138 | 数据科学团队必备的工程流程三部曲

2018-08-20 洪亮劼

AI技术内参 进入课程 >

讲述：初明明
时长 08:05 大小 3.71M

今天，我们继续来聊一聊数据科学团队的一些基础构建思路，讨论一些日常的在“工程流程”方面所需要注意的问题。和我们上一次分享的项目管理不一样，工程流程没有很多可以直接借鉴的经验，需要从业人员进行更多的思考和创新。

什么是工程流程

我们首先来看一看什么是“工程流程”。一般来说，工程流程指的是我们有什么制度或者说是策略来保障所做的项目能够达到一定的质量标准。

那工程流程和项目管理流程有什么区别呢？我们说，项目管理流程是从宏观上把握项目的进展，而工程流程则主要是**在微观上**，定义和掌控具体的每一个步骤上的输入、输出和过程。

从另外一个角度来说，这两者之间并不存在一个必然的关联关系，一个项目在细节的工程流程上成功与否和一个项目自身的最终成功与否，并不能完全划等号。

你是不是有疑问，既然如此，那我们为什么还要关注工程流程呢？原因是虽然一个好的工程流程并不一定带来项目的成功，但是可以增加成功的可能性或者说是概率。同时，一个好的工程流程可以帮助一个团队在日常的运作中减少问题的发生，从而能够达到事半功倍的效果。

那么，工程流程究竟包含哪些方面呢？

我们在今天的分享里讲三个方面。第一，代码管理的流程；第二，开发部署环境的流程；第三，数据管理的流程。这三个流程可以说是涵盖了一个人工智能项目发展和成功所必不可少的三个重要方面。

代码管理流程

人工智能项目一个很重要的环节就是开发代码。然而，因为数据科学、人工智能项目的一些特殊性，从业人员对于代码的管理普遍存在不够重视的情况。

我们在上一期的分享里提到过，数据科学和人工智能的很多项目，往往会被当作学术界的研究项目来进行开发。如果是研究项目，代码开发有哪些特点呢？我简单归纳了两大特点。第一，代码的主要目的是完成学术文章发表所需要的实验结果；第二，在绝大多数情况下，代码很容易变成无人维护和不能继续更新的情况。如果是当作研究项目来开发，那研究人员很可能并不在意代码的可读性、可维护性以及可扩展性等软件工程非常重视的方面，那么这样开发出来的代码就无法真正扩展为一个大型项目的代码库。

那么，对于一个人工智能项目的代码管理，我们需要去关注哪些因素呢？

第一，**所有的代码一定要保存在代码版本管理工具中**（而不是某一个数据科学家或者工程师自己的电脑上），这也是一个先决条件。在当今的软件开发的工具中，Git（或者是企业级的 GitHub）已经成为了一个标准的工具，用于代码版本管理、追踪和分享。任何项目以及任何人只要开始进行开发，都需要把所有的代码，包括核心的文档存放在代码版本管理工具中。这保证了代码能够被追踪并得到及时的备份。

第二，**如果我们利用 Git 或者类似的代码版本管理工具，代码的开发一定要尽量遵循这个版本管理工具所提倡的某种流程。**比如，我们有两个工程师在同一个项目中一起工作，在这

样的情况下，大部分的代码版本管理工具都可以允许这两个工程师在不同的“分支”（Branch）进行开发。这两个分支和项目当前的“主分支”（Master）又不同，因此，项目目前代码的运行不会受到这两个分支的影响，而这两个分支之间也不会互相影响。尽管进行分支开发已经算是软件开发的一个标准流程了，但是在人工智能项目中，依然有很多开发人员不遵循这个方法来进行代码不同开发进度之间的隔离。

开发部署环境流程

了解了代码版本管理的重要性之后，我们来看一看开发环境流程的控制。

从代码到可以被运行和部署的软件包，成为一个大型互联网产品或者人工智能产品中的一个组成部分，往往还有很多路要走。

首先，开发部署环境中一个重要的组成就是**流畅的从代码到部署软件包的“直通流程”**。目前在软件开发领域，有诸如“持续集成”（CI，Continuous Integration）和“持续交付”（CD，Continuous Delivery）这样的方法，来帮助工程师能够相对方便地对代码进行包装和部署。我们在这里并不去展开讨论 CI/CD 的内涵，但是要意识到，对于人工智能项目，我们也需要有能够流畅部署的思想。

那么，具体来说，哪个部分需要有流畅部署的思想呢？如果我们从大的角度来看，一个数据科学项目最需要动态更新的部分，往往是**模型的产生**，也就是说我们希望能够用最新的数据来训练和测试模型，让模型能够考虑到最新的用户行为。因此，就可以说，只要是对模型的产生流程有影响的步骤，都需要能够达到流畅部署。假如我们更新了模型产生的代码，这些代码必须能够快速地反应到生产系统中。

除了快速和持续部署，人工智能项目的另外一个重要需求就是**可以对代码的不同分支进行测试和运行**。也就是说，我们不仅仅需要能对“主分支”（Master）进行部署，还需要能够对不同的其他分支进行部署，从而能够无缝运行这些不同的分支。

这一点我们在开发环境中往往很容易忽视。举个例子，人工智能项目需要做大量的 A/B 测试，而这些测试中的“控制组”（Control）和“待遇组”（Treatment），往往就对应着代码中不同分支的开发成果。因此，在我们不清楚“待遇组”所对应的代码是不是能够真正带来好处之前，最好不要把这个分支和主分支进行合并。我们首先需要在线测试这个分支的效果，这就带来了运行不同分支的一个需求。

数据管理流程

最后我们来简单聊一聊数据管理流程。这也是从事数据科学项目我们最容易忽视的一个部分。

对于绝大多数的人工智能产品来说，**代码，也就是项目的业务逻辑，是和数据是密不可分的。**从某种意义上说，**我们应该把对数据的关注程度排在第一位。**因为即便有正确的代码，如果数据出现偏差，有时候哪怕是一点点小的偏差，都有可能对最后的结果（例如模型）产生重大的影响。因此，我们需要不断地强调数据质量的重要性。

那么，对于一个项目的数据管理，又有哪些方面需要注意呢？

在绝大多数的项目或者是产品中，数据的产生者、数据的运营者以及数据的使用者往往是不同的团队，这是数据管理的最大挑战。这种角色上的差别往往导致了对于数据质量的忽视。

举一个例子，如果数据的产生者是一个产品团队，在最近的一次软件更新中，一个工程师把旧代码中对数据进行追踪的部分主观臆断地删除了一些字段，或者是为了变量名好看，更改了字段的名字。如果这个更新没有通知数据的运营者或者是使用者，这往往会带来什么后果呢？后果是下游整个软件线的流程中，数据可能发生重大变化。严重的时候，这样的问题会导致模型发生完全异常，产生不可控的后果。

因此，**从“端到端”的思维来考虑数据链路是非常有必要的。**在数据的产生、运行和使用的链条上，所有的用户必须达成某种数据的 API，或者说是“共识”。任何对于数据的改变都需要在满足这种共识的基础上来沟通和进行。同时，数据的检测也是非常重要的，否则就是“垃圾进入导致垃圾输出”。

小结

今天我为你讲了数据科学团队的另外一个核心问题，那就是如何对工程流程进行管理。

一起来回顾下要点：第一，我们简单介绍了什么是人工智能项目的工程流程，以及这个概念和项目管理流程的区别；第二，我们分析了人工智能项目工程流程的三个主要方面，包括代码管理的要素，如何开发和部署环境，以及数据管理的要素。

最后，给你留一个思考题，除了我们所提及的工程流程的这三个方面，你还能想到什么其他的方面，在工程流程中也是至关重要的，需要我们在开发中注意呢？

欢迎你给我留言，和我一起讨论。

上一篇　137 | 如何做好人工智能项目的管理?

下一篇　139 | 数据科学团队怎么选择产品和项目?

精选留言 (2)

写留言

张胜斌
2018-09-13

我觉得还有文档的管理

展开

廉明
2018-08-31

数据这块管理能否展开讲一下

展开