**Other Keyword Extracting Methods**

**LDA (Latent Dirichlet allocation)**

* Highly dependent on the choosing of stop words
* Fixed K (the number of topics is fixed and must be known ahead of time)
* Uncorrelated topics (Dirichlet topic distribution cannot capture correlations)
* Non-hierarchical (in data-limited regimes hierarchical models allow sharing of data)
* Static (no evolution of topics over time)
* Bag of words (assumes words are exchangeable, sentence structure is not modeled)
* Unsupervised (sometimes weak supervision is desirable, e.g. in sentiment analysis)
* Limitations: LDA can be used to identify topic information in document collection or corpus. It uses the bag of words model, which treats each document as a word frequency vector, and translates text information into digital information. However, bag of words model does not consider the order between different words, which provide an opportunity to improve the model. The computational complexity of LDA model is high, and the speed is slow while running large-scale data set. Moreover, time limits problem is considered. The researcher does not have much time to achieve LDA implementation of keyword extraction

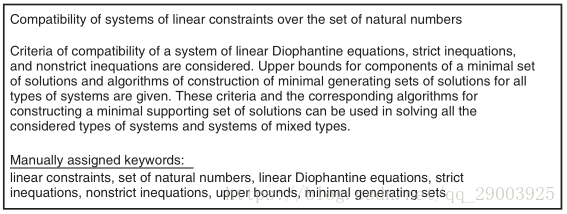
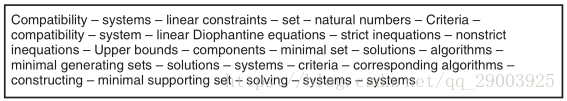
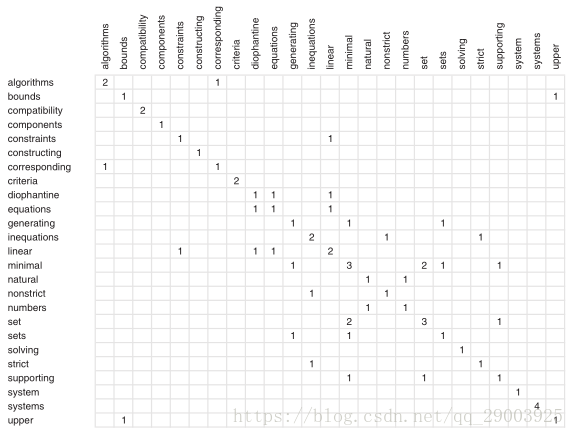
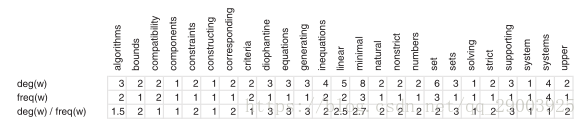
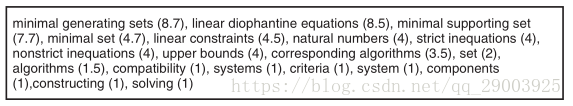
**Probabilistic latent semantic analysis (PLSA)**

* Limitations: PLSA appeared the Overfit phenomenon. It is a phenomenon in which a hypothesis is better fit on training data than the other assumptions, but the data are not well fitted in addition to training data. In order to solve this problem, LDA introduced dirchlet distribution. LDA is more advanced than PLSA due to LDA is given a prior dirchlet distribution for each document topics. PLSA mainly focuses on semantic emotional analysis. Hence, it is not suitable for keyword extraction by using PLSA model. The following content will indicate three other expected algorithms with different advantages, disadvantages and limitations

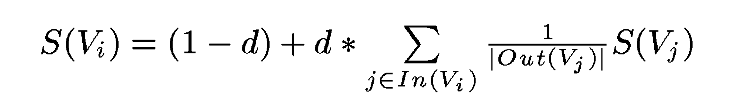
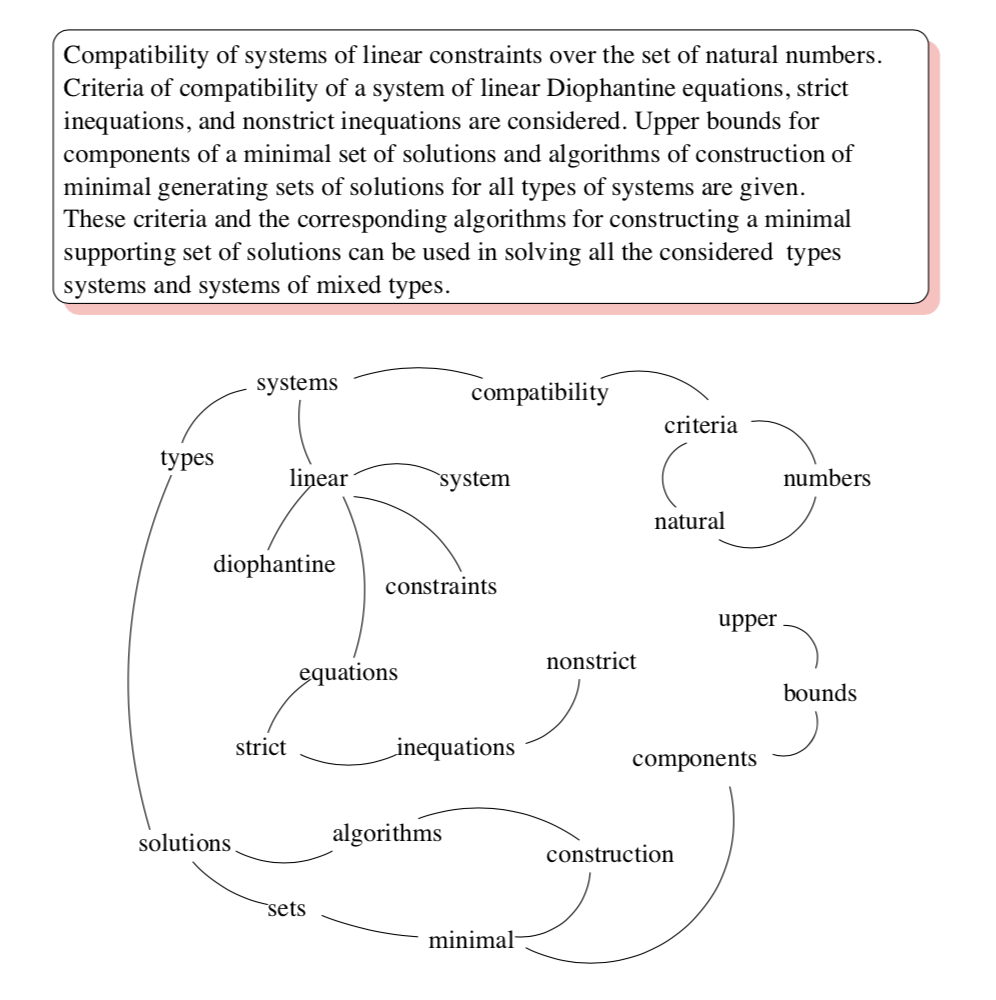
**TF-IDF**

* Limitations: When the document ratio of keywords appeared is large. The low frequency words will be highlighted. These words are not equal to meaningless words, such as some public figures, hot events, etc.. These low frequency words will be regarded as high-weight keywords, which over extend the importance of uncommon words.

**Rapid Automatic Keyword Extraction (RAKE)**

* The advantage is mostly “rapid”
* Able to select multiple words (as phrases) as keywords
* First divide the document into several sentences by the punctuation
* Divide the sentences further into several phrases by stop words
* wordScore = wordDegree(w) / wordFrequency(w)
* 
* 
* 
* 
* RAKE makes key phrases to acquire higher weight, so the accuracy is inadequate.

**Text Rank**

* Based on Page Rank
* Graph based ranking algorithm
* 
* 
* Disadvantages: 1. Compared with TF-IDF, the relation between the effect of keyword extraction of TextRank and the effect of segmentation is extremely large. 2. TextRank performance is not better than TF-IDF based on 50 documents. 3. Due to TextRank needs to achieve words construction and repeated iterations, the extraction speed is slower than TF-IDF.

References:

https://pats.cs.cf.ac.uk/@archive\_file?p=855&n=final&f=1-Final\_report\_Weinan\_Fan\_1631086.pdf&SIG=b37b7a5bb246ef882ec304ad10b55012ba05dd65f3ae166a60c50d2084b3438e