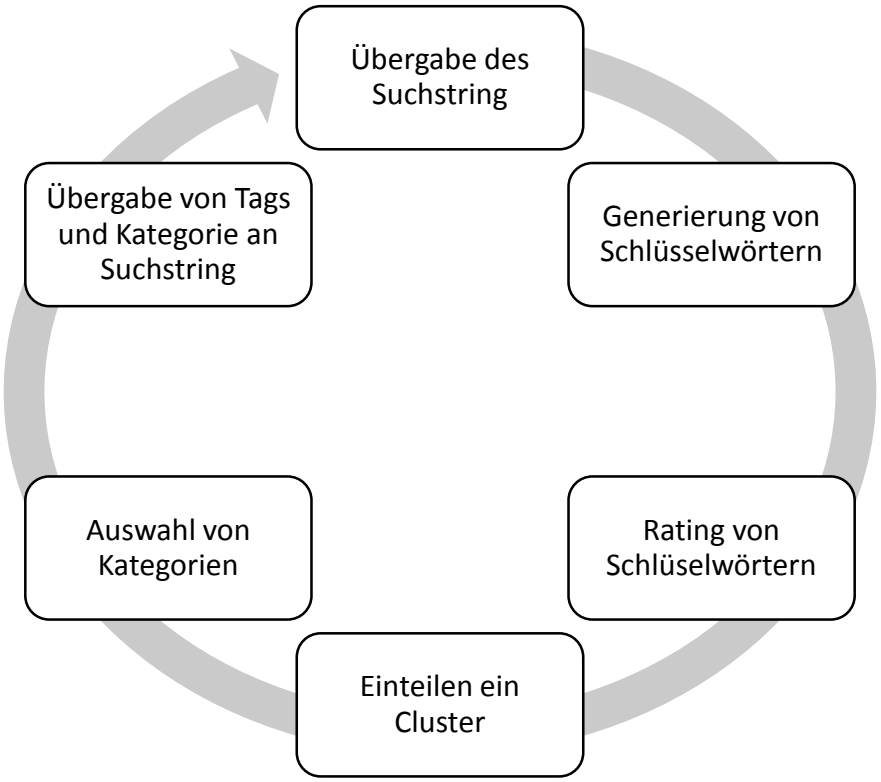
Sortieralgorithmus Kognitive Suche



1. **Übergabe des Suchstrings**
   - Der Suchstring wird an die Websuche übergeben bzw. Eingelesene PDFs werden nach dem Suchstring durchsucht
2. **Generierung von Schlüsselwörtern**
   - Abfangen der Schlüsselwörter der PDF-Box bzw. Generierung von Schlüsselwörtern aus den gefundenen Websites (z.B. mit OpenSource Software wie Hunspell)
3. **Rating von Schlüsselwörtern**
   - Jedes Schlüsselwort bekommt ein Rating
   - Je höher das Rating, desto Relevanter ist dieses Schlüsselwort
   - Zu jedem Schlüsselwort wird mitgespeichert auf welchen Seiten es gefunden wurde (Index-Nummern)
   - Ausschluss von unwichtigen Tags wie Präpositionen, Artikel, Pronomen etc ..
   - Rating von Tags (mit möglicher Gewichtung):
     - Nähe zum Suchwort 30%
     - Länge 10 %
     - Häufigkeit 30%
     - (Auftauchen in META-Tags oder META-Description) → Nur bei Websites (20%)
     - Vorkommen in Überschriften 30%

     ➔ Höheres Rating wenn Synonyme dieses Wortes vorkommen

     ➔ Nutzung von OpenThesaurus

Vermutlich nicht einfach realisierbar:

- Ist das Wort ein Subjekt oder Prädikat im jeweiligen Satz?
- Handelt es sich um ein Verb oder Substantiv?
- Wie viele Suchergebnisse gibt es wenn man nur nach diesem Wort sucht?

4. **Einteilen in Cluster**
    - Sehr komplexes Thema
    - Einfacher Ansatz:
        - Wörter die auf der gleichen Seite gefunden wurden, gehören zusammen
        - Bei ausreichend großer Anzahl durchsuchter Seiten ist ein einteilen in Cluster möglich
5. **Auswahl von Kategorien**
    - Jedes gefundene Cluster ist eine Kategorie
    - Die Überschrift des Clusters ist das Keyword mit dem höchsten Rating aus dem jeweiligen Cluster
    - Nur die z.B. 5 Schlüsselwörter mit dem besten Rating werden angezeigt
6. **Übergabe von Tags und Kategorie an Suchstring**
    - Der Ursprüngliche Suchstring wird um die Kategorie und die zugehörigen Schlüsselwörter ergänzt
    - Im nächsten Durchgang werden die bereits gefundenen Schlüsselwörter aus nicht gewählten Kategorien ignoriert