IR-Grundlagen - Worksheet 5

Dieses Arbeitsblatt behandelt Probabilistische Retrieval Methoden und das Binary Independent Model.

1. Wie wir in der Vorlesung gesehen haben können wir für das Ranking der Dokumente die Warscheinlichkeit ihrer Relevanz verwenden. Um diese Warscheinlichkeit zu schätzen wird für jeden Term der Collection ein Gewichtswert (c_t) errechnet. Die Formel hierfür lautet:

$$c_t = log\frac{p_t}{1-p_t} - log\frac{u_t}{1-u_t}$$

   mit:

   $p_t = P(x_t = 1|R = 1, q)$
   $u_t = P(x_t = 1|R = 0, q)$

   a. $\frac{p_t}{1-p_t}$, schätzen wir durch die Anzahl der relevanten Dokumente die den Term enthalten geteilt durch die Anzahl aller relevanten Dokumente die den Term nicht enthalten. Implementieren sie in der Klasse `Index` die Methode

      ```
      float getPTQuotient(String term)
      ```

      die für einen gegebenen Term diesen Quotient errechnet.

   b. $\frac{u_t}{1-u_t}$, schätzen wir durch die Anzahl der nicht-relevanten Dokumente die den Term enthalten geteilt durch die Anzahl aller nicht-relevanten Dokumente die den Term nicht enthalten. Implementieren sie in der Klasse `Index` die Methode

      ```
      float getUTQuotient(String term)
      ```

      die für einen gegebenen Term diesen Quotient errechnet.

2. Jetzt sind wir in der Lage zur Zeit der Indexierung, also im Konstruktor der `Index` Klasse für jeden Term einen Gewichtswert zu errechnen und zu speichern. Entscheiden sie in welcher Form diese Werte gespeichert werden sollen und implementieren sie den Konstruktor

   ```
   public Index(ArrayList<Document> collection)
   ```

3. Zuletzt muss eine Suchfunktion implementiert werden. Die Methode

```
ArrayList<Integer> probSearch(String query, int k)
```

errechnet für jedes Dokument einen RSV_{doc} Wert der sich nach folgender Formel zusammensetzt:

$$RSV_{doc} = \sum_{x_t=q_t=1} c_t$$

Der RSV_{doc} Wert ist also die Summe aller Gewichtswerte derjenigen Terme die sowohl im Dokument als auch in der Query vorkommen. Implementieren sie diese Methode.
