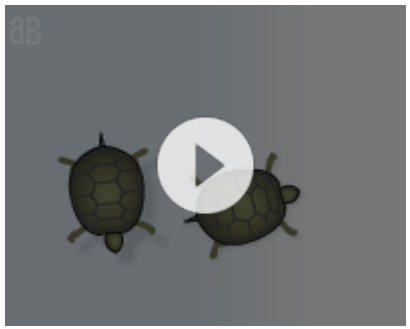
星空守望者--jkmiao

命运如同手中的掌纹，无论多曲折，始终掌握在自己手中...



Please feed the hungry pet turtles,thx!

昵称：星空守望者--jkmiao
园龄：2年8个月
粉丝：10
关注：3
+加关注

< 2017年12月 >

日	一	二	三	四	五	六
26	27	28	29	30	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

搜索

找找看

博客园 首页 新随笔 联系 订阅 管理

随笔-218 评论-15 文章-4

fastext 中文文本分类

1. 输入文本预处理, 通过jieba分词, 空格" "拼接文本串. 每行一个样本, 最后一个单词为双下划线表明label, __label__'xxx' . eg:

```
邱县 继刚 家庭 农场 小麦 、 玉米 、 棉花 、 大豆 、 蔬菜 、 苗木 种植 、 销售 ( 依法 须 经 批准 的 项目 , 经 相关 部门 批准 后方 可 开展 经营 活动 ) __label__A
江苏 嘉利欣 农业 科技 有限公司 农业 科技 研发 、 转让 、 咨询服务 展览 展示 服务 现代农业 休闲 观光 种植 、 销售 粮食 、 果蔬 、花卉 、 苗木 种植 中草药 销售 本 公司 种植 的 中草药 ( 特殊 中草药 除外 ) 养殖 、 销售 鱼 、 虾 、 螃蟹 ( 依法 须 经 批准 的 项
目 , 经 相关 部门 批准 后方 可 开展 经营 活动 ) __label__B
赞皇县 和谐 家庭 农场 农作物 果树 蔬菜 种植 销售 需 有关 部门 审批 的 审批 后 经营 __label__C
深圳市 修元 农业 开发 有限公司 农业 开发 、 绿化 工程 、 苗圃 种植 __label__A
```

```python
# 文本预处理
df2 = pd.read_csv('./industry_dalei_train.txt', encoding='utf-8')
df3 = pd.read_excel('./industry_standard.xlsx', encoding='utf-8')

# 映射转换
dalei2label_dict = dict((x, y) for x, y in zip(df3[u'大类名称'], df3[u'大类编号']))
```

```
df2['dalei_label'] = df2['sub_industry_name'].apply(lambda x: dalei2label_dict[x])
# 切割
df2['cut_name'] = df2['name'].apply(lambda x: ' '.join(jieba.cut(x)))
df2['cut_business'] = df2['business'].apply(lambda x: ' '.join(jieba.cut(x)))
df2['cut_train'] = df2['cut_name'] + ' ' + df2['cut_business'] + ' __label__' + df2['dalei_label']
df2['cut_train'].to_csv('industry_dalei_train.txt', index=None, header=None, encoding='utf-8')
```

2. pip install fasttext, 利用fasttext 的python 包进行分类.

```
# 训练和保存模型
da_clf = fasttext.supervised('./industry_dalei_train.txt', 'models/dalei_clf')

# 加载模型
da_clf = fasttext.load_model('./models/dalei_clf.bin')

# 测试
res = da_clf.test('./industry_dalei_test.txt')
print res.precision
print res.recall

# 预测使用, data为['cut document1' ,'cut document2']
da_clf.predict(data ,k=1)  # 预测标签
da_clf.predict_proba(da_df.iloc[:5],k=3)  # 预测标签概率
```

简单高效, 结果也不差. good luck~

参考:

https://pypi.python.org/pypi/fasttext/

http://www.41443.com/HTML/Python/20160909/449360.html

http://www.360doc.com/content/17/0427/02/20558639_648968041.shtml

php(7)
python(54)
生活感悟(11)
杂七杂八(6)

随笔档案

2017年7月 (8)
2017年6月 (6)
2017年5月 (7)
2017年4月 (2)
2017年3月 (2)
2017年2月 (7)
2017年1月 (7)
2016年12月 (5)
2016年11月 (10)
2016年10月 (8)
2016年8月 (3)
2016年7月 (4)
2016年6月 (4)
2016年5月 (6)
2016年4月 (2)
2016年3月 (5)
2016年2月 (9)
2016年1月 (12)
2015年12月 (7)
2015年11月 (8)
2015年10月 (11)
2015年9月 (13)
2015年8月 (14)
2015年7月 (9)
2015年6月 (7)

每天一小步，人生一大步！Good luck~

分类: NLP,python

星空守望者--jkmiao
关注 - 3
粉丝 - 10
+加关注

0 0

« 上一篇：django 多线程下载图片
» 下一篇：ubuntu14.04 安装jdk1.8及以上

posted on 2017-06-17 19:27 星空守望者--jkmiao 阅读(278) 评论(0) 编辑 收藏

刷新评论 刷新页面 返回顶部

2015年5月 (14)
2015年4月 (28)

最新评论

1. Re:ubuntu crontab python 定时任务备记
@焦距嗯，是的 ~ 应该考虑用notify-send命令...
--星空守望者--jkmiao
2. Re:ubuntu crontab python 定时任务备记
单纯的echo在屏幕上应该是看不到输出的吧？因为cron会把任何输出都email到root的信箱。可以输出到文件：*/30 * * * * echo '30 minutes, take a break.....
--焦距
3. Re:模板匹配法验证码识别
高产似母猪啊。最近又开始学习用 tornado 框架了。
--Haruhi0
4. Re:kddcup2015
哪位大神可以说一下，为什么source会有两种server和browser，它们有什么区别呢，或者说不同的source分别表示着什么呢，谢谢！

--liweier0958
5. Re:kddcup2015
亲，你这个代码没有真实运行过吧？我仔细看了一下，发现了一些不对劲的地方，比如：那个step1的gen_object_dict方法中，不管是在log_train.csv还是log_test.csv ，.....
--irisDataMaster

阅读排行榜

最新IT新闻:
· SpaceX再创历史 成首个执行NASA任务的私有公司
· 在暴涨暴跌中，持有比特币一周是一种怎样的体验？
· 郭台铭详解鸿海工业互联网战略 拟分拆在上海上市
· 扎克伯格休假照片曝光 配娃娃吃喝玩乐
· 金刚狼死侍回归漫威，迪士尼收购福克斯让好莱坞「变天」
» 更多新闻...

最新知识库文章:
· 以操作系统的角度述说线程与进程
· 软件测试转型之路
· 门内门外看招聘
· 大道至简，职场上做人做事做管理
· 关于编程，你的练习是不是有效的？
» 更多知识库文章...

1. python&pandas 与mysql 连接(4784)
2. linux下批量修改文件名之rename(3534)
3. [原创博文] 用Python做统计分析（Scipy.stats的文档）(2584)
4. python 之 决策树分类算法(2513)
5. 深度学习性能提升的诀窍(2389)

评论排行榜

1. kddcup2015(10)
2. ubuntu crontab python 定时任务备记(2)
3. 模板匹配法验证码识别(1)
4. doc2vec 利用gensim 生成文档向量(1)
5. python&pandas 与mysql 连接(1)

推荐排行榜

1. 深度学习性能提升的诀窍(2)
2. 正式进驻博客园(1)
3. 实用黑科技(1)