TextBlob情感分析调研

刘焕勇

2018-05-16

主要内容

- Textblob简介
- 情感分析接口
- 情感分析算法流程
- 实验
- 结论

TextBlob

- TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

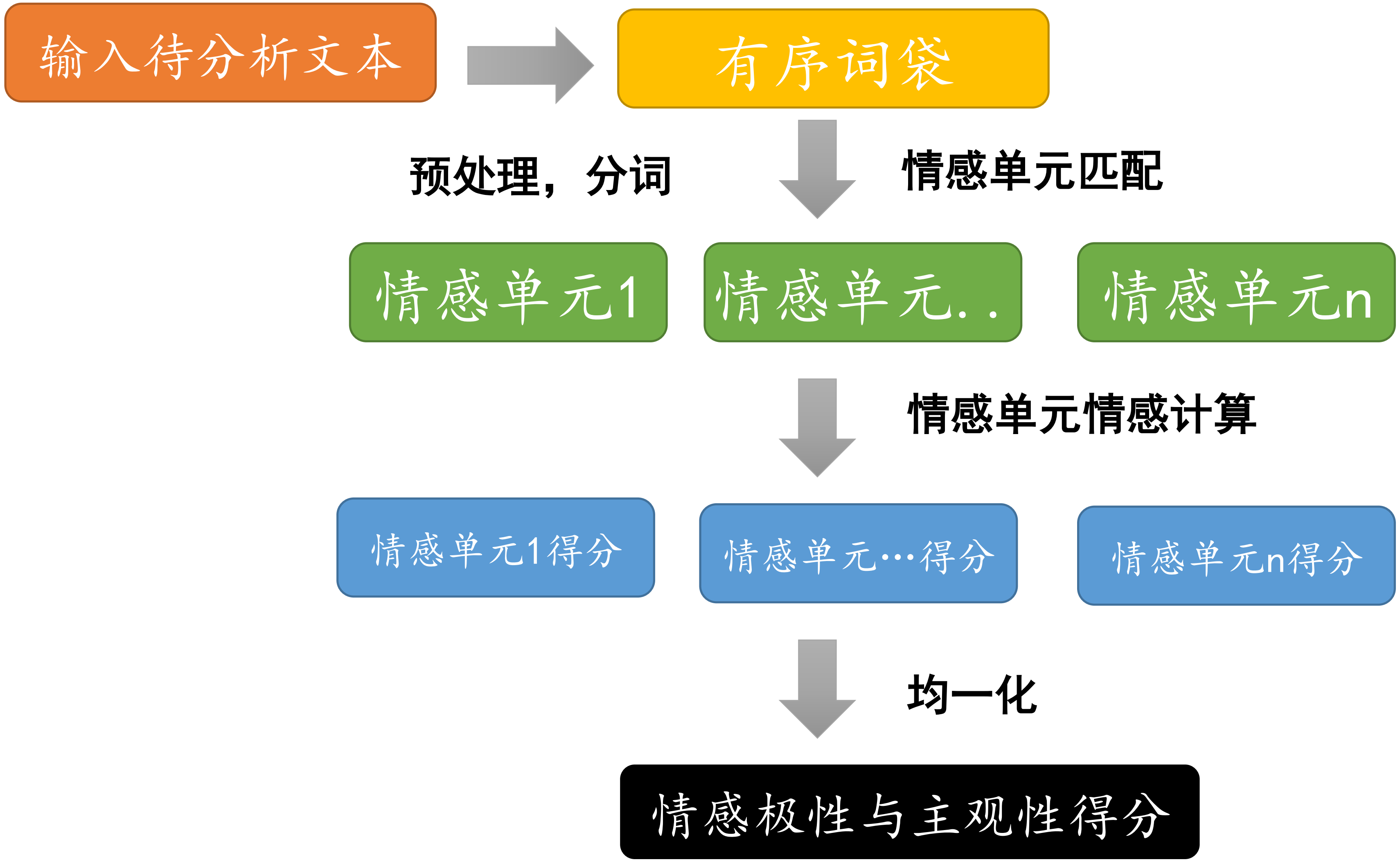
情感分析接口

- 引入： from textblob import TextBlob
- 使用:
  - >> S = ("not a very great calculation"
  - >> TextBlob(S).sentiment
  - >> Sentiment(polarity=-0.3076923076923077, subjectivity=0.5769230769230769)
- 说明:
  - polarity: negative vs. positive (-1.0 => +1.0) ：情感极性
  - subjectivity: objective vs. subjective (+0.0 => +1.0)： 情感主观性

算法步骤

- step1: given sentence S
- step2: SenenceTokenize
- step3: remove single word, where len(word)==1
- step4: search sentiment-chunck, [sentiword],[modifyword, sentiword], [denyword, sentiword], [modifyword, denyword, sentiword]
- step5:avearge ploraity, subjectivity in sentiment-chunck
- step6:finished

算法流程



输入待分析文本

有序词袋

预处理，分词

情感单元匹配

情感单元1

情感单元. .

情感单元n

情感单元情感计算

情感单元1得分

情感单元…得分

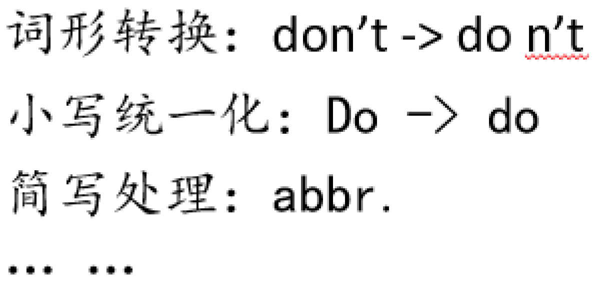
情感单元n得分

均一化

情感极性与主观性得分

预处理与分词

- 去除单字词
- 以空格为分隔符
- 标点符号为分隔符
  - 普通标点符号
  - 情绪符号：

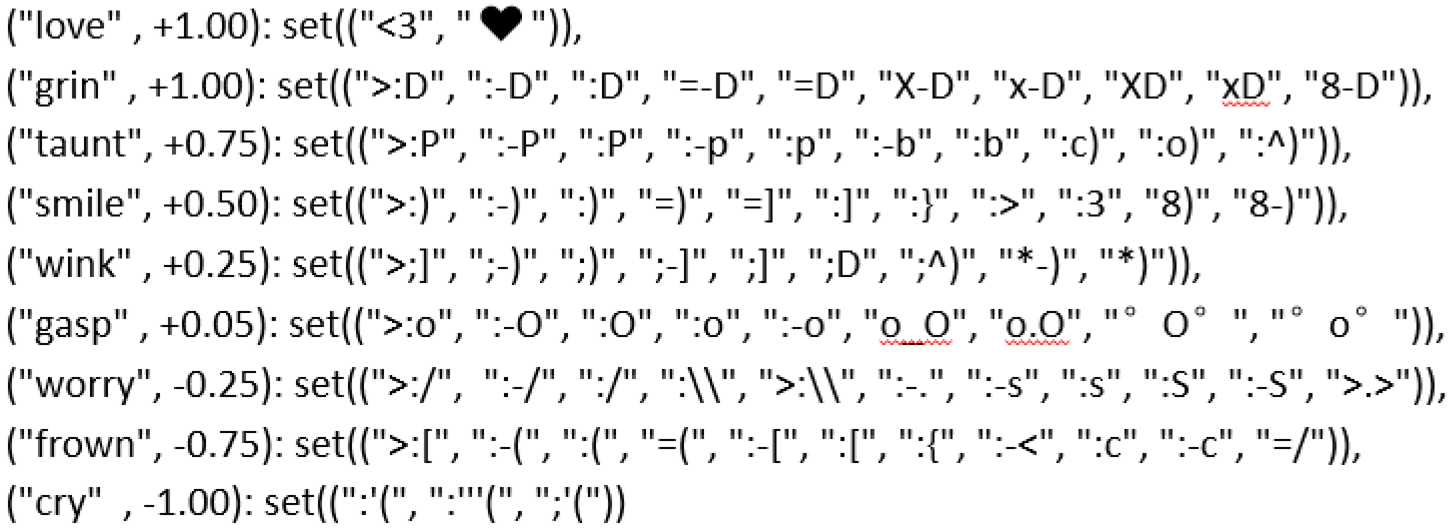


词形转换：don’t -> do n’t

小写统一化：Do -> do

简写处理：abbr.

… …



```
("love" , +1.00): set(("<3", " ♥ ")),
("grin" , +1.00): set((">:D", ":-D", ":D", "=-D", "=D", "X-D", "x-D", "XD", "xD", "8-D")),
("taunt", +0.75): set((">:P", ":-P", ":P", ":-p", ":p", ":-b", ":b", ":c)", ":o)", ":^)")),
("smile", +0.50): set((">:)", ":-)", ":)", "=)", "=]", ":]", ":}", ":>", ":3", "8)", "8-)")),
("wink" , +0.25): set((">;]", ";-)", ";)", ";-]", ";]", ";D", ";^)", "*-)", "*)")),
("gasp" , +0.05): set((">:o", ":-O", ":O", ":o", ":-o", "o_O", "o.O", "°  O°  ", "°  o°  ")),
("worry", -0.25): set((">:/",  ":-/", ":/", ":\\", ">:\\", ":-.", ":-s", ":s", ":S", ":-S", ">.>")),
("frown", -0.75): set((">:[", ":-(", ":(", "=(", ":-[", ":[", ":{", ":-<", ":c", ":-c", "=/")),
("cry"  , -1.00): set((":'(", ":'''(", ";'("))
```

情感单元匹配

- 约定：
  - m->modifyword: 程度副词, very
  - n->denywords:否定副词， not, no, never
  - S->sentiwords: 情感词， happy, horrible…
- 匹配规则：
  - Pattern1: “n?m*S”
  - Pattern2:“n?m*S[^S]*!$”

情感单元匹配举例

- "I am not no happy today"
  - ['no', 'happy']
- "this is never very good day but very much good in the world“
  - [['never', 'very', 'good’], ['very', 'much', 'good']]
- "this is never very very good day but very much good in the world"
  - [['never', 'very', 'very', 'good’], ['very', 'much', 'good']]
- "this is not a good great day"
  - [['not', 'good'], ['great']]
- "I am not no happy today"
  - [['no', 'happy']]

情感词

- 情感词表：
    - 文件名： Sentiment-en.xml
    - Author： Tom de smedt ,walter Daelemans
    - 2918个语义synset词条
    - 1528个唯一词
- 词条：
    - <word form="great" cornetto_synset_id="n_a-525317" wordnet_id="a-01123879" pos="JJ" sense="very good" polarity="1.0" subjectivity="1.0" intensity="1.0" confidence="0.9" />
- 释义：
    - polarity: negative vs. positive (-1.0 => +1.0) --情感极性
    - subjectivity: objective vs. subjective (+0.0 => +1.0) --情感主观性
    - intensity: modifies next word? (x0.5 => x2.0) -- 情感强度

Sentiment-en.xml

情感单元情感计算

- 单个情感词情感计算
- 否定词+情感词
- 修饰词+情感词
- 否定词+修饰词+情感词

单个情感词

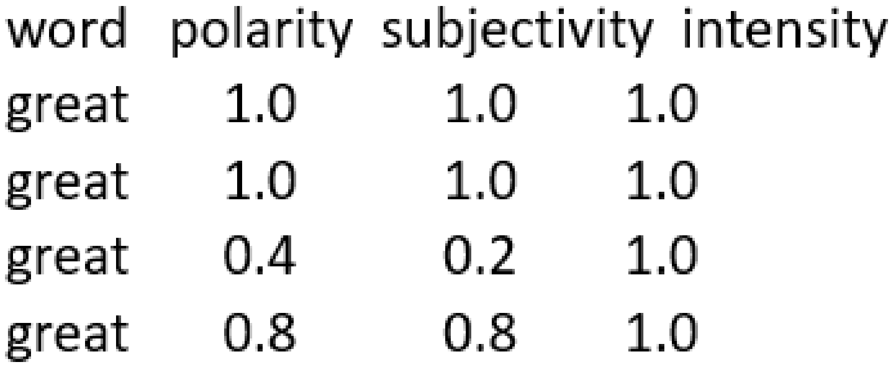
- 单个情感词的语义信息，一词多义，one to many：

word	polarity	subjectivity	intensity
great	1.0	1.0	1.0
great	1.0	1.0	1.0
great	0.4	0.2	1.0
great	0.8	0.8	1.0

- 单个情感词的情感得分
  - 一个词语下多个synset情感极性、主观性进行平均化
  - TextBlob("great").sentiment
  - Sentiment(polarity=0.8, subjectivity=0.75)

否定词+情感词

- TextBlob("great").sentiment
  - polarity=0.8
  - subjectivity=0.75



word	polarity	subjectivity	intensity
great	1.0	1.0	1.0
great	1.0	1.0	1.0
great	0.4	0.2	1.0
great	0.8	0.8	1.0

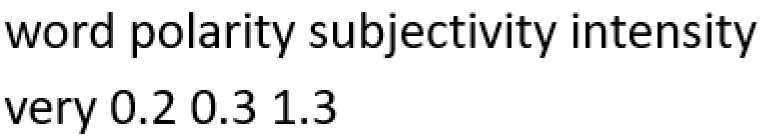
(Sigma(polarity)/n,
Sigma(subjectivity)/n

- TextBlob("not great").sentiment
  - polarity=-0.4
  - subjectivity=0.75

polarity*-0.5, subjectivity *1

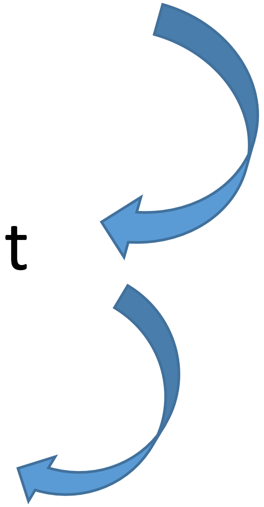
修饰词+情感词

- TextBlob("great").sentiment
  - polarity=0.8
  - subjectivity=0.75
- TextBlob(“very great").sentiment
  - polarity=1.0
  - subjectivity=0.975
- TextBlob(“very very great").sentiment
  - polarity=1.0
  - subjectivity=0.975



word polarity subjectivity intensity

very 0.2 0.3 1.3



Polarity:min(1.0, 0.8*1.3)
Subjectivity:min(1.0, 0.75*1.3)

否定词+修饰词+情感词

- TextBlob("great").sentiment
  - polarity=0.8
  - subjectivity=0.75

word	polarity	subjectivity	intensity
very	0.2	0.3	1.3

1、极性：否定对正向程度副词形成反比例逆转
2：主观性：否定只逆转程度，不逆转值

- TextBlob(“not very great").sentiment
  - polarity=−0. 3076923076923077
  - subjectivity=0. 5769230769230769

Polarity = −0.5*1/1.3*0.8≈−0.31

Subjectivity = 1/1.3*0.75≈0.58

情感均一化

- 对所有chunk进行得分平均化：
  - 给定每个chunk的权重weight, 默认都是1，即每个都同等重要
  - polarity = avg([(w, p) for w, p, s, x in chunks], weight),
  - subjectivity = avg([(w, s) for w, p, s, x in chunks], weight))

实验1-无情感词

- Chunks: ([words], polarity, subjectivity, label:None/profanity])

- s = "i don't want to share with you"
- Chunks:
  - []
- Result:
  - polarity=0.0,
  - subjectivity=0.0

实验2–只包含情感词

- s = "hello this is my favorite food and i don't want to share with you"
- Chunks:
  - [(['favorite'], 0.5, 1.0, None)]
- Result:
  - polarity=0.5
  - subjectivity=1.0

实验3–包含情感词与否定词

- s = "this is not a good great day"
- Chunks:
  - [(['not', 'good'], -0.35, 0.6000000000000001, None),
  - (['great'], 0.8, 0.75, None)]
- Result:
  - polarity=0.22500000000000003
  - subjectivity=0.675

实验4–包含情感词与否定词

- s = "this is not a good and not great day"
- Chunks:
  - [(['not', 'good'], -0.35, 0.6000000000000001, None)
  - (['not', 'great'], -0.4, 0.75, None)]
- Result:
  - polarity=-0.375
  - subjectivity=0.675

实验5–包含情感词、否定词、单个程度词

- s = 'this is not a good but never very a bad day’
- Chunks:
  - [(['not', 'good'], -0.35, 0.6000000000000001, None)
  - (['never', 'very', 'bad'], 0.2692307692307692 0.5128205128205128, None)]
- Result:
  - polarity=-0.0403846153846154
  - subjectivity=0.5564102564102564

实验6–包含情感词、否定词、多个程度词

- s = 'this is never very good day but very much good in the world'
- Chunks:
  - [(['never', 'very', 'good'], -0.26923076923076916, 0.46153846153846156, None)
  - (['very', 'much', 'good'], 0.7, 0.6000000000000001, None)]
- Result:
  - polarity=0.2153846153846154
  - subjectivity=0.5307692307692309

实验7–包含情感词、多个否定词

- s = 'i am not no happy today'
- Chunks:
  - [(['no', 'happy'], -0.4, 1.0, None)]
- Result:
  - polarity=-0.4
  - subjectivity=1.0

没有解决双重否定的类型.

总结

- TextBlob中的情感分析是基于规则与语义词表相结合的方法
- ・算法三要素：
  - 否定词表，提供逆转信息
  - 程度词表，提供强度修饰信息，修正极性与主观性
  - 情感词表，提供词语极性信息，主观性信息，强度信息，wordnet
- 局限：
  - 实际为ordered Bag of words，只考虑相近窗口修饰关系，没有考虑句法信息
  - 对双重否定类型的句子失效
  - 对情感单元的极性与主观性得分进行平均化的算法，可以改进
  - 不支持中文
- TextBlob机器学习方法：
  - 基于Bayes的情感分类：
    - 接口textblob.classifier(),只返回正正负信息
    - 训练语料：电影评论数据集，neg，pos
- 中文文本主观性评判
  - 就目前调研结果来看，还没有主观性语义知识库可以使用
  - 主观性的界定问题，尤其对新闻文本的主观性如何量化？

参考

- https://planspace.org/20150607-textblob_sentiment/
- http://textblob.readthedocs.io/en/dev/