数据仓库第一次作业汇报
—ETL练习

课程名称：　数据仓库

授课教师：　朱宏明

小组成员：陈丹怡 1851475

李一凡 1852144

杜玫 1853316

谭忠煜 1853898

软件学院　软件工程 专业　2018　级

数据仓库第一次作业汇报

—ETL练习

工作简介

作业要求

1）获取用户评价数据中的253,059个ProductID
2）从Amazon网站中利用网页中所说的方法利用爬虫获取253,059个Product页面
3）挑选其中的电影页面
4）分析其中不同的电影一共有多少部

解决方案概述

1）获取用户评价数据中的253,059个ProductID

使用movie.txt文本文件作为数据来源的电影评论信息，每条评论对应了一个Amazon用户和一个电影，评论包括了打分和评论内容等信息，共计七百余万条，文本大小超过9G。我们通过编写运行python脚本的方式，抽取每个评价对应的ProductID，之后使用python自带的Set数据结构进行去重，便得到了253059个不重复的ProductID。

2）从Amazon网站中利用网页中所说的方法利用爬虫获取253,059个Product页面

将上一步获得的ProductID，加载入我们用python编写的爬虫脚本，运行访问Amazon网站下该电影对应的信息页面，并从中爬取Amazon网站下这些电影的详细信息，并保存为包含Product特征数据的csv格式文件。

3）挑选其中的电影页面

首先，为了使Product特征数据便于后期处理，我们将上一步获得的CSV文件导入到PDI中进行数据清洗，去除特征中包含的中文字符和乱码、去除Product名称中的各种括号和其中的内容、合并导演和作者。

其次，将PDI清洗好的Product特征数据加载入我们编写的python脚本中进行电影和非电影的分类。为了区分一个Product是否是一部电影，我们依据电影与非电影和movie.txt中的评论数据以及Product的四个特征：名称、类型、格式、时间的关联关

系，生成了一套划分电影的详细标准。使用这套划分标准将Product特征数据分成了非电影Product特征和电影Product特征两个csv格式文件

4）分析其中不同的电影一共有多少部

为了将电影Product进行分类，我们依旧编写了python脚本。脚本会将每两个ASIN对应的电影Product进行相似度分析，生成相似系数。然后根据相似系数是否大于一个预设值来确定两个Product是否应归为一个电影。相似度分析是体现我们分类思想的核心，之后会有详细说明。

之后我们将Product之间形成的“同一电影”关系用无向图表示，邻接表存储。为获得所有的电影和其下属的Product，我们使用BFS对邻接表进行搜索，从而获得所有的连通分量，即所有的电影。

Product特征数据爬取

・初期我们首先考虑了爬虫软件，在试用过市面上的几种产品后发现，绝大多数爬虫软件都不可避免地有着功能收费、内存泄漏、速度缓慢等情况，于是我们转向了编写python爬虫脚本。

・在写脚本时我们首先尝试了request库进行爬取，通过更改浏览器代理头和http代理头等相关头部参数来尝试规避验证码，然而发现无效，验证码问题无法解决。

・最后我们使用了Selenium库，一个自动化网页测试库，通过控制浏览器来爬取数据，验证码弹出的频率大幅度降低，有效提高了爬虫的成活率。我们的验证码处理策略是通过URL队列来实现的，如果访问一个URL返回验证码，便跳过该URL，并将其入队。脚本循环访问该队列，直到访问到队列末尾，程序退出。

・脚本使用bs4库构建HTML解析树来分析爬取到的网页HTML文件，我们选择了数据缺失率小于10%或者对后期数据仓库构建必要的特征进行了提取。

・提取的特征有：Name, ASIN, ReleaseTime, Time, Actor, Director, Writer, Type, Comments, Format, Cost, Grade

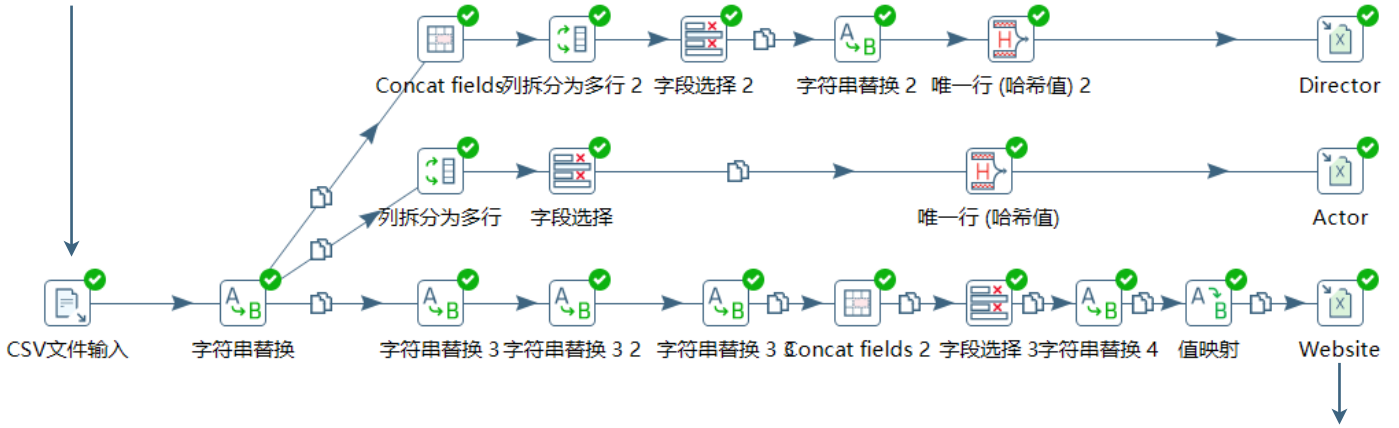
Order	URL	Name	ASIN	ReleaseTime	Time	Director	Writer	Actor	Type	Comments	Format	Cost	Grade
1	amazon.com/d	Dora's Christmas Carol Adver	B006IVJYH0	2009	48min	George Chialtas,Allan Jacobsen,Henry		Caitlin Sanchez,Caitli	Kids	183	Prime Video (str	$9.99	4.6
2	amazon.com/d	Mc5: Sonic Revolution: A Cel	B001J31PAG	2003	56min	Dom Phillips		Dennis Thompson,Mi	Music Videos and (	1	Prime Video (streaming on		2
3	amazon.com/d	Gb: Perfect School	B000G6BLCE	5-Sep-06	22 minutes		R.L. Stine	R.L. Stine, Kathryn Sl	Studio Specials	92	DVD	$7.55	4.7
4	amazon.com/d	Crime & Punishment - The Cc	B0000A5A1L	30-Sep-03	3 hours and 45 minutes			Francesca Gerrard, F	Genre for Featured	25	DVD	$16.99	3.3
5	amazon.com/d	Goosebumps: My Best Frienc	B000G6BLCY	5-Sep-06	22 minutes		R.L. Stine	R.L. Stine, Kathryn Sl	Studio Specials	9	DVD	$10.48	
6	amazon.com/d	The Wonderful Wizard of Oz	B0001O3YUI	1-Jun-04	1 hour and 30	Tim Reid	Don Arioli, L. Franl	Margot Kidder, Morga	Science Fiction & F	3	DVD		4.5
7	amazon.com/d	Tiny Planets: Bing Bong Bell [	B0000A5A1I	14-Oct-03	52 minutes	Alastair McIlwain		Carrigan van de Merv	Genre for Featured	28	VHS Tape	$9.49	4.4
8	amazon.com/d	Tiny Planets: Bing Bong Bell a	B0000A5A1J	14-Oct-03	52 minutes	Alastair McIlwain		Carrigan van de Merv	Genre for Featured	28	DVD	$8.84	4.4
9	amazon.com/d	Guitar Instruction Delta Blues	B0012GSYL8	1-Jan-09	1 hour and 43 minutes			Guitar Instruction Del	Musicals & Perform	1	DVD		3
10	amazon.com/d	Tiny Planets: Making Rainbov	B0000A5A1F	14-Oct-03	52 minutes	Alastair McIlwain		Carrigan van de Merv	Genre for Featured	12	VHS Tape	$9.49	4.8

- 注：253059个URL爬取到了251039条数据

电影类型Product提取

• 首先，为了使Product特征数据便于后期处理，我们将上一步获得的CSV文件导入到PDI中进行数据清洗，去除特征中包含的中文字符和乱码、去除Product名称中的各种括号和其中的内容、合并Director和Writer为SingleDirector。同时输出另外两个表哥包含所有的Actor和Director 。PDI处理完成后使用python脚本规范化了ReleaseTime和Time，并增加了Year、Month、Day 三个特征。

Order	URL	Name	ASIN	ReleaseTime	Time	Director	Writer	Actor	Type	Comments	Format	Cost	Grade
1	amazon.com/d	Dora's Christmas Carol Adver	B006IVJYH0	2009	48min	George Chialtas,Allan Jacobsen,Henry		Caitlin Sanchez,Caitli	Kids	183	Prime Video (str	$9.99	4.6
2	amazon.com/d	Mc5: Sonic Revolution: A Cel	B001J31PAG	2003	56min	Dom Phillips		Dennis Thompson,Mi	Music Videos and (	1	Prime Video (streaming on		2
3	amazon.com/d	Gb: Perfect School	B000G6BLCE	5-Sep-06	22 minutes		R.L. Stine	R.L. Stine, Kathryn Sl	Studio Specials	92	DVD	$7.55	4.7



Order	URL	Name	ASIN	ReleaseTime	Time	Actor	Type	Comments	Format	Cost	Grade	Year	Month	Day	Director
1	amazon.con	Gumnaam	B000IZJR3I	6/4/2010	143	Dhumal,Helen,	Genre for Featured	33	DVD	$6.98	3.4	2010	4	6	Raja Nawathe,
2	amazon.con	Venus Versus Virus: T	B0023S4A12	16/6/2009	290	Joanne Bonass	Genre for Featured	45	DVD	$24.99	4.3	2009	6	16	Kyle Jones, Scott McCle
3	amazon.con	Show Off! How to Be	6300157423	1/1/1998	30	Malcolm Jama	Genre for Featured	1	VHS Tape	$64.98	5	1998	1	1	Jim Hirschfeld,

◆ 注：因为重新排序的关系，案例中的order相同的行没有对应关系。

• 为了区分一个Product是否是一部电影，我们依据电影与非电影和movie.txt中的评论数据以及Product的四个特征：名称、类型、格式、时间的关联关系，生成了划分电影的详细标准。

• Step1 根据Product的一级目录（额外爬取的数据）划分出一级目录非“Movie&TV”的ISBN，认为这些ISBN对应的Product一定不是电影。

• Step2 根据movie.txt中的产品评论信息，统计一个产品中电影关键字和非电影关键字的出现次数，如果电影关键字出现次数总和小于非电影关键字出现次数总和，则认为这些ISBN对应的Product大概率不是电影。

电影关键字：Movie(s), theater(s), Hollywood, Bollywood。
非电影关键字：
（教程）to teach you, instructor(s), instruction(s), technique(s)
（连续剧）episode(s), season, series
（纪录片）documentary, doc, BBC, series

• Step3 根据Product名称，如果Product名称中含有 Analysis of , technique ,Collection 则认为该Product大概率不是电影

• Step4 根据Product时长，如果Product时长小于30min或者大于300min，则认为该Product大概率不是电影

• Step5 根据Product类型（二级目录），通过分析Amazon Product目录结构可知，如果Product类型为Bollywood ， Movies ，则认为该Product一定为电影。如果Product类型字段中包含 Boxed Sets，TV，Exercise，Special Interest，PBS以及CDs&Vinyl下的类型，则认为该Product一定不是电影。

• Step6 根据Product格式，如果Product 格式为Audio CD，则认为该Product一定不是电影

• Step7 根据之前的6个步骤，对于一个Product，可能同时存在6种判断情况和相应的处理方法
[ ]：没有判断结果，Product是电影
[可能不是]：Product不是电影
[可能不是，一定是]：Product是电影
[一定不是]：Product不是电影
[一定不是，一定是]：Product不是电影（矛盾情况，默认去除该Product）
[一定是]：Product是电影

Order	URL	Name	ASIN	ReleaseTime	Time	Actor	Type	Comments	Format	Cost	Grade	Year	Month	Day	Director
1	amazon.com	The Virgin of Juarez	B003AI2VGA	XX/XX/2006	89	Minnie Driver,A	Drama,Suspense	16	Prime Video (streaming onli		3.2	2006			Kevin James Dobson,
2	amazon.com	Far from Home: The	B00004CQT3		81	Jesse Bradforc	Genre for Featured	557	VHS Tape	$6.95	4.6				Phillip Borsos,Phillip Bors
3	amazon.com	Who's the Man?	B00004CQT4		85	Ed Lover, Doct	Genre for Featured	276	VHS Tape	$2.99	4.8				Ted Demme,Doctor Dr, E
4	amazon.com	My Kingdom	B0078V2LCY	XX/XX/2011	99	Geng Han,Bart	Drama	5	Prime Video (streaming onli		3.7	2011			Xiaosong Gao,
5	amazon.com	Reaper	B003ZG3GAM	XX/XX/2008	0		Drama	242	Prime Video (streaming onli		4.7	2008			
⋮															
204648	amazon.com	D-Day: Code Overlor	B000NQRR0M		0	NULL	Movies	4	DVD	$10.99		0	0	0	NULL
204649	amazon.com	The Star Packer	B000HL2PR8	XX/XX/1934	55	John Wayne,Ve	Western	46	Prime Video (strea	$19.99	3.9	1934	0	0	Robert N. Bradbury,
204650	amazon.com	King of the Ants	B002ZBPPE8	XX/XX/2003	102	Chris McKenna	Horror	55	Prime Video (strea	$0	3.5	2003	0	0	Stuart Gordon,
204651	amazon.com	Happiness Runs	B004D8EW3Q	XX/XX/2010	89	Hanna Hall,Jes	Drama	12	Prime Video (strea	$19.99	2.8	2010	0	0	Adam Sherman,
204652	amazon.com	The Band's Visit	B001DNCBPU	XX/XX/2008	87	Ronit Elkabetz,	Comedy,Internatior	224	Prime Video (strea	$12.99	4.3	2008	0	0	Eran Kolirin,
204653	amazon.com	Zelig	B00006BT6B		0	NULL	Movies	188	DVD	$13.84	4.6	0	0	0	NULL

- 电影类型Product流程共提取出204653个电影类型Product，除去46386个非电影类型Product

电影Product聚类分析

・Python脚本会将每两个ASIN对应的电影Product进行相似度分析，生成相似系数。然后根据相似系数S是否大于一个预设值来确定两个Product是否应归为一个电影。

・$S = S(N) * S(D) * S(A) * S(T)$

・S(N) 表示两个Product名称的相似度。采用了 Jaro-Winkler距离作为两个字符串相似度的评判指标。Jaro-Winkler距离适合于如名字这样较短的字符之间计算相似度。

$$S(N) = Jaro - WinklerDistance(str1,str2)$$

・S(D) 表示两个Product导演的相似度。采用枚举导演姓名然后统计导演重合人数Sd，Product1的导演人数为N1，Product2的导演人数为N2，C1表示除C2外的情况，C2表示Product1或Product2的导演为空，或一方是另一方的子集。

$$S(D) = \begin{cases} \frac{Sd * 2}{N1 + N2} & \text{C1} \\ 1 & \text{C2} \end{cases}$$

・S(A) 表示两个Product演员的相似度。采用枚举演员姓名然后统计演员重合人数Sa，Product1的演员人数为N1，Product2的演员人数为N2，C1表示除C2外的情况，C2表示Product1或Product2的演员为空，或一方是另一方的子集。

$$S(A) = \begin{cases} \frac{Sa * 2}{N1 + N2} & \text{C1} \\ 1 & \text{C2} \end{cases}$$

・S(T) 表示两个Product时长的相似度。Product1和Product2的时间长度分别为T1和T2。C1表示除C2外的情况，C2表示Product1或Product2的时间长度为0。

$$S(T) = \begin{cases} 1 - \frac{|T1 - T2|}{Max(T1,T2)} & \text{C1} \\ 1 & \text{C2} \end{cases}$$

• 根据相似系数S是否大于一个预设值Sense来确定两个Product是否应归为一个电影。

• 之后我们将Product之间形成的“同一电影”关系用无向图表示，邻接表存储。为获得所有的电影和其下属的Product，我们使用BFS对邻接表进行搜索，从而获得所有的连通分量，即所有的电影。