This is a sample Final Report template.

Abstract / Summary

* A synopsis of the proposed project, including the subset of GDELT data, the research question (you only need one question or hypothesis), which algorithm(s) will be used, the intended result, and the significance of the result. Suggested length: one paragraph.
* No abbreviations, subheadings, or citations.

Part1：

Research question: What is the difference between the topics of articles in China, the USA, and Russia in 2021? To investigate this question I used the shopping basket algorithm to analyze the topics of each country, and the topics of the articles were taken from the GKG Table of Gdelt2.0. The expected results, due to the epidemic and economic decline in 2021, I guessed that the topics of the articles would be related to taxes, income, and epidemics. Throught getting this result we can have some idea of what the above three countries will be concerned about in the year 2021.

疫情开始前后的中国社会关注的主题区别

研究问题：在疫情发生前，疫情开始流行与经历一段时间的疫情后，中国社会讨论所关注的主题有什么区别？研究这个问题这个项目对不同时期中国的主题进行市场篮子分析，文章的主题来自于Gdelt 的Global Knowledge Graph 2.0。预期结果，在疫情开始后，中国社会会更关注健康与流行病的主题。这个研究结果有助于了解疫情开始前后和发展一段时间后中国社会关注主题的变化，帮助研究疫情对中国社会的影响以及研究疫情的变化趋势。

##################################################################

Introduction

* Background. Explain the background information necessary to understand the features of the GDELT data you are using. Explain background information necessary to understand the algorithm(s) you will use. Suggested length: 1-2 paragraphs
* Motivation. Why is the research question relevant to you? What of your experience connects to the data, algorithms, or result of this research? How could this research help share your interests with others? Suggested length: 1-2 paragraphs
* Research Question or Hypothesis. You only need one question or hypothesis. Make sure to explain how the research question is relevant for the data set. Also explain how implementing the algorithm will help answer the research question. Suggested length: 1-2 paragraphs

研究数据背景：研究数据是来自Gdelt数据库中的Global Knowledge Graph 2.0，里面包含65种语言的世界新闻与文章，每一行有事件的链接，发生地点，文章主题，文章情绪等变量。这个项目将GKG Table中的Location与V2Themes字段提取出来，他们分别代表事件发生的地理位置与文章主题。下载得到三个时期的数据后再根据CountryCode对数据进行归类，把CountryCode为CH（中国） 的行过滤出，然后分别组成三个数据集对其进行分析。

算法背景：1.市场篮子分析是通过研究用户消费数据，将不同商品之间进行关联，并挖掘二者之间联系的分析方法，叫做商品关联分析法，即“购物篮分析”。说起购物篮分析，就不得不提到那个经典的案例：啤酒与尿不湿：其实这是营销里一个非常经典的案例。讲的是在美国沃尔玛超市里，经理发现很多啤酒和尿不湿一起购买的订单，经过调查发现是男性来给孩子买尿不湿时，会顺带给自己再买几瓶啤酒。后来经理把啤酒和尿不湿摆在一起，大大提高了两者的销量。购物篮分析的目的在于，通过对用户消费记录数据的挖掘和分析，找出用户购买习惯的一些潜在规律，从而可以为用户提供他们想要的搭配或套餐。套餐销量的提升一般会带来客单价的提升，从而提高公司收益。

2.关联规则挖掘可以让我们从数据集中发现项与项（item 与 item）之间的关系，“市场篮子分析”就是一个常见的场景，这个场景可以从消费者交易记录中发掘商品与商品之间的关联关系。The form of an association rule is I → j, where I is a set of items and j is an item. 关联规则研究的是集合I对J的影响。分析关联规则的support，confidence与interest，可以对某条关联规则有更多了解，从而发现有趣的关联规则。

3.Apriori算法： Apriori 算法其实就是查找frequent itemset 的过程。频繁项集是支持度大于等于Min Support阈值的项集，小于最小值支持度的itemset 被称为非频繁项集。它是一种逐层迭代的方法，先找出频繁1项集L1，再利用L1找出频繁2项集，以此类推……由于考虑到运行时间的缘故，这个项目只运行了一次Aprior算法。

4.在应用算法的途中，本项目大量运用Data301课程中教授的知识。运用Spark， 一个开源集群运算框架作为项目的研究地点，运用弹性分布式数据储存（RDD）进行存储数据，以及运用Spark中的算子，如map/reduce对数据进行运算。在完成基本的程序框架搭载后，项目将会把程序以及数据上传到could platform进行多核运行。

动机：从疫情开始流行的一刻起，人们的生活方式受到了极大的影响与限制。由于我来自中国，所以研究这个问题对于我而言可以更多地了解不同时期的中国社会热衷讨论的话题。我个人从2019年来到新西兰留学，由于疫情缘故，到现在都还没回家一次，这促使我通过这个项目了解中国所面临的挑战与社会话题有哪一些。 由于我生活在被疫情限制生活的时代，能够亲身感受到疫情对生活的影响，所以这些数据与我息息相关。在DATA301的Lab3中，我被要求完成对用户购买数据的市场篮子分析，那一次lab让我了解了市场篮子分析的步骤，这种分析方法被运用到这个项目当中。 研究这个topic，通过我的reserch，我能向身边的人介绍中国社会所热衷的主题在这几年见的变化。这是有意义的，能让我身边的朋友或者留学生更了解中国社会的变化。

问题：疫情开始前，疫情刚刚开始与疫情持续一段时间后的中国社会的主题有什么变化？假设：中国社会主题会与税，个人收入与健康相关。这个项目采用的数据来自Gdelt项目的Global Knowledge Graph 2.0，项目将Graph 里的Location字段与V2Themes字段取出。Location字段提取出CountryCode之后将带有CH（China）的Country取出作为研究的集合，因此数据集的选择是与研究问题相关的。

项目将市场篮子分析运用在数据集上。市场篮子分析分析是对Basket的集合进行分析，而GKG Table的每一行是对文章信息的整合，其中V2Themes字段包含文章的主题（一个以上而且远远小于总篮子数），因此我将每一篇文章的主题作为一个Basket。把CountryCode代表中国的Basket整合并对他们进行购物篮分析。应用市场篮子分析的步骤是: 第一步项目运用Apriori算法找出频繁项集。项目拥有函数Frequent\_Itemsets(Country\_And\_Theme, countryCode)，输入数据集与需要查找的国家代码，输出包含两个主题频繁项集，只包含单个主题与主题出现次数的集合（individual）与数据集的长度。

第二步项目运用函数support(Frequent\_pairs , n)计算出数据集元素support值。confidence(individual\_pairs, Frequent\_Themes, count\_of\_total\_pairs)计算出数据集元素的confidence。interest\_rate(individual, confidence)计算出数据集元素的interest。

报告对研究问题的分析运用上面两步产出的support，confidence与interest。对support，condifence与interest进行排序，找出数据明显的主题进行分析。support是指A和B两个主题同时出现的概率，代表了这个组合的可靠程度。confidence是指主题A出现之后主题B出现的条件概率,是出现A和B主题的文章占所有包含主题A文章的占比。interest绝对值高且为正数表示两个主题有趣，拥有研究的价值。

########################################################

Experimental Design and Methods

* Discuss in detail the algorithm, data flow, program design, and the steps you have followed to accomplish the specific objectives of the project. Suggested length: two paragraphs.
* Describe any specific code or libraries you have created or used to implement the methods - a simple list of modules or function names and a short comment on each

步骤0：设置使用Spark的库以及设置dbg（）函数用于debug。

步骤1：首先防止下载栏目中有多余的文件，项目先删除下载栏目中的所有CSV文件。然后从Gdelt数据库中下载2019， 2020与2021年2月到5月的的Global Knowledge Graph 2.0（每一天的数据是一个CSV文件）。程序先构造出需要下载的日期列表，再对下载日期列表进行迭代尝试下载。如果下载成功则返回日期，下载失败则返回None作为标记，并将日期保存在Before\_Epidemics，Start\_Epidemics， Still\_Epidemics中。

步骤2： 提取成功下载的日期，为将下载好的数据读入RDD做准备。具体操作为去除Before\_Epidemics，Start\_Epidemics， Still\_Epidemics中的None得到所有成功下载的日期，将他们读入sqlContext中再转化为RDD储存。

步骤3：在存储着GKG table的第二步得到的RDD，提取出存有Location中的CountryCode字段和将V2Themes字段转化成List。这一步得到三个代表不同时期的RDD。

步骤4： 构造Frequent\_Itemsets(Country\_And\_Theme, countryCode)函数，实现运用Apriori算法寻找频繁项集。先设置50为Min Support，然后filter出CounrtyCode为CH（代表China）的行作为研究对象。然后运用flatMap算子找出包含单个主题的频繁项集并将其broadcast到每一个Working Node，这样可以加快程序的运行速度。之后通过迭代每一个basket，寻找pairs，其中pairs的两个元素都必须在频繁项集中。使用reduceByKey整合结果，得到两个元素都在频繁项集的pairs与出现次数的集合。

步骤5：运用运用用Apriori算法找到频繁项集后项目可以开始计算关联规则的support，confidence与interest。support衡量的是两个频繁的主题共同出现的概率，程序运营support函数计算support值。support(Frequent\_pairs , n)函数通过输入频繁出现的pairs与代表对应数据集大小的n，运用map算子计算每一个pairs出现的次数与总数据集大小的比例得到support。

步骤6：在关联规则 A->B 中， 置信度是指主题A时主题B出现的条件概率,是主题A与B同时出现的文章占所有包括主题A的文章里的占比，confidence函数用于计算confidence值。confidence(individual\_pairs, Frequent\_Themes,count\_of\_total\_pairs)函数通过输入individual\_pairs，这代表的是单个主题出现的次数。求出individual\_pairs与所计算数据集的大小的比例作为主题A出现的概率。Frequent\_Themes代表两个频繁主题同时的次数，可以理解为主题A与B同时出现的次数.项目将Frequent\_Themes的两个主题分别作为Key，构造出A->B与B->A的形式，他们都代表A与B同时出现的概率，只是不同形式。之后把A->B,B->A与单个主题出现的概率运用union方法组成一个大的RDD，再利用Map算子计算出每一组A->B与B->A的confidence。

步骤7：关联规则A->B的interest被定义为A->B的confidence与B在整个数据集中出现概率的差，这能反映该关联规则是否有趣。项目运用interest\_rate(individual, confidence, n)来计算interest。individual代表单个主题出现的次数，先算出单个主题出现的概率，再利用map算子计算所有关联规则的interest。具体方法是将包含单个主题出现概率的RDD与第六步求出的confidence数据集利用join函数合并。之后用关联规则的confidence减去所对应主题的出现频率得到interest。

步骤8：分析得到的数据结果

用到的函数有：

1. intofile(filename), get\_filename(x)这两个函数来自learn页面提供的Sample GDELT project。他们是用于下载文件的函数，我对intofile函数作出了一点修改，使他能够返回成功下载的日期以及下载Global Knowledge Graph 2.0。
2. find\_country(string)， to\_list(string)用于将已经读入RDD的并取出Locations字段与V2Themes字段的数据集进行加工。find\_country函数将CountryCode从Locations字段中提出，to\_list函数将V2Themes字段转化为列表。

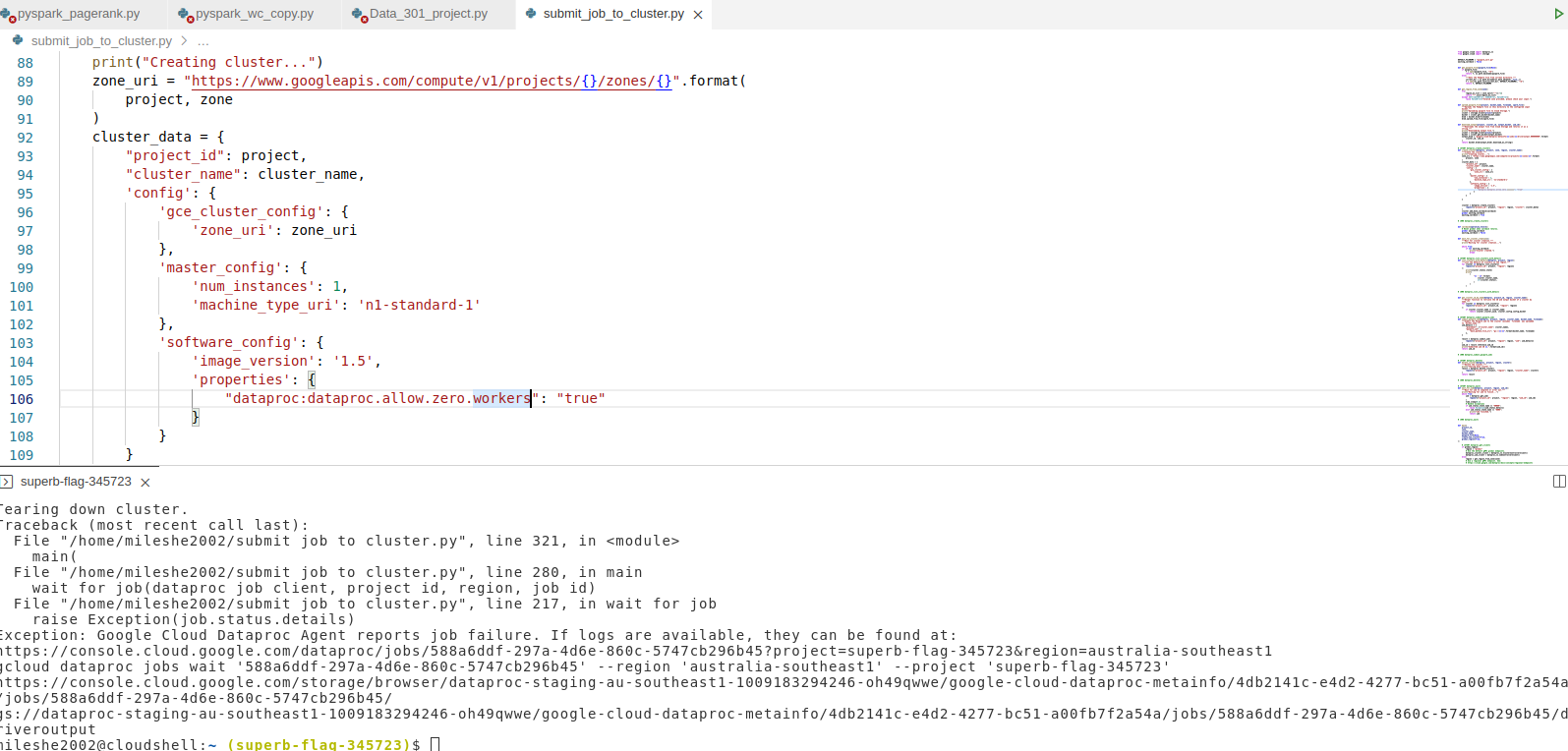
## Frequent\_Itemsets(Country\_And\_Theme, countryCode)函数输入需要寻找的国家名称与数据集，运行Apriori算法找出频繁项集并返回。

1. support(Frequent\_pairs , n)， confidence(individual\_pairs, Frequent\_Themes, count\_of\_total\_pairs)， interest\_rate(individual, confidence) 这三个函数分别求出所有关联规则的support， confidence与interest。

Results

* Supporting graphs or other figures
  + Note: at least two required graphs are of scalability. If you have access to google cloud these graphs should be of strong scalability (i.e. fixed problem size with an increasing number of processors) and weak scalability (problem size increases at the same rate as number of processors).   
      
    If you do not have access to google cloud and are using either your local computer or google colab, please generate two graphs:
    - 1: a “traditional” asymptotic algorithmic scalability graph -- problem size increases and the number of processors remains constant. You should test at least 4 problem sizes.
    - 2: A bar chart with two bars. Run your code once with the PySpark context using 1 process and once with \* processes. At the beginning of initialization, change the setting in local[?]:  
        
      conf = SparkConf().setMaster("local[1]").set("spark.executor.memory", "1g")  
        
      vs  
        
      conf = SparkConf().setMaster("local[\*]").set("spark.executor.memory", "1g")
* Answer and discuss the hypothesis or research question as best as you can with data. Suggested length: 1-2 paragraphs

sequential takes 20 min+ and goes error



2core 313.77718925476074s



4core 154.28740191459656s



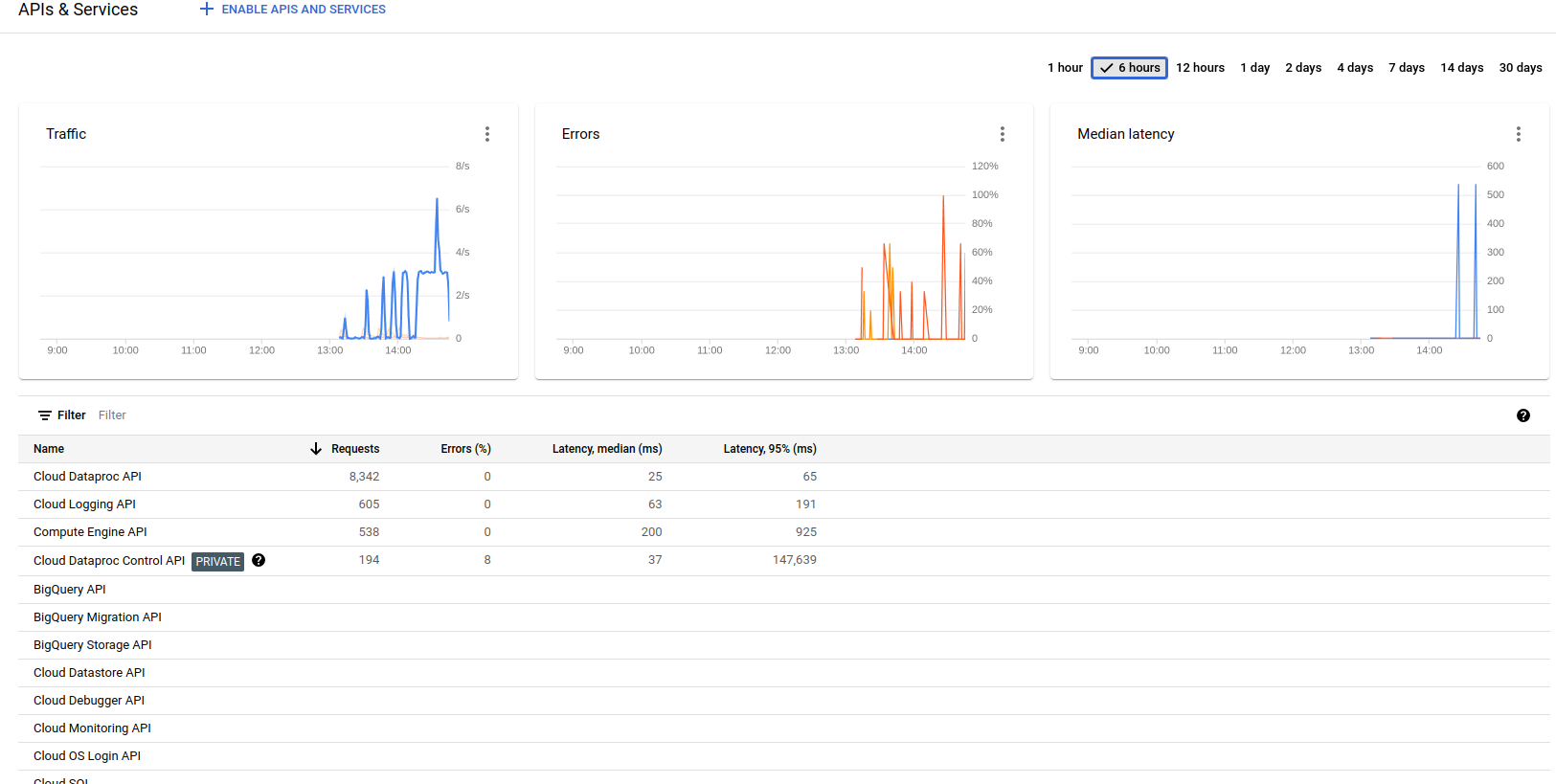
8 core 90.13733100891113s



16 core 60.23396849632263s



graph



result：

| individual\_Pairs | | |
| --- | --- | --- |
| Before | Start | Still |
| ('EPU\_ECONOMY\_HISTORIC', 1902) | ('WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION', 5018) | ('WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION', 1394) |
| ('TAX\_ETHNICITY\_CHINESE', 1634) | ('TAX\_DISEASE\_CORONAVIRUS', 4831) | ('GENERAL\_HEALTH', 1382) |
| ('TAX\_WORLDLANGUAGES\_CHINESE', 1634) | ('GENERAL\_HEALTH', 4789) | ('USPEC\_POLITICS\_GENERAL1', 1260) |
| ('USPEC\_POLITICS\_GENERAL1', 1554) | ('WB\_635\_PUBLIC\_HEALTH', 4383) | ('CRISISLEX\_CRISISLEXREC', 1243) |
| ('LEADER', 1494) | ('WB\_2165\_HEALTH\_EMERGENCIES', 4100) | ('TAX\_ETHNICITY\_CHINESE', 1224) |

这是三个不同的时期出现次数top5的单个主题。由于数据集合的偏差2019年的数据集大小是4025，2020年的数据集大小是7008,2021年的数据集大小是2984，因此第二年的主题看起来会比较多。从2019年的数据来看，出现最多的主题与经济，税收，国际政治较为相关。2020年与2021最大的主题也许就是新型冠状病毒与公共卫生事件了。

| support | | |
| --- | --- | --- |
| Before | Start | Still |
| (('TAX\_ETHNICITY\_CHINESE', 'TAX\_WORLDLANGUAGES\_CHINESE'), 0.4057142857142857) | (('WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION', 'WB\_635\_PUBLIC\_HEALTH'), 0.6254280821917808) | (('GENERAL\_HEALTH', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 0.4195710455764075) |
| (('LEADER', 'USPEC\_POLITICS\_GENERAL1'), 0.311055900621118) | (('GENERAL\_HEALTH', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 0.6242865296803652) | (('TAX\_ETHNICITY\_CHINESE', 'TAX\_WORLDLANGUAGES\_CHINESE'), 0.4101876675603217) |
| (('EPU\_POLICY\_GOVERNMENT', 'GENERAL\_GOVERNMENT'), 0.3003726708074534) | (('TAX\_DISEASE\_CORONAVIRUS', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 0.6090182648401826) | (('UNGP\_HEALTHCARE', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 0.3930965147453083) |
| (('LEADER', 'TAX\_FNCACT\_PRESIDENT'), 0.2732919254658385) | (('WB\_2165\_HEALTH\_EMERGENCIES', 'WB\_635\_PUBLIC\_HEALTH'), 0.5850456621004566) | (('GENERAL\_HEALTH', 'UNGP\_HEALTHCARE'), 0.3900804289544236) |
| (('EPU\_ECONOMY\_HISTORIC', 'USPEC\_POLICY1'), 0.26881987577639754) | (('WB\_2165\_HEALTH\_EMERGENCIES', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 0.5850456621004566) | (('WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION', 'WB\_635\_PUBLIC\_HEALTH'), 0.36896782841823056) |

这是三个不同的时期出现次数top5的成对的主题。在疫情开始前，也就是2019年，中国社会的文章经常讨论税收，国际政治，美国等话题，两个话题间的相似程度较大，同为税收或者同为政府。2020年与2021年成对出现的话题在疫情，营养，人口间组合，与少量税务问题，一定程度上说明新型冠状病毒对中国社会产生了影响。

| confidence | | |
| --- | --- | --- |
| Before | Start | Still |
| (('ECON\_FREETRADE', 'WB\_698\_TRADE'), 1.0) | (('EPU\_POLICY\_CENTRAL\_BANK', 'USPEC\_POLICY1'), 1.0) | (('WB\_1427\_NON\_COMMUNICABLE\_DISEASE\_AND\_INJURY', 'WB\_621\_HEALTH\_NUTRITION\_AND\_POPULATION'), 1.0) |
| (('TAX\_FNCACT\_CHIEF\_FINANCIAL\_OFFICER', 'TAX\_FNCACT\_OFFICER'), 1.0) | (('EPU\_POLICY\_CENTRAL\_BANK', 'WB\_1920\_FINANCIAL\_SECTOR\_DEVELOPMENT'), 1.0) | (('WB\_1427\_NON\_COMMUNICABLE\_DISEASE\_AND\_INJURY', 'WB\_1406\_DISEASES'), 1.0) |
| (('TAX\_FNCACT\_CHIEF\_FINANCIAL\_OFFICER', 'TAX\_MILITARY\_TITLE\_OFFICER'), 1.0) | (('EPU\_POLICY\_CENTRAL\_BANK', 'EPU\_CATS\_MONETARY\_POLICY'), 1.0) | (('TAX\_FNCACT\_VICE\_PRESIDENT', 'USPEC\_POLITICS\_GENERAL1'), 1.0) |
| (('TAX\_FNCACT\_CHIEF\_FINANCIAL\_OFFICER', 'TAX\_FNCACT\_CHIEF'), 1.0) | (('WB\_2934\_COPPER', 'WB\_507\_ENERGY\_AND\_EXTRACTIVES'), 1.0) | (('TAX\_FNCACT\_VICE\_PRESIDENT', 'TAX\_FNCACT\_PRESIDENT'), 1.0) |
| (('EPU\_CATS\_TAXES', 'USPEC\_POLICY1'), 1.0) | (('WB\_2934\_COPPER', 'WB\_895\_MINING\_SYSTEMS'), 1.0) | (('TAX\_FNCACT\_VICE\_PRESIDENT', 'LEADER'), 1.0) |

这是三个不同的时期出现次数top5的成对主题的confidence。观察表格发现confidence都是1.0，说明每当pairs中第一个主题出现第二个主题也会出现。2019年主题中，每当TAX\_FNCACT\_CHIEF\_FINANCIAL\_OFFICER出现'TAX\_FNCACT\_OFFICER， TAX\_MILITARY\_TITLE\_OFFICER， 'TAX\_FNCACT\_CHIEF都会出现。2020年中当EPU\_POLICY\_CENTRAL\_BANK出现'WB\_1920\_FINANCIAL\_SECTOR\_DEVELOPMENT， 'EPU\_CATS\_MONETARY\_POLICY与WB\_507\_ENERGY\_AND\_EXTRACTIVE都会出现。2021年每当非传染性危害被提及，人口与疾病主题都会被提及。同时当'TAX\_FNCACT\_VICE\_PRESIDENT被提及，USPEC\_POLITICS\_GENERAL1， 'TAX\_FNCACT\_PRESIDENT和LEADER也会被提及。

| interest | | |
| --- | --- | --- |
| Before | Start | Still |
| (('WB\_707\_POVERTY\_MEASUREMENT\_AND\_ANALYSIS', 'WB\_1152\_ECONOMIC\_AND\_SOCIAL\_MOBILITY'), 0.9873291925465838) | (('TAX\_MILITARY\_TITLE\_SOLDIERS', 'TAX\_FNCACT\_SOLDIERS'), 0.992579908675799) | (('TAX\_WORLDLANGUAGES\_ENGLISH', 'TAX\_ETHNICITY\_ENGLISH'), 0.9829088471849866) |
| (('WB\_1152\_ECONOMIC\_AND\_SOCIAL\_MOBILITY', 'WB\_707\_POVERTY\_MEASUREMENT\_AND\_ANALYSIS'), 0.9873291925465838) | (('TAX\_MILITARY\_TITLE\_COMMANDER', 'TAX\_FNCACT\_COMMANDER'), 0.9924372146118722) | (('TAX\_ETHNICITY\_ENGLISH', 'TAX\_WORLDLANGUAGES\_ENGLISH'), 0.9829088471849866) |
| (('TAX\_FNCACT\_CAPTAIN', 'TAX\_MILITARY\_TITLE\_CAPTAIN'), 0.986832298136646) | (('TAX\_WORLDLANGUAGES\_GERMAN', 'TAX\_ETHNICITY\_GERMAN'), 0.9921518264840182) | (('WB\_2601\_TRADE\_LINKAGES\_SPILLOVERS\_AND\_CONNECTIVITY', 'WB\_772\_TRADE\_FACILITATION\_AND\_LOGISTICS'), 0.9819034852546917) |
| (('TAX\_MILITARY\_TITLE\_CAPTAIN', 'TAX\_FNCACT\_CAPTAIN'), 0.986832298136646) | (('TAX\_ETHNICITY\_GERMAN', 'TAX\_WORLDLANGUAGES\_GERMAN'), 0.9921518264840182) | (('TAX\_ETHNICITY\_KOREAN', 'TAX\_WORLDLANGUAGES\_KOREAN'), 0.9815683646112601) |
| (('TAX\_RELIGION\_MUSLIM', 'TAX\_ETHNICITY\_MUSLIM'), 0.986583850931677) | (('WB\_2120\_SATELLITES', 'WB\_2329\_ACCESS\_AND\_CONNECTIVITY'), 0.9920091324200914) | (('TAX\_WORLDLANGUAGES\_KOREAN', 'TAX\_ETHNICITY\_KOREAN'), 0.9815683646112601) |

这是三个不同的时期出现次数top5的成对主题的interest。表格中的interest普遍大于0.98小于1，这说明这些主题都是有趣的。比如2019年值得关注的有财产管理与经济这两个主题。以上三年有趣的话题普遍是税务话题， 而没有出现于疫情相关的关键词。

Conclusion (suggest 3 paragraphs total, one for each prompt)

* Were you able to answer your hypothesis / research questions? Explain how and why (or why not).
* What implications do your results have?
* What future questions or directions would you take with your project?

项目认为目前研究得出数据能够证明在2020年疫情开始后，中国社会最关心的主题由经济与政治主题转移到了健康与疫情主题。这足以说明疫情出现对中国社会的影响。从单个主题的出现次数来看，2019年的出现次数top5的主题以经济与政治为主。到了2020年与2021年，出现次数最多的5个话题都与疫情，健康，人口相关。观察support值top5的话题，发现税收，国际政治，美国这三个主题的组合常常两两出现在2019年。在2020年与2021年成对出现的话题在疫情，营养，人口间组合。尤其是2020年的主题，top5主题组合的support值都大于0.5，这证明人们经常将疫情，营养与人口放在一起讨论。

通过观察中国社会热衷的主题，这份报告能够推断目前疫情依然对中国社会产生了不小的影响。这个结果的影响是：中国社会依旧面临疫情的影响，这给其他国家的提醒是重视新型冠状病毒，尽快研究出对抗病毒的方法。

这个项目还对confidence与interest进行了计算，通过观察interest结果发现，不同税务主题间常常有联系。在2020年与2021年分别出现了德国与韩国的关键词，这也是值得研究的方向。confidence表示一个主题的出现往往会引出另一个主题，比如EPU政策会引出经济问题与税务问题的相互蕴含。这使得未来的项目可以对这些高confidence的主题进行相关性分析，进一步挖掘他们间的关系。

Critique of Design and Project

* Choose one part of your design or methods (section 4 of your proposal, the algorithm, data flow, and software) that you feel could have worked better with a different approach. Describe what did not work with your design or method. Why did it not work? What changes could you have made that might have been better?  
  Suggested length: two paragraphs.

Reflection

* List course concepts and tools you found useful for completing the project
* What did you learn from the project? Suggested length: one paragraph

1.市场篮子分析。项目运用市场篮子分析作为分析主题的思路，将市场篮子分析方法贯通整个项目。

2.关联规则。关联规则是一种在大型数据库中发现变量之间的有趣性关系的方法。

3.Apriori算法： Apriori 算法其实就是查找frequent itemset 的过程。频繁项集是支持度大于等于Min Support阈值的项集，小于最小值支持度的itemset 被称为非频繁项集。它是一种逐层迭代的方法，先找出频繁1项集L1，再利用L1找出频繁2项集，以此类推……由于考虑到运行时间的缘故，这个项目只运行了一次Aprior算法。

4.在应用算法的途中，本项目大量运用Data301课程中教授的知识。运用Spark， 一个开源集群运算框架作为项目的研究地点，运用弹性分布式数据储存（RDD）进行存储数据，以及运用Spark中的算子，如map/reduce对数据进行运算。在完成基本的程序框架搭载后，项目将会把程序以及数据上传到could platform进行多核运行。

References

* Provide any citations and/or links to notebooks, datasets, etc
* Provide a list of student names that you have worked directly with on the project.
* You do not need to list code you have gotten from the class project forum on learn whether it was posted by an instructor or another student.