华东师范大学数据科学与工程学院实验报告

课程名称：社会计算　　**年级：**19 级

指导教师：钱卫宁　　**姓名：**庞瑞洋 卞思頔

实验名称：基于 github 数据的机器人与人类分类识别

一、 实验引言

本实验通过分析 github 数据集，差异化 github 用户的特征，并且将用户分为两类 ：人、机器人，并发现两者属性方面的差异。

我们使用从未知用户提取的特征关系分析来确定作为人，机器人的可能性，最后经过实验评估证明了所提出的分类系统的功效。

二、 问题描述

从 2016 年特朗普当先美国的总统，人们开始关注机器人在社交网络上对于人们的影响，以及开始关注机器人对我们的社会影响如何。同时这也引出了一个问题，我们应该采取什么样的方法来控制社交机器人的广泛传播。截止到 2020 年的美国大选，这个问题变得比以前更加严峻。

通过系统的分析，我们讨论机器人控制的相关研究趋势，并希望能够为检测和其他的相关努力提供一些信息。

三、 实验方法

1. 选取 2021 年来自 github 的 100 万条数据作为实验分析基础。实验重点分析名称中带有“bot”的机器人用户，且发帖多于十条的用户。
2. 文本相似度判断依据：

计算每个实验对象的平均的 Levenshtein-distance 和 jeccard-distance 值，并将其置于图中观察。分析 human 与 bot 所处位置的特征，根据计算用户每次提交的那个距离去判断它是不是近是一个机器人。

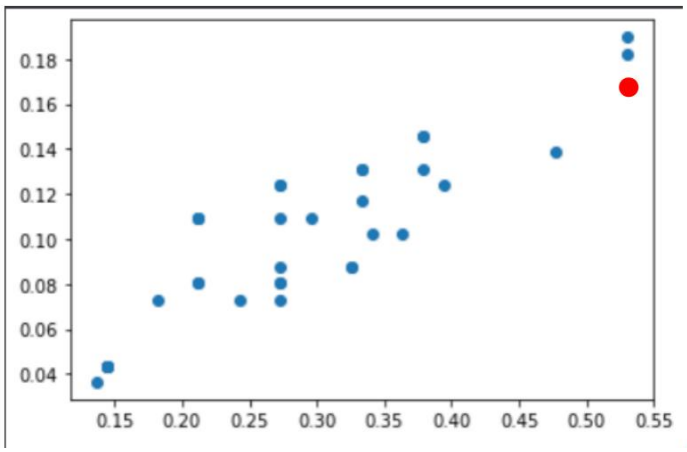
jeccard 距离：

$$\mathcal{J}(C_1, C_2) = 1 - \frac{|\, words(C_1) \cap words(C_2) \,|}{|\, words(C_1) \cup words(C_2) \,|}$$

levenshtein 距离：$$\mathcal{L}(C_1, C_2) = \frac{lev(C_1, C_2)}{max(|C_1|, |C_2|)}$$

3. 临界点寻找依据：

根据计算提交的 message 间的两个距离去判断它是不是近是一个机器人，需要找到一个（近似）临界点，使得能够大体上将 bot 与 human 进行分类。寻找临界点，寻找依据：通过计算临界点两边的机器人与人类的比例。最优的点保证小于临界点时的机器人比例最大，人类比例最小。




```
at point (0.60,0.65),botper = 0.212,huper = 0.080
at point (0.60,0.70),botper = 0.212,huper = 0.109
at point (0.60,0.75),botper = 0.212,huper = 0.109
at point (0.60,0.80),botper = 0.212,huper = 0.109
at point (0.65,0.50),botper = 0.242,huper = 0.073
at point (0.65,0.60),botper = 0.273,huper = 0.080
at point (0.65,0.65),botper = 0.273,huper = 0.080
at point (0.65,0.70),botper = 0.273,huper = 0.109
at point (0.65,0.75),botper = 0.273,huper = 0.124
at point (0.65,0.80),botper = 0.273,huper = 0.124
at point (0.70,0.50),botper = 0.273,huper = 0.073
at point (0.70,0.60),botper = 0.326,huper = 0.088
at point (0.70,0.65),botper = 0.326,huper = 0.088
at point (0.70,0.70),botper = 0.333,huper = 0.117
at point (0.70,0.75),botper = 0.333,huper = 0.131
at point (0.70,0.80),botper = 0.333,huper = 0.131
at point (0.75,0.50),botper = 0.273,huper = 0.088
at point (0.75,0.60),botper = 0.341,huper = 0.102
at point (0.75,0.65),botper = 0.364,huper = 0.102
at point (0.75,0.70),botper = 0.379,huper = 0.131
at point (0.75,0.75),botper = 0.379,huper = 0.146
at point (0.75,0.80),botper = 0.379,huper = 0.146
at point (0.80,0.50),botper = 0.295,huper = 0.109
at point (0.80,0.60),botper = 0.394,huper = 0.124
at point (0.80,0.65),botper = 0.477,huper = 0.139
at point (0.80,0.70),botper = 0.530,huper = 0.168
at point (0.80,0.75),botper = 0.530,huper = 0.182
at point (0.80,0.80),botper = 0.530,huper = 0.190
```

左图的横坐标表示机器人的比例，纵坐标表示人类的比例，因为要保证小于临界点时的机器人比例最大，人类比例最小，所以选取最靠近右下角的点，则对应到原来的距离为(0.8,0.7)。

4. 识别机器人：

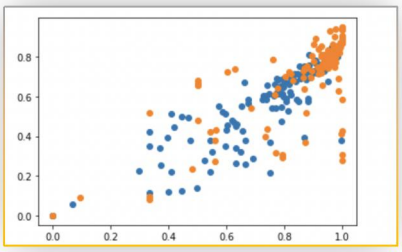
通过两种方法去识别两类机器人：

（1）1.一直进行类似的重复的更新等工作的机器人：每次 commit 的 message 的格式是类似的，文本相似度极高。

（2）判定和他人有相同行为的用户：即 commit 时的 message 和别人的完全相同，然后计算这些相同的 message 之间的距离，距离大的可能为机器人。在大部分情况下，message 相同可能都是 update 同一个文件等的 message，所以是 human 的可能性较大。但如果有多条毫不相干的 message 且都和别人（多个个体用户）的 message 相同，则可能为机器人，即复制不同人的操作。

四、 实验评价

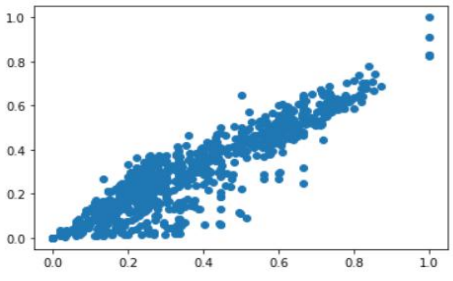
对于第一类的情况，在训练阶段，将 id 中带有 bot 字段的用户视为机器人，其余视为人类，则根据计算每次提交的那个距离去判断它是不是近是一个机器人（蓝色为姓名中有“bot”的机器人，橘色为人类，横坐标为 jeccard 距离，纵坐标为 levenshtein 距离）。可以清晰地看出两者距离分布的不同，人类主要出现在右上角而机器人主要出现在左下角。

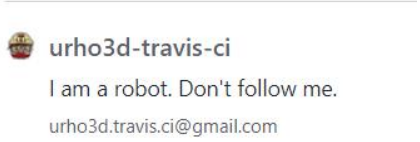


图中会发现在人类中有处于左下角极端值的情况，于是去查看该点，发现字段中不带有 bot 但是确实是机器人，一直在重复 autocommit 的工作，所以极端值出现原因：因为只是将 id 中带 bot 的视为机器人，所以产生异常点。

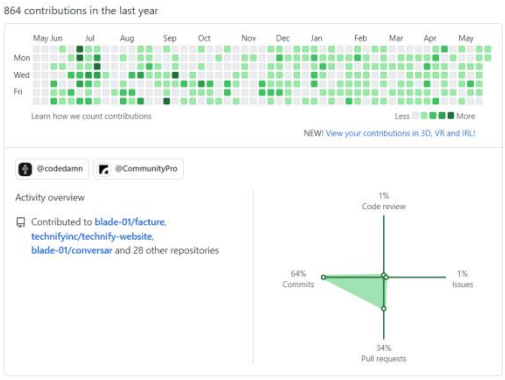
随后根据找到的临界点去对新的数据进行机器人的查找判别可以发现：在判定为机器人的用户中确定为 bot 的个数为 7，确定为 human 的个数为 11 ，无法确定的个数为 8。去除无法确定的，分类出来的 True Positive 值为 7/18,False Positive 值为 11/18。这是在采用了近似最优临界点的情况。如果采用小于该点作为临界点，则 True Positive 值会下降；如果采用大于该点作为临界点，则 False Positive 值会上升。因为训练时只是将带 bot 字段的用户视为机器人，并没有完整的机器人用户等的数据，所以效果不是很好，但是也可以作为一个分类的手段。

对于第二类的情况，寻找了有重复 message 的记录，然后计算这些 message 间的平均距离（如下左图，横坐标代表 jeccard 距离，纵坐标代表 levenshtein 距离），然后寻找距离大的用户，即靠近图中右上角的用户。如下最右图，确实可以分类出来机器人。相应的，判别出来的 True Positive 值为 3/8,False Positive 值为 5/8。



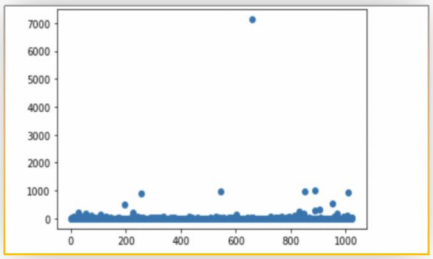


然后查看其中被误判的 human 用户，发现这些用户的活跃度都是非常高的。

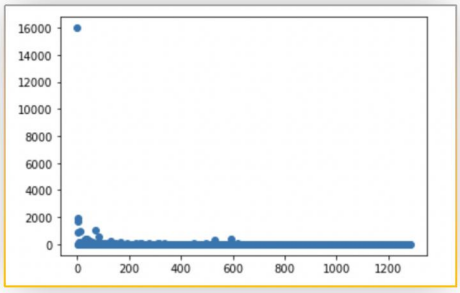


再结合采用的方法分析，因为选择了和其他人有重复 message 的用户，但是被无法判断该用户是复制其他人的 message 还是被别人复制了 message 导致的重复情况。而因为这些用户的活跃度非常高，所以 commit 的 message 也很多，进而被复制的情况也会很多，所以被误判为了机器人。所以对于此类机器人，仅仅用 message 的方法去判断效果并不是很好，可能还需要其他的特征一并判断。

除了机器人和人类的差别之外，机器人与机器人的区别也值得探究，观察机器每年 commit 的次数，可以看出除去个别机器人达到较高的超过 7000 的数目，大部分都在 1000 以下或附近。



除了提交的 message 数，观察机器人参与的仓库数，也可以看出除去个别少数的达到较高的 16000 的数字，大部分机器人都在 2000 以下。



五、 后续工作

分析了当前工作结束后，仍然有很多待解决的问题需要考虑：

1. 由于时间原因实验没有进行太多优化，重复计算可能较多，导致运行慢。
2. 需要进一步的实验优化，比如 bot 之间的差异需要更加明显，还需进一步探讨(如参与的项目数量等)；
3. 分析的范围也需要扩大到 id 不含有“bot”的用户身上，防止对于机器人的误认；
4. 需要更细节地观察寻找机器人行动的目的，依照所需特征更准确地进行分析预测；
5. 需要采用更多的特征进行机器人的判别比较，仅仅从单一方面去进行比较是不够的，最后呈现的效果也不是特别好。

六、 结论

本实验通过分析 github 数据集，差异化 github 用户的特征，我们使用从未知用户提取的特征关系分析来确定作为人，机器人的可能性，最后经过实验评估证明了所提出的分类系统的功效。最大的收获是体会了真正的数据分析的过程。从数据的获取，预处理，再到算法的设计，debug 的过程，然后最终跑出结果。

通过系统的分析，我们也可以讨论机器人控制的相关研究趋势，并希望能够为检测和其他的相关努力提供一些信息。

附录：

小组分工：

庞瑞洋：实验思路设计，思路总结修正，代码实现，小组汇报

卞思頔：实验思路设计，思路总结修正，汇报制作，论文编写