CSE431
FInal Exam

Tasnim Sakib Apon

ID : 20241068
Old ID : 18301297

1. Naive bayes algorithm in unkown words:

A simple way to deal with unkown words is simply to add an extra word to the vocabulary.

Let W_U - "unkown word"

$$\hat{P}(W_U|c) = \frac{Count(W_U,c)+1}{\left(\sum_{w \in V} count(w,c)\right)+|V+1|}$$

$$= \frac{1}{\left(\sum_{w \in V} count(w,c)\right)+|V+1|}$$

Unkown words will be modeled with the equation above.

And normally stop words are removed bfore applying naive bayes. However if stops words are found, naive bayes simply ignore the words. Stop words are removed both from test set and train set.

Theme:

Date: / /
Sat Sun Mon Tue wed Thu

2 5 usages of text catagorization

⇒ Language identification
⇒ Authorship ~~ind~~ identification
⇒ Age/Gender identification
⇒ Spam detection.
⇒ Assigning subject catagories, topic or genres.

3

	catagory	Document	Words	Words in same catagory
Train	Positive	I like walking	3	1
	Negative	This is not good	4	1
Test	? ?	I like drinking coffee	4	

4 prior probability: Total No.=2; Positive=1; Negative=1.

$$p(-) = \frac{1}{2} \quad \text{(Negative)}$$

$$p(+) = \frac{1}{2} \quad \text{(positive)}$$

5 Vocabulary size $|V| = 7$ {I, like, walking, This, is, not, good}

Theme:

Date: / /
□Sat □Sun □Mon □Tue □wed □Thu □Fri

6.

$$P(\text{like}|\text{negative}) = \frac{0+1}{4+2} = \frac{1}{11}$$

$$P(\text{like}|\text{positive}) = \frac{1+1}{3+2} = \frac{2}{10} = \frac{1}{5}$$

$$P(\text{drinking}|\text{negative}) = \frac{0+1}{4+2} = \frac{1}{11}$$

$$P(\text{drinking}|\text{positive}) = \frac{0+1}{3+2} = \frac{1}{10}$$

$$P(\text{coffee}|\text{negative}) = \frac{0+1}{4+2} = \frac{1}{11}$$

$$P(\text{coffee}|\text{positive}) = \frac{0+1}{3+2} = \frac{1}{10}$$

$\therefore P(\text{negative})(\text{test}|\text{negative})$

$$\Rightarrow \frac{1}{11} \times \frac{1}{11} \times \frac{1}{11}$$

$$\Rightarrow 7.51 \times 10^{-4}$$

$P(\text{positive})(\text{test}|\text{positive})$

$$\Rightarrow \frac{1}{5} \times \frac{1}{10} \times \frac{1}{10}$$

$$\Rightarrow 2 \times 10^{-3}$$

$\therefore$ So our test document was positive as $P(\text{positive}) > P(\text{negative})$.

Theme:

Date: / /
Sat Sun Mon Tue Wed Thu Fri

Document 1 ⇒ Dialouge system is computer system intended to converse with a human.

Document 2 ⇒ SIRI is a very common chatbot now a days

Document 3 ⇒ Chatbot is also known as dialouge system.

	Document 1	Document 2	Document 3
Chatbot	0	1	1
Dialouge Systen.	1	0	1