Final project presentation

Using naïve bayes classifier

With

contents

1

How it started?

So many live comment!

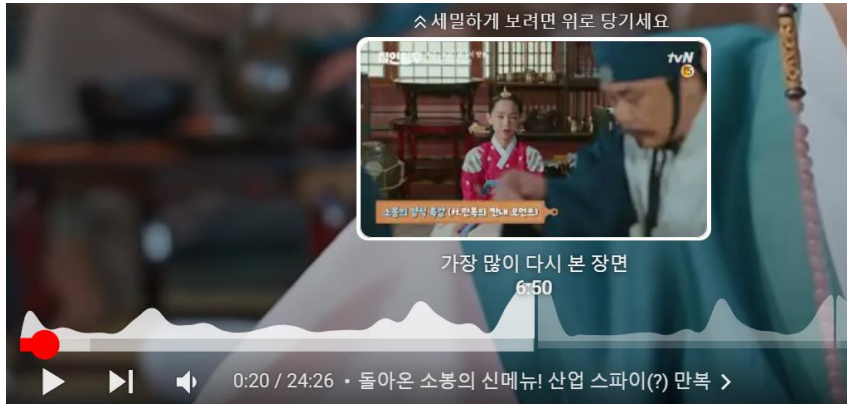
Can I get specific data on viewers' responses?

Advantages

- WangHongs can intuitively look specific data on viewers' good/bad ratio in real time.

- It provides an important indicator for editing after the end of the live broadcast.

Example : youtube



2

Preparation

Part 2

Requirements

Speed

- Naïve Bayes classifier
- MySQL

Versatility

- Build REST API
  With Django

Dataset

- Live comments,
  With labels

Expected challenges

- If compute a word not exist in database, it will decrease accuracy.
- Don’t know how much time it takes to compute, need to be less then 1s

3

Naïve Bayes classifier

Part 3

Practical example

Example: input “哈哈”

Positive		Negative	
哈	10	哈	1
…	…	…	…
…	…	…	…
…	…	…	…
Total	60	Total	40

P(P)*P(哈|P) *P(哈|P) = (60/100) * (10/60) * (10/60) **= 0.0166**

P(N)*P(哈|N) *P(哈|N) = (40/100) * (1/40) * (1/40) **= 0.00025**

Part 3

Problem

Example: input “啊哈哈”

Positive		Negative	
哈	10	哈	1
啊	0	啊	1
…	…	…	…
…	…	…	…
Total	60	Total	40

P(P)* P(啊|P) *P(哈|P) *P(哈|P) = (60/100) *(0)* (10/60) * (10/60) **= 0**

P(N)* P(啊|N) *P(哈|N) *P(哈|N) = (40/100)*(1/40) * (1/40) * (1/40) **= so small**

Part 3

Laplace smoothing

Example: input “啊哈哈”

Positive		Negative	
哈	10+1	哈	1+1
啊	0+1	啊	1+1
…	…	…	…
…	…	…	…
Total	60+10	Total	40+10

P(P)* P(啊|P) *P(哈|P) *P(哈|P) = (70/120) *(1/70)* (11/70) * (11/70) = **2.05*10^-4**

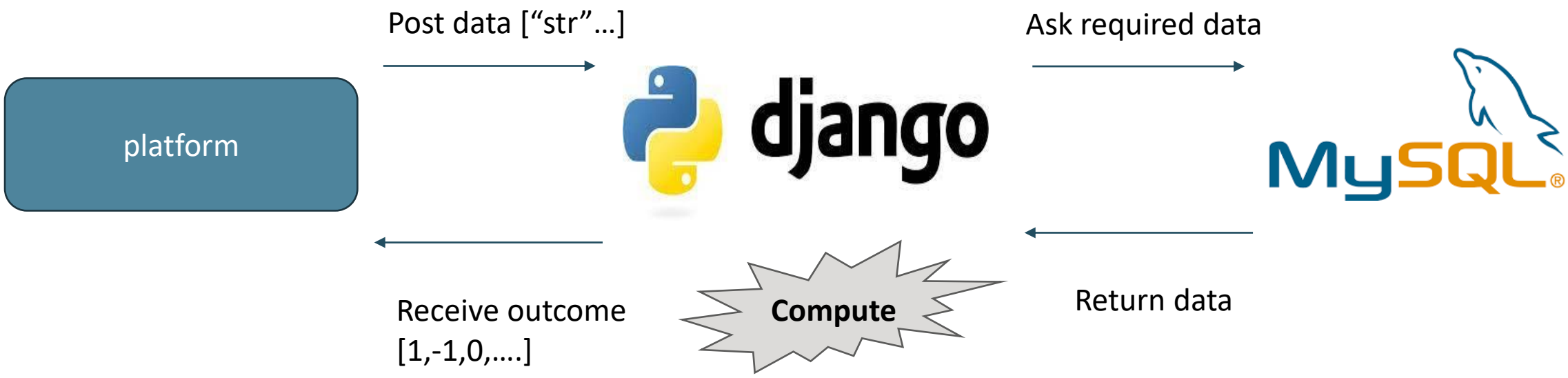
P(N)* P(啊|N) *P(哈|N) *P(哈|N) = (50/120)*(2/50) * (2/50) * (2/50) = **2.66*10^-5**

4

API demonstration

Part 4

Api working process



Part 4

Test result

Input	Label	Result	
实在	1	1	O
全是假的	-1	-1	O
拍到	1	1	O
666	1	0	X
挺好的	1	1	O
好吃吗	0	0	O
什么都带货	-1	0	X
已拍了	1	1	O
还不错	1	1	O
我已经下单了	1	1	O
好多人	1	1	O
还没抢到	-1	-1	O
全部退单	-1	-1	O
没有	-1	-1	O
.....	....	.....	.....

40 Test data
=>34correct
Accuracy : 85%

5

Limitation

Limitation

- This API only work for Chinese, does not work on other languages.

  why?

  1. There is no meaning on alphabet itself!
  2. The form of word is not fixed. (e.g : “do-did-done-does…”)

- Do not care of relationship between Hanzi (e.g: “不好”)

- Since Data is not enough, only work for particular category

Thanks!

contributions

3210300330金铉洙 : **develop api, make ppt, presentation**

3210103196姜郎山 : **produce datasets, write report**

3210104909张佳骏 : **produce datasets, write report**