Data Analytics Challenge – Recommender System

Was haben wir bis jetzt?

Items-Datensatz // Transactions Datensatz

1. Text Mining Similarity auf Basis der Klappentexte
2. Author-MainTopic-Matrix (zum Reduzieren der Items für TM-Similarity)
3. Ludmilas Association Rule Mining: Überprüfung und Ausgabe der Items, welche gemeinsam in Transactions waren
4. Matrizen, welche:
   1. Autoren gemeinsam in einer Session waren
   2. Main Topics gemeinsam in einer Session waren
   3. Items gemeinsam in einer Session waren
5. Sebastians Subtopic Smiliarities
6. Items-Datensatz, nach Sprachen aufgeteilt
7. Titel-Übereinstimmung auf Basis der Buchstabenfolgen (Q-Grams)
8. Ludmilas Filter (Autor, Main Topics, Publisher) gemeinsam

Was für weiter Informationen liegen vor?

* Wir haben 6475 Items (von 78334 Items in Items 6) welche Duplicate sind, bezogen auf die Titel (100%ige Übereinstimmung der Titel)
* Für 13732 Items in den Crawler-Daten liegen keine Klappentexte vor
* Für 33634 Items liegen keine Subtopics vor (ob Subtopic und Main Topic übereinstimmt, ist nicht überprüft)
* 315 von 360 Main Topics (gestutzt auf 3 Stellen) (ca. 87,5%) kommen in Transactions vor
* 10655 von 35727 Autoren (ca. 29,82%) kommen in Transactions vor
* 24909 von 78334 Items (ca. 31,8%) kommen in Transactions vor
* Ludmilas Algorithmus ist für 300 von 1000 Items in Evaluation anwendbar
* 27849 Items haben sowohl Subtopics als auch einen Klappentext (ca. 35,56%)

Welcher Ansatz deckt am meisten ab? Womit können wir die besten Ergebnisse erzielen?

1. Text Mining Similarity auf Basis der Klappentexte (ca. 82.5% der Items haben eine Beschreibung)
2. Sebastians Subtopic Smiliarities (ca. 57% der Items haben Subtopics)
3. 27849 Items haben sowohl Subtopics als auch einen Klappentext (ca. 35,56%)
   1. Wir könnten also erst die Klappentext-Similarity berechnen, dann die similarity über die Subtopics und wenn dort in den Recommendations Überschneidungen vorkommen, dann sind das für uns “save Empfehlungen” (wie auch mit Ludmilas Algorithmus und TM-Similarity auf Blatt)
   2. Können hier noch schauen, welche haben Transactions (das sind die “super save Recommendations”)

Wie gehen wir bisher vor?

Zunächst lassen wir Ludmilas ARM laufen, können damit 300 von 1000 Ziel-Items abdecken.