Modul IE1: Semesterkurzbeitrag

Ausgangslage

Im Rahmen Ihres Semesterkurzbeitrags werden Sie 50 Anfragen auf einer Kollektion von über 20‘000 Dokumenten auswerten. Die Anfragen stellen Informationsbedürfnisse dar, welche manchmal in Stichworten, aber auch als vollständige Sätze formuliert sind. Sie werden mittels eines Experiments Information Retrieval (IR)-Systeme auf ihre Fähigkeit hin untersuchen, diese Informationsbedürfnisse mittels relevanter Information zu "erfüllen". WICHTIG: Ihr Semesterkurzbeitrag soll NICHT eine weitere, "kleine" Projektarbeit sein. Sie sollen vielmehr eine Forschungsfrage formulieren und bearbeiten. Falls Sie diesbezüglich Unsicherheiten haben, wenden Sie sich bitte an den Dozenten.

Aufgaben

Ihre Aufgabe ist es, eine spannende Forschungsfrage zu formulieren, und darauf aufbauend ein IR-Experiment auszugestalten, welches für die Anfragen möglichst gute oder interessante Suchresultate in der Kollektion findet. Sie erstellen mindestens eine Datei, welche für jede Anfrage eine Rangliste von Suchresultaten enthält. Dazu schreiben Sie einen kurzen(!) Bericht, welcher Ihren Lösungsweg aufzeigt.

Rangliste

Sie geben eine Ranglistendatei im trec_eval Format ab, welche je maximal 1000 Suchresultate für alle 50 Anfragen enthält (d.h., maximal 50000 Zeilen). Falls Sie mehrere Experimente durchführen, wählen Sie die in Ihren Augen beste oder interessanteste Rangliste aus (Sie können maximal 3 Ranglisten einsenden).

Das trec_eval Dateiformat enthält eine Liste von Zeilen mit folgendem Aufbau:

Spalte	**Query ID**	**Iteration**	**Dok.Nummer**	**Rang**	**Score**	**System**
Inhalt	ID der Anfrage	Konstant auf Q0 (Qnull)	ID eines gefundenen Dokuments	Rang des Dokuments	RSV	Name des Systems, verwenden Sie Ihren Gruppen-namen
Datentyp	Integer	String	String	Integer	Float	String

Die Felder werden jeweils durch ein einzelnes Leerzeichen getrennt. Beispielzeilen:

```
30 Q0 ZF08-175-870 0 4.238 irg1
30 Q0 KS09-937-830 1 4.097 irg1
30 Q0 OW01-739-219 2 3.951 irg1
```

Nach der Abgabe aller Beiträge werden die Ranglisten verglichen und **öffentlich** publiziert.

Bericht

Verfassen Sie einen Bericht, welcher lesbar auf maximal zwei A4 Seiten Platz hat. Darin erklären Sie Ihr Experiment, welches zur abgegebenen Rangliste geführt hat. Führen Sie aus, welche Systeme (Beispiele: Eigenbau, Laborumgebung, Google Desktop Search, Microsoft Search Server, etc.) Sie eingesetzt haben und was die Motivation war. Beschreiben Sie den Aufbau des Index, welchen Sie erzeugt haben. Zeigen Sie, wie die Anfragen an die Systeme gestellt wurden (manuell, automatisch, iterativ, etc.) und wie Sie die Resultate interpretieren. Wie die Ranglisten werden auch die Berichte **öffentlich** aufgeschaltet.

Die Qualität des Experiments resp. die fachliche Individualität der Arbeit ist die Grundlage für Ihre Bewertung. Dabei sind Sie nicht darauf beschränkt, nur möglichst „gute“ Suchresultate zu erhalten. Auch "freakige" Ansätze wie zum Beispiel Optimierung auf Antwortgeschwindigkeit sind willkommen, sofern die Resultate interessant sind und brauchbar bleiben. Ihre Suchresultate validieren die Aussagen Ihres Berichts.

Rahmenbedingungen

- Arbeit in **Zweiergruppen**: Organisieren Sie sich frühzeitig, damit Sie bald loslegen können.
- Ausgabe am Donnerstag, 17.9.2020
- Ein **kurzes "Konzept"** (ca. 10 Zeilen) per **Email an** bram@zhaw.ch **bis 11. Oktober 2020**
- In den **Praktika vom 22.10. oder 29.10.** findet eine kurze **Fragestunde** zum Verlauf der Semesterkurzbeiträge statt (max. 1 Lektion).
- In den Praktika wird immer wieder auf "Überlappungen" mit der Aufgabenstellung der Semesterkurzbeiträge hingewiesen. Sie sind jederzeit willkommen, mit dem Dozenten neue Anknüpfungspunkte, Bezüge zum Theoriestoff etc. zu diskutieren.
- **Abgabe** bis am **Mittwoch, 2. Dezember 2020 um Mitternacht** (eine pünktliche Abgabe ist auch Teil der Aufgabe!). Ein OLAT-Ordner wird für den Upload zur Verfügung gestellt.
    - Überprüfen Sie das **Format der Rangliste** sorgfältig
    - Grösse des **Berichts maximal 2 A4 Seiten**
- Die **Bewertung** erfolgt nach diesem Massstab (Punkte werden an Semesterendprüfung angerechnet):
    - 0 Punkte für keine Abgabe
    - 5 Punkte für ein minimalistisches Experiment, z.B. ein vorgefertigtes System zum Laufen gebracht – aber keine interessantes, "originelle" Hypothese/Idee
    - 10 Punkte für eine vernünftige Auseinandersetzung mit dem Thema. Es ist eine klar erkenntliche Forschungsfrage formuliert und bearbeitet.
    - 15 Punkte für ein klar überdurchschnittliches Experiment – interessante Idee
    - 20 Punkte für ein sehr aufwändiges und interessantes Experiment, das einen aussergewöhnlichen Grad an Originalität beinhaltet

  Achtung:

- **Die Bewertung erfolgt nicht auf Grund des Zeitaufwands der Studierenden, oder der Quantität des Outputs. Es sei nochmals betont: Sie bearbeiten KEINE kleine Projektarbeit. Das Ziel ist das Formulieren und Bearbeiten einer interessanten**

Forschungsfrage. Dieser Aspekt, und NICHT die Anzahl der Codezeilen o.ä. wird bewertet.

Lernziele

Sie sollen sich neben den regulären Praktikumsaufgaben auf eigene und kreative Weise mit dem Thema Information Retrieval auseinandersetzen. Mit Ihrem Wissensaufbau im Modul lernen Sie jede Woche wieder neue Ansätze kennen, welche Sie sukzessive in den Semesterkurzbeitrag einfliessen lassen können. Durch die systematische Betrachtung verschiedener Systeme und die konkrete Erstellung von Suchresultaten sollen Sie unter anderem die Fähigkeit erlangen, in einer Firma Information Retrieval Systeme für den Gebrauch zu implementieren resp. evaluieren zu können – eine zentrale Aufgabe in diesem Gebiet (z.B. Erschliessen eines vorhandenen Dokumentenarchivs)

Ressourcen

Die Kollektion und Anfragen beziehen Sie im OLAT System.