浅析慕课教学视频中的学习行为规律
——以《计算机文化基础》课程为例

经56 王思萍 2015012527

1 研究方案设计

1.1 研究主题

本研究以计算机文化基础课程数据中的视频播放情况数据为基础，聚焦单个方面的数据并进行深入挖掘，探寻从慕课教学视频的数据中反映出的学生学习行为规律，并基于对学生观看视频的基本情况与学习规律的分析，为未来该课程教学视频的改进提出建议。

1.2 可视化算法

本研究用到的可视化图像共有三类。首先是直方图(非条形图)，用于描述各学期观看视频的基本情况；其次是折线图，用于描述单个视频每千分比时刻的点击量，同时也用于描述各学期所有视频平均每千分比时刻的点击情况；最后是词云图，使用 `PCA` 算法与中文预训练集，将视频关键词绘制在二维坐标系的同时使得表意相近的词语距离亦相近。具体算法设计如下：

1.2.1 直方图

- 所用数据读入到字符串中并以换行符进行分割
- 使用**正则表达式**找出学堂号的起始位置并把用户名的部分去掉
- 通过 `set` 与 `list` 的相关操作统计出每学期观看视频的总人数
- 使用 `pandas` 库中的相应方法创建 index 为视频编号的 dataframe 并按点赞次数排序，并计算出每个视频的人均点击次数
- 根据上述信息绘制直方图。其横坐标为人均点击次数，纵坐标满足：

  $$\text{纵坐标} \times \text{区间值} = \text{人均点击次数在对应区间的视频数量}$$

1.2.2 慕课曲线

- 所有数据读入到字符串中并以换行符进行分割
- 使用**正则表达式**找出学堂号的起始位置并把用户名的部分去掉
- 获取视频名称、时长、每次点击的开始与结束时间，并创建字典

- 用函数获取每个视频每千分点时刻的点击次数
- 根据字典内容绘制图片。横坐标代表每千分时刻的千分比刻度，纵坐标代表该刻度时在观看(即开始和结束时间包含了这个时刻)的人次

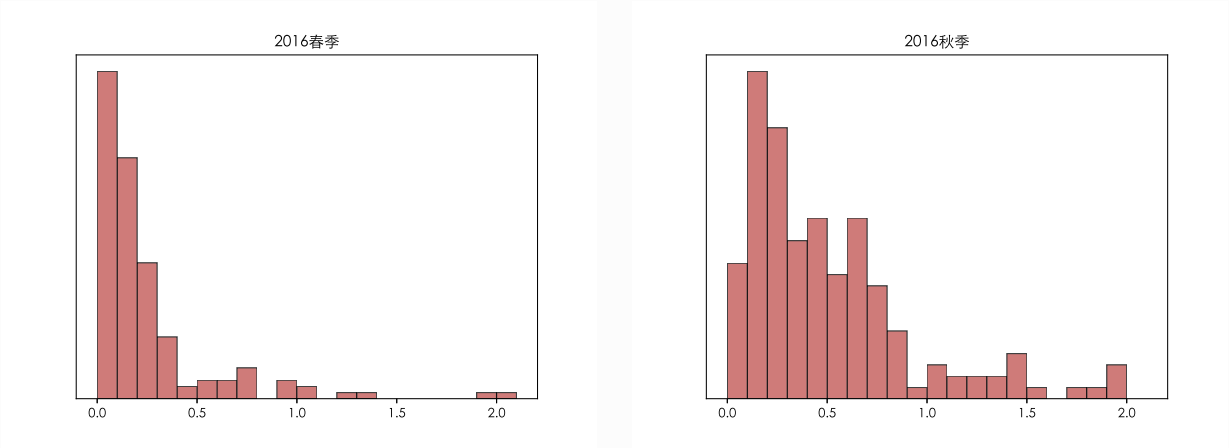
1.2.3 PCA词云图

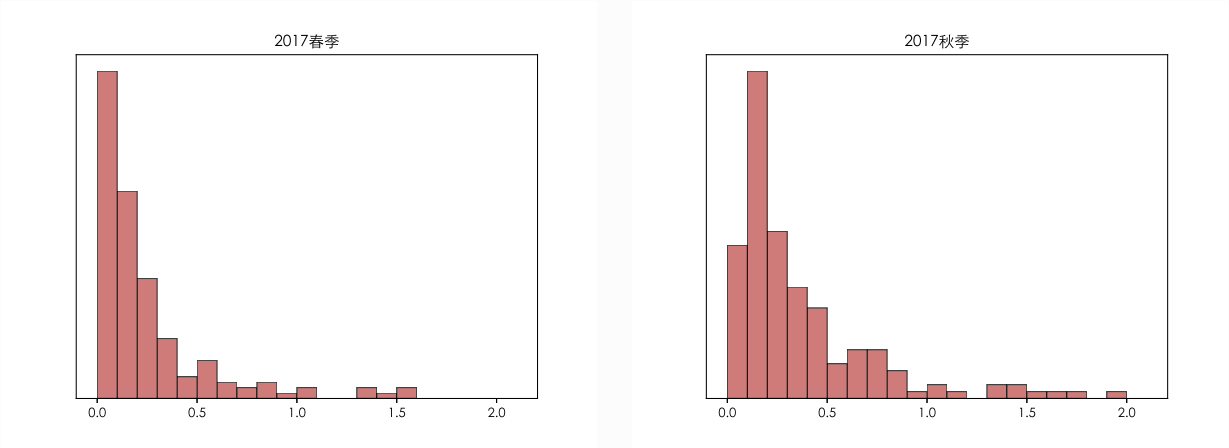
- 对于各个学期的视频名称，获取词向量，并通过调用 `ChineseWordVector` 中的词向量数据生成各个学期视频名称的词向量文件
- 使用 `PCA` 算法将 300 维度的词向量降至 2 维，从而可以绘图
- 按照 `1.2.1` 和 `1.2.2` 中处理与获取数据文件的方法对原始数据进行处理和提取，整合以上函数，绘制PCA词云图
- 绘图时，对于不同点击量区间的关键词，赋以**不同字体大小**(随点击量增大而增大)、**不同透明度**(随点击量增大而减小)。

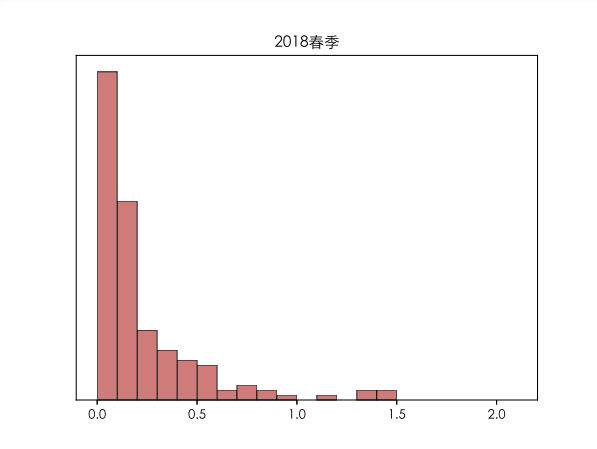
2 描述性统计

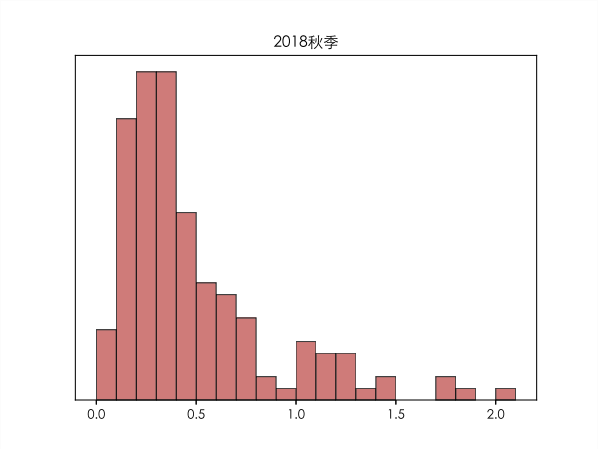
2.1 不同学期视频的人均点击量统计

基于 `1.2.1` 中的实现过程，绘制六张直方图如下所示:









本小节使用了2016-2018年春、秋季学期 `学生行为` 中的 `video_study_export` 数据，共计六个数据文件。本小节绘制的图片类型为直方图。其横坐标为视频人均点击量，即：$\text{点击量}/\text{该学期观看过视频的总人数}$，并在 0.0 - 2.0 之间按照 0.1 的区间值划分成 20 个区间。纵坐标满足：$\text{纵坐标} \times \text{区间值}(0.1) = \text{人均点击次数在对应区间的视频数量}$

从以上六张图片中可以发现，2016-2018年春季学期的整体视频观看情况不容乐观。绝大部分视频的人均观看次数都在 (0.0, 0.1) 这个区间中。而对于秋季学期，整体视频观看情况稍好于春季，但绝大部分视频的人均观看次数仍然在 (0.0, 0.5) 区间中。对于所有学期，人均观看次数大于1的视频数量都非常少。这应该是由于不同学期的课程模式不同导致的。对于春季学期，该课程为自主模式，社会学生为主，助教简单运营，缺少强制性，因此学生坚持学习该课程的意愿不够强烈；而相比秋季学期，则是混合模式，线下课程学生为主，助教投入较大，课程具有较高的强制性，因此学生观看视频的意愿也更强一些。可以观察到右侧三张图表中人均播放量超过1次的视频 (尤其是2016年与2018年) 还是占有一定比例的，而左侧三张图表 (即春季学期) 人均播放量超过1次的视频则非常少。

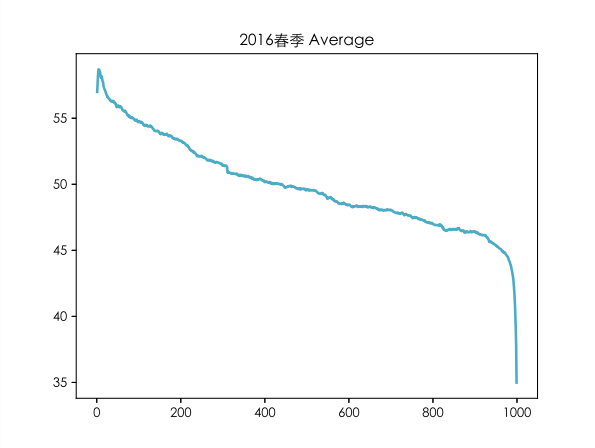
按照年份对比，发现从2016年《计算机文化基础》课程在慕课平台上线以来至2018年的三年中，春季、秋季学期分别的变化并不大，但三年的春季学期视频的人均播放量是呈缓慢下降趋势的，可以观察到到2017、2018两年的春季，人均播放量小于1次的视频占比高于2016年春季。而对于秋季学期，也可以发现2017、2018两年的视频播放情况较2016年有下降。这可能是2016年课程上线时的宣传产生的正向作用导致的。

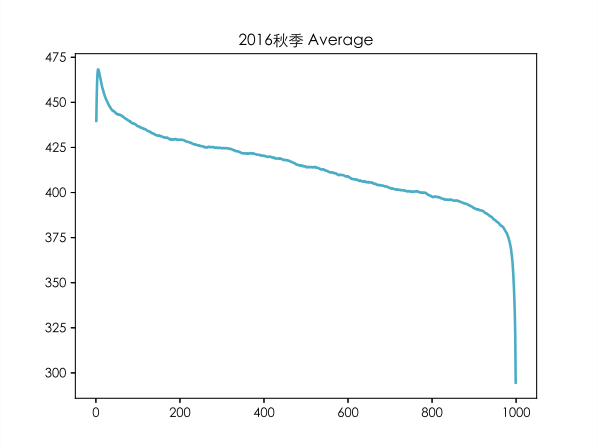
2.2 各视频每千分比时刻观看人数统计曲线

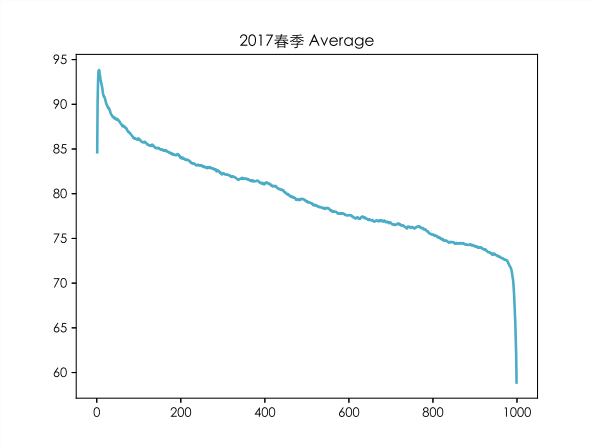
本小节使用了2016-2018年春、秋季学期 `学生行为` 中的 `video_study_export` 数据，共计六个数据文件。本小节绘制的图片类型为曲线图。横坐标代表每千分时刻的千分比刻度，纵坐标代表该刻度时在观看(即开始和结束时间包含了这个时刻)的人次。

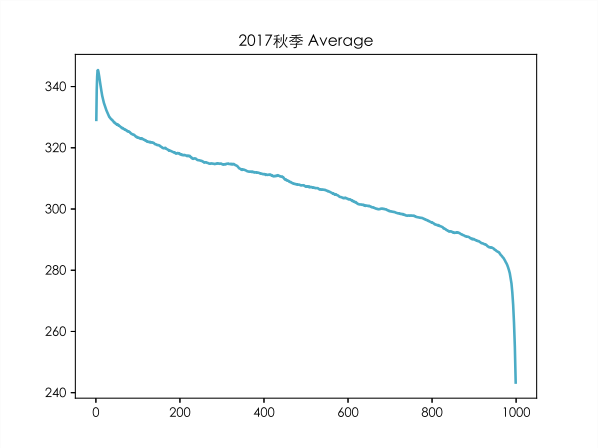
2.2.1 各学期平均"慕课曲线"

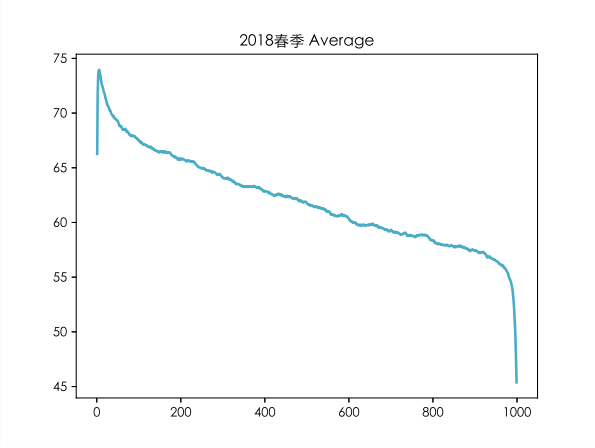
基于 `1.2.2` 中的实现过程，绘制六张曲线图如下所示：

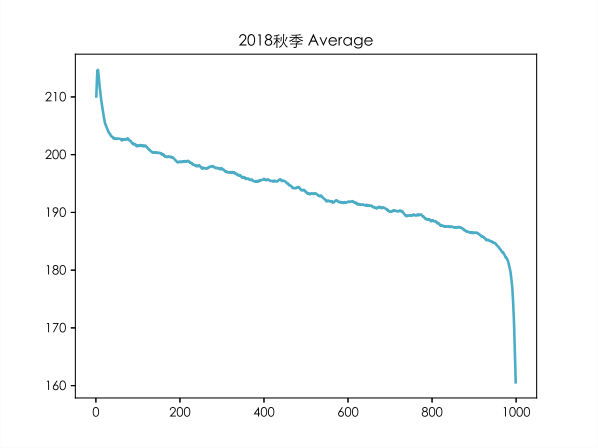








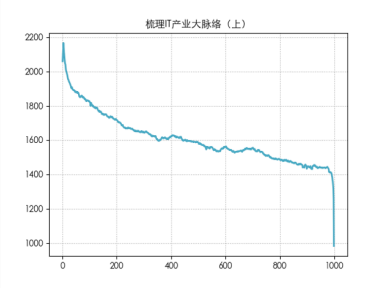




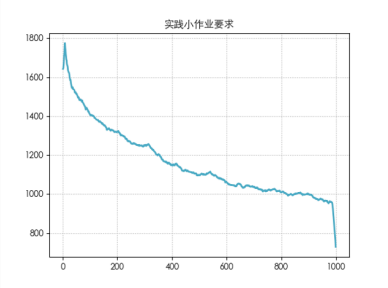
可以观察到，六个学期的课程在视频的每千分比时刻平均观看次数上变化趋势相似，均为视频的开始阶段观看人数较多，随着视频播放至结束，观看次数依次下降，在最后一小段突然出现急剧的下降。同时，这一下降趋势的强弱 (即曲线斜率) 随着视频的播放而减小。可以推测出，大部分学生都会在视频播放的中间选择放弃观看剩余部分的视频，而放弃观看的趋势会随着视频播放至结束先减小后增大。之所以在视频的最后放弃比例增大，推测是同学们会倾向于跳过片尾或总结部分。

2.2.2 单独视频的"慕课曲线"

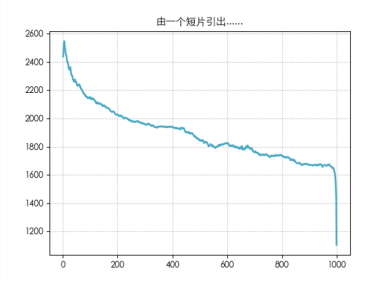
对于每个播放量超过1000次的视频，也绘制了代表该视频的"慕课曲线"。对于大部分这样播放量较多的视频，其"慕课曲线"的形状与 `2.2.1` 中的形状相似。下图简单列举了三个具体视频的例子，分别是《梳理IT产业脉络 (上) 》《实践小作业要求》和《由一个短篇引出……》，它们的"慕课曲线"均为开始时观看人数较多，随着视频播放至结束观看人数逐渐下降，并在结尾阶段迅速递减到0。



例子1



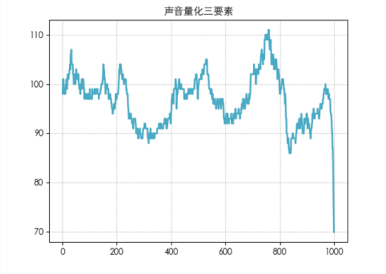
例子2

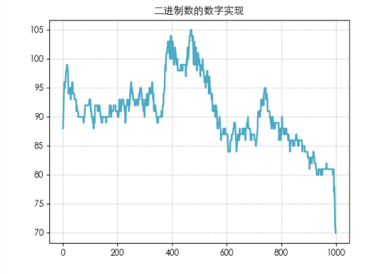


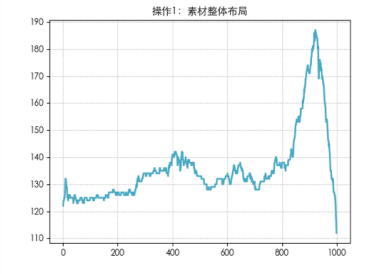
例子3

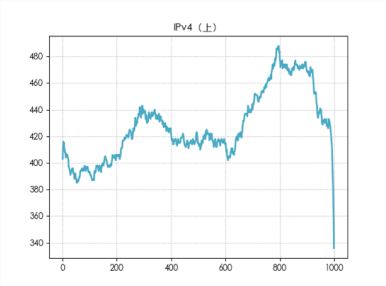
3 特殊曲线分析

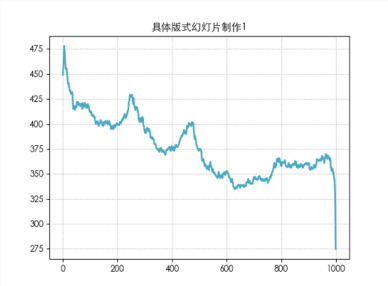
然而，也有少量播放量较多的视频，它们的曲线与"慕课曲线"的形状不符。下图列举了六张特殊曲线图，第一行的三张来自2016年春季的视频，分别是《声音量化三要素》《二进制数的数字实现》和《操作1：素材整体布局》，第二行的三张来自2016年秋季的视频，分别是《IPV4（上）》《具体版式幻灯片制作1》《像素、分辨率和色深》。

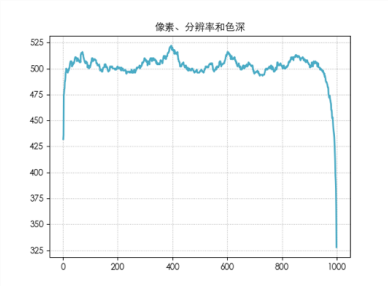












猜想这些具有特殊曲线的形成可能和同学们在习题中遇到了难度较高的题目从而回看视频有关，同时也可能与视频内容本身吸引人，例如《具体版式幻灯片制作1》中的曲线有两个"峰"，说明这两个时间点视频阐述的内容是较为重要的。

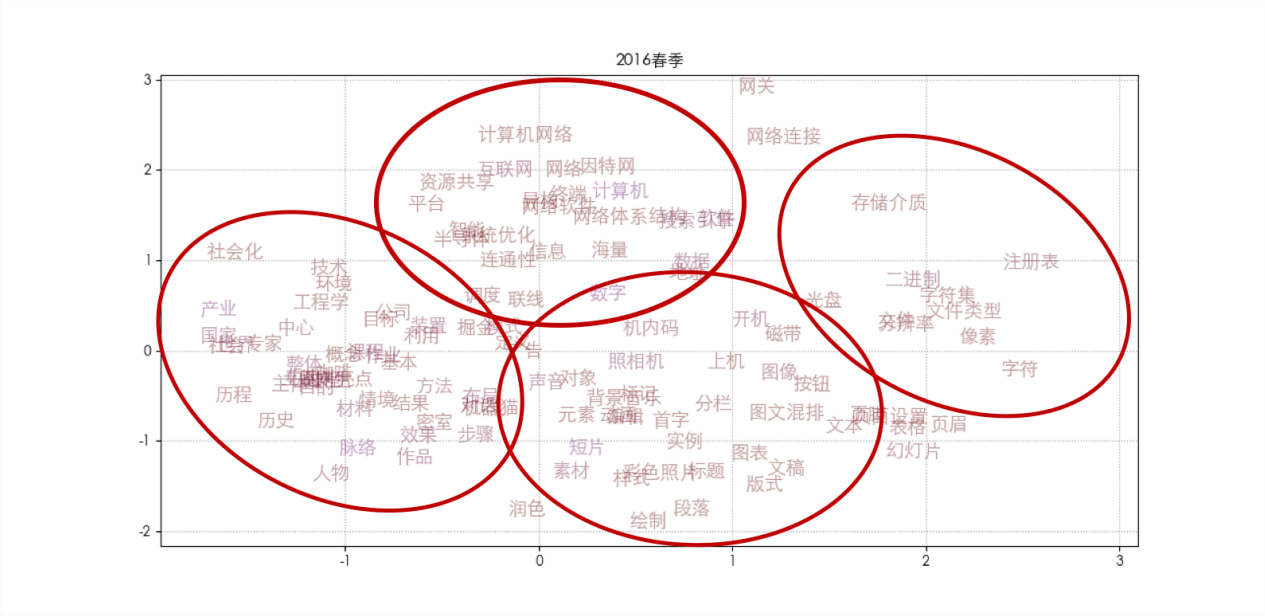
针对第一个猜测，我们统计了2016年春、秋季学期课后习题的争取率并进行了排序，发现2016春季中有一道题："下列声音文件格式中，_不是波形声音文件格式"，是一道与声音三要素有关的题，这道题的正确率只有 0.241379；同时还有一道与进制有关的题目，"下列数中最小的一个是（括号外数字代表进制）"，这道题的正确率也只有 0.483660 。

同时，在2016年秋季学期的课后题 (绝大部分正确率均在80%以上) 中，也有题目为"黑白数字图像中的每个像素占___bit存储空间。"，其正确率只有 0.649293；同时，另一道题，"一张分辨率为72PPI的图像，其每平方英寸区域上包含___个像素点。" 正确率也只有 0.690246。这两道题都与像素有关。因此，可以发现同学们确实会有因为习题原因而回看视频的行为。

4 关键词分析

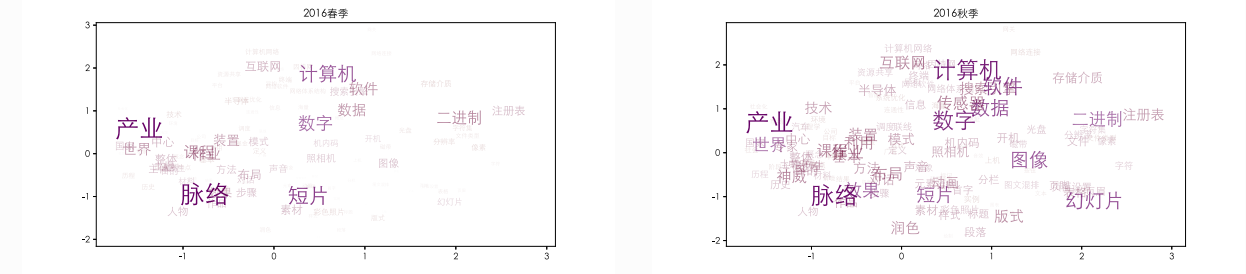
本小节使用了2016-2018年春、秋季学期 `学生行为` 中的 `video_study_export` 数据，共计六个数据文件。本小节绘制的图片类型为词云图。使用 `PCA` 算法与中文预训练集，将视频关键词绘制在二维坐标系的同时使得表意相近的词语距离亦相近。

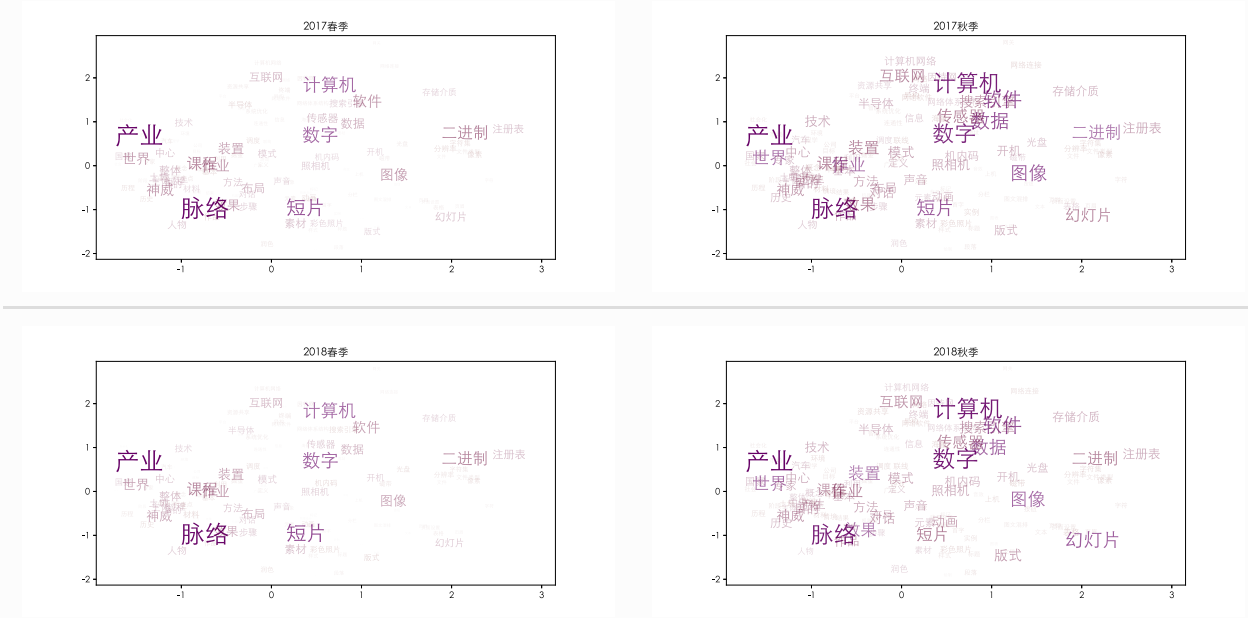
基于 `1.2.3` 中的实现过程，首先在不改变视频名称关键词的字号和透明度时，可以通过词云图发现该课程内容设计的四个板块。如下图所示：



总体来说，板块可以分为"计算机"与"文化产业"两大部分，而其中"计算机"部分又可分为系统与互联网、多媒体、逻辑名词 (如"二进制"、"字符"等) 三个小板块。

基于 `1.2.3` 中的实现过程，并针对视频的点击量赋予不同字号和透明度之后绘制的六张PCA词云图如下所示：





上述词云图可以反映同学们对于不同授课内容的感兴趣程度。从春季和秋季学期的角度看，可以发现春季学期受到关注的关键词明显多于春季学期，这也是由于不同的教学模式以及不同的学生群体导致的。可以发现，秋季学期比春季学期明显多出了"幻灯片"、"图像"等多媒体板块的词汇，可以推测的原因可能是：1）混合教学中线下课堂会留与多媒体制作有关的作业；2）线下课堂（大学生）与社会人士相比，对幻灯片制作与图像处理更加感兴趣。

同时还可以发现，从2016年秋季开始，每张图中的"文化产业"板块都会出现"神威"一词。推测是在2016年秋季向视频中新增了"神威"部分的视频，同时由于是我国的前沿科技，吸引了同学们的兴趣。经登录慕课平台并查阅目录，验证了这一推测。

5 结论

5.1 小结

通过以上分析，我们可以得出以下结论：

- 多数教学视频的人均播放量较低；
- 通过"慕课曲线"可以发现，有相当数量的同学会在观看视频的中途退出视频，仅有一般左右的比例会相对完整地观看视频；
- 两种教学模式下，同学们都有回看某些视频的习惯，原因是多方面的，视频对应的课后习题有难度是原因之一；
- 对于不同的教学模式，参与"混合模式"的同学会比"自主模式"的同学对于观看视频更加重视，其关注的话题也更加多元；
- 新增前沿科技的介绍会受到同学们的关注。

5.2 建议

5.2.1 针对已有的视频

- 提高视频趣味性与难度
- 适当提高课后题目的难度

5.2.2 针对课程内容

- 增设前沿技术介绍的相关内容

6 文件目录

根目录

- pics：存放全部生成的图片的文件夹
    - hist：存放各学期直方图文件
    - range：存放各学期慕课曲线文件
    - word：存放哥2学期词云图文件
- data：存放原有数据与经python语言处理后生成的数据文件
- src：存放全部源代码
- docs：存放全部说明文档

代码目录

- Paint_Hist.py：绘制直方图的代码
- Paint_Range.py: 绘制慕课曲线的代码
- Paint_Words.py：绘制词云图的代码
- problem.py：计算各学期课后习题正确率的代码