UNIVERSITÄT LEIPZIG

Institut für Informatik
Lehrstuhl für Rechnernetze und Verteilte Systeme

Praktikum Computational Advertising
Prof. Dr.-Ing. Christoph Lindemann
Sommersemester 2011

Betreuer: **Dipl.-Inform. Jan Friedrich, friedrich@rvs.informatik.uni-leipzig.de**
Dipl.-Inform. Michael Petrifke, petrifke@rvs.informatik.uni-leipzig.de

Aufgabe 4: Latent Dirichlet Allocation (LDA)

Übersicht

Eine weitere Möglichkeit Query Rewriting zu betreiben ist die Verwendung von klassifizierten Topics. Hierbei werden verschiedenste Eigenschaften (z.B. Kookkurrenzen) der Terme verwendet, um diese in Cluster einzuteilen. Es wird angenommen dass Terme im gleichen Cluster im Query Rewriting Prozess äquivalent zueinander sind (i.e. das gleiche Thema behandeln) und somit untereinander austauschbar sind.

Eine Möglichkeit der Berechnung dieser Cluster bietet LDA [1]. Jedem Term wird die bedingte Wahrscheinlichkeit zugeordnet zu den verschiedenen Clustern zu gehören.

Aufgabe

Ersetzen Sie ihre Implementierung des Query Rewriting aus Aufgabe 3 durch LDA. Dafür clustern Sie Terme in 40 Topics und generieren Rewrites, indem Sie alle Wörter der Query durch die jeweils 10 besten Terme des gleichen Clusters, dem auch der zu ersetzende Term angehört, ersetzen. Die „besten Terme“ werden hier als die Terme mit der höchsten Wahrscheinlichkeit zum jeweiligen Cluster zu gehören definiert und die „Angehörigkeit zu einem Cluster“ als das Cluster für das der Term die höchste Wahrscheinlichkeit besitzt verwendet zu werden.

Auf der Praktikumswebseite finden Sie ein Archiv mit Texten das Sie als Korpus betrachten sollen. Verwenden Sie *lda-c* [2], die Referenzimplementierung von David M. Blei, um die entsprechenden Wahrscheinlichkeiten zu bestimmen.

Als Eingabe für lda-c dient ein Textfile, in dem jede Zeile ein Dokument repräsentiert und folgendes Format besitzt:

```
[M] [term_1]:[count] [term_2]:[count] ...  [term_N]:[count]
```

Dabei steht `M` für die Anzahl der unterschiedlichen Terme, die im Dokument vorkommen, `term_n` ist die ID eines Termes (jedem Term wird eine eindeutige ID zugewiesen beginnend bei 0 und durchnummeriert) und `count` die Häufigkeit des Termes im Dokument.

Alle restlichen Parameter werden in der `readme.txt` erläutert. Die Ausgabe befindet sich schlussendlich in der Datei `final.beta`. Jede Zeile steht für ein Cluster c und enthält den Logarithmus der Wahrscheinlichkeit, dass der Term t (in der Zeile nach seiner ID sortiert) bei einem Generierungsprozess für die Topic k, die durch Cluster c repräsentiert wird, Verwendung findet.

Sie können nun verschieden an die Aufgabe heran gehen:

- Sie verwenden den Quelltext von lda-c und integrieren diesen in Ihr Programm
- Sie rufen lda-c aus ihrem Programm heraus auf (z.B. über `system()`)

Schlussendlich müssen Sie ihre bisherige Konsolenanwendung um einen Befehl erweitern, der einen Ordner als Parameter übergeben bekommt, in dem die Dokumente des Korpus liegen.

```
performLDA <folder>

    LDA-Berechnung
```

Es werden mithilfe von lda-c die Cluster berechnet (dabei müssen die Dokumente natürlich in das entsprechende Eingabeformat konvertiert werden) und in einer Datenbank gespeichert (die Datenbank muss die Cluster enthalten und die Zuordnung der Terme zu den Clustern mit der jeweiligen Wahrscheinlichkeit). Diese Datenbank wird dann online für das oben beschriebene Rewriting verwendet.

Überlegen Sie sich selbst einen sinnvollen Wert für die Alpha Variable von LDA und informieren Sie sich über deren Bedeutung.

Abgabe

Bitte schicken Sie pro Gruppe genau eine Lösung bis zum **03.07.11, 23:59 Uhr** per Email an Jan Friedrich (*friedrich@rvs.informatik.uni-leipzig.de*). Die Abgabe sollte den gesamten Quellcode enthalten sowie ihre Build-Skripte (z.B.. autotools, cmake, qmake, Custom Makefile, vcproj, ...). Geben Sie auch an in welcher Umgebung die Aufgabe angefertigt wurde (Betriebssystem, Compiler, inkl. Version). Wir empfehlen die Aufgabe mit GCC unter einem Unix-kompatiblem Betriebssystem oder der CygWin-Umgebung [3] zu bearbeiten. Sollten Sie dennoch unter Windows arbeiten wollen, können wir bei Problemen eventuell nur begrenzt Hilfe anbieten. Zudem müssen Sie zum Reviewgespräch einen Laptop mitbringen, auf dem der Code sowohl kompiliert als auch ausgeführt werden kann.

Die Termine für die Reviewgespräche sind am **04.07.**, **05.07** und **06.07** jeweils von **15 bis 17 Uhr**. Kommen Sie bitte innerhalb dieser Zeiten, wenn möglich mit eigenem Laptop, unangemeldet mit beiden Gruppenmitgliedern zu Jan Friedrich (Raum 04-06). Die Reviews werden ca. 15 Minuten dauern.

Referenzen

[1] Latent Dirichlet Allocation, D.M. Blei, A.Y. Ng, M.I Jordan, JMLR, 2003

[2] http://www.cs.princeton.edu/~blei/lda-c/

[3] http://www.cygwin.com/

Informationen zur Abgabe

- **Siehe Webseite: http://rvs.informatik.uni-leipzig.de/de/lehre/SS11/praktika/ca/**

Viel Erfolg!