利用垂直网站抽取结构化数据构建知识图谱

利用豆瓣构建影视领域知识图谱：

主要有以下几大类本体：

1. 电视剧（只考虑已上映的电视剧，因为未上映的电视剧很多属性关系缺失），含有属性：类型，制片国家/地区，语言，首播，集数，单集片长，评分，主页链接（主码）
2. 电影（只考虑已上映的电影，因为未上映的电影很多属性关系缺失），含有属性：类型，制片国家/地区，语言，上映日期，片长，评分，主页链接（主码）
3. 演员，含有属性：性别，星座，出生日期，出生地，影人简介，主页链接（主码）
4. 导演，含有属性：性别，影人简介，主页链接（主码）
5. 编剧，含有属性：性别，影人简介，主页链接（主码）
6. 制片人，含有属性：性别，影人简介，主页链接（主码）

实体间的关系有：

实体1---关系--->实体2

1. 导演---导演--->电视剧
2. 导演---导演--->电影
3. 编剧---编剧--->电视剧
4. 编剧---编剧--->电影
5. 制片人---制片--->电视剧
6. 制片人---制片--->电影
7. 演员---主演--->电视剧
8. 演员---参演--->电视剧
9. 演员---主演--->电影
10. 演员---参演--->电影

几个关键问题：

1. 如何区分主演和参演？

在演员一栏，取前四为演员为主演，其余则为参演。如下图，琅琊榜之风起长林，黄晓明、刘昊然、佟丽娅、张慧雯为主演，其余为参演



1. 如何区分同名实体？

利用实体主码区分，即主页链接

1. 如何区分演员、导演、编剧、制片人？

利用实体主页的职业属性值区分，若实体的职业属性值存在多个职业，则取第一个属性值作为其本体，如下图，孔笙，同时存在“导演 / 演员 / 摄影 / 制片人 / 剪辑”多个职业，取导演作为他的本体



因为以导演作为孔笙的本体，所以在抽取本体属性时，只抽取导演的属性，即性别、影人简介、主页链接。同时，因为孔笙又是演员和制片人，所以在将该实体存入neo4j数据库时，除了为其赋予导演label外，也为其赋予演员label和制片人label。

1. 如何抽取实体属性？

利用实体主页的结构化数据抽取，如下图，琅琊榜之风起长林，类型为“剧情/古装”，语言为“汉语普通话”。



1. 如何抽取关系？

利用电影电视剧实体的“全部演职员”页的结构化数据抽取，如下图，则值琅琊榜之风起长林的导演为孔笙、李雪，主演为黄晓明、刘昊然、佟丽娅、张慧雯。



1. 电影电视剧的演职员中有些演员、编剧等没有主页，怎么办？

没有主页，说明不存在该实体，则删除这条关系（关系承载的实体必须存在，不能指向空实体）

1. 有些实体的属性缺失，怎么办？

直接使用“无”

构建知识图谱流程：

1. 抽取实体，相同类的实体存储在一个csv文件中，形成如 actor.csv、director.csv、writer.csv、movie.csv等文件。以actor.csv文件为例，实体csv文件的格式如下图所示：



因为周星驰同时既是演员，又是导演、编剧，所以actor.csv文件的:LABEL列有3个值：演员;导演;编剧。因为周星驰职业属性的第一个值是演员，所以将周星驰这个实体存在了actor.csv文件中。

1. 抽取关系，想同类的关系存储在一个csv文件中，形成如：direct.csv、lead.csv等文件。以lead.csv文件为例，关系csv文件的格式如下图所示：



1. 对关系文件中的每条关系中的每个实体，根据实体主码，判断是否存在于实体文件中，若该条关系中有一个实体不存在于实体文件中，则删除该条关系。注意：在判断关系中的实体是否存在于实体文件中时，可将所有的实体主码存储在一个字典中，然后判断关系中的实体是否存在于字典中，不可将实体主码存储在列表中，因为字典的底层存储结构是哈希表，所以判断关系中的实体是否存在于字典中，复杂度为O(1)，而判断关系中的实体是否存在于列表中，复杂度为O(n)。
2. 将所有实体csv文件和关系csv文件导入到了neo4j数据中，形成知识图谱。