NATURAL LANGUAGE PROCESSING

Designed by Sun Menghan

0

CONTENT

http://111.160.216.4:1780/wiki/metasolo/wiki/10.自然语言分析

1

INTRODUCTION

目标：

从一段商品信息中提取出商品的品牌，款号，颜色，尺码以及种类。

例子：

vaude 1115051 A.ba（沃德）男款2.5层冲锋衣 赭石色 2XL特价

Brand=vaude, Number=1115051, Size=2XL, Color=赭石色, Style=男款 冲锋衣

CHALLENGES

名词堆叠导致

语法缺失

缺乏上下文信息推断

产品属性

印刷错误/缩写

必须归一化到

标准值

产品信息

不完整

2

RELATED WORK

通过收集有关资料，构成原始思路：

1. 中文分词
2. 分析数据
3. 提取信息

中文分词
(Chinese Word Segmentation)

中文分词是指将一个汉字序列切分成一个一个单独的词。

在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符。

故中文比之英文要复杂的多。

Python中文分词组件jieba

http://www.oschina.net/p/jieba/

java中文分词工具包IKAnalyzer

http://www.oschina.net/p/ikanalyzer/

支持用户词典扩展定义

Demo

导入 sougou 颜色细胞词库，并加入其他关键词

1. 下载 颜色名称.scel
2. 使用 scelToTXT.py 转换成TXT

    {195}kongquelan 孔雀蓝

    {194}kongquelv 孔雀绿

    {477}kuanghui 矿灰

    {193}lajiaohong 辣椒红
3. 提取颜色词放入自定义词库

    userdictionary.dic

IKanalyzer 分词效果在编号方面优于 jieba，例如：

SCARPA 63039-200深灰40

IKanalyzer: IKAnalyzer.tokenStream()

scarpa,63039-200,深灰,40

Jieba: jieba.cut()

SCARPA, ,63039,-,200,深灰,40

分类器
(Classifier)

分类是数据挖掘的一种非常重要的方法。

分类的概念是在已有数据的基础上学会一个分类函数或构造出一个分类模型。该函数或模型能够把数据库中的数据纪录映射到给定类别中的某一个，从而可以应用于数据预测。

Spark MLlib

Apache Spark is a fast and general-purpose cluster computing system. It provides high-level APIs in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It also supports a rich set of higher-level tools including Spark SQL for SQL and structured data processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming.

中文词语特征提取
(Feature Extractors)

TF-IDF

(词频 term frequency–逆向文件频率 inverse document frequency)

主要思想：

如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

朴素贝叶斯分类器
(Naive Bayes Classifier)

朴素贝叶斯分类的正式定义如下：

1、设$x = \{a_1, a_2, ..., a_m\}$为一个待分类项，而每个a为x的一个特征属性。

2、有类别集合$C = \{y_1, y_2, ..., y_n\}$。

3、计算$P(y_1|x), P(y_2|x), ..., P(y_n|x)$。

4、如果$P(y_k|x) = max\{P(y_1|x), P(y_2|x), ..., P(y_n|x)\}$，则$x \in y_k$。

3

SOLUTION

命名实体识别
(Named Entity Recognition, NER)

命名实体识别
(Named Entity Recognition, NER)

是指从文本中识别具有特定类别的实体（通常是名词），例如人名、地名、机构名、专有名词等。命名实体识别是信息检索，查询分类，自动问答等问题的基础任务，其效果直接影响后续处理的效果，因此是自然语言处理研究的一个基础问题。

Bootstrapped Named Entity Recognition for Product Attribute Extraction

D11-1144.pdf

NEXT Blue Petite Bootcut jeans size 12 BNWT
B C NA NA G S S NA

Paul Smith Osmo White Plimsoll Trainers – UK 6 RRP : £ 100
B B NA C NA G NA S S NA NA NA NA

Figure 1: Example listings and their corresponding labels from the clothing and shoes category.

主要算法
Algorithms

HMM (Hidden Markov model 隐马尔科夫模型)

http://www.nlpr.labs.gov.cn/2005papers/gjhy/gh71.pdf

SVM (Support Vector Machine支持向量机)

Biomedical named entity recognition using two-phase model based on SVMs

CRF (Conditional random field条件随机场)

http://psb.stanford.edu/psb11/conference-materials/proceedings%201996-2010/psb08/leaman.pdf

MaxEnt (Maximum Entropy Models 最大熵模型)

	SVM	MaxEnt	HMM	CRF
w/ Viterbi	**89.47%**	**88.13%**	**83.82**	**93.35%**

Table 2: Classification accuracy (%) on 9-class NER on men's clothing dataset, comparing SVM, MaxEnt, supervised HMM, and CRF.

NER工具

HanLP: Han Language Processing
http://hanlp.linrunsoft.com/

LingPipe

http://alias-i.com/lingpipe/index.html

OpenNLP

http://opennlp.apache.org/

Stanford Named Entity Recognizer (NER)

http://nlp.stanford.edu/software/CRF-NER.shtml

斯坦福NER采用Java实现，可以对命名实体进行识别

基于条件随机场(Conditional Random Fields, CRF)

3 class:	Location, Person, Organization
4 class:	Location, Person, Organization, Misc
7 class:	Time, Location, Organization, Person, Money, Percent, Date

Demo

Stanford NER CRF FAQ

http://nlp.stanford.edu/software/crf-faq.shtml

How can I train my own NER model?

The documentation for training your own classifier is somewhere between bad and non-existent. But nevertheless, everything you need is in the box, and you should look through the Javadoc for at least the classes **CRFClassifier** and **NERFeatureFactory**.

4

IMPLEMENTATION

traindata.tsv:

kawadgarbo B
alpine O
jacket G
88282 N
深红 C
s S

Remark:

B=Brand, G=Garment type/Style,

N=Number, C=Color, S=Size, O=Other

austen.prop

Once you have such a properties file, you can train a classifier with the command:

```
java -cp stanford-ner.jar
edu.stanford.nlp.ie.crf.CRFClassifier -prop austen.prop
```

An NER model will then be serialized to the location specified in the properties file:

ner-model.ser.gz

Demo

使用model分析文件

Result:

kailas凯乐石 KG620164 A.ba(凯乐石）女款弹力超薄风衣 冰雪蓝 L

Brand=kailas, Number=kg620164, Size=L, Color=冰雪蓝, Style=女款

Demo

5

FUTURE PLAN

01

印刷错误/缩写
必须归一化到
标准值
（纠错功能）

02

自动生成新的
训练数据
训练新模型

03

自动识别
品牌

Q & A

THANKS

Thank you for your attention!