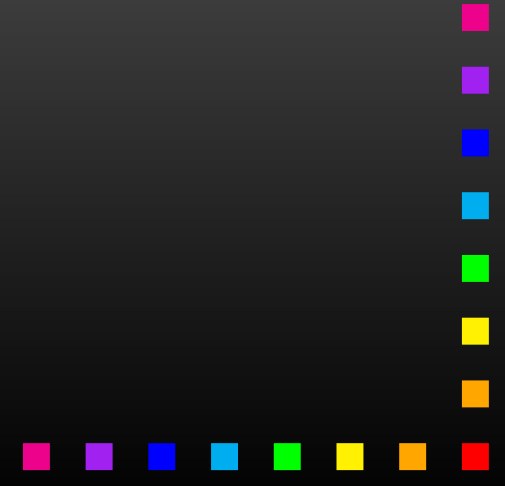
# Messungen zur Performanz der Dublettenerkennung mittels Trigrammen

Dr. Harald Jele

harald.jele@uni-klu.ac.at

Universität Klagenfurt



# Kontext des Projekts

- Phase 2: Integration der Datensätze des ISIS-Bibliothekssystems der IFF-Fakultät der Uni Klagenfurt (Schottenfeldgasse, 1090 Wien) in das Verbundsystem. Ca. 10.000 Titeldatensätze
- Phase 1: Titeldublette = gleiche ISBN + gleiches Jahr + gleiche Auflage (abgeschlossen 2005, ca. 6.300 Titeldatensätze + Items + Holdings)

# Die einzelnen Schritte (1)

- Datenentladen aus den zwei Quellsystemen (Aleph und ISIS)
- Entscheidungslogik über die heranzuziehenden Kategorien und Laden der Daten in die Datenbank
- Generieren der Stoppwort-Tabelle

# Die einzelnen Schritte (2)

- Normalisierung der Daten (UTF8 $\rightarrow$ ASCII-7-bit, Rückführungsregeln auf „Grundformen“ mglst. ohne in die Semantik der Daten einzugreifen, Eliminierung von recherchierten Angaben etc.) $\hookleftarrow$ unterschiedliche Katalogisierungsregeln oder deren Auslegung
- Auffinden der zurückgebliebenen UTF8-Zeichen
- Datensatzvergleich

# Verwendete Hard- und Software

- AMD Athlon64 3800+ Sockel AM2, 2GB Hauptspeicher
- SATA-Festplatte 80GB, 7500rpm, 8MB Cache (Western Digital)
- Betriebssystem Linux Ubuntu Dapper (32bit)
- Perl v5.8.7, MySQL v5.0.22

# Datenentladen

- Möglichkeit, die Daten aus einem Full-Table-Export zu generieren (schnell, aber möglicherweise nicht aktuell)
- Online-Export (wesentlich langsamer, aber aktuell. Zeitdauer für 400.000 Datensätze ca. 48 Stunden)

# Entscheidungslogik und Datenladen (1)

```
kat_100  kat_100b kat_100c kat_100f kat_359
kat_200  kat_200b kat_200c
kat_331  + kat_335 in einem String
kat_403
kat_410  kat_410a
kat_412  kat_412a
kat_425a kat_425b kat_425c kat_425
kat_433  kat_433a kat_433b
kat_540a kat_540b kat_540
Zusätzlich für Reihenwerke
kat_453m
kat_453r
kat_455
```

# Entscheidungslogik und Datenladen (2)

- ingesamt waren 29 Spalten in 437.533 Sätzen zu berücksichtigen (72MB flache TXT-Datei) → Ladedauer: 1m44s
- Entfernen von 40.688 gelöschten Sätzen (DEL=Y) → 1m02s
- Entfernen der lokalen 4.299 Sätze (ID=LL anstelle ID=AC) → 5s

# Entscheidungslogik und Datenladen (3)

- Daraus Erstellen der Daten zum weiteren Vergleich (`10 Spalten`, `392.616 Sätze`) → `5m30s` („`create_import_data.pl`“)
- Laden der erstellten Daten `32s` (`53MB`)

# Generieren der Stoppwort-Tabelle (1)

- Durchsuchen der Datensätze nach markierten („» «“) Begriffen.
  Dauer `4m27s` → ergibt `684` Sätze („`collect_stopwords.pl`“)

# Generieren der Stoppwort-Tabelle (2)

- Jene Begriffe, die zu mehr als `2.000` Treffern führen, wurden im Laufe der Dublettenprüfung markiert. Jene, die keine markierten Stoppwörter darstellten, sich jedoch dafür qualifizierten, wurden zusätzlich in die Stoppwort-Tabelle mit aufgenommen (vgl. dazu z.B. die Stoppwort-Tabelle von allegroC) $\rightarrow$ Laden von 15 zus. Sätzen (1s) (insg. 699 Stoppwörter)

# Normalisieren der Daten

- Erfolgt nach den Vorgaben von Beate Rusch 1999 (KOBV). `392.616` Sätze, `46m47s` („`normalize_me.pl`“)

# Auffinden von zurückgeblie-benen UTF8-Zeichen (1)

- Scheint zunehmend wichtig zu sein, da anscheinend viele KatalogisiererInnen Sonderzeichen nicht über die Tastatur eingeben oder aus der Aleph-Zeichentabelle übernehmen, sondern aus anderen Quellen (MS Word??) wählen ↯

# Auffinden von zurückgeblie-benen UTF8-Zeichen (2)

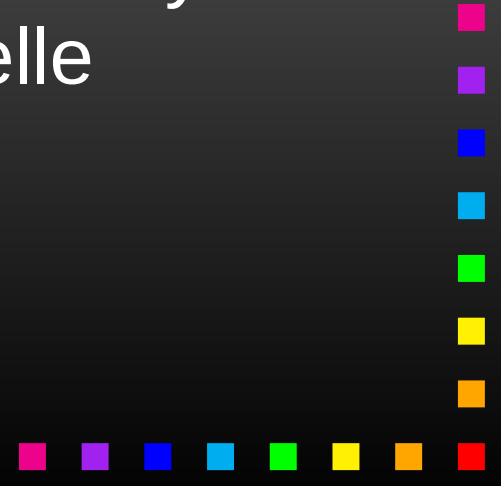
- Bsp. Komma, Minus oder Apostrophe werden bei der Katalogisierung durch typographisch ähnliche Zeichen ersetzt → dieser Umstand plagt uns auch bei jedem Link-Checking
- Dauer `392.616` Sätze, `6m2s` („`check_left_utf8-characters.pl`“)

# Ähnlichkeitsprüfung (1)

- Entladen der normalisierten Daten (`43MB in 10 Spalten`): `3s` (Dem gegenüber: geladen wurden aus Aleph `72MB in 29 Spalten`)
- Laden der normalisierten Daten in eine Vergleichstabelle (`392.616` Sätze): `31s`

# Ähnlichkeitsprüfung (2)

- Problem des Suchbegriffs: zuerst 1. Begriff aus dem Personennamen; wenn dieser leer ist, dann 1. Begriff aus dem Titel, der nicht in der Stoppwort-Tabelle vorkommt
- Problem großer Treffermengen $\longrightarrow$ Analyse und evtl. Begriff in Stoppwort-Tabelle aufnehmen??



# Ähnlichkeitsprüfung: Performanz-Ergebnisse (1)

- 392.616 Anfragesätze und 392.616 Abfragesätze
- im Schnitt ergeben sich dabei 75 Abfrageergebnisse pro Datensatz
- ergibt insgesamt 29.414.802 Ergebnissätze (mit je 21 Einzelberechnungen (Spalten)) $\longrightarrow$ das ist eine 4.3GB flache ASCII-7bit-Textdatei
- Berechnungsdauer: 258 Stunden

# Ähnlichkeitsprüfung: Performanz-Ergebnisse (2)

- 3.100 Anfragesätze und 392.616 Abfragesätze
- im Schnitt ergeben sich dabei 90 Abfrageergebnisse pro Datensatz
- ergibt insgesamt 279.000 Ergebnissätze (mit je 21 Einzelberechnungen (Spalten)) $\longrightarrow$ das sind 40.8MB flache ASCII-7bit-Textdatei
- Berechnungsdauer: 2.4 Stunden

# Ähnlichkeitsprüfung: Performanz-Verbesserung

- Verteilung auf mehrere Rechner (virtuelle Instanzen). Verteilung auf 4 Rechner des verwendeten Typs $\longrightarrow$ 2.5 Tage Rechenzeit
- Vorherige Eliminierung von „eindeutigen“ Dubletten (ISBN + Jahreszahl + Bandang.?) $\longrightarrow$ massive Reduzierung der Anfragesätze

# Vielen Dank für die Aufmerksamkeit