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“

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,d 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 relevante 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.