Item-Based Collaborative Filtering

Yang LI

2017 年 11 月 9 日

School of Software Engineering, Tongji University

Chinese Text Segmentation

中文分词框架选择

- Jieba
- THULAC
- IK Analyzer
- Stanford

分词算法的分类

基于词典

- 正向最大匹配
- 逆向最大匹配
- 双向匹配

基于统计

- HMM
- CRF
- SVM
- DL

分词结果

```
康师傅/ 面霸/ 煮/ 面上/ 汤/ 排骨面/ 五入/ 100g/ */ 5
百事可乐/ 330ml
南孚/ 七号/ 碱性/ 2/ 只/ 挂/ 卡装/ LR03/ -/ 2B
农夫山泉/ 550ml
妙洁/ 抽取/ 式/ 保鲜袋/  / 大号/ MBL/ -/ A
南孚/ 五号/ 碱性/ 4/ 只/ 挂/ 卡装/ LR6/ -/ 4B
南孚/ 五号/ 碱性/ 2/ 只/ 挂/ 卡装/ LR6/ -/ 2B
美年达/ 橙/ 600ml
宏腾/ 麻辣/ 香/ 280g
南孚/ 七号/ 碱性/ 4/ 只/ 挂/ 卡装/ LR3/ -/ 4B
果珍/ 甜橙/ 味/ 400g
红星/ 特制/ 二锅头/ 56/ 度/ 500ml
西麦/ 核桃/ 高钙/ 燕麦片/ 700g
太太/ 乐加鲜/ 味精/ 100g
宏腾/ 油炸/ 辣子/ 300g
西麦/ 脑维/ 燕麦片/ 700g
德芙/ 香黑/ 巧克力/ 43g
超霸/ GP1.5/ V/ 扣式/ 电池/ A76
南孚/ 五号/ 碱性/ 6/ 只/ 挂/ 卡装/ LR6/ -/ 6B
```

一些具体的改进

- 根据词性删除了没有太多意义的副词、数量词等
- 后题考虑增加关键词的筛选 (TextRank Algorithms, 基于 PageRank)
- 与多个信息 (商品品牌、商品单位、包装类型) 结合处理

Similarity

计算相似度

原始

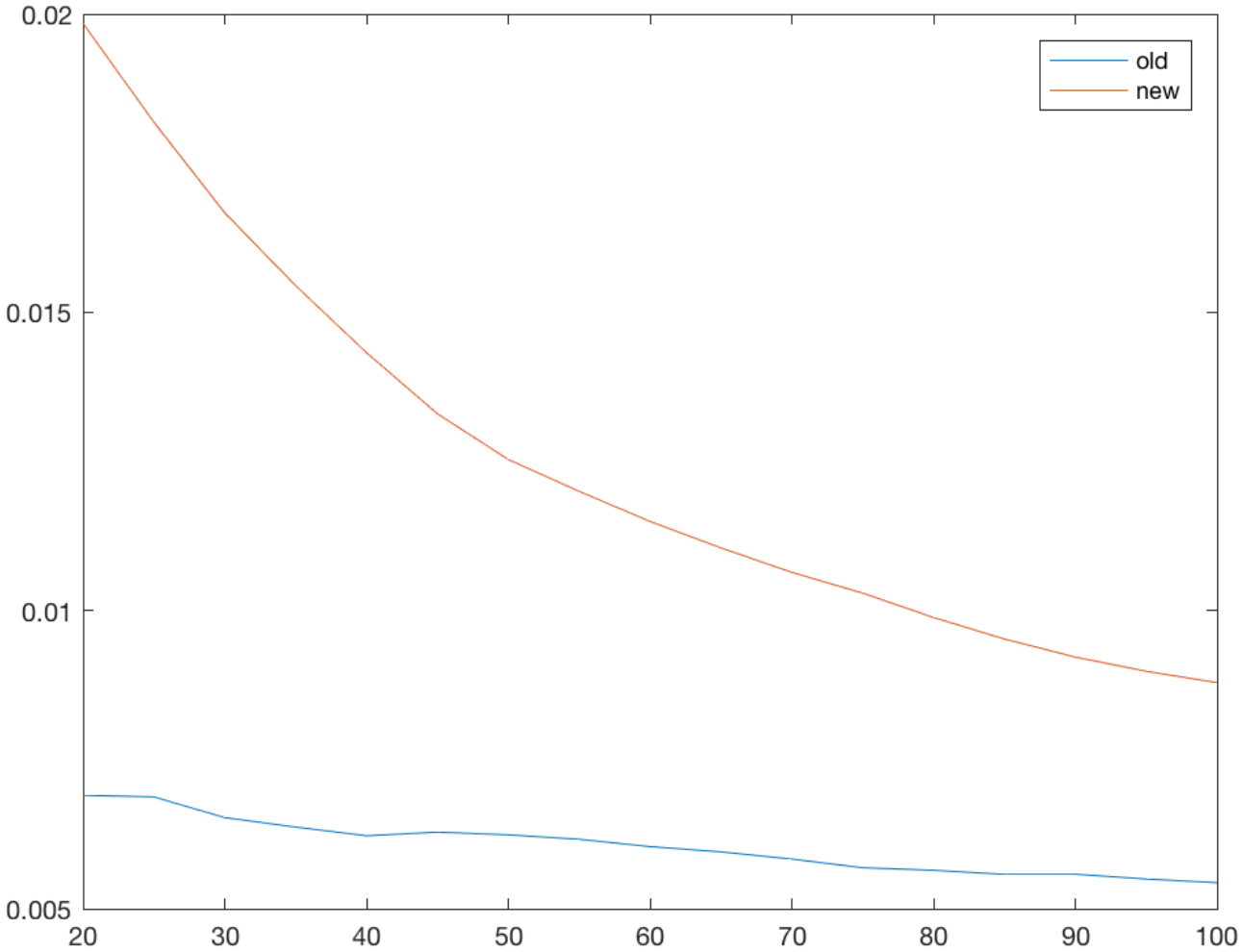
- 对每一个用户每次购买的东西增加一个相似度
- 最后除以总量 (归一化)

现在

- 对于同一大类的商品
- 对商品名进行分词
- 分词结果计算 Levenshtein Distance(Smith-Waterman algorithm, Needleman–Wunsch algorithm)
- 对距离进行处理 (取倒数, 归一化, 高斯函数)

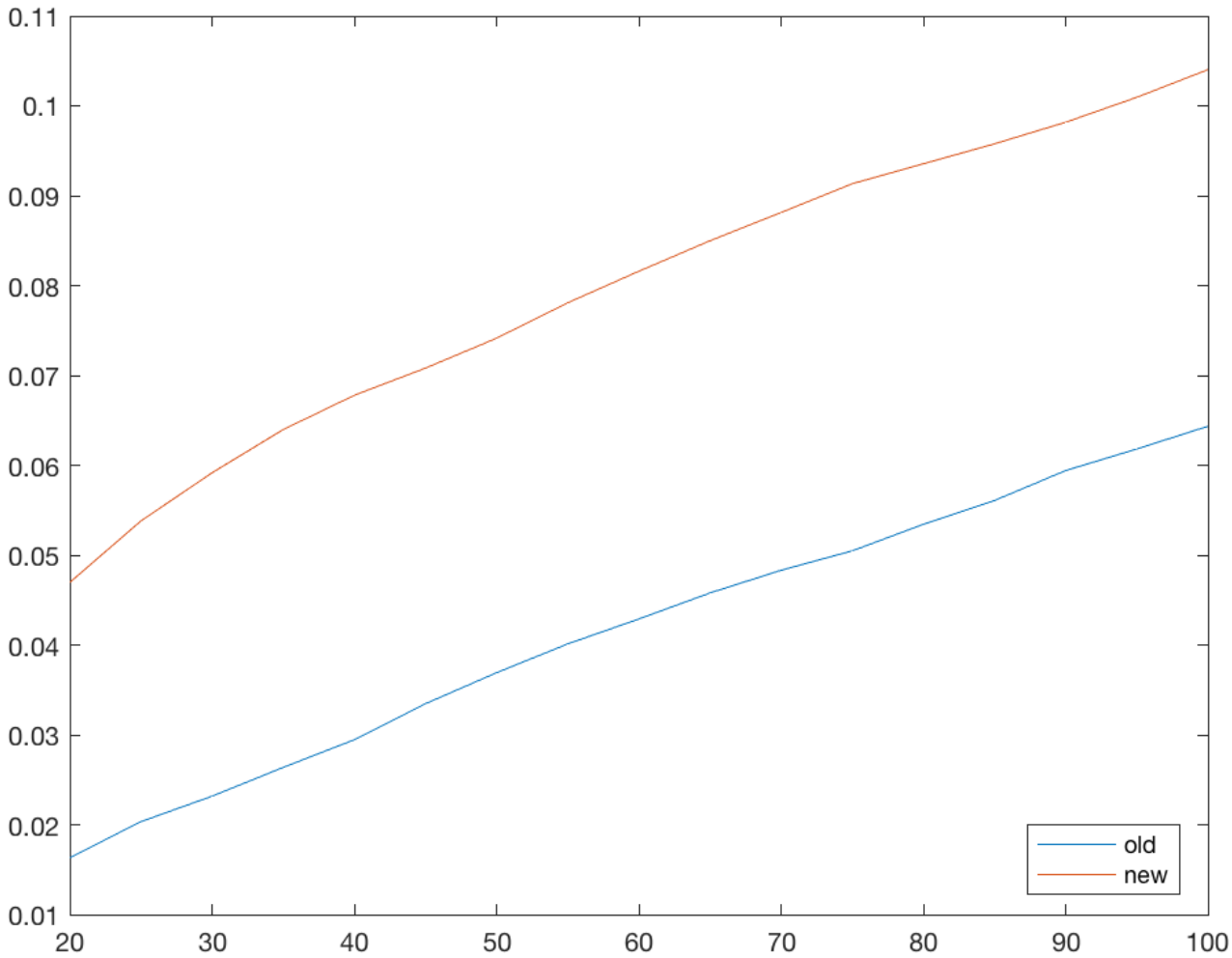
结果

precision



结果

recall


0.11
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
20
30
40
50
60
70
80
90
100
old
new

问题

- 评价指标 (precision, recall?)
- 分词结果的准确性
- 对距离的处理