SIMPSONS SEARCH – ZWISCHENBERICHT

INFORMATION RETRIEVAL SS 2019

DER DATENSATZ

- 4 CSV Dateien
  - Script_Lines
  - Locations
  - Episodes
  - Characters
- Datenbankstruktur

	Zeilen/ Einträge	Spalten	Inhalt
Script_Lines	158.315	13	u.a. Spoken Words inkl. Normalisierung, Character Referenz, Location Referenz, Word Counts
Locations	4.459	3	Liste aller Locations
Episodes	600	12	u.a. Episodenname, Views, Ratings
Characters	6.722	4	Liste aller Charaktere

DER DATENSATZ

- **Script Lines** 158.315 Zeilen (bis Gesamt-Episode 568; Season 26; Episode 16)
- **Locations** 4.459 Zeilen
- **Episodes**: 600 Zeilen
- **Characters**: 6.722 Zeilen

STAND DER SUCHMASCHINE

- .net Core Webanwendung
- Indexierung und Suche ermöglicht durch lucene.net
- Lucene.net: c# clone der Apache.Lucene Search Library

- Indexierung des Datensatzes und der Suche funktionieren bereits

DERZEITIGES VERHALTEN DER SUCHMASCHINE

- Stichwortsuche gesprochener Wörter/Sätze von Charakteren
- Die Suchanfrage wird mit dem string in dem Feld „raw_text“ verglichen
- Rating der Ergebnisse anhand exakter Übereinstimmung mit Suchanfrage

TOPICS UND TOPICKLASSEN - DATENBANKBASIERT

- [character name]
  - Out: Folgen, in denen der Charakter vorkommt
- [character name] [character name]
  - Out: Szenen, in denen beide Charaktere vorkommen / interagieren
- [character name] catchprhase
  - Out: Sätze, die der Charakter am häufigsten benutzt
- [character name] [location]
  - Out: Szenen mit dem Charakter an angegebenem Ort *oder* direkt Rede des Charakters an dem Ort

TOPICS UND TOPICKLASSEN – INTERPRETATIVE SUCHE

- [technology]
  internet, computing
  AI, artificial intelligence

- [politics]
  abortion
  gun control
  Trump
  Bush Jr
  politics (in general)
  Elections
  immigrants

- [violence]
  crime
  military
  Iraq War
  terrorism

- [culture]
  travel
  super heroes
  sports
  musicals, singing
  fantasy (genre)

TOPICS UND TOPICKLASSEN – INTERPRETATIVE SUCHE

- [social issues]
  feminism
  alcohol, alcoholism
  vegetarians, vegans
  patriotism
  racism
  LGBT
  marriage
  mobbing, bullies
  (illegal) drugs

- [Holidays]
  Halloween
  Christmas
  Valentine's day
  St. Patricks Day
  Thanks Giving

AKTUELL/TO DO:

- Topics ausformulieren
- Indexing von Lucene.Net
- Datenbankorientierte vs interpretative Suche

ABSTRAKTE TOPICS

- → naiver Ansatz: Wortlisten
- Bsp. Topic ‚feminsim‘:
  - Women, empowering, equality, gender, pay gap, sexism, social justice,... [erweiterbar]
- Modellierung zwischen Query und Index
- Effizienz?
- Alternativen

ERGEBNISSTRUKTUR

- Verschiedene Ergebnisebenen implementieren:
  - Episode
  - Scene
  - Quote
  - Default?
- Bsp. quote:
  - Well you got that right thanks for your vote girls

WEITERE FRAGEN

- Must-haves vs nice-to-haves (Checkliste?)
- Baseline und Optimierungen
- Known-item search