Information Retrieval and Text Mining

http://informationretrieval.org

Practical exercise 3 (IIR 5, IIR 6 & IIR 9)

Max Kisselew

Institute for Natural Language Processing, University of Stuttgart

2012-11-20

Inhalt

- Was ist was?
- Besprechung der Aufgaben
- Eure Fragen zum bisherigen Inhalt

Was ist was? (IIR 5)

- Zipf's law
  - Zipf's law: The i^{th} most frequent term has frequency proportional to $1/i$.
  - $\mathrm{cf}_i \propto \frac{1}{i}$
  - cf is collection frequency: the number of occurrences of the term in the collection.
  - So if the most frequent term (*the*) occurs cf_1 times, then the second most frequent term (*of*) has half as many occurrences $\mathrm{cf}_2 = \frac{1}{2}\mathrm{cf}_1$ ...
  - ...and the third most frequent term (*and*) has a third as many occurrences $\mathrm{cf}_3 = \frac{1}{3}\mathrm{cf}_1$ etc.
- Heaps law
  - $M = kT^b$
  - M is the size of the vocabulary, T is the number of tokens in the collection.
  - Typical values for the parameters k and b are: $30 \leq k \leq 100$ and $b \approx 0.5$.

Besprechung der Aufgaben 1 + 2

Was ist was? (IIR 6)

- document frequency df_t
  - Number of documents in the collection that contain a term t
- inverse document frequency idf_t
  - $\log \frac{N}{df_t}$
- vector space model
- document vector

Binary → count → weight matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
ANTHONY	5.25	3.18	0.0	0.0	0.0	0.35	
BRUTUS	1.21	6.10	0.0	1.0	0.0	0.0	
CAESAR	8.59	2.54	0.0	1.51	0.25	0.0	
CALPURNIA	0.0	1.54	0.0	0.0	0.0	0.0	
CLEOPATRA	2.85	0.0	0.0	0.0	0.0	0.0	
MERCY	1.51	0.0	1.90	0.12	5.25	0.88	
WORSER	1.37	0.0	0.11	4.15	0.25	1.95	
...							

Each document is now represented as a real-valued vector of tf-idf weights $\in \mathbb{R}^{|V|}$.

Binary → count → weight matrix

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Anthony	5.25	3.18	0.0	0.0	0.0	0.35	
Brutus	1.21	6.10	0.0	1.0	0.0	0.0	
Caesar	8.59	2.54	0.0	1.51	0.25	0.0	
Calpurnia	0.0	1.54	0.0	0.0	0.0	0.0	
Cleopatra	2.85	0.0	0.0	0.0	0.0	0.0	
mercy	1.51	0.0	1.90	0.12	5.25	0.88	
worser	1.37	0.0	0.11	4.15	0.25	1.95	
...							

Each document is now represented as a real-valued vector of tf-idf weights $\in \mathbb{R}^{|V|}$.

Was ist was? (IIR 6)

- document frequency df_t
  - Number of documents in the collection that contain a term t
- inverse document frequency idf_t
  - $\log \frac{N}{df_t}$
- vector space model
- document vector
- cosine similarity

Besprechung der Aufgaben 3 + 4

Eure Fragen

Noch Fragen?

Ende

Bis zum nächsten Mal!