META KNOWLEDGE 库介绍

汇报人：李晓敏

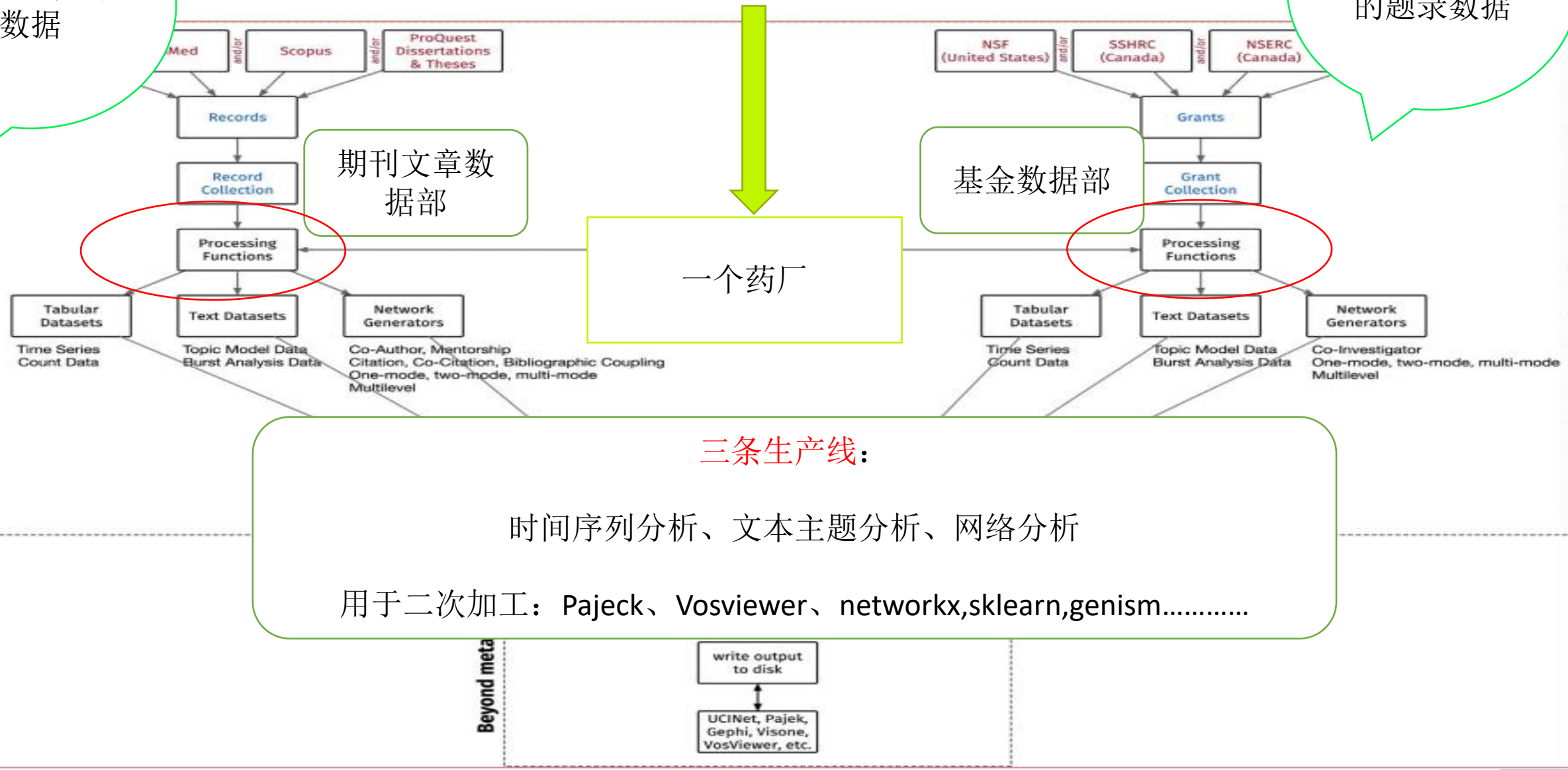
METAKNOWLEDGE简介

Metaknowledge 包是由John McLevey基于python语言开发的用于分析文献数据的包，这个分析包能对Scopus、web of science、 PubMed,以及部分资助基金相关的题录数据进行初步处理，其处理后的文献数据能与其它数据分析包和软件进行无缝衔接，对文献数据进行引文年谱分析（RPYs），网络分析（networkx、pajeck）以及文本分析（LDA，genism）

用一个比喻来理解MK的工作流程

药材1：文献数据

药材2：基金的题录数据



期刊文章数据部

基金数据部

一个药厂

三条生产线：

时间序列分析、文本主题分析、网络分析

用于二次加工：Pajeck、Vosviewer、networkx,sklearn,genism……

导入相关的包

未安装→命令行： **pip3 install metaknowledge==3.3.2**

Jupyter: ! pip3 install metaknowledge==3.3.2

```
import metaknowledge as mk
#网络分析包
import networkx as nx
#作图用
import matplotlib.pyplot as plt
import seaborn as sns
#数据分析
import pandas as pd
import os
```

METAKNOWLEDGE之时间序列分析

相关概念——RPYS(参考文献出版年谱分析)

指以某一个学科领域的全部相关文献所引用的全部参考文献的“出版年份”为横轴，以每年全部被引参考文献的“总被引频次”为纵轴而形成的分布图。

其原理是: 在一个学科领域的相关文献所引用的所有参考文献中，只有很小比例的参考文献是在该学科领域产生之前发表的; 且在这小部分参考文献中，通常存在几篇文献的被引用频次远高于同年或前后几年内发表的其他参考文献，那么这几篇文献很可能是对该学科领域的产生具有重要作用的根源文献。同时，满足上述条件的参考文献在图谱上出现的位置一定是图谱上的某个峰值。因此，通过对参考文献出版年图谱在学科领域产生之前的峰值进行分析，就有可能找到该学科领域的历史根源文献。

[1]李信,陆伟,李旭晖.一种新兴的学科领域历史根源探究方法:RPYS[J].图书情报工作,2016,60(20):70-76.

METAKNOWLEDGE之时间序列分析

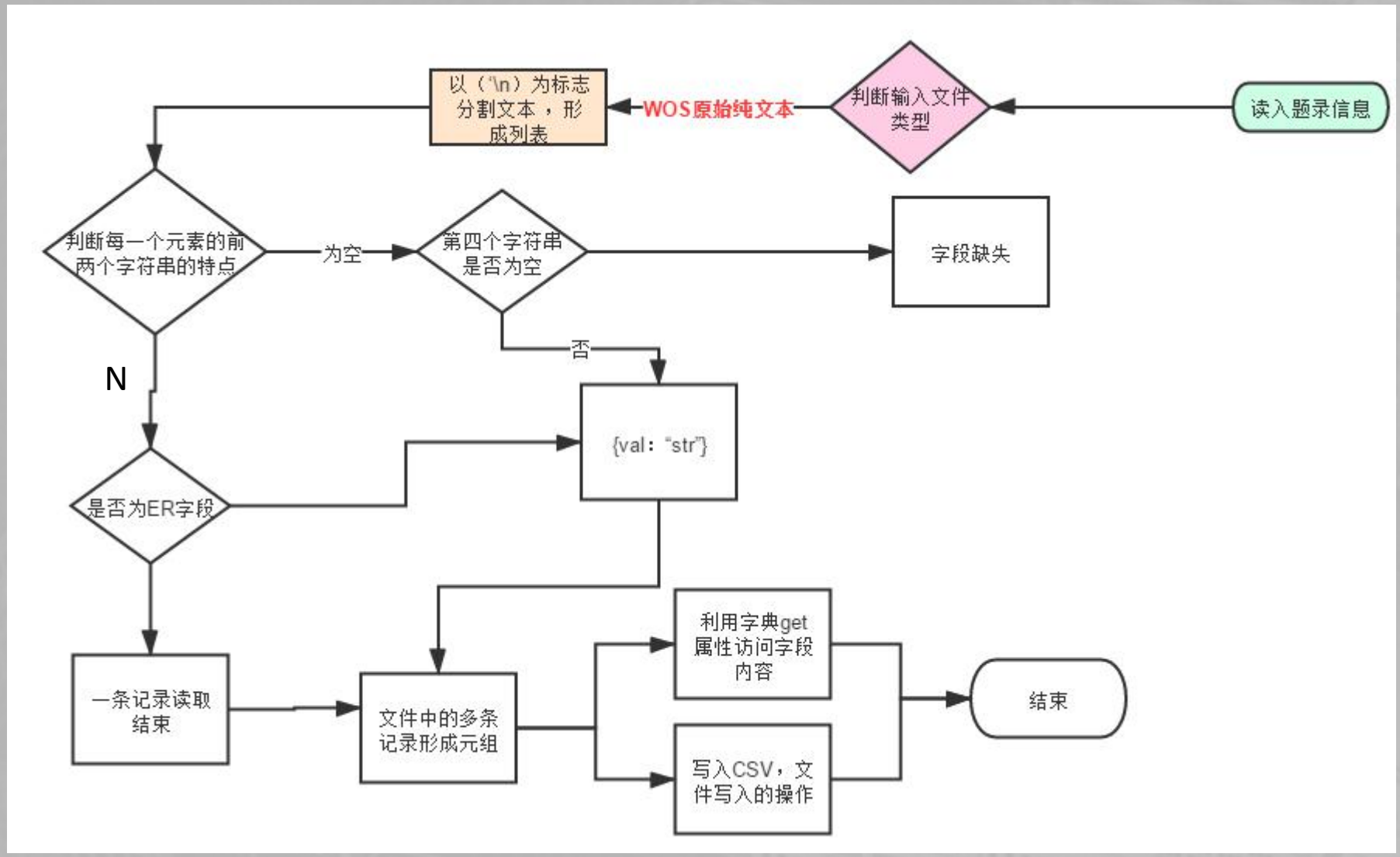
相关概念——标准化RPYS(参考文献出版年谱分析)

通过计算每一出版年度的引用文献数量偏离五年中值的程度，这些与中位数的偏差可以用计数或百分比来表示，在重要的年份会出现明显的峰值出版的书籍或文章。

源码分析
——题录数据结构
——记录解析流程

```
FN Clarivate Analytics Web of Science
VR 1.0
PT J
AU Huang, H
   Cao, KM
   Kong, YQ
   Yuan, SM
   Liu, HK
   Wang, YC
   Liu, YZ
AF Huang, Hai
   Cao, Kaiming
   Kong, Yaqiong
   Yuan, Siming
   Liu, Hongke
   Wang, Yucai
   Liu, Yangzhong
TI A dual functional ruthenium arene complex induces differentiation and
   apoptosis of acute promyelocytic leukemia cells
SO CHEMICAL SCIENCE
LA English
DT Article
ID PML-RAR-ALPHA; BINDING; ONCOPROTEIN; LINES; DNA
AB Human acute promyelocytic leukemia (APL) is the most malignant form of acute leukemia. The fusion of PML and RAR alpha genes is responsible for over 98% of cases of A
C1 [Huang, Hai; Cao, Kaiming; Yuan, Siming; Liu, Yangzhong] Univ Sci & Technol China, Dept Chem, CAS Key Lab Soft Matter Chem, Hefei 230026, Anhui, Peoples R China.
   [Kong, Yaqiong; Liu, Hongke] Nanjing Normal Univ, Coll Chem & Mat Sci, Jiangsu Key Lab Biofunct Mat, Nanjing 210046, Jiangsu, Peoples R China.
   [Wang, Yucai] Univ Sci & Technol China, Sch Life Sci, Hefei 230027, Anhui, Peoples R China.
   [Wang, Yucai] Univ Sci & Technol China, Med Ctr, Hefei 230027, Anhui, Peoples R China.
RP Liu, YZ (reprint author), Univ Sci & Technol China, Dept Chem, CAS Key Lab Soft Matter Chem, Hefei 230026, Anhui, Peoples R China.
EM liuyz@ustc.edu.cn
OI Wang, Yucai/0000-0001-6046-2934
FU National Key R&D Program of China [2017YFA0505400]; National Natural
   Science Foundation of ChinaNational Natural Science Foundation of China
   [21877103, 21573213]; Collaborative Innovation Center of Suzhou Nano
   Science and Technology
FX This work was supported by the National Key R&D Program of China
   (2017YFA0505400), the National Natural Science Foundation of China
   (21877103, 21573213) and the Collaborative Innovation Center of Suzhou
   Nano Science and Technology. A portion of this work was performed at the
   Steady High Magnetic Field Facilities, High Magnetic Field Laboratory,
```

源码分析——流程


读入题录信息
判断输入文件类型
WOS原始纯文本
以（'\n）为标志分割文本，形成列表
判断每一个元素的前两个字符串的特点
为空
第四个字符串是否为空
字段缺失
否
N
是否为ER字段
{val：“str”}
一条记录读取结束
文件中的多条记录形成元组
利用字典get属性访问字段内容
写入CSV，文件写入的操作
结束

METAKNOWLEDGE之时间序列分析

mk核心代码总结：

```
import metaknowledge as mk
folder_collec = mk.RecordCollection(r'filepath')
folder_collec.writeCSV(r"F:\metaknow\example data\folder.csv")
df = pd.DataFrame(folder_collec.makeDict())
growth_by_journal = pd.DataFrame(folder_collec.timeSeries('journal', outputFile =
r'F:\metaknow\example data\growth_journals.csv'))
RC1314 = folder_collec.yearSplit(2013, 2014)
folder_collec.rpys(2000,2019)
```

METAKNOWLEDGE之网络分析

社会网络分析：

“社会网络 ”指的是社会行动者及其间关系的集合,社会网络分析 就是要建立关系模型，力图描述群体关系结构，研究这种结构对群体功能或者群体内部个体的影响 。

科学文献之间的引证关系自然形成了引文网络，它表现出科学文献之间纵向继承和横向关联的交流态势 。在文献的引用关系中,除了单一引用外，还存在两篇或两篇以上文献同时引用同一篇文献的 “文献耦合（bibcoupling） ”关系，或两篇文献同时被别的文献共同引用的“文献同引 （cocitation）”关系 。

[1]宋歌.社会网络分析在引文评价中的应用研究[J].图书情报工作,2010,54(14):16-19+115.

METAKNOWLEDGE之网络分析

mk核心代码总结：

```
coauth_net = folder_collec.networkCoAuthor()

mk.graphStats(coauth_net)

mk.dropEdges(coauth_net, minWeight = 2, dropSelfLoops = True)

journal_cocite = folder_collec.networkCoCitation(coreOnly = True)
```

METAKNOWLEDGE之文本分析

LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。

LDA是一种非监督机器学习技术，可以用来识别大规模文档集（document collection）或语料库（corpus）中潜藏的主题信息。它采用了词袋（bag of words）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

METAKNOWLEDGE之词袋模型

Bag-of-words model (BoW model) 忽略掉文本的语法和语序等要素，将其仅仅看作是若干个词汇的集合，文档中每个单词的出现都是独立的。BoW使用一组无序的单词(words)来表达一段文字或一个文档.。近年来，BoW模型被广泛应用于计算机视觉中。

例子：

We are having data analysis and visualization class.

Do you like having data analysis and visualization class ?

生成词典（语料库）：｛‘We’：0，‘are’：1，‘having’：2，‘data’：3，‘analysis’：4，‘and’：5，‘visualization’：6，‘class’：7，‘Do’：8，‘you’：9，‘like’：10｝

每个句子都可以用11维的向量来表示：

例如第2句：

[0,0,1,1,1,1,1,1,1,1,1]

METAKNOWLEDGE之文本分析

mk核心代码总结：

```
folder_collec = mk.RecordCollection(r'F:\metaknow\example data')

topic = folder_collec.forNLP(r'F:\metaknow\example data\LDA_folder.csv',

dropList=stopwords,lower=True,removeNumbers=True)

document = topic['abstract']

docs = np.asarray(document)
```

谢谢！