



The true alchemists do not change lead into gold;
they change the world into words.
(William H. Gass)

Praktikum 10

Textanalyse

[20 Punkte]

In diesem Praktikum „bauen“ Sie ein System zur automatischen Text-Analyse, mit dem Sie Produkt-Reviews analysieren können. Dabei können Sie die verschiedenen Algorithmen und Datenstrukturen einsetzen, die Sie bisher bereits in der Vorlesung kennengelernt haben, und auch neue anwenden.

Aufgabe 1: Grundfunktionen

Sie implementieren die Grundfunktionen eines Text-Analyse-Systems.

Teilaufgaben:

- a) Lesen Sie die gegebenen Produkt-Reviews ein.
- b) Geben Sie grundlegende statistische Informationen über die Texte aus, z.B.
 - Anzahl Dokumente
 - Gesamtgrösse der Dokumente (z. B. in KB oder MB)
 - Anzahl Wörter
 - Anzahl unterschiedliche Wörter
 - Durchschnittliche Textlänge (Zeichen pro Dokument)
- c) Erstellen Sie einen Index, in dem alle Wörter aus allen Texten mit ihren Häufigkeiten aufgelistet werden
- d) Implementieren Sie eine rudimentäre Suchmaschine, mit der Sie nach einzelnen Wörtern suchen können.

Abgabe: keine

Aufgabe 2: Feature-Ausbau

Erweitern Sie Ihr Text-Analyse-System um neue, sinnvolle, spannende Funktionalität. Was Sie genau implementieren ist Ihnen überlassen. Hier einige Vorschläge, was man alles machen könnte:

- exakte Suche nach mehreren Wörtern, die mit UND oder ODER verknüpft sind



- exakte Suche nach zwei Wörter, wobei das erste Wort vorkommen muss und das zweite Wort NICHT vorkommen darf
- Filtern nach Dokumenten mit einer bestimmten Länge
- Filtern nach Dokumenten mit positivem oder negativem Aussagen (-> Sentiment)
- Approximative Suche, die auch gewisse Schreibfehler berücksichtigt (-> Levenshtein-Distanz)
- Sortierung der Suchergebnisse nach „Relevanz“
- Suche mit regulären Ausdrücken
- Suche nach unvollständigen Wörtern (z. B. „Student“ liefert auch „Studenten“, „Studentin“, „Studentinnen“ usw.
- Index, der alle Wörter auflistet und in dem man per Klick auf die entsprechenden Dokumente kommt
- Ein Glossary, das nur die „wichtigsten“ Wörter enthält (-> Topic Detection, Stopword Liste)
- Visuelle Darstellung, welche Themen in einer Dokumentmenge „wichtig“ sind (-> Topic Detection, Tag-Cloud)

Für die Implementierung von einzelnen Features können Sie bei Bedarf auf existierende Software-Pakete verwenden.

Dokumentieren Sie, was Ihr System kann – am besten mit geeigneten Screenshots.

Abgabe: Quelltext + *aussagekräftige* Doku in OLAT bis 28.5.2017 (Semesterende).

Gruppenarbeit: Sie können in Teams bis zu 3 Personen zusammenarbeiten.

Demo: In der letzten Vorlesung findet eine Demo statt, in der Sie Ihr System vorstellen.

Bewertung: Bewertet wird die Funktionalität Ihres Systems sowie die Qualität von Implementierung und Dokumentation.