Digital Libraries WS 2018/2019
Übungsblatt 8

Aaron Winziers - 1176638; Michael Wolz - 1195270

6. Januar 2019

Aufgabe 1

a/b)

1. Universität Trier **(Sehr relevant)**
2. Liste der Studentenverbindungen in Trier **(relevant)**
3. Hochschule Trier **(relevant)**
4. Hamline University **(relevant)**
5. Universität **(relevant)**
6. Trier **(relevant)**
7. Trierer Arbeitsstelle für Künstlersozialgeschichte **(relevant)**
8. Karl-Marx Universität **(relevant)**
9. Deutsches Wörterbuch **(relevant)**
10. Universität Triest **(nicht relevant)**
11. Werkzeug **(nicht relevant)**
12. Lars von Trier **(nicht relevant)**
13. Liste deutscher Palindrome **(nicht relevant)**
14. Keule **(nicht relevant)**
15. Stövchen **(nicht relevant)**
16. Sumpf **(nicht relevant)**
17. Schrein **(nicht relevant)**
18. Fertigkeit **(nicht relevant)**
19. Junge **(nicht relevant)**
20. Schlägel **(nicht relevant)**

c)

Präzision := $\frac{9}{20} = 45\%$

d)

```
<top>
<num> 1
<title> Programmiersprache
<desc> Informationen über Programmiersprachen
<narr> Relevante Dokumente definieren was eine Programmiersprache ist,
wie diese entstandensind und wofür sie gebraucht werden. Außerdem sind
Artikel zu expliziten Programmiersprachen relevant. Historische Ergeinisse
zur Entstehung von Programmiersprachen sind ebenfalls relevant.
</top>
```

1. Programmiersprache **(Sehr relevant)**
2. Liste von Programmiersprachen **(Sehr relevant)**
3. Python (Programmiersprache) **(relevant)**
4. C (Programmiersprache) **(relevant)**
5. Java (Programmiersprache) **(relevant)**
6. R (Programmiersprache) **(relevant)**
7. Kotlin (Programmiersprache) **(relevant)**
8. Objektorientierte Programmierung **(relevant)**
9. Scratch (Programmiersprache) **(relevant)**
10. C++ **(relevant)**
11. Esoterische Programmiersprache **(relevant)**
12. Computerprogramm **(relevant)**
13. Höhere Programmiersprache **(relevant)**
14. Pascal (Programmiersprache) **(relevant)**
15. Rust (Programmiersprache) **(relevant)**
16. Erlang (Programmiersprache) **(relevant)**
17. D (Programmiersprache) **(relevant)**
18. BASIC **(relevant)**
19. Haskell (Programmiersprache) **(relevant)**
20. Go (Programmiersprache) **(relevant)**

Präzision := $\frac{20}{20} = 100\%$

Aufgabe 2

a)

Alle Dokumente in einer (großen) Kollektion nach Relevanz zu bewerten ist ein sehr ressourcenintensiver Prozess. Außerdem wird es bei zunehmender Größe einer Kollektion schwieriger eine sinnvolles Verhältnis zwischen Effektivität, Effizienz und Kosten zu bilden.

b)

Pooling ist ein Prozess der versucht die Anzahl der Dokumente die nach ihrer Relevanz bewertet werden müssen stark reduziert und dadurch die Kosten für die Bewertung senkt. Um die Menge der zu bewertende Dokumente zu bilden wird die gleiche Anfrage von mehrere Suchmaschinen durchgeführt und deren besten Resultate in den Pool hinzugefügt. Der Einfluss von unterbewerteten Dokumenten ist deswegen vernachlässigbar weil der Pool die Menge der *besten* Resultate der Suchmaschinen ist. Das heißt dass alle weitere Dokumente die zum Pool hinzugefügt werden könnten zwar relevant sein könnten, jedoch würde die Qualität der Ergebnisse immer weiter senken mit zunehmender Größe des Pools.

c)

Eine Suchmaschine könnte ggf. relevante Dokumente, die nicht im Pool berücksichtigt werden, finden und würde dadurch sehr schlecht evaluiert werden.

Dieses Problem könnte z.B. dadurch gelöst werden indem man den Pool durch eine Größere Menge der besten Resultate der anderen Suchmaschinen erweitert. Diese Lösung würde aber zu eine Verringerung der Qualität des Pools führen da zunehmend irrelevante Dokumente betrachtet werden.

Alternativ ist es möglich die beste Resultate der neuen Suchmaschine zu den Pool hinzuzufügen um nicht nur die neue, sondern auch die andere Suchmaschinen neu zu evaluieren mittels des erweiterten Pools.

d)

Welches Testszenario besser ist, ist stark davon abhängig was man von den Benchmarks lernen will.

Mittels des ersten Szenarios ist es möglich mehr Recall orientiert zu testen, da die Menge der zu durchsuchende Dokumente viel Größer ist. Mittels der größeren Mengen der gefundenen Dokumente

Das zweite Szenario bietet eine bessere Möglichkeit um die Präzision der Systeme zu testen, da durch die höhere Anzahl der Topics mehr Suchen durchgeführt werden müssen.

e)

Dass der Nutzer auf den zweiten Link klickt deutet darauf dass das Ergebnis relevant ist weil Klicks zur Relevanz korrelieren. Der Nutzer wird auch natürlicherweise auf den Link klicken der er/sie für am relevantesten hält.

Dabei ist nicht zu vergessen dass Nutzer sehr unberechenbar sein könne und mehr Klicks nicht unbedingt höhere Relevanz bedeuten. Der Text in einem Link kann nicht immer die Relevanz perfekt abbilden für ein Nutzer, was dazu führen kann dass ein Nutzer auf ein Link klickt der in der Realität weniger relevant ist.

f)

In diesem Fall treffen die Begründungen aus Teil (e) noch zu, jedoch kommt hier noch dazu dass ein Nutzer in der Regel ein schnelles Resultat haben will und dadurch nimmt die Wahrscheinlichkeit je weiter ein Link nach unten in einer Liste gestuft wird ab, unabhängig von der Relevanz des Dokuments.

Aufgabe 3

a)

Rangpositionen	A	E	G	B	P	C	J	H	R	D
Präzision	1	0.5	0.33	0.5	0.4	0.5	0.43	0.38	0.33	0.4

$\Rightarrow AP = \frac{1+0.5+0.5+0.4}{4} = 0.6$

b)

Rangpositionen	A	E	G	B	P	C	J	H	R	D
Präzision	1	0.5	0.67	0.5	0.6	0.5	0.43	0.38	0.33	0.3

$\Rightarrow AP = \frac{1+0.67+0.6+0}{4} = 0.57$

c)

$MAP = \frac{0.6+0.57}{2} = 0.59$

$GMAP = \sqrt{0.6 * 0.57} = 0.58$