

## Term Extraction on ancient Texts

Text Mining for Classical Studies  
Leipzig, 2009/10/01

**Marcus Puchalla**  
Natural Language Processing Group  
Department of Computer Science  
University of Leipzig

---

- introduction
- what's term extraction
- term extraction strategies:
  - TFIDF
  - frequency class method
- preliminary work status
- look-out

introduction

- Bachelor Thesis at the NLP Group, Leipzig
  - “Term Extraction an ancient Texts using modern natural language processing methods”
- main concern is efficiency analogy of different term extraction methods

what's term extraction

- extract/collecting relevant terms for a given domain
- constituting the linguistic surface manifestation of domain concepts
  
- linguistic, statistical and hybrid appendage
  - no real knowledge about verbal features
  - concentrate on statistical methods
  
- progress of term extraction
  1. calculate weights for words in corpus
  2. define a marginal value for possible term candidates (tc)
  3. create list of possible tc
  4. experts needed for final decision if term is relevant

term extraction strategies

TF-IDF

term frequency – inverse document frequency

- importance of a single word in document of a collection
- product of
  - TF – term frequency
    - how often occurs the word in a given document

$$ft_{ij} = \sum \text{occurrences}(word_i, document_j)$$

- IDF – inverse document frequency
  - how important in general is the word

$$tfidf_i = \log \left( \frac{\sum \text{documents}}{\sum (\text{documents} - \text{containing} - word_i)} \right)$$

- tf – prefer words occur very often in a single document
  - often ~ important
  - sometimes infrequent words can be very important
- idf – prefer words occur seldom in the whole collection
  - subject specific words may only appear in some documents
  - words occur in all documents get a zero score, (mathematical reason)

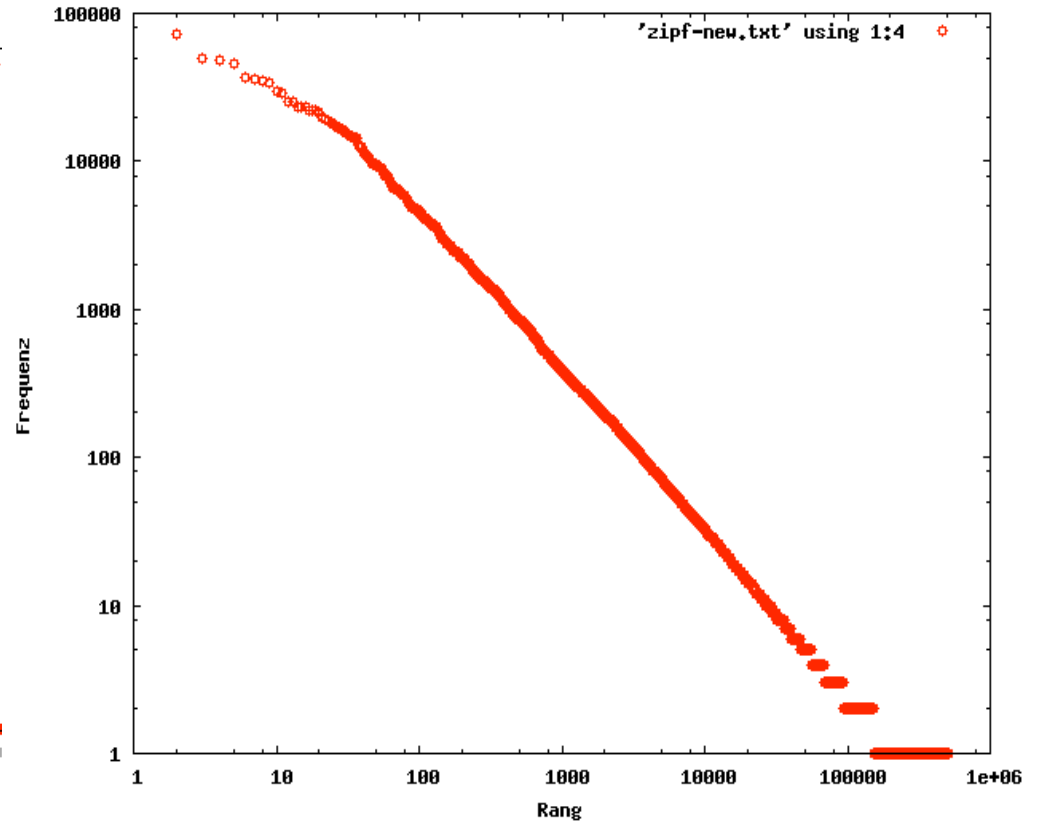
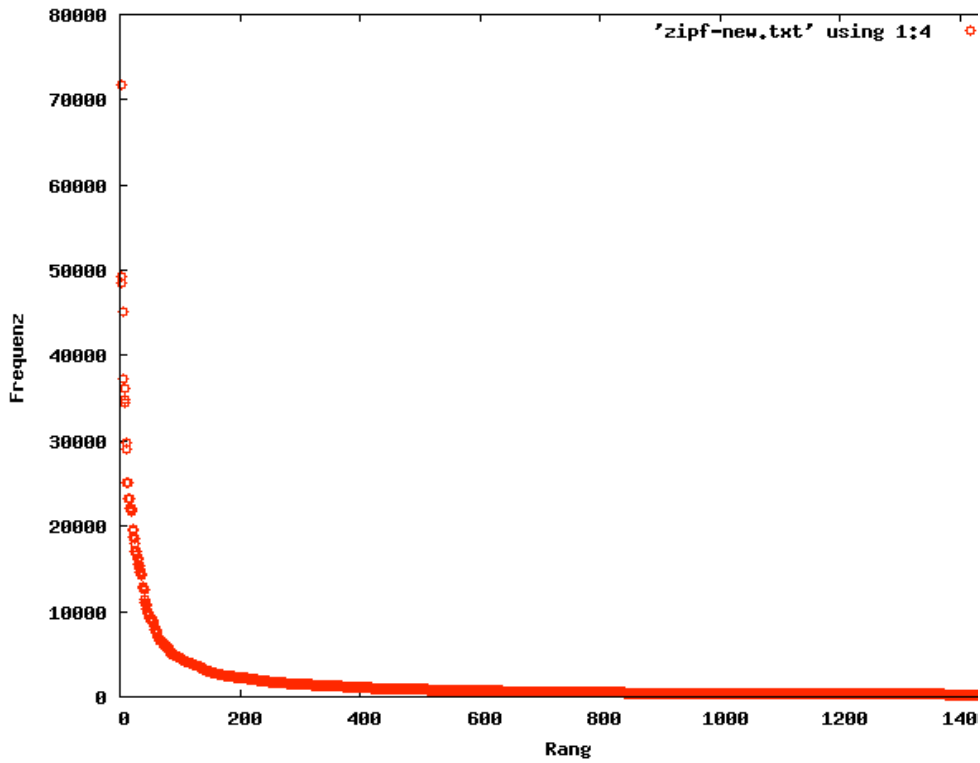
tf \* idf prefers words that occur often in a some document but relatively seldom in the collection of documents

## frequency class method

- based upon Zipf's law  
'few words are used often but most words are used rarely'
- how often the most common word of a corpus occurs in comparison to the analysed word
- most common words e.g.
  - duke db - "και"
  - german - "der"
  - english - "the"

preliminary work status

zipf distribution for duke database



log

look-out

TLG to come