2020秋 可视化与可视计算概论 期末项目报告

主题：中美关系可视化

组员：杨昊翔（1800012969）、吴宇恒（1800012947）、刘云蛰（1800013123）、易普（1800013116）

（一）数据描述

本部分相比于演示PPT，补充了数据采集和处理的详细细节。

可视化国际关系有诸多方法：可以直接使用编年史，可以清晰地看到国家关系的关键节点，但存在数据量少、特征多为文字的缺点；可以使用国家间的贸易数据，其数量可观，容易统计分析，但局限于经济一个角度。因此，我们提出用报纸数据进行中美关系的可视化。相比于编年史或经济数据，报纸数据具有丰富、历史久远的特点，并且可以从侧面反映中美的关系。我们采集的数据如下：

我们从一共 9 份报纸的网站上爬取了数据，这些报纸包括：$new\ york\ times$，$china\ daily$、$christan\ science\ monitor$、$global\ times$、$los\ angeles\ times$、$scientific\ america$、$the\ sun$、$wall\ street\ journal$、$washing\ ton\ post$。为了尽量**避免数据集中存在的偏见**，我们的数据中有 $china\ daily$ 和 $global\ times$ 两家在中国发行的报纸，也有 $scientific\ america$ 这样科学类的杂志。

我们采集数据的方法是：在每个网站的资料库（$archive$）中搜索 "$China\ America$" 关键词，爬取所有的和 $China$ 或 $America$ 有关的文章。对于部分会将 $China$ 或 $America$ 无关信息也显示出来的网站，我们手工对爬虫数据进行二次筛选，强制要求每一条数据中出现 $China$ 或 $America$。在数据处理的过程中，U.S. 被视为 $America$ 的同义词。

我们所用的数据一共约 300000 条，每条均包括（来源报纸 $source$，标题 $headline$，摘要 $abstract$，时间 $date$，链接 $href$）。我们的数据中没有正文，因为这些报纸的正文大多是收费的。值得注意的是，我们获得的数据集中**没有标签**，比如每条新闻的政治、经济、文化的分类是缺失的。由于大部分报纸的数据集中于近 20 年（部分报纸早期的数据也只有图片），为了获得 1900 年左右的较早期数据，我们电邮了美国杂志 $New\ York\ Times$。所用的数据放在代码同一目录下。

（二）项目设计

本部分相比于演示PPT，补充了每张图的数据筛选的详细方法。

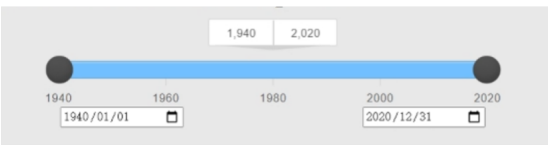
我们遵守**先概览，再细节筛选**的可视化原则。

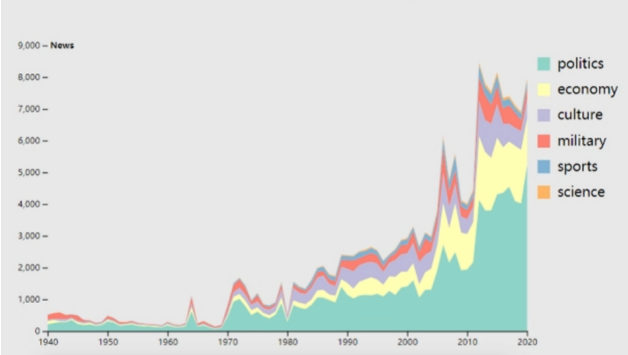
（1）时间滚轮和面积堆叠图

首先，如果读者想对中美关系的整体发展有一个把握，那么他可以使用**时间滚轮**和**堆叠面积图**模块。

时间滚轮可以让读者选择某一起始和终止年份，也可以在两个**日历模块**上具体的到日期的起始和终止时间点。

面积堆叠图对时间滚轮选中的时间内的所有数据条（每一条新闻）进行政治、经济、文化、军事、科技、体育的分类。





分类的规则如下：我们收集了六个类别的字典，字典里包含这个类别里常见的词，比如政治类别等的常用词等。如果某条新闻中出现了政治相关的词，我们便统计其出现的词频。这样，我们可以统计六个类别的词汇在每条新闻中出现的词频。我们在六种类别中选取词频最高的类别，认为其是这条新闻的类别。

该分类的必要性和合理性：限于人力和时间等资源，我们不可能对数十万条数据进行逐一手工标记，所以我们无法取得标签，无法进行有监督学习；经过手工二次确认，我们收集的高频词确实能较好地筛选每条新闻的类别。我们直接忽略那些没有出现任何关键词的新闻，不将其统计进堆叠面积图。

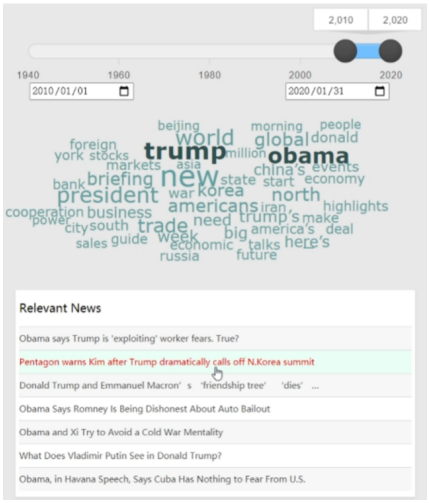
（2）大事件模块与词云模块

在**大事件模块**，我们综合网络上关于 中美关系事件的文字报道，将其整理为时间线（Timeline）。

在**词云模块**部分，我们使用**TF-IDF**来统计词频，筛选出较为关键的词。经过实际测试，TF-IDF比简单词频统计的效果好不少，但是其仍然产生了大量无关词语。因此，我们反复更新停用词，将无意义的词语筛选出去，最终得到一个比较好的效果。

（3）详细信息模块

进一步，倘若读者想了解这段时间发生的具体事件，甚至进入每一条新闻，就可以点击详细信息模块。点选词云中的某些关键词，详细信息模块就会显示其对应的新闻。这一部分我们将所有新闻按照其中出现的点选关键词的总数量进行排序。我们也可以很容易地将其修改为根据词频加权比重来排序，但我们认为将每个词语等同看待更符合读者搜索的意图。例如，如果我们把时间选定为2010年-2020年，搜搜关键词“trump”“obama”，就能看到“奥巴马和习近平避免冷战”一条新闻出现在第四个。



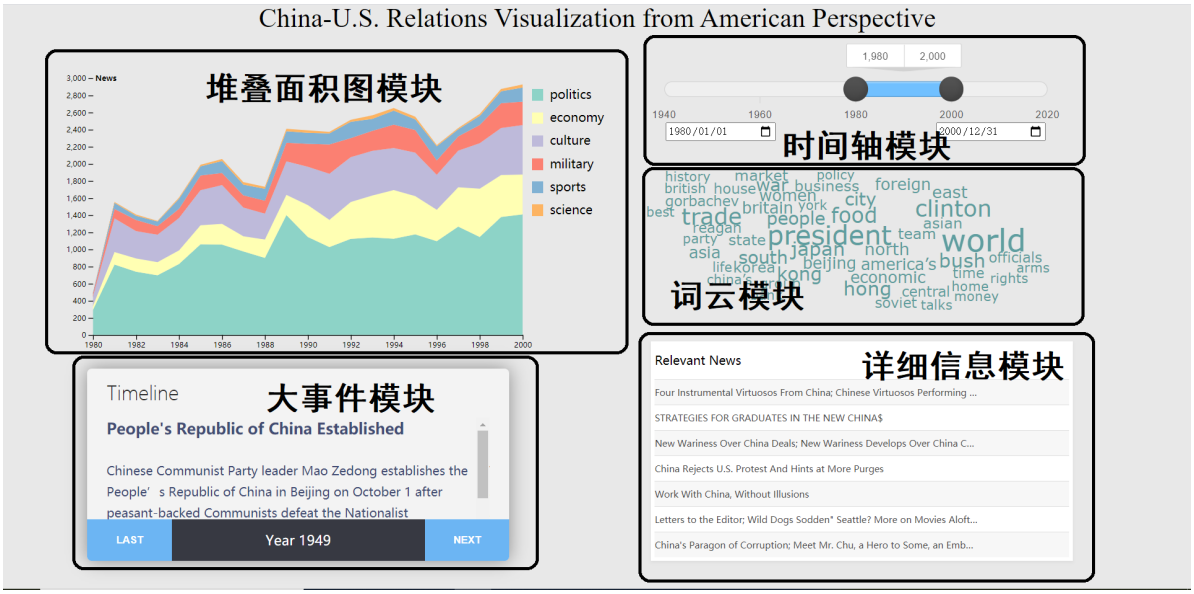
（4）问题难点与数据交互系统

本次实验过程中，数据量是一大难点。我们大约有30W条数据，每条需要处理的headline和abstract一般都大于10个词。如果对每个句子进行词语token化，那么总数据量大约在300W级别。每次**时间改变**，图中多个模块的结果都要重新计算；每次**拖动时间条**，都会触发多次事件时间改变（一般为3~5次左右）。这些事实导致了几个问题：我们不可能每次都从硬盘读取文件，而必须在预处理好后从内存读取文件；如果不进行更多处理，每次拖动所要处理的数据量高达3000W，若不采样，即便算法的时间复杂度为线性，也很难进行流畅的反应。

我们的解决方法是多层采样。首先，我们对全部数据进行一次1:10采样。这次采样丢失了很多信息，也确实是我们算法的一大问题。比如，起初我们能在词云数据中看到一些中国历史上的变革，但在筛选了1:10数据后，部分关键词从词云中消失了。但经过我们确认，较为重要的事件仍得到保留。我们认为这是因为关键事件的报道频率很高，不太可能被完全抹除。同时，我们也推荐读者在条件允许的情况下，用完整的数据进行可视化结果的深入研究。

然后，考虑到不同模块计算需要的事件并不完全相同，我们对不同的模块，采用不同的采样频率进行二次采样。注意，多层采样时，应该保证几个模块的采样频率的最小公倍数频率尽量小，最好是呈倍数关系，这样可以尽量避免信息缺失导致的问题，比如选择词云的某个词后，找不到对应的新闻。在调试的过程中，我们保证第一次数据读取进内存后不再从硬盘读取数据，并对每一步进行细致的时间分析，尽量确保系统的流畅性。

（三）最终系统



更具体的系统演示，烦请见演示视频和课堂展示PPT。

（四）发现

本部分与演示PPT相同。

我们在可视化结果中有以下发现：

（1）近百年的中美关系呈总体上升趋势，期间略有波澜。在有关中美关系的新闻讨论中，经济、文化的比重增加，逐渐取代政治，表现出两国交流的日切深入。

（2）中美关系的几个重要节点是：1971 年左右的乒乓外交事件，1979 年左右的中美建交、关系好转，以及 2012 年后逐步掀起的贸易竞争关系。期间新闻规模爆炸式增长。

除此之外，我们也发现了不少趣事，比如。

（1）苏联解体前，美国报纸常在提及中国时亦苏联，可视化词云结果中常出现“soviet”“communist”字样。

（2）各大报纸在 1964 年有一讨论高峰。经检查，此为 1964 年我国原子弹爆炸成功，报纸中大幅增长的为军事新闻，这说明了原子弹爆炸成功带来的令人惊讶的影响力。

（3）越战前后，中国常常与越南“vietnam”一起被提及。

（4）几乎在每个时间段，提及中国时，都伴随着美国总统的名字，美国总统为词云的榜上常客。这一点从某种意义上反映了总统在美国的特殊舆论地位。

（五）讨论

在本次实践过程中，我们体验了收集数据、处理分析、利用可视化知识挖掘数据特征、做报告、展示海报的全过程，通过一步步的探索，把一个最初的想法变成一个切实的、可分享的成果。

我们的问题在于：

（1）展现的可视化结果在美学设计上有进步空间。

（2）交互的流畅感、过渡的平滑性还有待提升。

（3）我们希望能点击一个词时，将左上角的面积堆叠图，更换为对于这个词在选中年份之间的出现频率折线图，这样可以展现出一个词在时间段内的变化情况。但很遗憾由于时间关系，这部分的内容没能完成。

（六）分工

姓名：杨昊翔　学号：1800012969

提出使用报纸数据用于可视化的idea；写9个网站的爬虫；完成2个csv文件的处理；寻找**面积堆叠图**、**大事件模块**、**时间滚轮**、**词云**四个模块前端的可用模板；**详细信息模块**的后端部分；写可视化结果的发现；做ppt，剪视频，写报告，做poster，更新Wiki。

姓名：吴宇恒　学号：1800012947

完成2个csv文件的处理；整合组员收集的报纸数据，观察总体特征，去除无效的新闻后输出为统一的格式；完成**词云模块**的后端部分，设计后端**TF-IDF**等模块。

姓名：刘云蛰　学号：1800013123

前端的整体排版布局设计；交互响应模块设计；**时间滚轮**和**日期选择器**的整理调试；**大事件模块**的整理调试；**详细信息模块**的前端部分搜集、整理与调试；完成2个csv文件的处理。

姓名：易普　　学号：1800013116

词云模块的前端整理调试，及后端代码的对接工作；**堆叠面积图**的前端模块整理调试；完成2个csv文件的处理；向New York Times发送邮件取得了格式最好、含有早年完整信息的一批数据。

注：“csv文件的处理”指从爬取的原始html数据中提取结构化的信息。