Abschlussvortrag

Softwareprojekt SoSe 2013
Gruppe 2: Answer Ranking

Bastian Beyer, Lauritz Brandt

Universität Heidelberg

22. Juli 2013

Gliederung

Aufgabenstellung

- Question-Answer-Ranking parallelisieren
- zwei ähnliche Korpora vergleichen
- **Können mit vielen Daten, die verrauscht sind, bessere Ergebnisse erzielt werden als mit wenigen sauberen Daten?**

Korpora

Überblick

- L5: kleines Korpus, gesäubert
- L6: großes Korpus, ungesäubert
  $\rightarrow$ Was führt zum besseren Modell?

- **Aufteilung der Korpora:**

Training Set	Development Set	Test Set
80%	10%	10%

Korpora

Korpusstatistik

	L5-Korpus		**L6-Korpus**	
	Fragen	Antworten	Fragen	Antworten
Anzahl:	142.627	962.232	3.895.407	30.740.332
durchschnittliche Länge (Token):	27,9	136,6	28,5	114,9
durchschnittliche Anzahl Antworten pro Frage:	6,7		7,9	

meiste Antworten auf eine Frage: 4.623

Beispiele aus den Korpora

L5

- **Frage:** „How can video games actually be good for you?“
- **beste Antwort:**
  „Video games offer something rare in the terms of education. They cause people to want to learn. Now, there is an exception to every rule but it's usually and exception and not a rule. Myst is one of the best selling computer games of all time and the entire basis is problem solving puzzels that ëxcerciseöur minds. The Sims is another highly succesful game where one [...]“
- **weitere Antwort:** „BY IMPROVING EYE HAND COORDINATION“

Beispiele aus den Korpora

L6

- **Frage:**
  „WILL HE EVER MARRY ME? i luv dis guy so much,but d problem is dat,i have slept with is friends,2 preciesly,and i want 2 get serioua,do u think he would ever luv me,or think of given me a ring?“
- **beste Antwort:**
  „no sista, i don tink so. he be not giving out a ring soon. he be checking out other womens hoo hoo as long as he can. he dont wanna settle down. peace.“
- **weitere Antwort:**
  „THE ONLY WAY HE WILL GET OVER SOMETHING LIKE THAT IF HE REALLY LOVE YOU AND DON'T HAVE A EGO PROBLEM GOOD LUCK“

Beispiele aus den Korpora

Frage ohne beste Antwort (aus L6)

```
<uri>342495</uri>
<subject>when will the complete season of laguna beach
the second season be available to buy?</subject>
<nbestanswers><answer_item>I dont know but hopefully
soon! I love Laguna Beach and cant wait till it comes
on again in the summer.</answer_item></nbestanswers>
```

Features

Feature Group 1: TF-IDF

- term frequency - inverse document frequency
- für ranking functions genutzt
- Fragen $\rightarrow$ querys, Antworten $\rightarrow$ documents

$$\mathrm{TF}(q_i, D) = \frac{\mathrm{f}(q_i, D)}{\max\{\mathrm{f}(w, D) \mid w \in D\}}$$

$$\mathrm{IDF}(q_i, D) = \frac{N - \mathrm{n}(q_i) + 0,5}{\mathrm{n}(q_i) + 0,5}$$

$$\text{TF-IDF}(D) = \sum_{q_i \in Q} \mathrm{TF}(q_i, D) \cdot \mathrm{IDF}(q_i, D)$$

[nach: C. Manning, P. Raghavan, H. Schütze: Introduction to Information Retrieval]

Features

Feature Group 1: BM25

- ähnliches Maß wie TF-IDF
- Fragen $\rightarrow$ querys, Antworten $\rightarrow$ documents

$$\mathrm{BM25}(D) = \sum_{q_i \in Q} \mathrm{IDF}(q_i, D) \; \frac{\mathrm{f}(q_i, D) \cdot (k+1)}{\mathrm{f}(q_i, D) + k\left(1 - b + b\,\frac{L_D}{L_{avg}}\right)}$$

$$b = 0.75$$

$$k = 2.0$$

[nach: C. Manning, P. Raghavan, H. Schütze: Introduction to Information Retrieval]

Features

Feature group 2: Translation Probability

- Wahrscheinlichkeit $P(Q|A)$, dass Frage Q die Übersetzung der Antwort A ist
- Fragen $\rightarrow$ target language, Antworten $\rightarrow$ source language
- IBM Model 1:

$$\mathrm{P}(Q|A) = \prod_{q \in Q} \mathrm{P}(q|A)$$

$$\mathrm{P}(q|A) = (1-\lambda)\,\mathrm{P}_{\mathrm{ml}}(q|A) + \lambda\,\mathrm{P}_{\mathrm{ml}}(q|C)$$

$$\mathrm{P}_{\mathrm{ml}}(q|A) = \sum_{a \in A} \mathrm{T}(q|a)\ \mathrm{P}_{\mathrm{ml}}(a|A)$$

$$\lambda = 0.5$$

- $\mathrm{T}(q|a)$ mit GIZA++ berechnet $\rightarrow$ Bottleneck

Features

Beispiel für Feature-Vektoren

```
1 1.9999999999999998 8.24786190515090 7.676593550828078E-54
0 4.0 4.136133181596938 4.7380691610031 49E-68
1 8.142857142857142 17.775751078190105 5.238548079520859E-110
0 3.6095238095238096 25.088142228444397 1.964811172336369E-140
1 0.0 20.393392588984945 5.756371212343492E-56
0 12.266666666666666 9.750517740965504 2.918606738290437E-68
0 2.2666666666666666 16.921289108445073 7.96886999152163E-99
1 21.377777777777778 13.483295491666304 1.1770417997074081E-57
0 2.9111111111111114 38.351459673539296 7.478472876890658E-75
```

Mehr Beispiele aus den Korpora

Frage mit den meisten Antworten

- **Frage** „Why was disco so successful?“ (4623 Antworten)
- **beste Antwort:**
  „BECAUSE IT WAS SOO HAPPY AND UPBEAT. SOME PEOPLE WOULDN'T WANT TO ADMIT IT BUT THEY LOVED IT. EVERYBODY BOOGIED TO DISCO!!!! AND IF IT WASN'T 4 DISCO WE WOULD'NT HAVE DANCE MUSIC TODAY & I KNOW YOU LOVE DANCE MUSIC. JUST LISTEN TO THE ROOTS OF DANCE MUSIC. ALL DANCE MUSIC IS DISCO MIXED WITH ELECTRONIC FUTURISTIC SOUNDS.“
- **weitere Antworten:**
  - „Because everybody was high during the 70's??“
  - „COCAINE!!!!“
  - „WHO CARES??“
  - „John Travolta“
  - ...

Noch mehr Beispiele

- **Frage:** „How do you tap into the resources of your soul?“
- **beste Antwort:**
  „Without my permission and willingness to allow you to do that, it can’t be done.“
- **weitere Antworten:**
  - „I smoke lots of weed and wake and bake.“
  - „peyote“
  - „prey“
  - „well“
  - „The HOLY GHOST and fervent prayer!!!!!!!!“
  - „The Humanities.“
  - ...

Data processing

Pairing

- **Input:**
  - Fragen mit Antworten
  - voted best answer
  - other answers
- **Output:**
  - gepaarte Vektoren als positive Samples
  - positive Samples: bestAnswer - otherAnswer
    (negative Samples: otherAnswer - bestAnswer)

Data processing

Standardisierung

- schnellere Konvergenz
- Vergleichbarkeit der Werte

$$\frac{\text{Wert} - \text{Durchschnitt}}{\text{Standardabweichung}}$$

Data processing

Binning

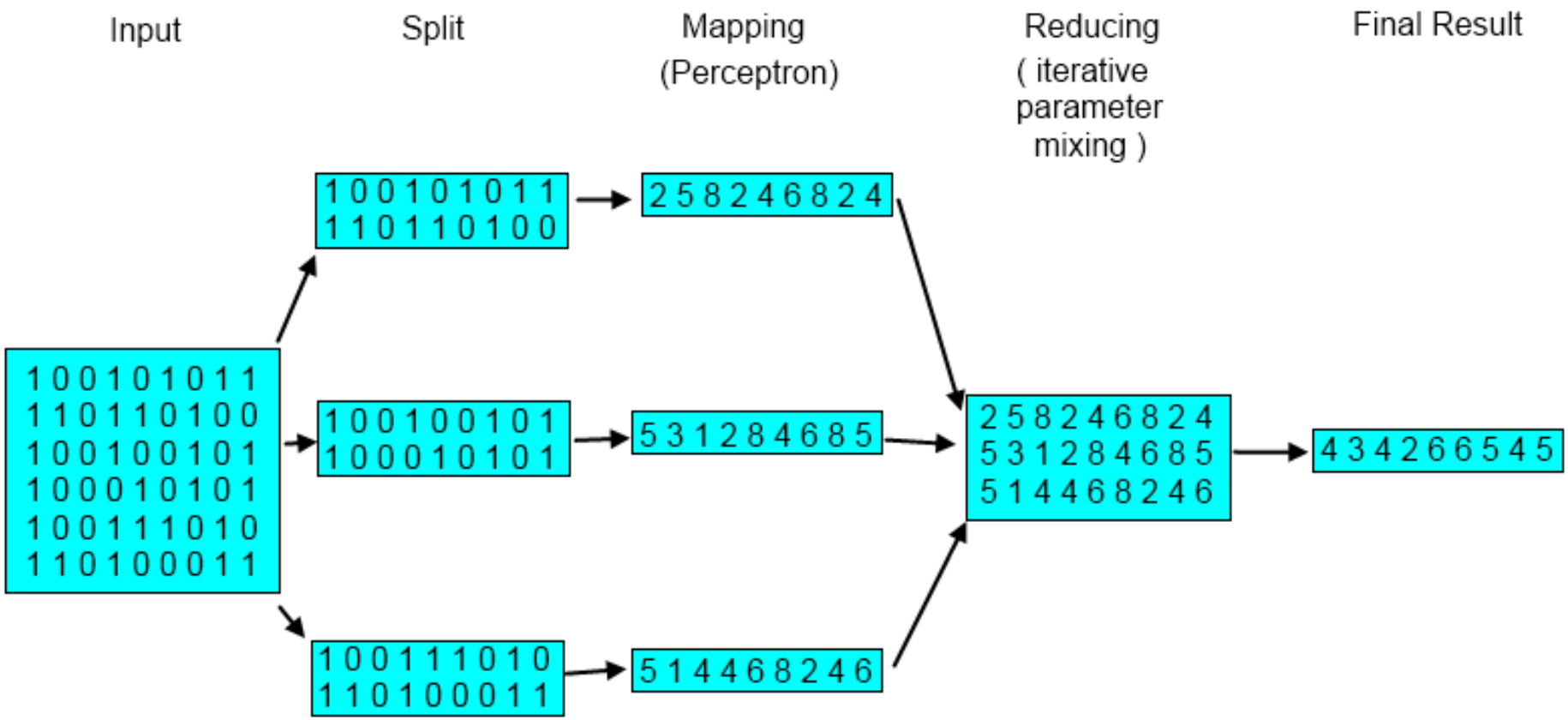
- gut für noisy data
- Einteilung des Wertebereichs in Intervalle
- mehrere binäre Features

Training model

- parallel Perceptron mit MapReduce
- Map: Feature-Vektoren mit Perceptron verarbeiten
- Reduce: Gewichtsvektoren sammeln und verarbeiten

Training model

Hadoop Perceptron



Training model

Data processing model

- Vereinfachung: nur positive Samples
- Algorithmus für iterative parameter mixing:

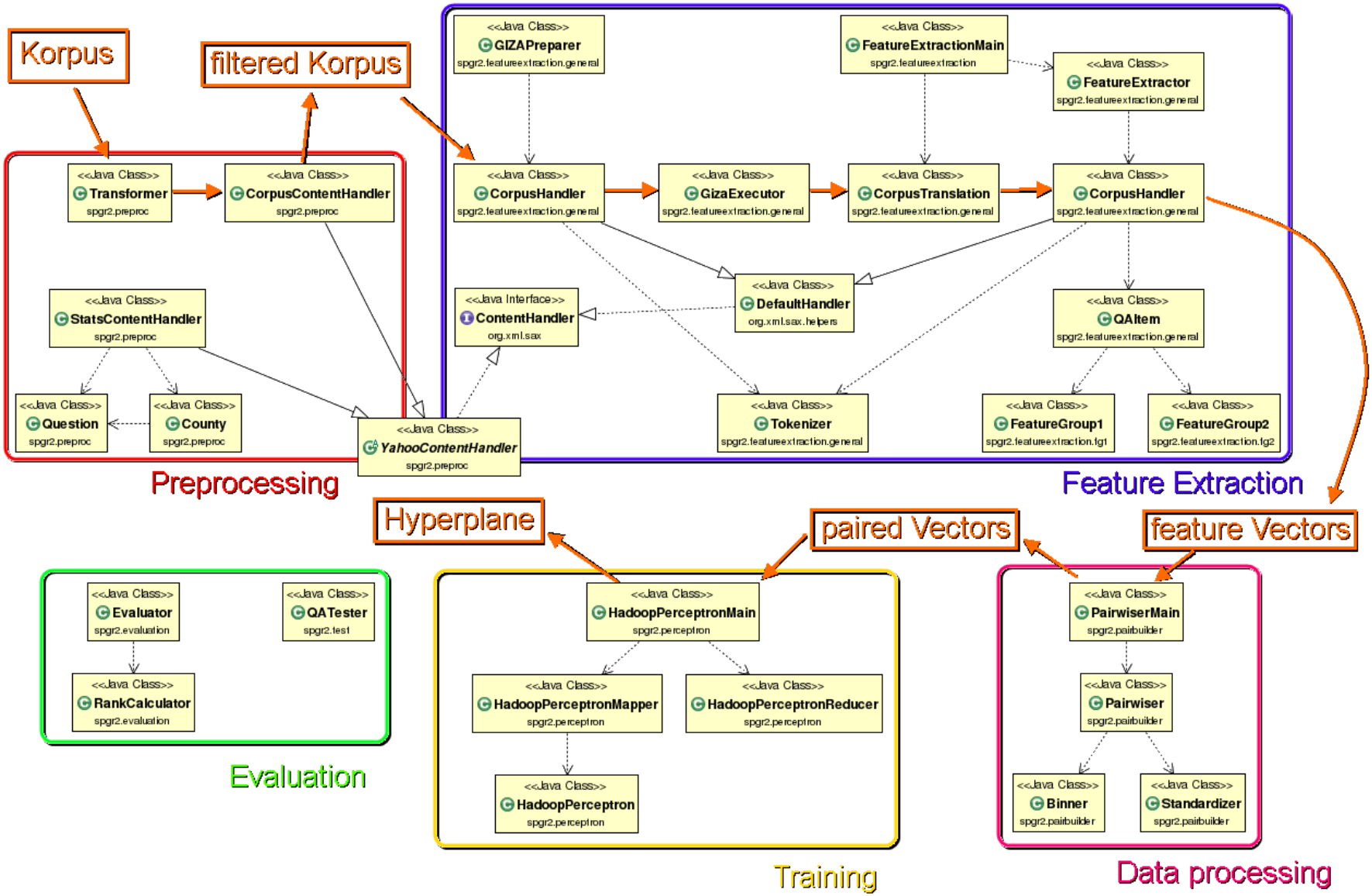
PerceptronIterParamMix($\mathcal{T} = \{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^{|\mathcal{T}|}$)

1. Shard $\mathcal{T}$ into S pieces $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_S\}$
2. $\mathbf{w} = \mathbf{0}$
3. for $n : 1..N$
4. $\quad \mathbf{w}^{(i,n)} = \text{OneEpochPerceptron}(\mathcal{T}_i, \mathbf{w})$ †
5. $\quad \mathbf{w} = \sum_i \mu_{i,n} \mathbf{w}^{(i,n)}$ ‡
6. return $\mathbf{w}$

OneEpochPerceptron($\mathcal{T}, \mathbf{w}^*$)

1. $\mathbf{w}^{(0)} = \mathbf{w}^*;\ k = 0$
2. for $t : 1..T$
3. $\quad$ Let $\mathbf{y}' = \arg\max_{\mathbf{y}'} \mathbf{w}^{(k)} \cdot \mathbf{f}(\mathbf{x}_t, \mathbf{y}')$
4. $\quad$ if $\mathbf{y}' \neq \mathbf{y}_t$
5. $\quad\quad \mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \mathbf{f}(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{f}(\mathbf{x}_t, \mathbf{y}')$
6. $\quad\quad k = k + 1$
7. return $\mathbf{w}^{(k)}$

Datenfluss


Korpus
filtered Korpus
<<Java Class>> Transformer spgr2.preproc
<<Java Class>> CorpusContentHandler spgr2.preproc
<<Java Class>> StatsContentHandler spgr2.preproc
<<Java Class>> Question spgr2.preproc
<<Java Class>> County spgr2.preproc
<<Java Class>> YahooContentHandler spgr2.preproc
Preprocessing
<<Java Class>> GIZAPreparer spgr2.featureextraction.general
<<Java Class>> FeatureExtractionMain spgr2.featureextraction
<<Java Class>> FeatureExtractor spgr2.featureextraction.general
<<Java Class>> CorpusHandler spgr2.featureextraction.general
<<Java Class>> GizaExecutor spgr2.featureextraction.general
<<Java Class>> CorpusTranslation spgr2.featureextraction.general
<<Java Class>> CorpusHandler spgr2.featureextraction.general
<<Java Interface>> ContentHandler org.xml.sax
<<Java Class>> DefaultHandler org.xml.sax.helpers
<<Java Class>> QAItem spgr2.featureextraction.general
<<Java Class>> Tokenizer spgr2.featureextraction.general
<<Java Class>> FeatureGroup1 spgr2.featureextraction.fg1
<<Java Class>> FeatureGroup2 spgr2.featureextraction.fg2
Feature Extraction
Hyperplane
paired Vectors
feature Vectors
<<Java Class>> Evaluator spgr2.evaluation
<<Java Class>> QATester spgr2.test
<<Java Class>> RankCalculator spgr2.evaluation
Evaluation
<<Java Class>> HadoopPerceptronMain spgr2.perceptron
<<Java Class>> HadoopPerceptronMapper spgr2.perceptron
<<Java Class>> HadoopPerceptronReducer spgr2.perceptron
<<Java Class>> HadoopPerceptron spgr2.perceptron
Training
<<Java Class>> PairwiserMain spgr2.pairbuilder
<<Java Class>> Pairwiser spgr2.pairbuilder
<<Java Class>> Binner spgr2.pairbuilder
<<Java Class>> Standardizer spgr2.pairbuilder
Data processing

Evaluation

- Goldstandard: Metadaten im Korpus: die jeweils als beste Antwort auf eine Frage markierte Antwort
- Evaluationssystem erhält Daten aus Testing Unit
- Besonderheiten beim Ranking:
  - jeder Rank nur einmal vergeben $\rightarrow$ Anzahl der FP und FN gleich
  - Recall, Accuracy, F-measure ungeeignet
- zwei Evaluierungsmaße:
  1. Precision
  2. Mean Reciprocal Rank

Evaluation

Precision

- **Precision** gibt an, wie oft die in Bezug auf den GS richtige beste Antwort vom erlernten System als solche erkannt wurde.
- Maß, wie häufig Fehler gemacht werden, egal wie weit sie vom GS abweichen

$$\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

Evaluation

Mean Reciprocal Rank

- **Mean Reciprocal Rank** (MRR) gibt das Inverse vom Rank der korrekten Antwort an
- gibt ergänzend zur Precision an, wie schwer die Fehler sind
- somit einerseits ein Maß für die Anzahl der Fehler und andererseits ein Maß für die Schwere der Fehler

$$\frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}$$

Evaluation

Ergebnisse

bei Training auf:	**L5-Korpus**	**L6-Korpus**
Precision	56.3652%	56.3682%
MRR	68.8143%	68.8173%

- Test-Ergebnisse bei Training auf großem Korpus geringfügig besser als bei Training auf kleinem Korpus, aber kaum merkliche Abweichung
- Abweichung signifikant?
- Bei diesem Versuch konnten mit einer größeren, aber verrauschten Datenmenge keine merklich besseren Ergebnisse erzielt werden.

Resümee

Schwächen der Implementierung

- GIZA++ läuft sehr lange → parallelisieren / Performanz verbessern
- zudem großer Speicherbedarf → größtes Bottleneck
- wenig modular programmiert, parallel Perceptron und feature extraction sehr auf Yahoo!Answers-Korpora ausgelegt
- Hadoop Job Management muss besser werden → Hadoop Interface schwierig
- viele Nullwerte bei feature extraction
- teilweise uneffizienter Code (Redundanz etc.) → Code überarbeiten

Resümee

Weiterführendes

- feature extraction / GIZA++ parallelisieren
- Feature groups 3 und 4 implemetieren, um sich auf unsauberen Daten zu verbessern
- ähnliche Fragen gruppieren
- mehr Samples durch Cross-pairing von Fragen und Antworten

Resümee

Probleme beim Projektablauf

- Terminplan muss eingehalten werden → Deadlines einführen und Status abfragen
- technische Schwierigkeiten (z.B. hohe Serverauslastung)
- Gruppenmitglieder sind nur teilweise verlässlich → vorher klare Ziele stecken, die alle akzeptieren und umsetzen
- Gruppenleiter sinnvoll
- Gegenseitige Motivierung zum Arbeiten hilft bei der Einhaltung von Deadlines!

Quellenangabe (1)

- Felix Hieber, and Stefan Riezler. Improved Answer Ranking in Social Question-Answering Portals. SMUC '11, October 28, 2011, Glasgow, Scotland, UK.
- Ryan McDonald, Keith Hall, and Gideon Mann. Distributed Training Strategies for the Structured Perceptron. In: Human Language Technologies. The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (June 2010), pp. 456-464.
- Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In: Proceedings of WSDM'08, Palo Alto, CA, 2008.

Quellenangabe (2)

- Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers on large online QA collections. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08), Columbus, OH, 2008.
- Libin Shen and Aravind K. Joshi. Ranking and reranking with perceptron. In: Journal of Machine Learning Research, 60(1-3): 73-96, 2005.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.

Vielen Dank für eure Aufmerksamkeit!

- **Frage:**
  „A farmer had 15 sheep, and all but 8 died. How many are left?“
- **beste Antwort:** „15, a dead sheep is still a sheep.“
- **weitere Antwort:**
  „8 duh. why do you keep asking these same weird ?s they are ridiculously weird“

- **Frage:** „How can childhood obesity be prevented?“
- **weitere Antworten:**
  - „liposuction“
  - „no“
  - „Get them to run around fast food joints, but never go in.“