Анализ на настроения /сантимети част I (sentiment analysis)

Д-р инж. Огнян Кабранов
Дипл. инж. Даниела Цветкова

2019 година

Проблем - автоматично да определяме сантимента на коментар : положителен или отрицателен

-

Днес ще се занимаем с малко теория анализ на сантиментите (sentiment analysis)

Примери

- Уеб страница трябва автоматично да брои позитивните и негативните коментари за ресторант и да прави класация на ресторанти
- Това ще помогне на мениджмънта да определи какво трябва да бъде направено.
- Или верига като Кауфланд трябва да определи как клиентите приемат ново кисело мляко

Примери за коментари на ресторант

“В ресторанта кебапчетата бяха студени и бирата топла” -

“Отлична храна, салатата беше много прясна”

“Бавно и нелюбезно обслужване”

“Любезен персонал и вкусна храна”

Какво искаме да постигнем

- На базата на вече известни и класифицирани коментари да приготвим данни да обучим нашата системи за машинно обучение.
- Тези системи трябва бъдат в състояние да определят сантимента на нови изречение/коментар.

Откъде да започнем ????

Нека започнем няколко нови термина

Векторизация: трансформация на изречение във вектор

Пример **изречение**: “'Иванчо получи нов компютър и Иванчо стана нов човек'”

Речник: *{'иванчо': 0, 'компютър':1, 'нов': 2, 'получи': 3, 'стана': 4, 'човек': 5}*

Вектор: [[2 1 2 1 1 1]]

Малко Python

```
from sklearn.feature_extraction.text import CountVectorizer
text= ['Иванчо получи нов компютър и Иванчо стана нов човек']

#Създаване на трансформация за векторизиране
vectorizer = CountVectorizer()

#Токенизация и създаване на речник
vectorizer.fit(text)
```

Малко Python

```
#Как изглежда речника

print(vectorizer.vocabulary_)

{'иванчо': 0, 'получи': 3, 'нов': 2, 'компютър': 1, 'стана': 4,
'човек': 5}

# Кодиране на документа като вектор:

vector = vectorizer.transform(text)
```

Малко Python

```
#Как изглежда вектора

print(vector.shape)

(1, 6)

print(vector.toarray())

[[2 1 2 1 1 1]]
```

Какво е корпус (corpus)

Корпус (corpus) е речникът с изречения, с които ще обучим системата, която може да анализира нормално изречение.

Какво е корпус (corpus)

```
corpus = [
      'това е документ номер едно',
      'а това е документ номер две',
      'а това е документ три',
      'дали това е документ три',
       ]
```

Речник, генериран от корпуса

```
vectorizer = CountVectorizer()
X = vectorizer.fit(corpus)
print(vectorizer.vocabulary_)
```

Като резултат с отпечатва речникът, който е генериран от корпуса

```
{'това': 5, 'документ': 2, 'номер': 4, 'едно': 3, 'две': 1,
'три': 6, 'дали': 0}
```

!!!!!! ПУНКТУАЦИЯТА СЕ ИГНОРИРА

Векторизиран корпус

```
vector = vectorizer.transform(corpus)
print(vector.shape)
print(vector.toarray())
```

Като резултат се отпечатва матрицата на корпуса

```
(4, 7)
[[0 0 1 1 1 1 0]
 [0 1 1 0 1 1 0]
 [0 0 1 0 0 1 1]
 [1 0 1 0 0 1 1]]
```

Векторизиран корпус

Речник:

```
{'дали':0,'две':1,'документ':2,'едно':3,'номер':4,'това':5,'три
':6}
```

```
corpus = [
      'това е документ номер едно',
      'а това е документ номер две',
      'а това е документ три',
      'дали това е документ три',
       ]
```

Векторизиран корпус

```
{'дали':0,'две':1,'документ':2,'едно':3,'номер':4,'това':5,'три
':6}
```

```
това(5) е документ(2) номер(4) едно(3)
[0 0 1 1 1 1 0]

а това(5) е документ(2) номер(4) две(1)
 [0 1 1 0 1 1 0]
```

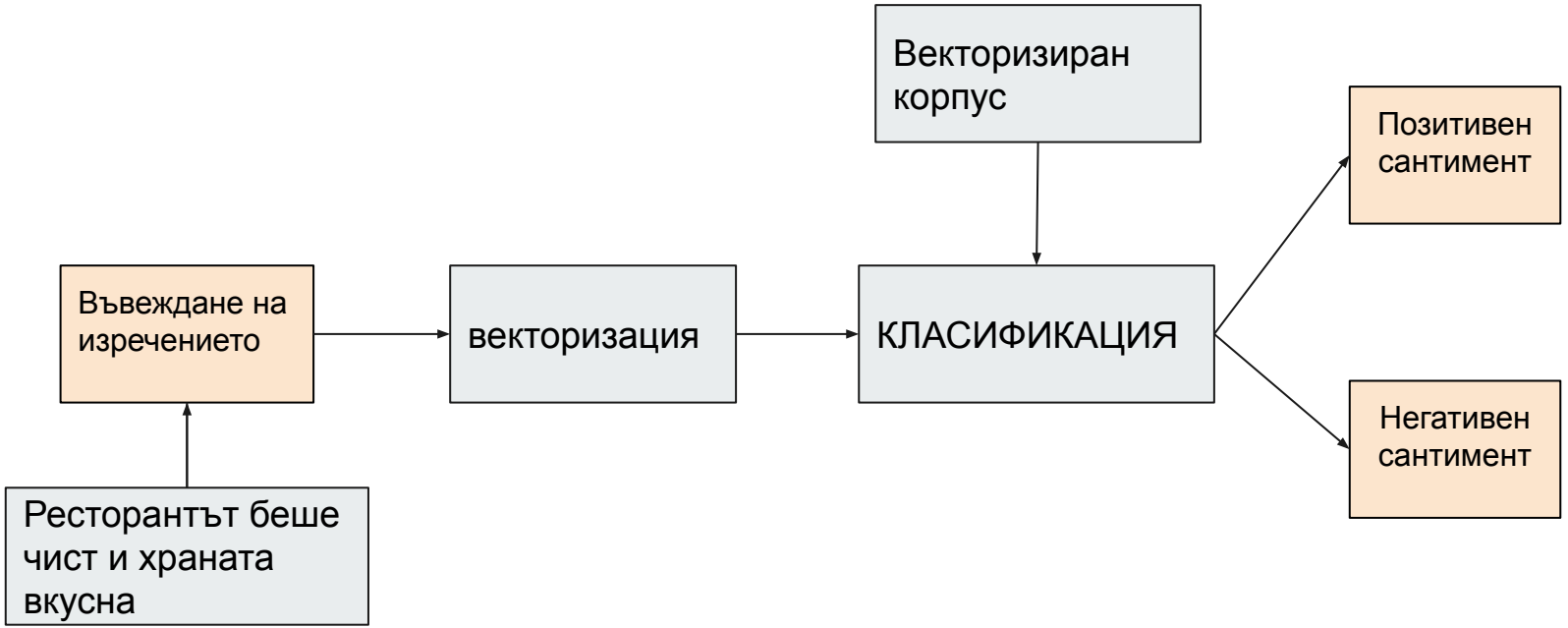
Векторизиран корпус

```
{'дали':0,'две':1,'документ':2,'едно':3,'номер':4,'това':5,'три
':6}


'а това(5) е документ(2) три(6)'
 [0 0 1 0 0 1 1]

'дали това е документ три',
 [1 0 1 0 0 1 1]]
```

Да дефинираме проблема за разпознаване на настроения/сантименти



БЛАГОДАРЯ И ДО НОВИ СРЕЩИ ?

Thank you! Danke ! Merci !