02｜Sparrow RecSys：我们要实现什么样的推荐系统？

你好，我是王喆。

上节课，我们明确了推荐系统要解决的基本问题，清楚了深度学习推荐系统的技术架构，这节课我们开始走进实战。

作为程序员，我相信你肯定听过，甚至可能还很认同 Linux 之父 Linus Torvalds 的那句话“Talk is cheap.Show me the code.”。我也一样，所以只讲解理论知识不是这门课的风格，我希望你通过这门课的学习，不仅能构建出一棵深度学习推荐系统的知识树，还能动手实现出一个看得见、摸得着、能操作、能修改的推荐系统。

所以今天，**你跟着我的讲解，只需要花三十分钟的时间，就能将一套完整的深度学习推荐系统，Sparrow RecSys（随着课程的进行，我们会逐渐补充新的模块），在你自己的电脑上运行起来**。这也是我们这门课最终要实现的深度学习推荐系统。

废话不多说，直接运行

废话不多说，我们先把 Sparrow RecSys 安装运行起来。因为我已经把项目相关的所有代码（代码还会随着课程进行持续更新）、数据都整理到 GitHub 的开源项目中，所以你不需要额外安装任何的支持软件，也不需要额外下载任何数据。

这样，整个安装过程就跟“把大象装进冰箱“一样，只需要三步，就是打开冰箱门，把大象装进去，关上冰箱门。“翻译”成咱们的过程就是，从 GitHub 中 clone 代码，在本地以 maven project 的形式安装，

运行 RecSysServer 主函数启动推荐服务器。接下来，我们详细地解释一下这三个步骤。

首先，从 GitHub 中 clone 代码。这里，我直接给出了 Sparrow Recsys 开源项目的地址：https://github.com/wzhe06/SparrowRecSys。点击之后，你需要使用`git clone https://github.com/wzhe06/SparrowRecSys.git`命令，或者从 Web 端下载的方式，把代码下载到本地。

然后，你可以在本地以 maven project 的形式安装，也就是导入项目到 IDE。我推荐你使用 IntelliJ IDEA 为本项目的 IDE。这样，我们直接使用 IDEA，打开本地的 Sparrow Recsys 项目根目录就能导入项目。不过有一点需要注意，如果项目没有自动识别为 maven project，你还需要右键点击 pom.xml 文件，选择将该项目设置为 maven project 才能进行后面的操作。

最后，运行 RecSysServer。等到所有库文件自动下载完毕，项目编译完毕后，我们找到项目的主函数`com.wzhe.sparrowrecsys.online.RecSysServer`，右键点击运行。因为推荐服务器默认运行在 6010 端口，所以我们打开浏览器，输入`http://localhost:6010/`，就能看到整个推荐系统的前端效果了。

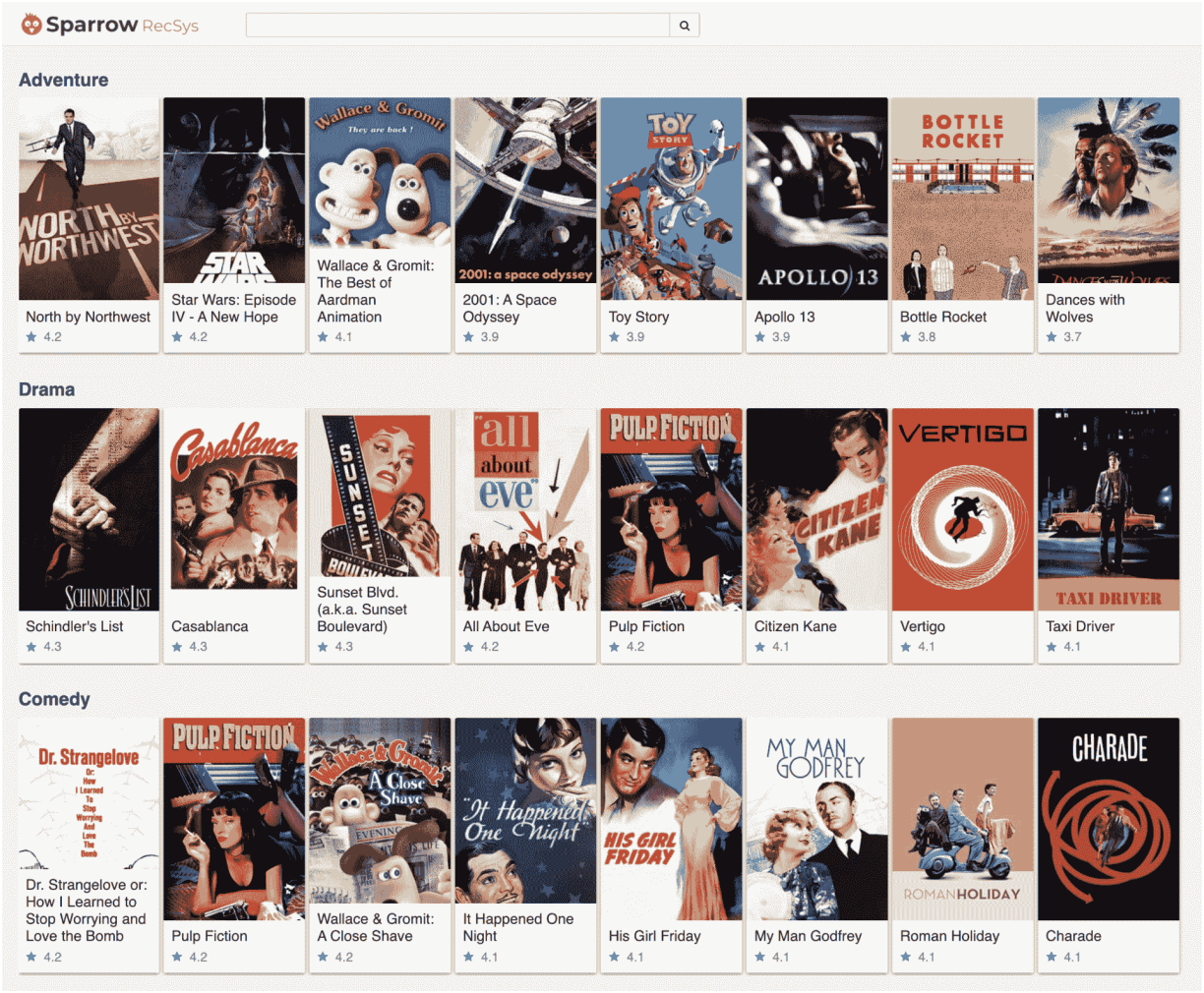

图1 Sparrow Recsys的主页

如果通过上面的步骤，你的浏览器显示出了由多个电影列表组成的 Sparrow Recsys 的主页，那么恭喜你，你已经拥有了这套深度学习推荐系统。

而且我相信，你把 Sparrow Recsys 这只“大象”装到自己冰箱里的时间，不会超过 30 分钟。但第一次见面的热情过后，你会不但想知其然，还想知其所以然，那接下来我就和你说说 Sparrow Recsys 的来历，以及功能和架构。而且在接下来的课程中，我会以它为例来给你讲透深度学习推荐系统。

“麻雀虽小，五脏俱全”的 Sparrow Recsys

Sparrow RecSys，全称 Sparrow Recommender System，中文名“**麻雀推荐系统**”，名字取自“**麻雀虽小，五脏俱全**”之意。

你第一眼见到它，可能认为它像个 Demo 或者玩具。虽然它不可能真正具备一个工业级深度学习推荐系统的全部功能，但我希望它是一颗能够成长为参天大树的种子，一只未来有可能大鹏展翅的雏鸟。在投入一定的精力改造、拓展之后，它甚至有可能支撑起一个规模互联网公司的推荐系统框架。这就是我设计 Sparrow RecSys 的初衷。我也希望你能够在实现 Sparrow RecSys 的过程中，快速领略深度学习推荐系统的主要模块和主流技术，并且找到乐趣、找到成就感。

那么 Sparrow Recsys 到底实现了哪些功能呢？它又包含了哪些深度学习推荐系统的关键技术呢？下面，我会为你一一讲解。

Sparrow Recsys 的功能有哪些

Sparrow RecSys 是一个电影推荐系统，视频推荐是我最熟悉的领域，这也是我以电影推荐作为切入点的原因。像所有经典的推荐系统一样，它具备“相似推荐”“猜你喜欢”等经典的推荐功能，在页面设置上，主要由“首页”“电影详情页”和“为你推荐页”组成。

首先，是 Sparrow RecSys 的首页。

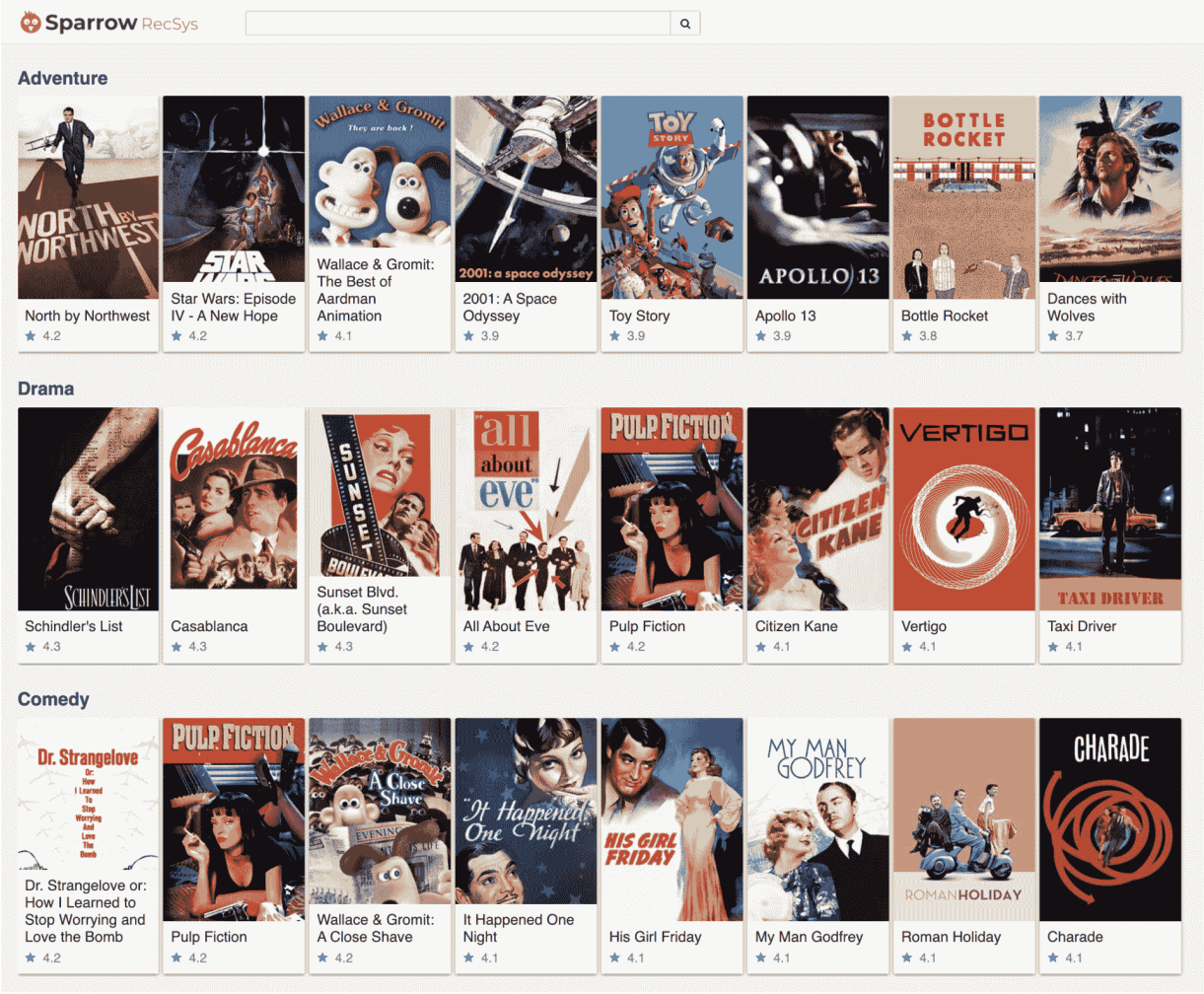

图2 Sparrow RecSys的首页

Sparrow RecSys 的首页由不同类型的电影列表组成，当用户首次访问首页时，系统默认以历史用户的平均打分从高到低排序，随着当前用户不断为电影打分，系统会对首页的推荐结果进行个性化的调整，比如电影类型的排名会进行个性化调整，每个类型内部的影片也会进行个性化推荐。

其次，是电影详情页。

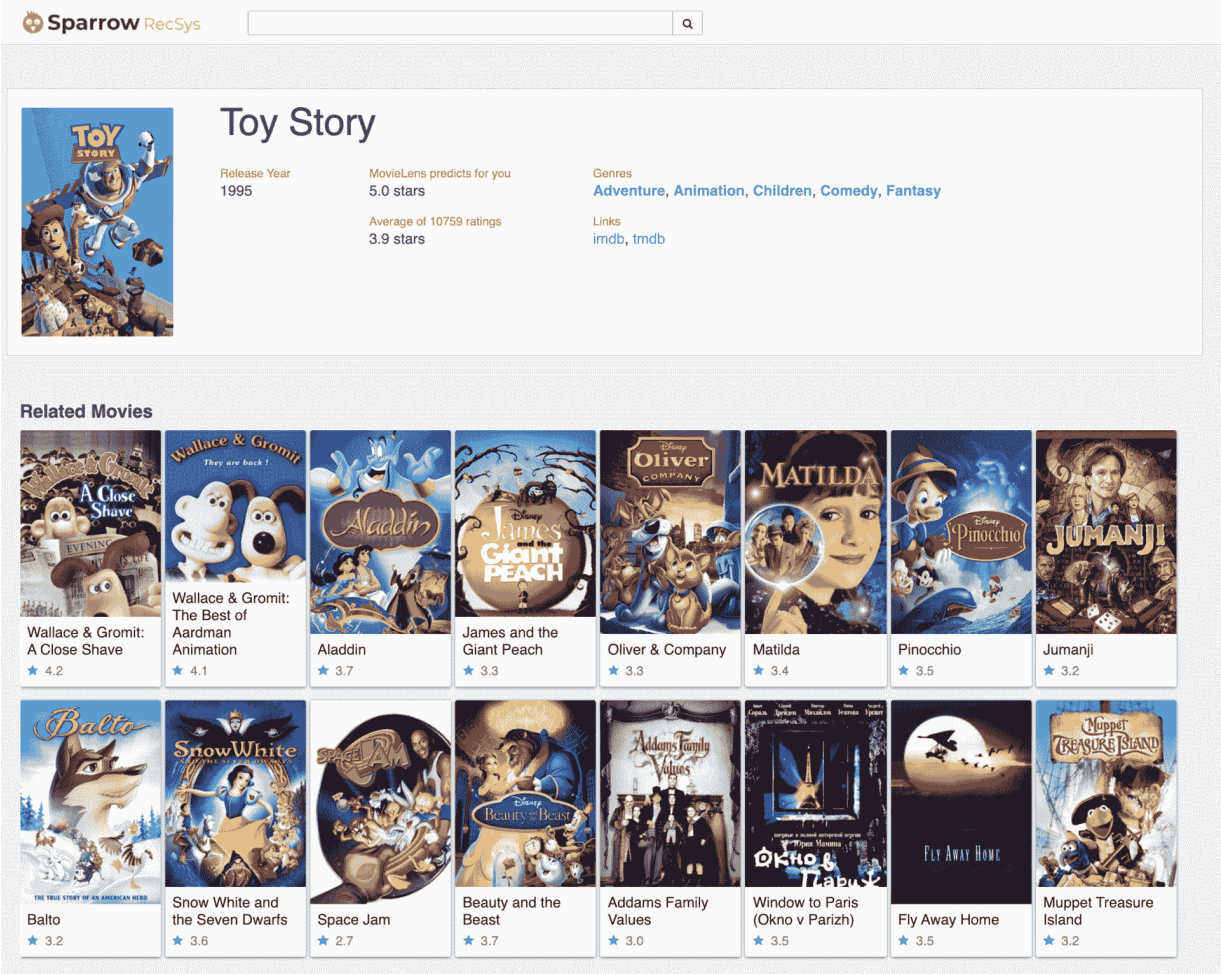

图3 电影详情页

你可以看到电影详情页除了罗列出电影的一些基本信息，最关键的部分是相似影片的推荐。相似内容推荐是几乎所有推荐系统非常重要的功能，传统的推荐系统基本依赖于基于内容（Content based）的推荐方法，而我们这门课程会更多地讲解基于深度学习 Embedding 的相似内容推荐方法。

最后，是为你推荐页。

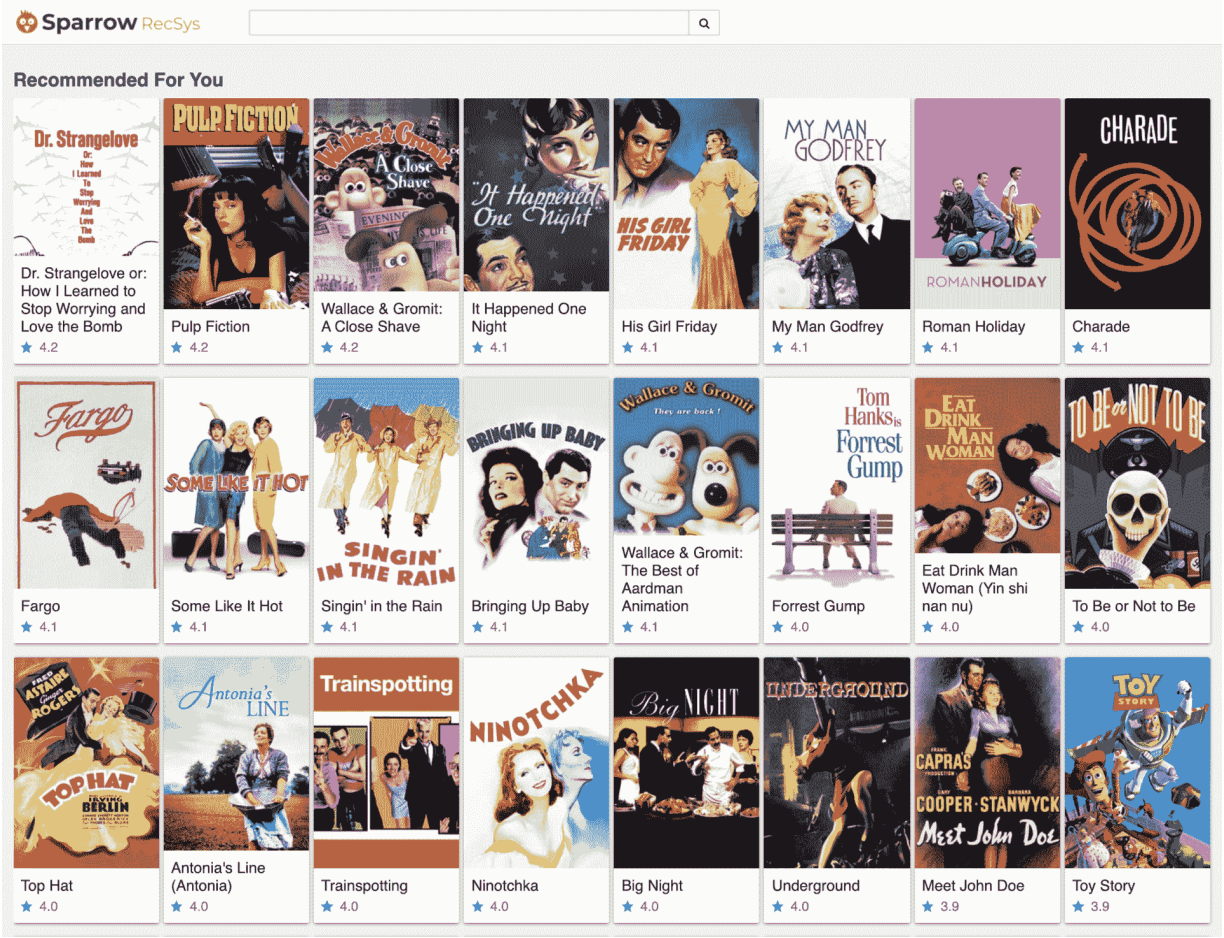

图4 为你推荐页

这一部分也是整个推荐系统中最重要的部分，是用户的个性化推荐页面。这个页面会根据用户的点击、评价历史进行个性化推荐。这几乎是所有推荐系统最经典和最主要的应用场景。我希望在这门课程中，你能够动手完成个性化推荐中的每个关键步骤，包括但不限于特征的处理、候选集的召回、排序层主要模型等等。

Sparrow Recsys 的数据从哪来？

知道了 Sparrow RecSys 的功能之后，你肯定想问，“老师，咱们的数据从哪来呀？”。既然 Sparrow RecSys 是一个开源项目，那么

Sparrow RecSys 的数据源肯定也是开源和免费的，它的数据源来自于著名的电影开源数据集movieLens。

为了方便你调试，咱们这门课程的教学数据集对 movieLens 数据集进行了精简，只留下了 1000 部电影。如果希望在全量数据集上进行推荐，你可以去 MovieLens 的官方网站下载全量数据，它一共包含了 27000 部电影。

MovieLens 的数据集包括三部分，分别是 **movies.csv（电影基本信息数据）、ratings.csv（**用户评分数据）**和 links.csv**（外部链接数据）。下面，我就具体说说它们分别长什么样。

1. movies.csv（电影基本信息数据）

movies 表是电影的基本信息表，它包含了电影 ID（movieId）、电影名（title）、发布年份以及电影类型（genres）等基本信息。

movieId	title	genres
1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama|Romance
5	Father of the Bride Part II (1995)	Comedy
6	Heat (1995)	Action|Crime|Thriller
7	Sabrina (1995)	Comedy|Romance
8	Tom and Huck (1995)	Adventure|Children
9	Sudden Death (1995)	Action
10	GoldenEye (1995)	Action|Adventure|Thriller

图5 电影基本信息数据

MovieLens 20M Dataset 包含了 2016 年前的约 13 万部电影，我们课程的实验数据集从中抽取了前 1000 部电影。电影数据集是我们推荐

的主体，其中分类、发布年份、电影名称等信息也将是推荐模型可以利用的重要特征。

2. ratings.csv（用户评分数据）

ratings 表包含了用户 ID（userId）、电影 ID（movieId）、评分（rating）和时间戳（timestamp）等信息。

userId	movieId	rating	timestamp
1	2	3.5	1112486027
1	29	3.5	1112484676
1	32	3.5	1112484819
1	47	3.5	1112484727
1	50	3.5	1112484580
1	112	3.5	1094785740
1	151	4.0	1094785734
1	223	4.0	1112485573
1	253	4.0	1112484940
1	260	4.0	1112484826

图6 用户评分数据

MovieLens 20M Dataset 包含了 2000 万条评分数据，我们课程的实验数据集从中抽取了约 104 万条评论数据。评论数据集是之后推荐模型训练所需的训练样本来源，也是我们分析用户行为序列、电影统计型特征的原始数据。

3. links.csv（外部链接数据）

links 表包含了电影 ID（movieId）、IMDB 对应电影 ID（imdbId）、TMDB 对应电影 ID（tmdbId）等信息。其中，imdb 和 tmdb 是全球最大的两个电影数据库。因为 links 表包含了 movieLens 电影和这两个数据库 id 之间的对应关系，所以，我们可以根据这个对应关系来抓取电影的其他相关信息，这也为我们大量拓展推荐系统特征提供了可能。

movieId	imdbId	tmdbId
1	114709	862
2	113497	8844
3	113228	15602
4	114885	31357
5	113041	11862
6	113277	949
7	114319	11860

图7 外部链接数据

此外，MovieLens 的数据集中还包含了 tags.csv，它用于记录用户为电影打的标签，由于课程中暂时没有使用标签数据，我就展开说了。

Sparrow Recsys 涵盖的技术点

清楚了 Sparrow Recsys 的功能和数据，你肯定迫不及待地想知道 Sparrow Recsys 会使用哪些技术，可以实现哪些模型。

那我们直接来看下面这张 Sparrow Recsys 的技术架构图。你会发现，它其实就是我们用具体的技术选型，把上节课的深度学习推荐系统架构图给填上得到的。所以，Sparrow Recsys 就是深度学习推荐系统架构的一个实现。

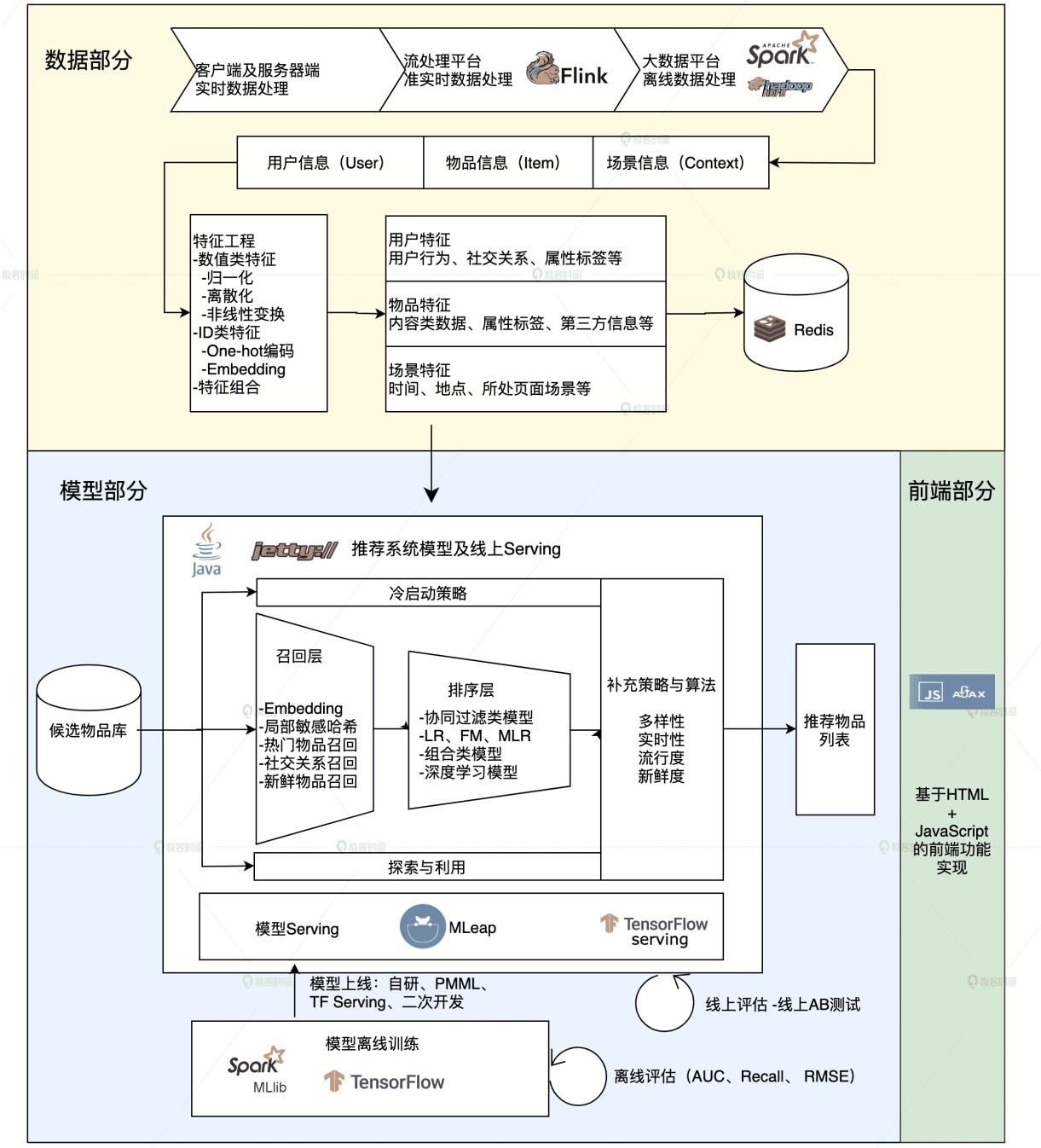

图8 Sparrow Recsys的推荐系统架构

你可以看到，它一共分为三个模块，分别是数据、模型和前端。其中每个部分都用业界推荐系统的主流技术，比如数据部分我们会用 Spark，Flink 进行样本和特征的处理，模型部分我们会使用 TensorFlow 训练深度神经网络、Wide&Deep、PNN 等模型。

模块	技术点	涉及工具
数据部分	Spark进行批量特征处理	Spark
	Flink进行流式数据处理	Flink
	Redis保存线上推荐服务器所需特征	Redis
	HDFS保存离线训练所需训练样本和特征	HDFS
模型部分	使用Spark MLlib训练Embedding和传统推荐模型	Spark MLlib
	使用TensorFlow训练深度学习推荐模型	TensorFlow
	使用MLeap、TensorFlow serving进行模型上线和在线推断	MLeap、TensorFlow serving
	使用Jetty搭建推荐服务器	Jetty
前端部分	使用简单的HTML和JavaScript实现前端用户体验部分	HTML、JavaScript

图9 Sparrow Recsys中的技术点

我想啊，你在看到这么多的技术点和技术平台之后，肯定想问，我们的课程能把它们都讲完、讲透吗？这是个好问题，我也有必要在这里说清楚。

从中，我们可以总结出 Sparrow Recsys 中具体用到的技术点。

其实推荐系统是一个应用属性很强的领域，想把推荐系统学好，我们就必须去学习各式各样相关的平台、技术，所以我们这门课的涉及面非常广。但你也不用因为要学这么多的技术而感到惊慌，因为我们没

有必要去深究每个平台内部的原理、优化的方法，我们当好一个使用者就好。

举个例子你就明白了，我们处理数据需要用到 Spark，但我们有必要成为 Spark 的专家吗？其实不用。因为即使你已经走上了工作岗位，也有平台架构部的同事能够提供 Spark 的很多技术支持。所以学习这门课程，我们大可抱着一个使用者，而不是开发者、维护者的心态去使用不同的技术平台。当然，如果你想成为某个细分方向的专家，比如 Spark 的专家、Flink 的专家等等，我相信极客时间上肯定还有很不错的课程供你学习。

所以希望你能够通过 Sparrow Recsys 认识到主流深度学习推荐系统都使用了哪些技术，让自己有一个全面的认识，建立自己的知识广度。如果还想深入钻研某个方向，也可以由此开始，努力成为一个领域的专家。

小结

这堂课，我带你熟悉了我们将要实现的推荐系统 Sparrow Recsys，它将是我们深度学习推荐系统这门课的落地项目和实现范例。希望有这个真实可用的推荐系统作为支撑，这门课可以同时兼顾概念讲解和代码实战，也让我们接下来的共同合作能够更好。

从开篇词到这一节课，我们从推荐系统要解决的核心问题，生发出深度学习推荐系统的技术架构，再到让技术架构实实在在地落地到 Sparrow Recsys 这个开源项目上。我想你已经可以感受到架构篇的学习过程，其实就是一个从抽象到具体，从形而上到形而下的过程。

那在搭建起这整门课程的框架之后，接下来我们将会一起深入到技术细节，以及深度学习的实践中，一起去体验深度学习浪潮之巅的推荐

系统知识，期待继续与你同行！

课后思考

1. 当你把 Sparrow Recsys 在自己的电脑上安装运行起来之后，对照着上节课的深度学习推荐系统架构图，你能试着说出每个模块的代码属于架构图中的哪一部分吗？
2. 你觉得对于一个电影推荐系统来说，什么数据对生成用户个性化推荐结果最有帮助？

好啦，快按照这节的方法把 Sparrow Recsys 运行起来吧！课后的两个问题也并不困难，相信你肯定可以回答出来。今天就讲到这里了，我们下节课见！