A Study on Emerging Topics Discovery by Text Minging

Team Ziqiang Liu, Haiyun Xu, Zenghui Yue, Guoting Yuan, Yan Qi, Zhengyin Hu and Ning Yang

Reporter: Yan Qi

Atlanta, 10.17

Step 1: Parsing and Rreprocessing of Raw Data

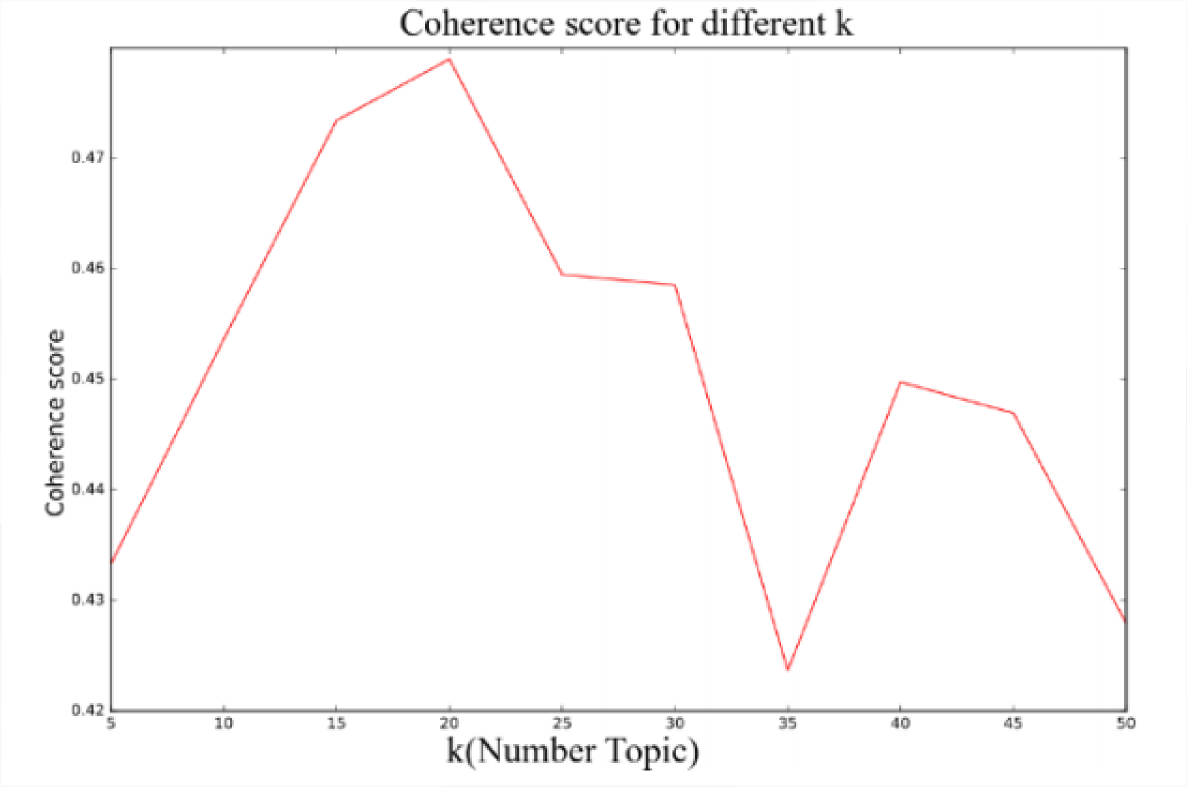
The XML data was parsed into CSV format with a total of 2,584 records. Then three fields of **the pubyear, title and abstract** were extracted into the CSV dataform and the text field was obtained by merging the title and abstract fields. Table1shows the basic format of the experimental data.

Id	Pubyear	Text（title+abstract）
1	T1	Text1
2	T2	Text2
3	T3	Text3
……	……	……
2584	T_m	$Text_n$

NLTK toolkit was used in the data preprocessing of text field, including conversion of capitalization, removing of numbers and punctuation and filtering of pause words, to improve the accuracy and validity of topics recognition.

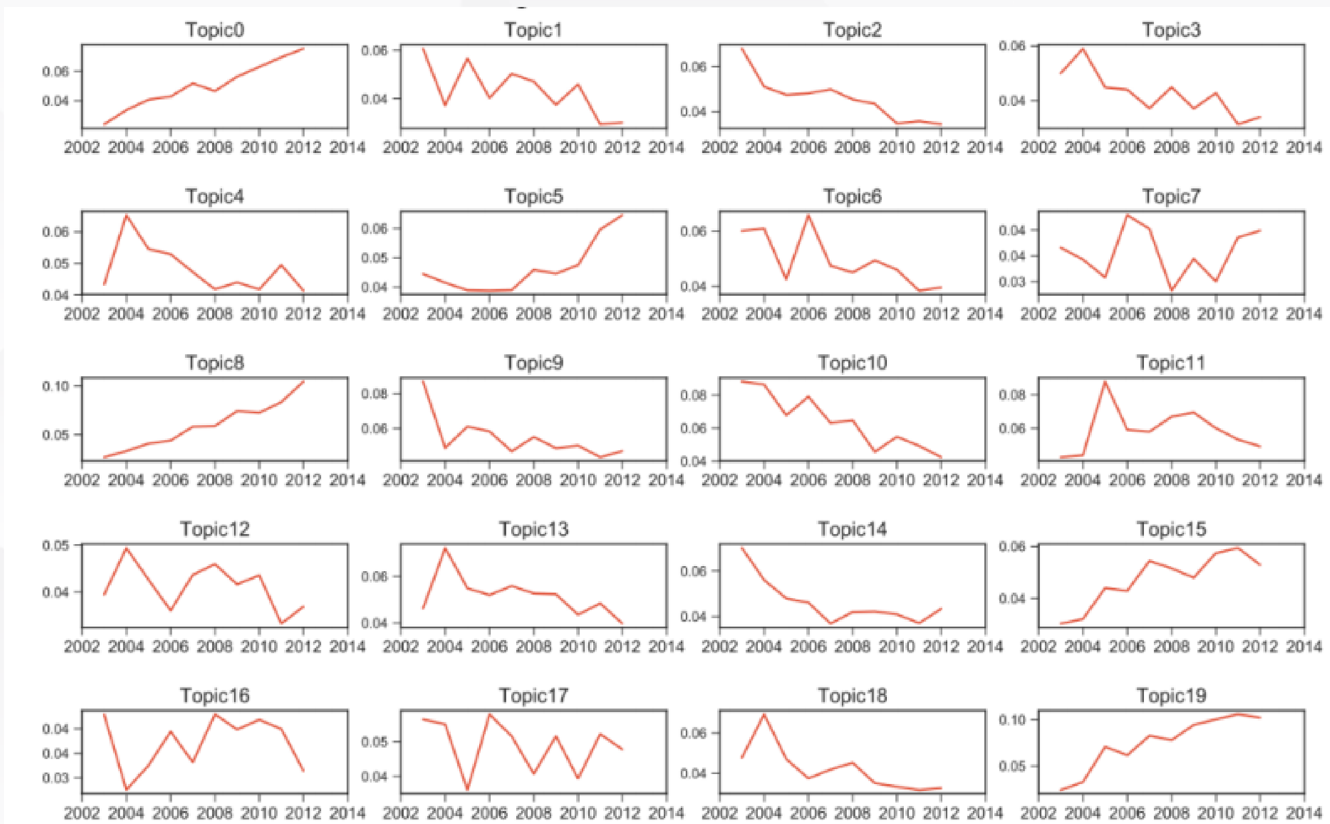
Step 2: LDA Topic Recognition with Optimized Parameters

Python's Gensim toolkit was used for topic recognition in the method. In the parameter optimization, the number of topics k was determined by calculating the consistence score of the topic model with alpha=50/k, beta=0.01.**The consistency score gets its highest value when k = 20. Therefore, the topics number k = 20 was determined with alpha = 2.5, beta = 0.01.**



Step 3: Building of Topic Time Intervals

Based on the results of the previous LDA topic recognition, the topic change time intervals data were further constructed with **the self-defined function def topic_dis_year()**. In order to observe the trend of the topic time intervals more intuitively and effectively, the topic time intervals line chart was drawn by using the Matplotlib toolkit.



Step 4: Emerging Topics Recognition

BasedThe function def gram_dis_year() was defined to construct the 2-gram vocabulary time intervals data by extracting 2-gram vocabulary with NLTK tools. Next, **the average growth rate of each 2-gram vocabulary** in the past three years was calculated, and the top300 topic words were obtained through a sorting. Finally, the vocabulary of **top300 topic words were mapped as a schedule** to our emerging topics identification results.

	A	B	C	D
1	rank	Topic-id	Topic words	emerging(2-gram)
2	1	Topic8	biology	"synthetic biology", "biology a
3	1	Topic8	synthetic	"synthetic biology", "syntheti
4	1	Topic8	engineere	"metabolic engineere", "engine
5	1	Topic8	system	"biological system", "genetic s
6	1	Topic8	biological	"biological system", "syntheti
7	1	Topic8	genome	"synthetic genome", "genome wi
8	1	Topic8	assemle	
9	1	Topic8	recent	
10	1	Topic8	field	"field synthetic"
11	1	Topic8	development	"development synthetic"
12	2	Topic5	enzyme	"restriction enzyme"
13	2	Topic5	production	"protein production"
14	2	Topic5	strain	"coli strain"
15	2	Topic5	substrate	"substrate specificity"

Thanks for your attention

Q&A?