Web Mining Übung

- www.ke.tu-darmstadt.de/lehre/ss14/web-mining/uebungen
  - zusätzliche Informationen, Registrierung, Upload, Übungsblätter
- Aufgaben
  - aus dem Bereich Data-, Text- und Web-Mining
  - Crawling, Textanalyse, Textklassifizierung, Clustering, Information Extraction, etc.
  - großer Spielraum bei Lösungsfindung
  - Programmierung ist notwendig
    - aber die Programme sind nur Mittel zum Zweck
- Umfang
  - Aufgaben teilweise zeitintensiv, je nach Programmiersprache, Vorkenntnisse, Programmierfähigkeit etc.
  - dafür praktischer Einsatz der Techniken aus der Vorlesung

Web Mining Übung

- Zeitplan
  - 5 Aufgaben
  - ca. alle 2 Wochen
    - Abgabe sonntags
    - Übungsstunde dienstags
    - neue Übungsblätter in der Regel Anfang der Woche
  - erste Übung Abgabe 11. Mai, Besprechung 13. Mai
- Beurteilung:
  - 10+2 Bonuspunkte je Übung, max. 50 Punkte
  - Verbesserungen bis zu einem Notengrad sind möglich
  - Punkte/5 ≈ Klausurpunkte
    - plus „Wissensbonus“ natürlich
  - nur bei bestandener Klausur!

Web Mining Übung

- Gruppenarbeit möglich
  - Gruppengröße max. 3
- Registrierung
  - über Upload-Seite
  - TU-ID-Login notwendig
  - Anmeldung in den nächsten Tagen möglich
  - eine Person muss eine Gruppe erzeugen
    - alle anderen treten dann bei
    - (Erzeuger teilt den anderen die Nummer bei)
  - am 9. Mai werden die Gruppen fixiert
  - Gruppenwechsel nachträglich nur über Veranstalter möglich

Zentraler Au

KE 201X

1

Willkommen bei Webmining

Du bist noch nicht registriert. Drücke bitte hier , um dich zu registrieren.

Bitte geben Sie Ihre TU-ID und Ihr Passwort an

TU-ID:

2

Passwort:

Ich möchte gewarnt werden, bevor ich mich in einen anderen Bereich einlogge.

ANMELDEN

Willkommen bei Webmining

3

Du bist noch nicht registriert. Drücke bitte hier , um dich zu registrieren.

KE 201X

Gruppe Upload Ergebnisse

4

Du bist noch keiner Gruppe zugeordnet.

Gruppe	Teilnehmer	
4	0	

5

Erstelle eine neue Gruppe.

Ich möchte nicht an der Übung teilnehmen und mich ABMELDEN .

Web Mining Übung

- Ablauf Übungsstunde
  - Durchbesprechen der abgegebenen Lösungen
  - Gruppen stellen abwechselnd ihre Lösungen vor
  - mindestens ein Teammitglied muss anwesend sein
    - und vorführen können!
- Abgabe
  - über Upload-Seite
  - ein Teammitglied lädt für alle hoch
    - letzter Upload wird gewertet
  - Upload einer Zip-Datei

Web Mining Übung

- Abgabe enthält
  - **Lösungsdokument**
    - **PDF**-Datei
    - **präsentierfähig** (siehe nächste Folie)
  - Ergebnisdateien
    - Tabellen, Grafiken etc.
  - Quelldateien und Programm
    - kompilierbar und ausführbar
  - Beispieldateien
    - benutzter Korpus, Webseiten, Wörterbücher, etc.

Web Mining Übung

- Form des PDFs
  - Lösung **ohne mündliche Erklärung** nachvollziehbar!
    - Benotung soll allein anhand PDF-Datei möglich sein

      → für Erklärungen wichtige Tabellen, Diagramme etc. müssen direkt im PDF sein
    - Generell: Lösungen müssen so ausführlich sein, daß man sie ohne Nachfragen nur durch Lesen des PDFs versteht
    - aber trotzdem in Form von Folien (keine *Berichte*)
  - zusätzliches Material nur als Nachweis oder für Detailfragen
  - Referenzen auf benutzte Beispieldateien angeben

Web Mining Übung

- Tabellen, Diagramme, Graphen:
  - vollständige Beschriftung der Achsen
  - Beschreibung

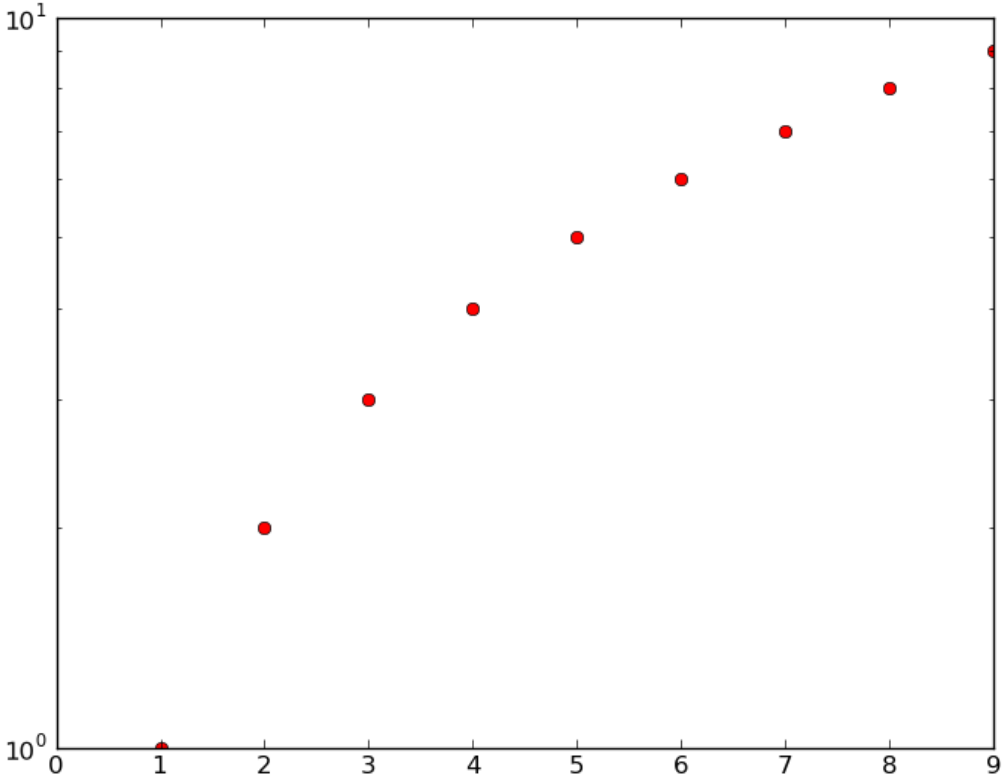

Abb. 2: Gewicht in Abhängigkeit der Flaschen in sinnlosem Datensatz A

Web Mining Übung

- Codelistings, Pseudocodes:
  - Aufzeigen der Funktion / Funktionsweise
  - nicht einfach vollständigen Source Code hineinkopieren!
  - unbedingt Erklärung!

```
perl -n -l -a -e 'BEGIN{$x=0} {$x+=$F[$#F]} END{print $x}' preisliste.txt
```

Abb. 1: Aufsummierung der Preise (letztes Wort einer Zeile) einer Preisliste

Web Mining Übung

- Programmierung
  - beliebige Sprache
  - besonders geeignet: Skriptsprachen
    - Python, Perl, Ruby, Groovy
    - aber auch Java, **Javascript**, etc.
  - Benutzung von Libraries erlaubt
    - doch Aufgabenstellung ist zu beachten
    - „implementieren“ heißt nicht „verwenden“
    - einige auf Homepage aufgezeigt
- Weitere Tools
  - matplotlib, jfree, gnuplot für Plots
  - r-project für statistische Berechnungen und Grafiken
  - graphviz fürs Zeichnen von Graphen
  - etc. etc. etc.

Web Mining Übung

- Weiterer Ablauf
  - nach Abgabe und Besprechungsstunde: Bewertung der Übungen normalerweise im Laufe der Woche
  - Bewertungen enthalten teilweise kurzes Feedback
  - über das Upload-System einsehbar
- Betreuung
  - Eneldo eneldo@ke.tu-darmstadt.de
    - inhaltliche Fragen, Übungsbetrieb, Übung
    - Sprechstunde: Mittwoch 16:30-17:30
  - Forum

Web Mining Übung

- Beispiel 1. Übung
  - Überlegen Sie sich eine neuartige, originelle Web Mining Anwendung, die mit Text-Klassifikationsverfahren gelöst werden könnte. Skizzieren Sie eine mögliche Umsetzung (Sammlung der Trainingsdaten, Klassifikation der Trainingsdaten, Einsatz des gelernten Klassifikators in der Praxis).
  - Schreiben Sie ein einfaches Programm, das eine sortierte Liste der in einem Text vorkommenden Worte (im weitesten Sinn alles was durch Leerzeichen begrenzt wird) mit den assoziierten Häufigkeiten (absolut und prozentual) erstellt und sortiert ausgibt.
    - Vergleichen Sie die 30 am häufigsten vorkommenden Worte in zwei oder mehreren längeren Texten der gleichen Sprache (z. B. E-books, Projekt Gutenberg, etc.). Sind diese Worte als Merkmale für Text-Klassifizierungs-Aufgaben geeignet? Warum?