实验报告

姓名：王菲

学号：10175501111

一、实验目的

探索项目可持续性的影响因素

二、实验任务

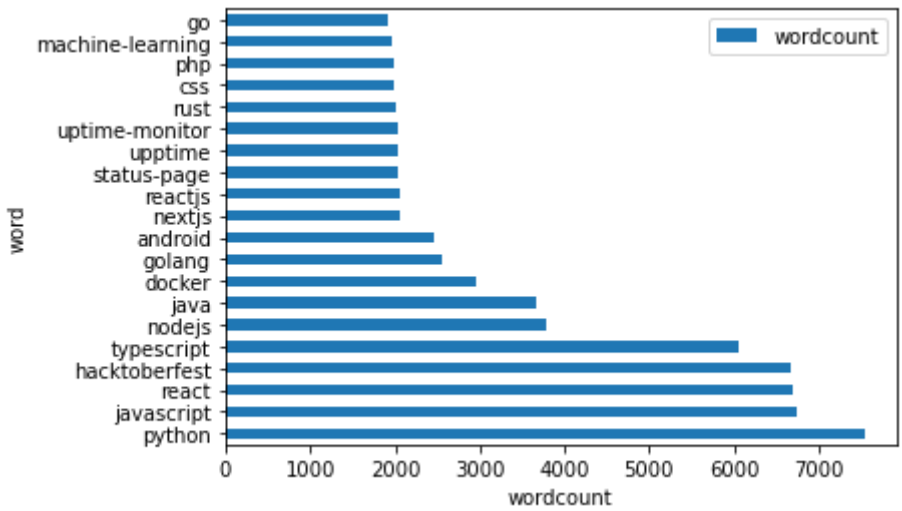
1. 筛选同类型项目
2. 探索顶级项目
3. 探索TOP3顶级项目生命周期
4. 定义活跃度指标
5. 研究项目生命周期的划分
6. 通过聚类进一步研究

三、使用环境

Jupyter notebook

四、实验过程

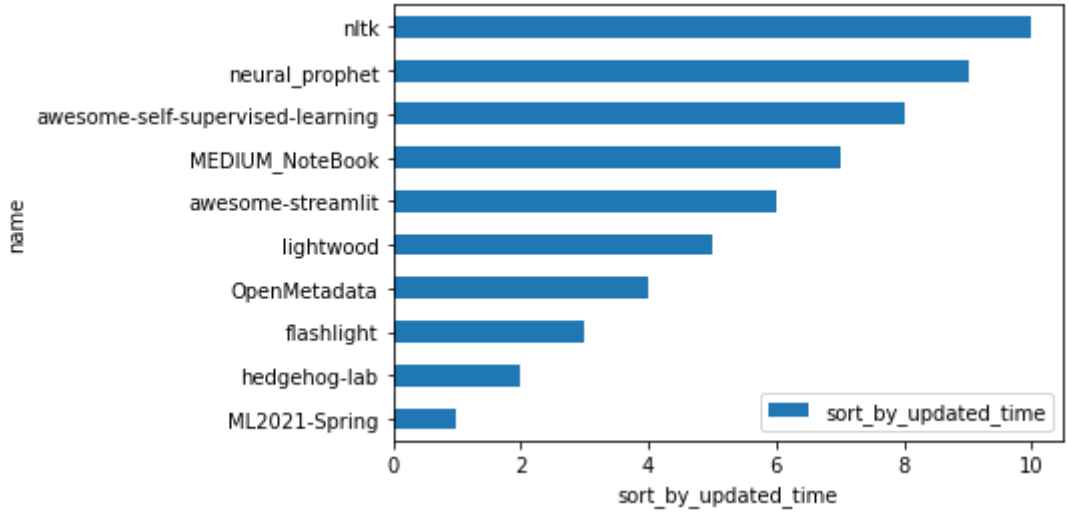
筛选同类型项目



图片中显示的是数据库中出现最频繁的前20个topic，其中machine-learning是一个比较贴近的topic，所以将筛选该主题的项目进行分析；

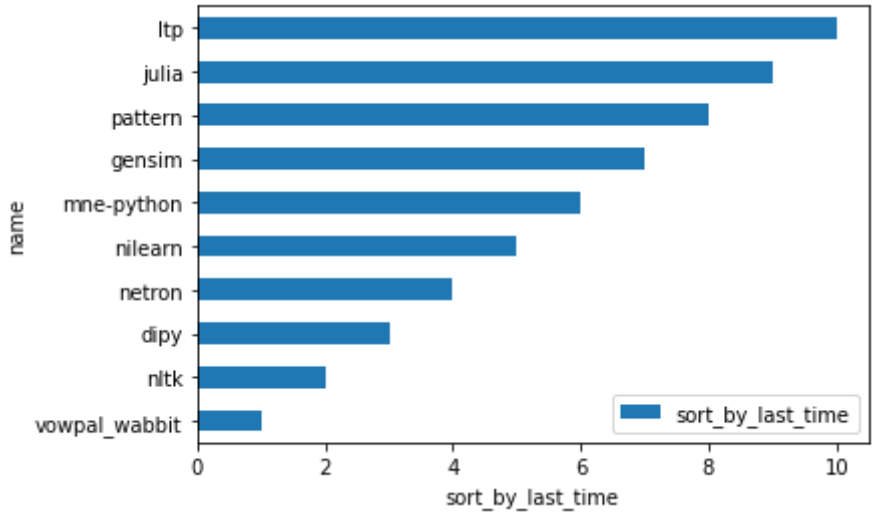
探索顶级项目

1. 最近更新的项目



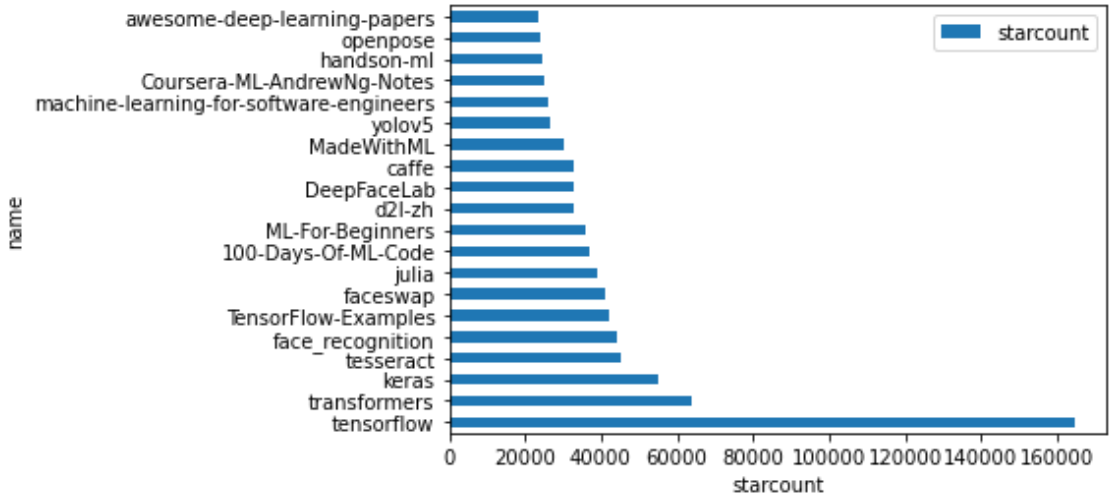
可见ML2021-Spring等项目最近仍然在更新；

2. 持续时间最长的项目（update和push取时间最近的）



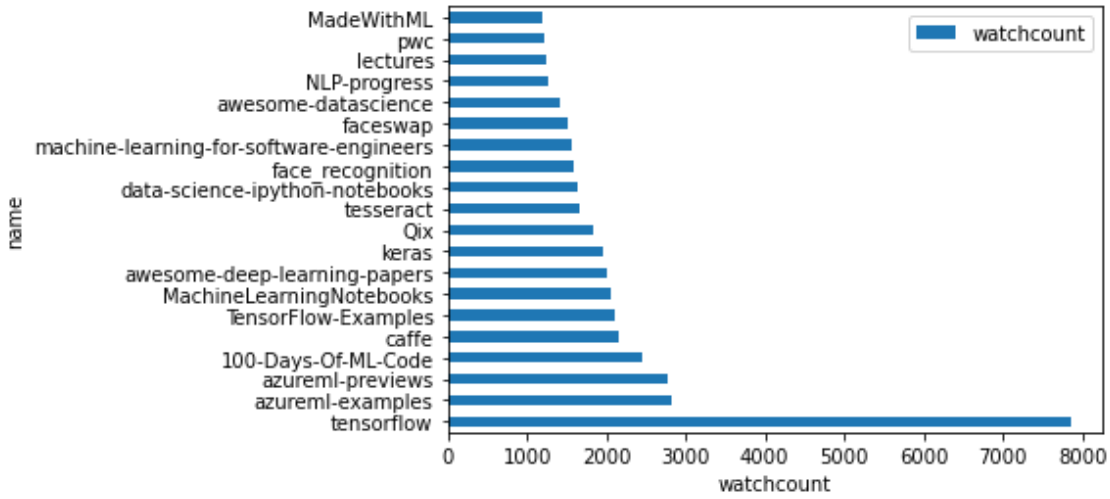
可见vowpal_wabbit是持续时间最长的项目；

3. STAR的数量



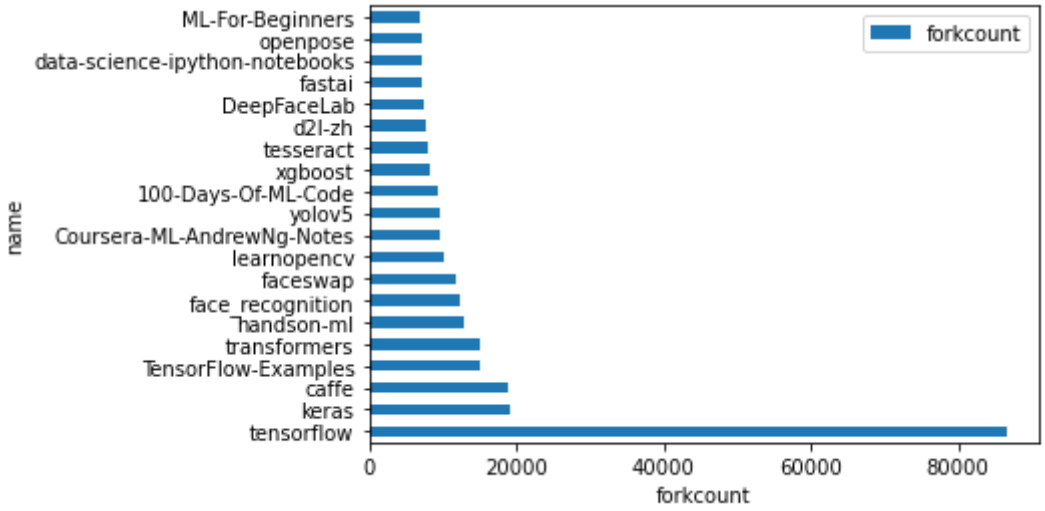
可见tensorflow是star最多的项目；

4. 关注的数量



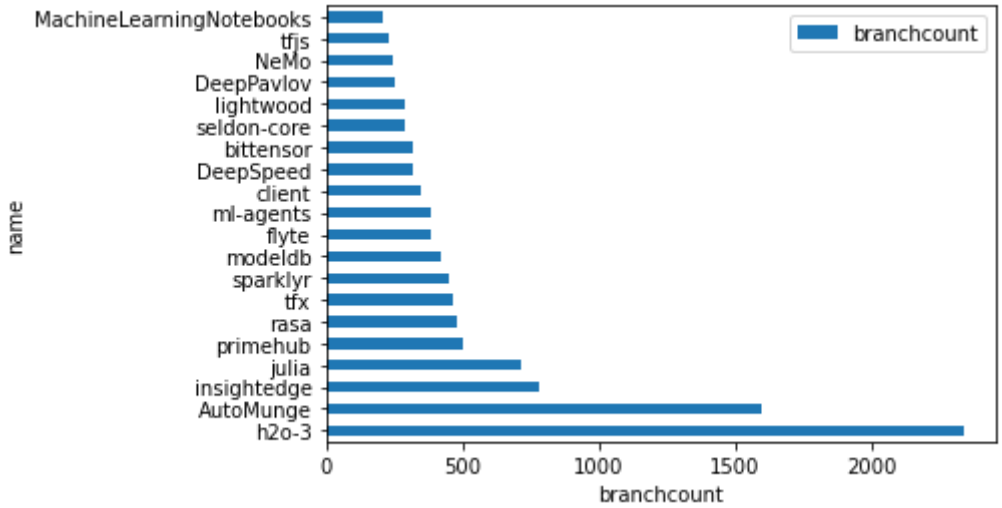
可见tensorflow被关注的人数最多；

5. 被FORK



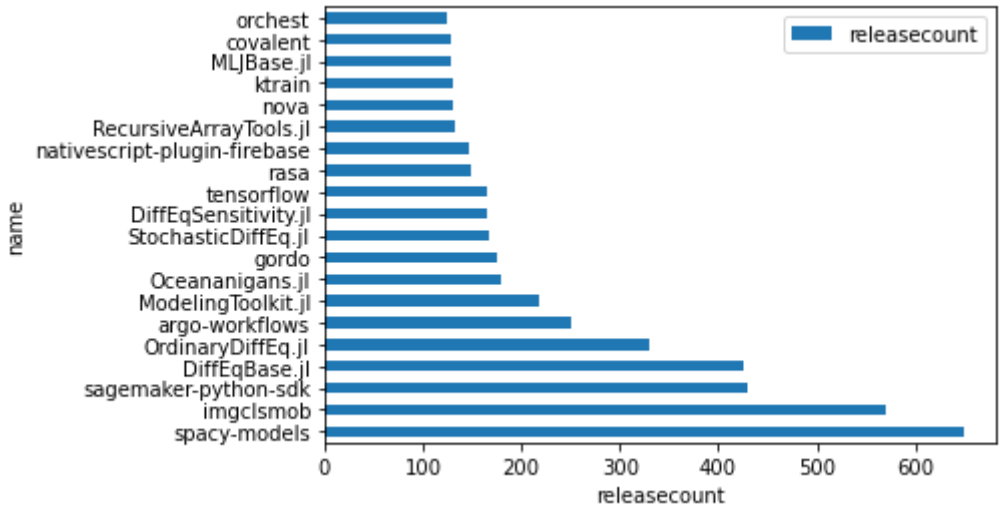
可见tensorflow被fork的数量最多；

6. 分支数量



可见h2o-3分支数量最多；

7. 发布数量



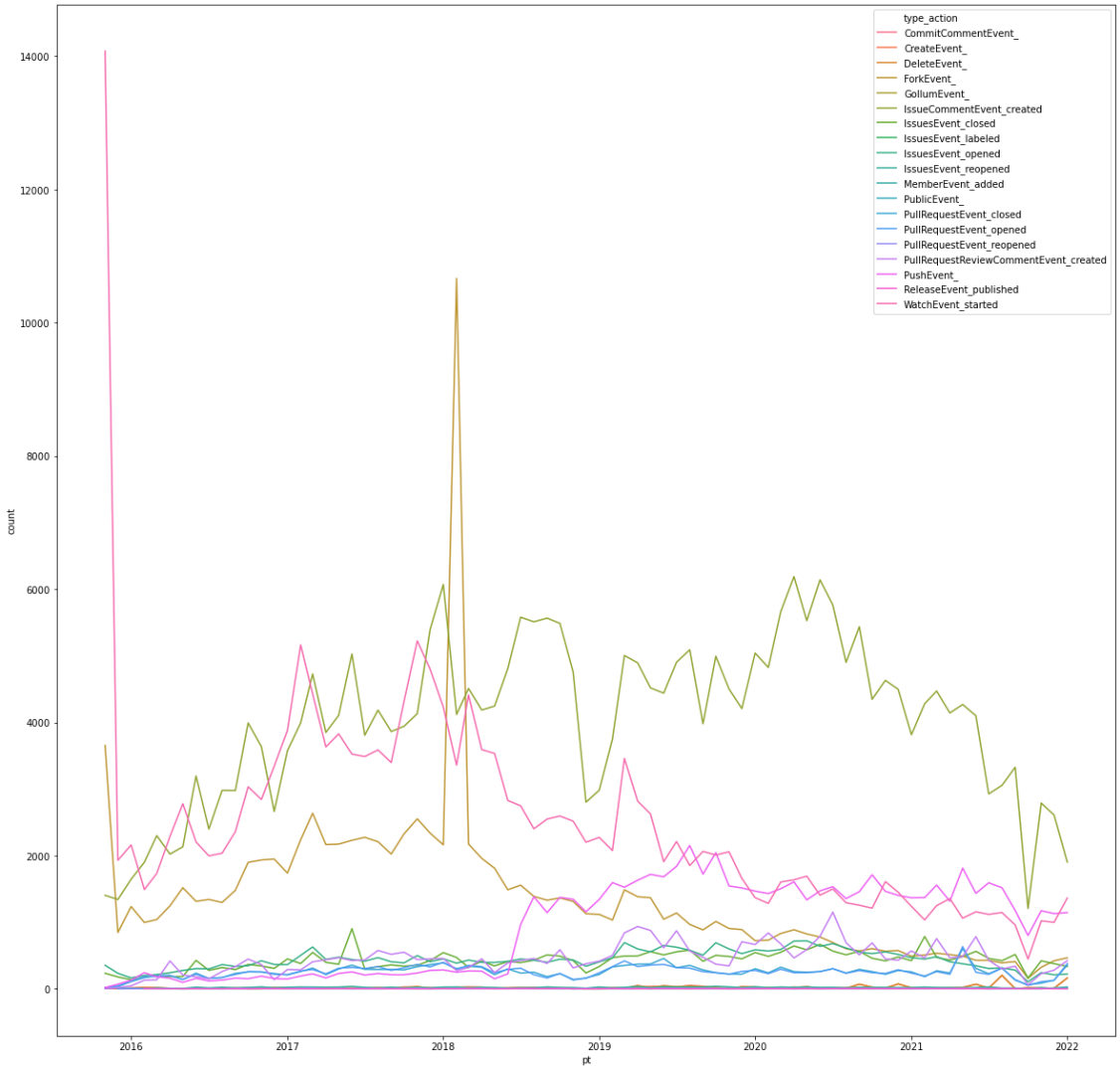
可见spacy-models发布数量最多。

8. 综合排序

   前三个顶级机器学习项目为tensorflow、transformers和keras.

探索TOP3顶级项目生命周期

1. tensorflow创建于2015-11-07，owner为tensorflow；



   可见tensorflow项目在建立之初就备受关注，然后经过了一段时间的增长，在2019年之后新增关注人数逐渐下降；

IssueCommentEvent一直在逐步增加，但在2019年左右新增数量突然下降，之后又逐步增加并在2020年初达到新高，又在2021年年中的时候突然下降；

Fork事件的数量在建立之初已经有一部分人对该项目Fork并进行开发，后来在2018年初突然骤升，又在之后突然骤降，并逐步下降；

值得关注的是从2018年初开始PushEvent开始逐步增加，之前的数量较少，应该是更多开发人员加入导致的；

所以tensorflow项目的生命周期可以从2018年初这个点开始划分。

2. transformers创建于2018-10-29，owner为huggingface；

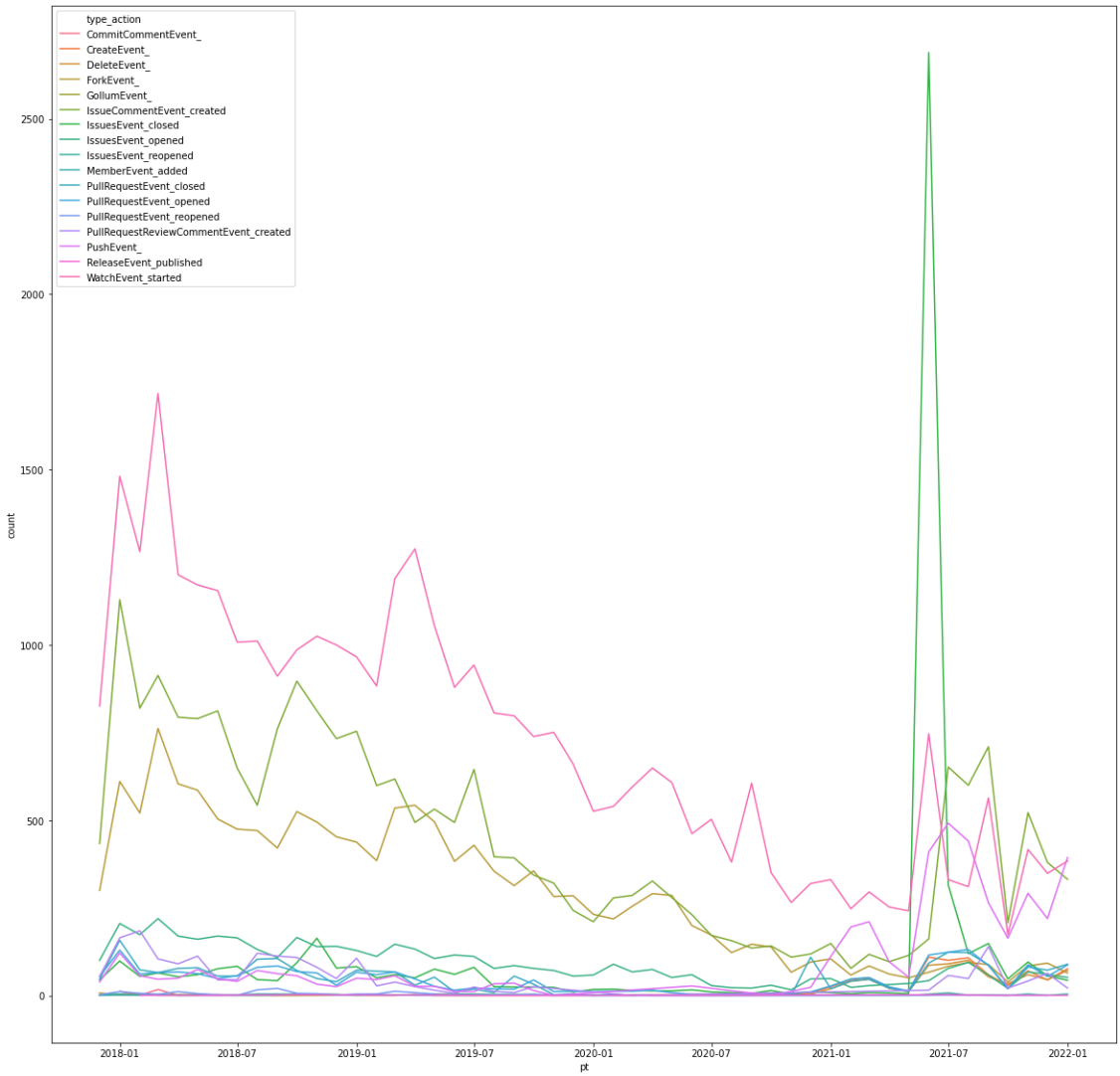
transformers项目在2018年10月末建立之后，一开始关注的人数不多，在2019年10月份大致一年的时间里开始发展；

项目一直到2020年7月份左右都发展十分迅速，之后到2021年5月份左右都属于平稳发展期；

之后在2021年10月份所有数据都骤降，不确定是否是数据异常，之后又保持在另一个数值较低的平稳发展期；

所以transformers项目的生命周期有两个节点，2020年5月份以及2021年5月份。

3. keras创建于2015-03-28，owner为keras-team.



keras项目的事件趋势较为明显，从项目初步发展数值增加到2018年3月份，之后新增数量便逐步递减，到2019年5月份有所动荡，之后便一直下降到2021年7月份，之后又小幅度上涨到一定水平；

而且在将近2021年6月份左右时Fork事件的数量才逐步上升；

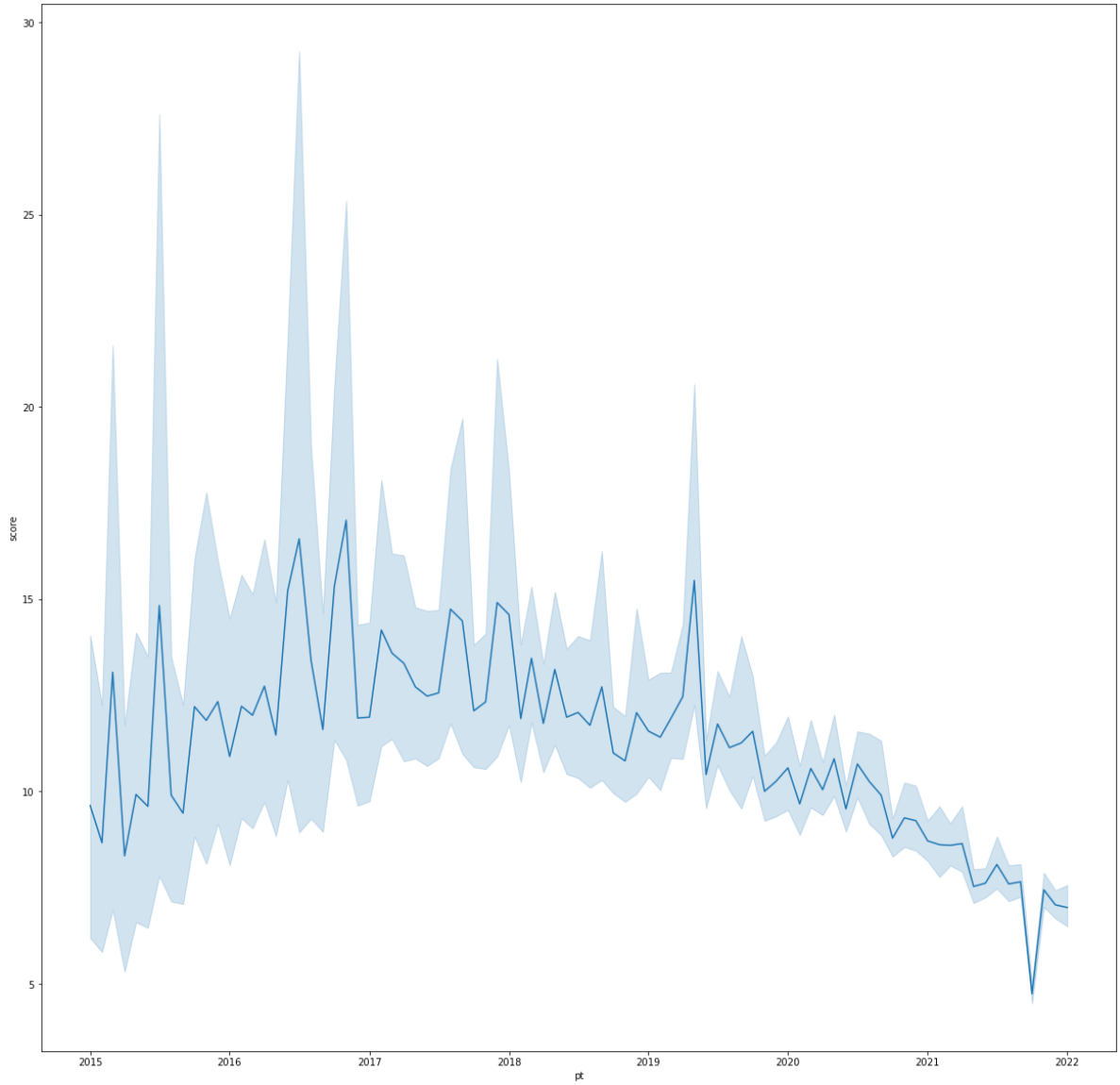
所以keras项目的生命周期节点应为2021年6月份。

4. 经过上述项目生命周期节点的分析，可见github中项目的主要事件为Watch、Fork以及IssueComment，主要能判断生命周期的事件是Push.

定义活跃度指标

设置关注的权重为1，Fork的权重为2，评论Issue的权重为1，定义score；

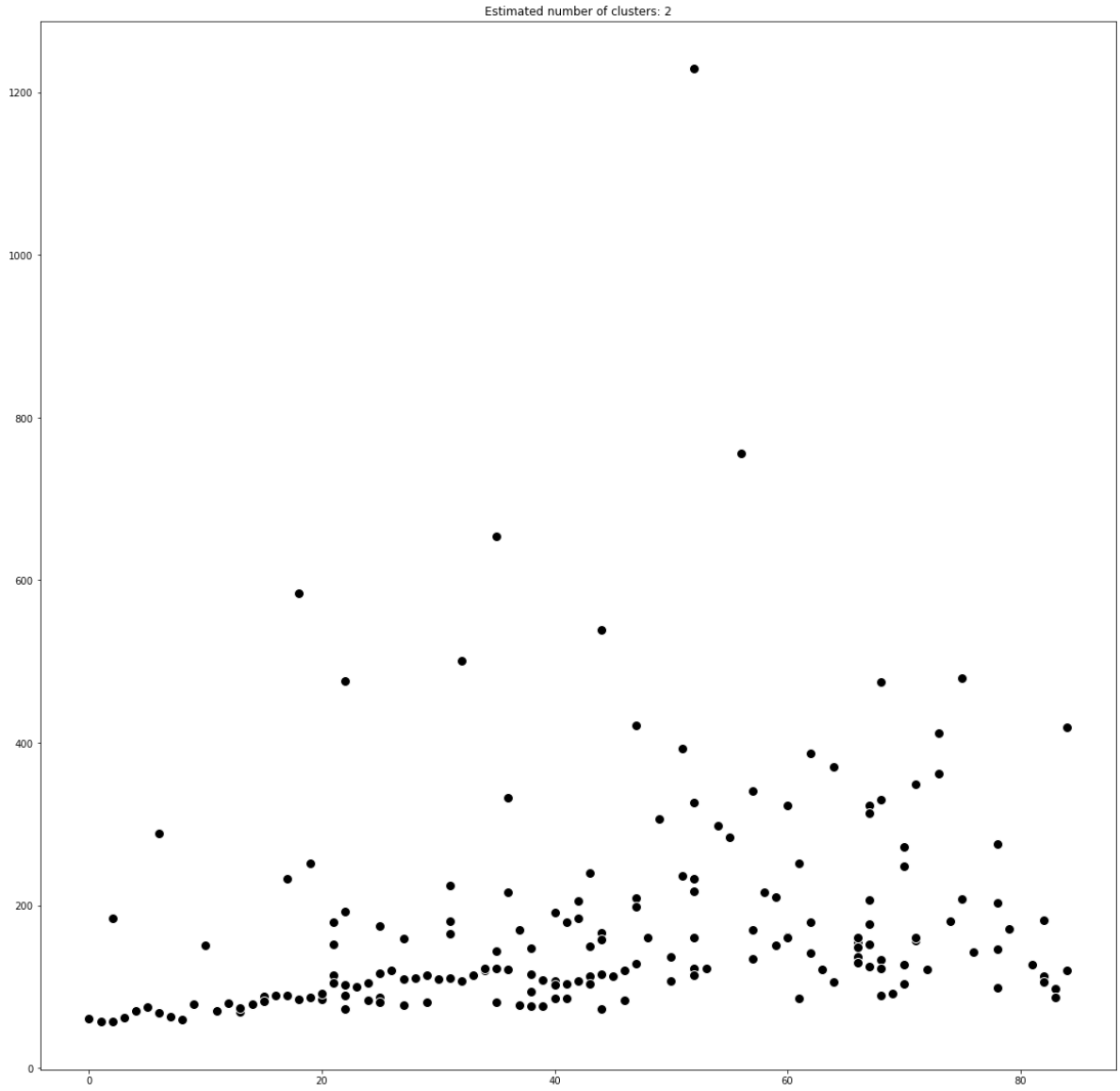
所有机器学习项目活跃度的趋势如下：



可见从2015年到2019年是机器学习项目的巅峰期。

根据活跃度聚类

聚类结果



因为按照月份对日期进行聚合后数据点较少，聚类结果较差，但是可以看出项目活跃度是逐渐先升再降的。

五、总结

通过这次课程实验的探究，研究了机器学习项目中各事件的次数，探究了TOP3顶级项目生命周期的划分以及影响划分的三个事件。根据探索性分析的结果对项目活跃度进行定义，得到所有机器学习项目活跃度的发展趋势，最后通过聚类算法进一步研究机器学习项目在活跃度上有什么特殊趋势。