公益基金会推荐系统

团队名称：C3

1. 摘要：

本方案结合项目型基金会固有的特征和资助型基金会对项目型基金会偏好（相互关系），使用协同过滤的方法，为资助型基金会推荐适合他们投资的项目型基金会。同时，新的基金会在输入相应信息之后，根据基本特征相近的办法，也可得到合适的推荐。

2. 背景：

现实中，投资型基金会常常面临有钱缺找不到合适项目的场景。不同的公益基金会由于自身建立的目标，基金会财务状况，甚至一些地理位置带来的“地缘”因素，会对投资的项目型基金会产生不同的偏好，亟需有个性化的推荐系统。资助型基金会通过他们对项目型基金会的合理投资，不仅更高效地将资本和执行方结合实现了自身的愿景，也对社会资源进行有效合理的分配。他们的作用，是在政府无法管控到或管控不足的领域内，由民间力量聚集起来的看不见的手，为社会问题提供了有效的解决方案。在基金会中心网在案的 6125 家基金会记录中，有 3813 家基金会都曾经投资过其他的基金会，占总数的 62.29%。 因而，解决这一问题不仅对项目型基金会本身有重要帮助，也具有深远的社会影响。

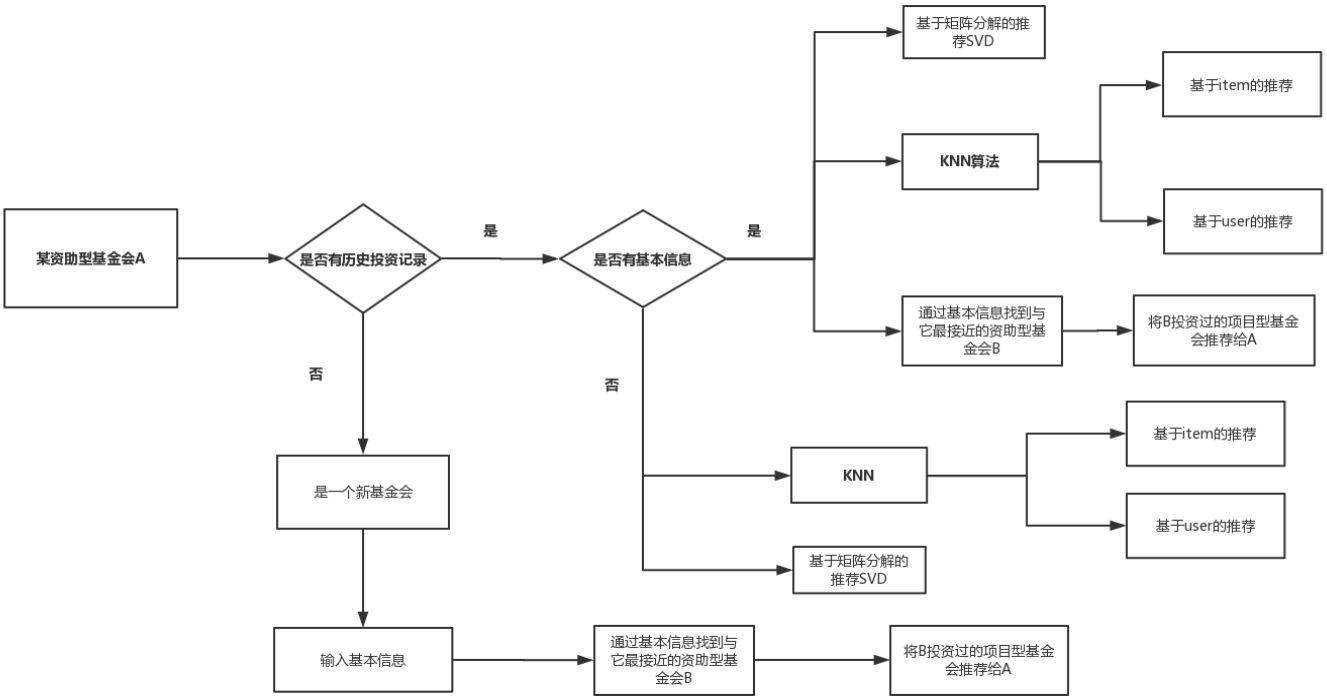
3. 需求：

我们的系统可以满足两类资助型基金会的需求。

对于已经纳入基金会中心网“基金会接受捐赠情况”的资助型基金会，我们可以通过它**历史投资**的数据，为它推荐符合它便好的项目型基金会。

对于未纳入基金会中心网“基金会接受捐赠情况”的资助型基金会，我们无法得知它历史投资的数据。但是我们会通过问卷的方式，要求基金会简要回答 17 项关于基本信息的问题。通过它的基本特征，为它寻找到和它最相似的已经在基金会中心网数据库中的基金会，进而做出推荐。

基金会投资推荐系统



4. 最终方案及数据来源：

最终方案：

a. 对于已经纳入基金会中心网“基金会接受捐赠情况”的资助型基金会推荐系统的设计，我们主要采纳了 SVD 和 KNN 算法。

SVD 算法，即是 Netflix 推荐系统中涉及的经典推荐方法。由于数据本身对应关系比较少，大部分投资型基金会只投资了一个受助型基金会，用 SVD 算法得到的结果集中在被投资多次的“热门”项目型基金会，我们结合了 KNN 算法 。KNN 基于资助型基金会对项目型基金会的打分进行了系统过滤，当中相似度的概念我们选取了 Cosine 距离 。两个算法我们都是都过 R 的”Surprise”包实现，分别取前 5 的结果（如果可以有结果）。

关于 SVD 和 KNN 算法用到的打分的数据清洗和具体计算如下：

数据清洗：

首先，通过对基金中心网“基金会接受捐赠情况”记录的爬虫，我们得到 2011 年到 2015 年的 32550 条数据；

其次，筛选出当中捐款方为基金会的，得到 32550 条数据；

再者，我们计算出 2011 年到 2015 年投资次数大于等于 1 的基金会共计 2010 家；对于有多次投资的基金会，我们选取 2011-2015 年间投资金额的平均值，筛选出有财务数据（净资产）的企业共计 4306 家；

最后，通过结合财务信息数据和捐赠情况数据，并对年份去重，我们最终得到 1564 个基金会数据。

数据清理使用软件：stata14 （详见 DataClean 文件夹）

请将输入数据和代码置于同一路径文件夹下，并在代码开头将 cd 后的内容修改为该路径

输入数据：

FundationList_final.xls 网络爬虫的 32550 条捐赠数据

Foundation_asset_2013.xlsx 2013 年的财务数据

Foundation_asset_2014.xlsx 2014 年的财务数据

Foundation_asset_2015.xlsx 2015 年的财务数据

输出数据：

"score_all.xls " 最终评分数据：其中 user 和 item 分别为资助型基金会和项目型基金会，各自对应相应的 userid 和 itemid。rate_t 为投资次数，rate_a 为投资金额占净资产的百分比，score_t 为投资次数打分，score_a 为投资金额打分，score_all 为最终得分。

打分参考了两个指标：

投资额占净资产的百分比：由于基金会本身财务状况不同，对项目型基金会的投资额自然会产生影响。我们通过基金会的净资产去除了这一影响，更加公允地表示了基金会投资意愿。我们的给分标准是：前 10%是 10 分，10%-20%是 20 分，以此类推到 100 分。对极端值的处理：<0.01 的是 5 分， >100 的是 100 分。

投资次数（2011-2015 年之间）：投资次数带有强烈的偏好暗示，因此我们在下面的最终给分中给了投资次数更大的权重。对于投资次数这一部分而言，我们的给分标准是：31 分为基地，多增加一次投资+3 分，最高 100 分（投资了 24 次）。

最终的分数 Score_all = 投资额占净资产的百分比的得分* 30% +投资次数的得分* 70%，分布如下：

最终分数 Score_all 的分布

	Percentiles		Smallest		Obs	1564
1%	24.7		23.2		Sum of Wgt.	1564
5%	24.7		23.2		Mean	40.4
10%	27.7		23.2		Std.Dev	9.6
25%	32.8		23.2			
50%	39.7		Largest		Variance	92.4
75%	48.7		78.1		Skewness	0.2
90%	51.7		79		Kurtosis	3
95%	53.8		80.5			
99%	62.5		82.9			

b. 对于未纳入基金会中心网“基金会接受捐赠情况”的资助型基金会推荐系统的设计，我们通过关键特征值，找到和它最相似的纳入系统的基金会，推荐最相似基金会所投过的项目型基金会。

当前面一种情况做完之后，我们面临一个新的问题，新的资助型基金会（未纳入基金会中心网“基金会接受捐赠情况“的资助型基金会）出来了怎么办？由于我们没有任何关于他的投资信息，我们可以通过用基金会本身的基本信息找到已经存在在基金会中心网官方资料中离它最近的基金会，把这个已经存在的资助型基金会的投资过的项目型基金会推荐给这个新的基金会。我们会当探索到用户输入的基金会为新的资助型基金会时，询问 17 个关键指标，利用这些特征值根据已有的数据进行相似度计算。这 17 个关键指标是我们整合了资助型基金会所有的基本信息，财务信息和透明度打分，运用 PCA 的方法提取了 17 个重要的特征值 。

补充一点，相信聪明的您已经从我们的流程表中看出，当已经存在的基金会（前面一种情况所讨论的）同样出现在官方那个指标的表格里，我们会帮它找到 最近的资助型基金会，然后把这个基金会投过的项目型基金会推荐给它。

整体数据来源补充说明：

由于数据量的大小对推荐系统的精准度有显著的影响，我们通过爬虫取得了基金会中心网上 6125 家基金会的全部记录 32 万多条投资信息，而没有直接采用比赛组委会给出的 605 家基金会的数据。

爬虫的代码在 spider 文件下：invest_inf.py 爬取了基金会中心网“基金会接受捐赠情况”所有资助型基金会对项目型基金会的投资情况数据；invest_url.py 爬取了所有在 invest_info.py 列表里的投资型基金会的详细地址；finance_info.py 是根据每个投资型基金会详细地址，爬取了他们 2013 年度到 2015 年度的财务数据。

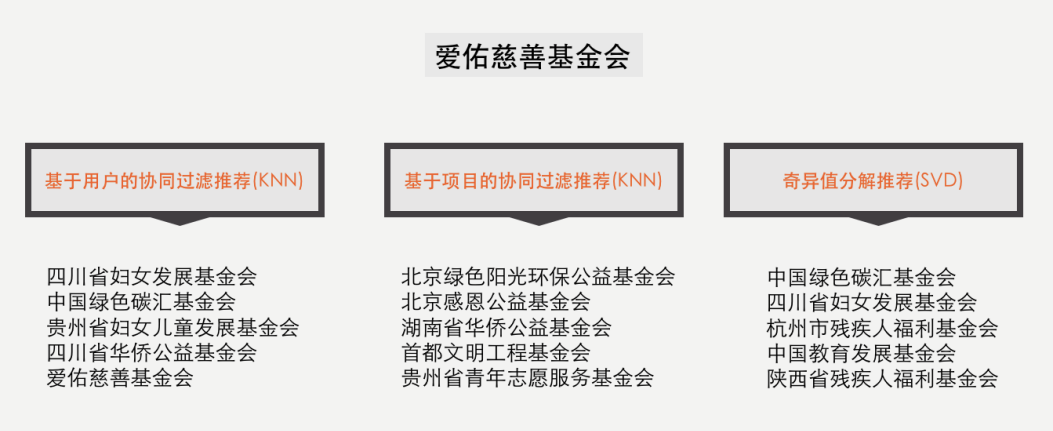
5. 研究方法

本方案结合项目型基金会固有的特征和资助型基金会对项目型基金会偏好（相互关系），通过这两个维度的信息， 为资助型基金会推荐适合他们投资的项目型基金会。

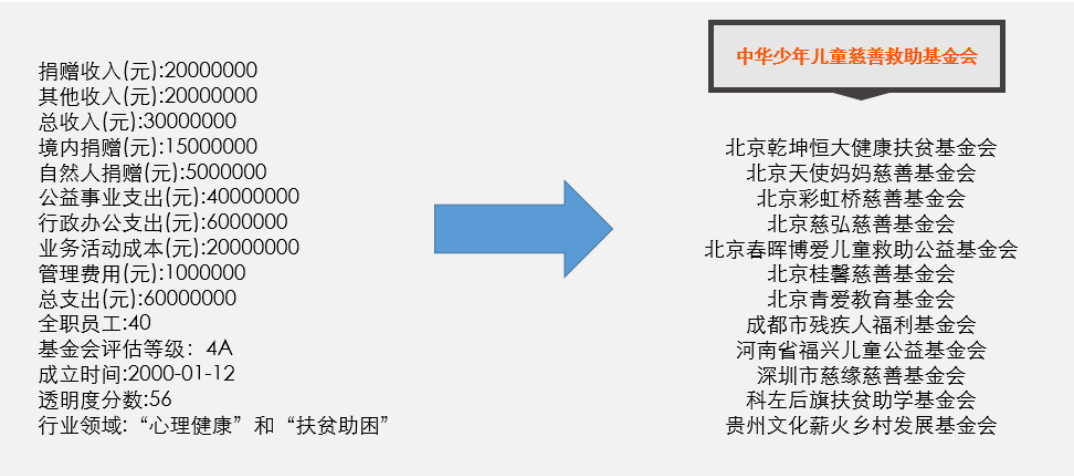
6. 研究用例

1. 代码需要在 python3 的环境下运行。需要安装一些 python 的包，具体参考 requirements.txt。也可以一键安装所需要的环境： pip install -r requirements.txt。
2. 在 shell 中执行 main.py。
3. 进行查询。

   对已经纳入基金会中心网“基金会接受捐赠情况”的资助型基金会，我们以“爱佑慈善基金会”和“上海同一苍穹下公益基金会”为例，可以查询到如下结果：



和未纳入基金会中心网“基金会接受捐赠情况”的资助型基金会，我们虚拟输入如下参数，可以得到推荐：



7. 讨论

a. 由于我们使用的算法需要大量的数据作为支撑，我们在投资信息上，选择爬虫的方式获取了共计 32550 条数据，相比于官方给出的 3200 条数据，为 a 类情况的资助型基金会研究极大地增加了样本量。但在 b 类情况的资助型基金会研究中，我们仅使用了官方给出的 605 数据作为支撑，未来可以考虑进行更全面的数据爬虫，为符合 b 类情况的资助型基金会，提供更完整的分析。

b. 使我们的推荐系统更加完善，我们需要有更多基金会领域的行业洞察和业务理解。后续我们希望通过问卷的形式了解不同基金会的需要，并根据问卷的结果对一些显著的结果和现象，做有针对性的面对面访谈，使得我们的推荐系统设计更加切中要害。