Fachprojekt
Data Mining – Datenanalyse und Sprache

Prof. Dr. Katharina Morik
Informatik LS8

Eine Fülle von Daten liegt in Form natürlicher Sprache vor und eine Vielzahl von Methoden und Werkzeugen gibt es, mit diesen unstrukturierten Daten umzugehen. Mit einem einfachen Schlagwort sind die Ansätze nicht per Suchmaschine oder bei Wikipedia zu finden. Damit ergibt sich eine Situation, die Ihnen später im Beruf oft begegnen wird:

- Ein Thema ist plötzlich in aller Munde, aber die Entwicklung dieses Gebiets haben Sie nicht verfolgt.
- Wie kann man sich einarbeiten? Man muss Fachliteratur lesen!
- Am Anfang versteht man nur Bruchstücke der Artikel – am Ende war es dann doch nicht so schwer.

Diese Situation wird mit dem Fachprojekt geübt. Das Fachprojekt soll es den Teilnehmern ermöglichen, sich innerhalb eines Themenkomplexes (in diesem Fall Datenanalyse und natürliche Sprache) sowohl auf theoretischer als auch praktischer Ebene zu orientieren.

Dazu gehören die Tätigkeiten wissenschaftlichen Arbeitens:

- Literatur-Recherche (ausgehend von einem Seminarthema/Paper)
- Tool-Recherche (Implementierung, Installation von Tools und Anpassung)
- kritische Evaluierung (anhand von Qualitätsmetriken, Test-Daten)
- Dokumentation (in Form einer wissenschaftlichen Ausarbeitung)

Natürlich ist der Zeitrahmen (siehe Zeitplan) für das Fachprojekt recht eingeschränkt, was bei der konkreten Umsetzung dieser Punkte berücksichtigt wird.

Das Fachprojekt soll einige *Softskills* vermitteln, die allesamt bei der Bearbeitung der Kernpunkte mit eingebracht werden:

- Zusammenarbeit im Team
- Wissenschaftliches Schreiben
- Präsentieren
- Umgang mit Tools
- Vertiefung von Programmiererfahrungen (hier in Java)

In Gruppen werden Themen bearbeitet, wobei die wissenschaftlichen Tätigkeiten gemeinsam mit den Softskills eingeübt werden.

Themen

In dem Fachprojekt werden vier Themen bearbeitet, jedes in einer Gruppe. Zu einem Thema gehören Grundlagentexte, Tools und weiterführende Texte. Ein Anwendungsbeispiel wird gemeinsam gesucht. Es kann für alle Gruppen die selbe Anwendung sein, für jede Gruppe eine eigene oder mehrere aber nicht alle Gruppen arbeiten an der selben Anwendung..

Textklassifikation Eine Menge von Dokumenten in vorgegebene Klassen einzusortieren, ist die Grundlage für viele Anwendungen. Dabei wird ein Dokument als Menge von Wörtern aufgefasst.

- Die Grundlage der Textklassifikation wird in den ersten Kapiteln des Buches [2] gegeben.
- Ein Tool zur Textbearbeitung ist das *Text Processing Plug-In* von RapidMiner.
- Sobald die Texte als Vektoren (bag of words) vorliegen, können sie klassifiziert werden. Zur Klassifikation gibt es verschiedene Verfahren. Eine Einführung findet sich in [8]. Die Verfahren sind verfügbar in RapidMiner unter Modeling/Classification and Regression.

Informationsextraktion Aus einem Dokument relevante Angaben zu extrahieren ist die Aufgabe der Informationsextraktion. Dabei werden zunächst Wortvorkommen im Text semantische Kategorien zugeordnet, z.B. *Person, Ort, Firma* (Named Entity Recognition) und dann Relationen zwischen solchen Kategorien im Text gefunden. Hier werden also nicht Dokumente als Gesamtes, sondern Sätze und Wortvorkommen in den Dokumenten betrachtet.

- Die Grundlage der Informationsextraktion ist zu finden in [5], vor allem Abschnitte 3.1, 4.1 und 4.2

  http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-16089/diss_final2007_DS.pdf

  Ein konkreter Ansatz ist beschrieben in [6].
- Ein Tool zur Informationsextraktion ist das *Information Extraction Plug-In* von RapidMiner [3].
- Als Lernverfahren werden vor allem Conditional Random Fields eingesetzt. Ein Artikel dazu ist [4]. Das Verfahren ist auch in dem *Information Extraction Plug-In* verfügbar.

Web Mining Das Web ist die größte Dokumentensammlung, die es je gab. Hier wollen wir nicht die Verweisstrukturen untersuchen oder gar soziale Netze analysieren, sondern bleiben bei den Texten. Die ungeheure Fülle an Sätzen wird z.B. von Google dazu benutzt, automatisch zu übersetzen. Man kann per Anfrage an das WWW überprüfen, ob ein Satz wohlgeformt ist, welche Wörter häufig zusammen vorkommen, welche Wörter

gemeinsam einen bestimmten Inhalt anzeigen (z.B. positives oder negative Beurteilung eines Produktes)...

- Ein neues, hocheffizientes Verfahren stellt nicht nur fest, welche Wörter häufig gemeinsam vorkommen, sondern will inhaltliche Bezüge zwischen Wörtern herausfinden [1].
- Ein Tool zum Crawling und zur Bereinigung von HTML-Seiten ist das WebMining Plugin von RapidMiner.
- Die semantischen Bezüge können genutzt werden, um eine Menge an Web-Seiten zu strukturieren.

Clustering von Ergebnislisten Anfragen an das WWW oder auch an eingeschränkte Sammlungen ergeben oft neben den gewünschten auch Treffer, die aus ganz anderen Gebieten stammen. Das kann z.B. an der Doppeldeutigkeit der Anfrage liegen (*Jaguar* – Auto oder Tier?). Wie können Ergebnislisten so strukturiert werden, dass man leicht die gewünschte Teilmenge findet?

- Ein neuer Ansatz stellt Artikel in einer *U-Bahn Karte* zusammen, um sie zu strukturieren [7].
- Bereits sorgfältig annotierte Wörter und Sätze gibt es im Wörterbuch der deutschen Sprache (DWDS):

  http://www.dwds.de

  Man kann dort Anfragen nach den Umgebungen von Wörtern stellen. Kann man etwas anderes rümpfen als die Nase? Glücklicherweise ist die TU Dortmund (Angelika Storrer) an diesem bundesweiten Projekt beteiligt, so dass wir Zugang zu der Sammlung annotierter Sätze haben. Arbeitsumgebung zur automatischen Annotation von eigenen Texten (WebLicht): https://weblicht.sfs.uni-tuebingen.de/ Hierzu können Sie sich über Shibboleth mit Ihrem TU-Dortmund-Account anmelden.
- Auch bei DWDS leiden die Ergebnisse unter Mehrdeutigkeiten. Können sie mit dem U-Bahn-Ansatz besser dargestellt werden?

RapidMiner kann kostenlos heruntergeladen werden. Zur Einführung in RapidMiner gibt es Webinars
http://rapid-i.com/content/view/189/198/lang,de/

Auch LaTeX ist frei verfügbar. Die offizielle Seite ist:
http://www.latex-project.org/
Es gibt eine Fülle von Umgebungen, z.B. TeXShop (für Mac) und TeX Live http://tug.org/texlive/ (http://www.dante.de/tex/tl-install-windows.html) für Windows-Systeme.

Literatur

Zeitplan

Das Fachprojekt gliedert sich wie folgt:

Vorstellung der Themen und Anwendungsgebiete 10.9.2012

Seminar 16.10. – 6.11.2012 Hier werden die Grundlagentexte und Tools präsentiert. Dazu gibt es ein Referat und eine Vorführung des jeweiligen Tools, die die anderen Teilnehmer direkt am Rechner mitverfolgen können.

Planung 13.– 20.11.2012 Die Anwendungsfälle und Fragestellungen zur Evaluierung werden ausgewählt und ein Vorgehen zu ihrer Realisierung geplant.

Implementierung und Evaluation 27.11. 2012– 8.1.2013 Gemäß der Planung wird unter Verwendung der Tools ein Anwendungsfall realisiert und damit eine festgelegte Menge an Fragen beanwortet.

Dokumentation 15.1. – 29.1.2013 Die Ergebnisse der eigenen Arbeit werden mit Bezug auf den internationalen Stand der Kunst dokumentiert. Ein gemeinsames kurzes Video könnte die Arbeit des Fachprojektes vergnüglich vermitteln.

Literatur

[1] Huy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. Large-scale learning of word relatedness with constraints. In *Proc. KDD 2012*, 2012.

[2] Thorsten Joachims. *Learning to Classify Text using Support Vector Machines*, volume 668 of *Kluwer International Series in Engineering and Computer Science*. Kluwer, 2002.

[3] Felix Jungermann. *Documentation of the Information Extraction Plugin for RapidMiner*, August 2011.

[4] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[5] Marc Roessler. *Korpus-adaptive Eigennamenerkennung*. PhD thesis, Universitaet Duisburg Essen, 2006.

[6] Marc Roessler and Katharina Morik. Using unlabeled texts for named-entity recognition. In Tobias Scheffer and Stefan Rüping, editors, *ICML Workshop on Multiple View Learning*, 2005.

[7] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. Metro maps of science. In *Proceedings KDD 2012*, 2012.

[8] Stefan Wrobel, Katharina Morik, and Thorsten Joachims. Maschinelles Lernen und Data Mining. In G. Görz, C.-R. Rollinger, and J Schneeberger, editors, *Einführung in die Künstliche Intelligenz*, pages 517–597. Oldenburg, 2000.