改后的实验计划

chenlin

to 蔡敏毅 2015-01-23 17:31 Hide

From：chenlin<chenlin@xmu.edu.cn>
To：蔡敏毅<mrvege@stu.xmu.edu.cn>
Date：January-23-15 (Fri) 17:31
Size：128 KB

Modification.docx (44 KB)

敏毅：

这是改后的整体框架图，你理解一下。不懂的问我。

要改动的地方比较多，你预估一下工作量和可行性，并给我反馈。在今晚之前确定所有实验方案，我好写文字部分。

一，每个评论的标签分类，目前的分类效果基本都很差，拟避开这一块，改为
1，不使用任何商品特征，只使用情境特征，即李培的时间、地点、对象、价格、菜式。即b向量的维度现在变小了，不包含餐馆特征了。
词库使用word2vec扩展
2，做一次过滤，不管是正例负例，只选择总体情感是正的评论。在实验中你可以选择跑一个情感分类器（对全文），也可以直接选择评分大于餐馆平均值的评论，或者用模式过滤如”适合/合适……“
”XXX还可以“，以时间定。
3，重点在这里：
不管评论是否有标签，我们的假设是：用户出去吃饭必定有且仅有一个context。因为我们要用评论来学a和b，所以就有4种情况
3.1 多个标签，不会影响a的学习，但会影响b
例如 “朋友请客来的 坐在了靠窗边的位子，这家店的采光比sogo店要好多了，但还是更喜欢sogo店的氛围。 这次点的小牛肋眼，也不错，很嫩 ”。标签是“情侣约会”、“朋友聚餐”
”情侣约会“这个标签会影响”朋友“这个特征的权重
3.2 没有标签，影响a和b的学习
3.3 一个标签，但是标错了，影响b，但不影响a
3.4 一个标签，也标对了，不影响a和b
结合不平衡问题，拟采用采样方法构造训练集。首先收集一个餐馆下所有评论的x分布，设其为background distribution p_b，而该评论当前的标签为empirical distribution p_e。显然，如果一个餐馆的评论越多，应该越接近于p_b。所以实际上一个评论被收集为x的正例的概率是\alpha p_b +(1-\alpha) p_e，\alpha=1/(1+e^{-\betaN})，N为该餐馆的评论数，\beta为参数可调
例如 “朋友请客来的 坐在了靠窗边的位子，这家店的采光比sogo店要好多了，但还是更喜欢sogo店的氛围。 这次点的小牛肋眼，也不错，很嫩 ”。标签是“情侣约会”、“朋友聚餐”。假设该餐馆所有评论中六个类别（情侣约会、朋友聚餐、……）的比例分别是5：1：1：1：1：1，则该评论被抽为情侣约会的概率是0.5\alpha +0.5 (1-\alpha).当然，对于每个标签都需要重新采样正例

学习过程不变
可以直接使用上次本科生标注的作为答案。如果结果不理想，请调一下答案。但是注意：这回不能用点评系统自带的标签了，因为我们已经声明该标签不可靠，必须是人工标注的。
你需要给我的实验结果：
每个场景下最好的情境特征

Table 1: Top contextual factors with largest b_k^x

Factor	Banquet	Party	Family	Dating	Leisure
Time	noon	afternoon	noon	weekend	weekend
Loc.	E.40th St.	Houhai	Zoo	Sanlitun	Houhai
Comp.	boss	friend	family	family	family
No. Comp.	5-10	3-4	3-4	2	3-4
Price	200-500	50-80	200-500	80-120	120-200
Cuisine	roast duck	hotpot	Russian	ryori	fastfood

比较实验（原来的基本都可以保留，因为我找到了相关文献，可以当作是复现了别人的实验。但需要换成现在的标注集重新评估一下。）
\beta取值对结果影响

二，商品特征的权重在不同情境下的学习
1，由于一，3描述的问题，首先对于第二部分的正例负例选择是基本按照原样的。多标签、一个标签的都可以放进去。没有标签的需要经过步骤一的学习得到标签后放入。
2，商品特征不等同于评论中出现的特征，并且商品特征是在不同特征极性上的概率分布。如下
餐馆c：菜品+ 0.8 菜品- 0.2
3，我们只考虑评论中商品特征，不考虑情境特征，且只考虑为正的特征。词库不必要做word2vec扩展。
4，具体学习过程我晚上写给你。公式推导会有一点不同。
你需要给我的实验是：
每个场景下最高的商品特征

Table 2: Top 10 commodity features with largest a_k^x

Banquet	Party	Family	Dating	Leisure
dish+	texture +	texture+	environment+	environment+
cuisine+	course+	course+	texture+	service+
private room+	waiter+	food container+	waiter+	texture+
design+	environment+	course-	service+	design+
menu+	service+	texture-	course+	dish+
dish-	sofa+	design+	dish+	course+
course+	quantity+	menu+	attitude+	location+
atmosphere+	food container+	food container-	quantity+	atmosphere+
appearance+	design+	appearance+	deal+	seat+
service+	dish+	quantity-	ingredients+	quantity-

比较试验（原来的基本都可以保留，因为我找到了相关文献，可以当作是复现了别人的实验。但需要换成餐馆重新评估一下。）

三，推荐模型（原文公式8）
扩展为考虑三种情况
1，考虑全label，即和原来一样
2，考虑multi-label，即在第二部分中p(x|c)》0.5的x才考虑，其他全部为0
3，考虑single-label，即第二部分中p(x|c)最大的那个x才考虑，其他全部为0
你需要给我的实验是
比较试验：这三种情况，以及过滤法+NMF（或者MF）的推荐系统
过滤法就是只选择相同的场景标注作为评分，MF就是矩阵分解。

总之，改动的目的都是避开比较试验和大规模的代码改动。但是仍然有2个必须要做的比较试验，即最后的推荐系统比较试验。代码改动主要是涉及训练集的生成（简单），以及一个优化过程（也相对简单）。你预估一下工作量看哪个没法做尽快告诉我。时间很紧，如果来不及，尽量用简单的策略替代，但不要删减内容。再删减文章就不够页码了。

Chen LIN
Associate Professor, Ph.D.
School of Information Science & Technology
Xiamen University, Xiamen, 361005, P.R.C.