Seite übersetzen
I am a PhD student and I want to know more softwares used by digital libraries. I know some like Greenstone, DigiBis or Fedora, but I would like to know other ...

A Study on the Open Source Digital Library Software's: Special ... - arXiv
[PDF]
<https://arxiv.org/pdf/1212.4935> ▼ Diese Seite übersetzen
von S Trambo - 2012 - Zitiert von: 16 - Ähnliche Artikel
1. A Study on the Open Source Digital Library Software's: Special Reference to DSpace, EPrints and Greenstone. Shahkar Trambo. Department of Library and ...

Digital libraries: Comparison of 10 software: Library Collections ...
<https://www.tandfonline.com/doi/full/.../14649055.2012.107663...> - Diese Seite übersetzen
von M Andro - 2012 - Zitiert von: 21 - Ähnliche Artikel
03.12.2013 - Abstract. This article is an English abstract (and not an extract), it is a synthesis of a study published, in French, in a book about software for ...

Digital libraries: Comparison of 10 software - ScienceDirect
<https://www.sciencedirect.com/science/.../S146490551200022X> - Diese Seite übersetzen
von M Andro - 2012 - Zitiert von: 21 - Ähnliche Artikel
This article is an English abstract (and not an extract), it is a synthesis of a study published, in French, in a book about software for building digital libraries: Andro ...

Digital Library Software - SlideShare
<https://de.slideshare.net/tudlis/digital-library-software> ▼ Diese Seite übersetzen
23.11.2017 - A presentation on Digital Library Software by Rupesh Kumar A, Assistant Professor, Department of Studies and Research in Library and ...

Greenstone digital library software 2.80: a software suite for building ...
www.unesco.org/.../greenstone-digital-library-software-280-a-sof... ▼ Diese Seite übersetzen
Greenstone is a suite of software for building and distributing digital library collections. It provides a new way of organizing information and publishing it on the ...

Brief Survey of Digital Library Software Systems | U-M Library
<https://www.lib.umich.edu/.../library.../brief-survey-digital-librar...> ▼ Diese Seite übersetzen
08.07.2010 - DLPS is currently (July 2010) exploring possible avenues for the future development of DLXS. DLXS is a mature and robust digital library ...

Digital Library Open Source Software: A Comparative Study ...
www.academia.edu/.../Digital_Library_Open_Source_Software_... ▼ Diese Seite übersetzen
Digital Library Open Source Software: A Comparative Study M S. Patil Satish Kanamadi M S Patil, Librarian, St. John Institute of Pharmacy and Research

b)

Die relevanten Ergebnisse wurden in dem Screen in a) grün markiert. Es ergibt sich eine Präzision von 100%.

c)

Für den Recall müssen alle Ergebnisse der Anfrage betrachtet werden. Da dies in unserem Fall bei Google „ungefähr 568000000“ Ergebnisse sind, wäre es mit einem **extremen** Arbeitsaufwand verbunden, diese alle durchzugehen. Um den Recall von zwei Suchmaschinen zu vergleichen, könnte man die relevante Menge z.B. auf die ersten 100 Ergebnisse einschränken.

Aufgabe 3

a)

F-Measure

Seien R der Recall und P die Precision dann ist

$$F_{\beta} = (1 + \beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R}$$

Dabei gibt β an wie Recall und Precision gewichtet werden sollen. Bei $\beta > 1$ wird Precision höher gewichtet, bei $0 < \beta < 1$ wird Recall höher gewichtet. Das F-Measure ist ein Evaluationsmaß, das Precision und Recall eines Systems verrechnet.

- $F_{\beta} = 0$ kann dann vorkommen, wenn Precision oder Recall = 0, d.h. wenn keine relevante Ergebnisse geliefert werden.
- $F_{\beta} = 1$ kann dann vorkommen, wenn alle relevanten Ergebnisse geliefert werden (Recall = 1) und alle Ergebnisse relevant sind (Precision = 1)

Fallout

Sei FP die Menge der False-Positives oder die Menge der gelieferten Ergebnisse die nicht relevant sind und TN die Menge der True-Negatives, oder die Menge der nicht relevanten Daten die tatsächlich nicht in der Ergebnismenge liegen, dann

$$Fall - out = \frac{FP}{FP + TP}$$

- **Fallout = 0** kommt vor, wenn alle Ergebnisse relevant sind

- **Fallout = 1** kommt vor, wenn alle Ergebnisse *nicht* relevant sind

Quelle: https://www.wikipedia.com/en/Precision_and_recall

b)

TP (True-Positives) - Die Menge der relevante Ergebnisse

TN (True-Negatives) - Die Menge der nicht relevante Daten die nicht in der Ergebnismenge geliefert werden

FP (False-Positives) - Die Menge der nicht relevante Ergebnisse

FN (False-Negatives) - Die Menge der relevante Daten die nicht in der Ergebnismenge geliefert werden

Relevante Menge: $\{d_1, d_4, d_5, d_6, d_9, d_{13}, d_{17}, d_{18}, d_{26}\} = 9$

$S_1: \{d_1, d_3, d_6, d_7, d_9, d_{13}, d_{15}, d_{16}, d_{17}, d_{24}\}$

TP = 5, FN = 4, FP = 4, TN = 17

- Recall = $\frac{5}{5+4} = \frac{5}{9}$
- Precision = $\frac{5}{5+4} = \frac{5}{9}$
- Fall-out = $\frac{4}{4+17} = \frac{4}{21}$
- $F_1 = \frac{2 \cdot 5}{2 \cdot 5 + 4 + 4} = \frac{5}{9}$

$S_2: \{d_2, d_5, d_6, d_9, d_{10}, d_{13}, d_{17}, d_{18}, d_{21}, d_{22}, d_{24}, d_{26}, d_{27}\}$

TP = 7, FN = 2, FP = 6, TN = 15

- Recall = $\frac{7}{7+2} = \frac{7}{9}$
- Precision = $\frac{7}{7+6} = \frac{7}{13}$
- Fall-out = $\frac{6}{6+15} = \frac{6}{21}$
- $F_1 = \frac{2 \cdot 7}{2 \cdot 7 + 6 + 2} = \frac{7}{11}$

$S_3: \{d_1, d_3, d_6, d_{10}, d_{13}, d_{17}, d_{18}, d_{26}\}$

TP = 6, FN = 3, FP = 2, TN = 19

- Recall = $\frac{6}{6+3} = \frac{2}{3}$
- Precision = $\frac{6}{6+2} = \frac{3}{4}$
- Fall-out = $\frac{2}{2+19} = \frac{2}{21}$
- $F_1 = \frac{2 \cdot 6}{2 \cdot 6 + 2 + 3} = \frac{2}{3}$

Je nachdem welches Maß man als am sinnvollsten sieht sind Verschiedene Resultate das "Beste". Nach dem Recall hat S_2 das beste Ergebnis geliefert, nach Precision, F_1 und Fall-out war S_3 das beste Resultat.

c)

Um Recall = 1 zu erreichen geben wir alle Dokumente zurück

- Precision: Gleich der prozentualen Menge der Relevanten Ergebnisse
- Fallout: 1, da alle nicht relevante Daten in der Ergebnismenge enthalten sind
- F_1 : Abhängig davon wie groß die Menge der tatsächlich relevanten Ergebnisse ist im Verhältnis zu der der nicht relevanten

Um Fall-out = 0 zu erreichen geben wir keine Dokumente zurück

- Precision: Nicht definiert wegen Division von 0
- Recall: 0, da es keine True-Positives gibt
- F_1 : 0, da es keine True-Positives gibt

Der Korpus muss mindestens ein Dokument beinhalten.