B6311506 นายกันตพงศ์ เสาเกลียวตีไกร

Assignment 6

Training data			Test data		
Doc#	Content	Category	Doc#	Content	Category
D1	He loves her badly	Novel	D7	He walks with her	?
D2	I slept	Diary	D8	I fight a dog	?
D3	A dog walks on a highway	Article			
D4	He slept	Novel			
D5	I and a dog walk on a highway	Diary			
D6	He fights near a highway	Article			

*คำที่ควรจะมีรากศัพท์เดียวกัน เช่น walk และ walks จะนับเป็นคำเดียวกัน

(3 คะแนน) 1 ค่าความน่าจะเป็นของ Category แต่ละตัวมีขนาดเท่าไร

(2 คะแนน) 2 ดูจาก Training set แล้วตอบคำถามต่อไปนี้

2.1 จำนวนคำของทุกเอกสารที่ถูกจัดในแต่ละ Category (รวมตัวซ้ำด้วย)

2.2 จำนวนคำของทุกเอกสารของทุก Category (ไม่รวมตัวซ้ำ)

(3 คะแนน) 3 ค่าความน่าจะเป็นของแต่ละคำทั้งหมดโดยอิงจาก Category มีขนาดเท่าไร

(2 คะแนน) 4 จงหาว่าเอกสาร D7 D8 ควรถูกจัดให้อยู่ในกลุ่มใด

1. ค่าความน่าจะเป็นของ Category แต่ละตัวมีขนาดเท่าไร

Ans $P(\text{Novel}) = \frac{2}{6} = \frac{1}{3}$

$P(\text{Diary}) = \frac{2}{6} = \frac{1}{3}$

$P(\text{Article}) = \frac{2}{6} = \frac{1}{3}$

2. ดูจาก Training set แล้วตอบคำถามต่อไปนี้

2.1 จำนวนคำของทุกเอกสารที่ถูกจัดใน แต่ละ Category (รวมตัวซ้ำด้วย)

Ans คำทั้งหมดใน Novel = 6

คำทั้งหมดใน Diary = 10

คำทั้งหมดใน Article = 11

2.2 จำนวนคำของทุกเอกสารของทุก Category (ไม่รวมตัวซ้ำ)

Ans มีทั้งหมด 14 คำ

3. ค่าความน่าจะเป็นของแต่ละคำทั้งหมดโดยอิงจาก Category มีขนาดเท่าไร

Ans ใน Novel

$P(\text{He} \mid \text{Novel}) = 0.15$

$P(\text{her} \mid \text{Novel}) = 0.1$

$P(\text{loves} \mid \text{Novel}) = 0.1$

$P(\text{badly} \mid \text{Novel}) = 0.1$

$P(\text{Slept} \mid \text{Novel}) = 0.1$

$P(\text{I} \mid \text{Novel}) = 0.05$

$P(\text{A} \mid \text{Novel}) = 0.05$

$P(\text{dog} \mid \text{Novel}) = 0.05$

$P(\text{walk} \mid \text{Novel}) = 0.05$

$P(\text{and} \mid \text{Novel}) = 0.05$

$P(\text{on} \mid \text{Novel}) = 0.05$

$P(\text{highway} \mid \text{Novel}) = 0.05$

$P(\text{fight} \mid \text{Novel}) = 0.05$

$P(\text{near} \mid \text{Novel}) = 0.05$

ใน Diary

P(He|Diary) = 0.05 P(her|Diary) = 0.05

P(loves|Diary) = 0.05 P(badly|Diary) = 0.05

P(Slept|Diary) = 0.1 P(I|Diary) = 0.15

P(A|Diary) = 0.15 P(dog|Diary) = 0.1

P(walk|Diary) = 0.1 P(and|Diary) = 0.1

P(on|Diary) = 0.1 P(highway|Diary) = 0.1

P(fight|Diary) = 0.05 P(near|Diary) = 0.05

ใน Article

P(He|Article) = 0.1 P(her|Article) = 0.05

P(loves|Article) = 0.05 P(badly|Article) = 0.05

P(Slept|Article) = 0.05 P(I|Article) = 0.05

P(A|Article) = 0.2 P(dog|Article) = 0.1

P(walk|Article) = 0.1 P(and|Article) = 0.05

P(on|Article) = 0.1 P(highway|Article) = 0.15

P(fight|Article) = 0.1 P(near|Article) = 0.1

4. จงหาว่าเอกสาร D7 D8 ควรถูกจัดให้อยู่ในกลุ่มใด

Ans (D7) $P(Novel|D_7) = P(He|Novel)\cdot P(walks|Novel)\cdot P(with|Novel)\cdot P(her|Novel)\cdot P(Novel)$

$= 0.15 \times 0.05 \times 0.05 \times 0.1 \times \frac{1}{3}$

$= 0.0000125$

$P(Diary|D_7) = P(He|Diary)\cdot P(walks|Diary)\cdot P(with|Diary)\cdot P(her|Diary)\cdot P(Diary)$

$= 0.05 \times 0.1 \times 0.05 \times 0.05 \times \frac{1}{3}$

$= 0.00000416$

$P(Article|D_7) = P(He|Article)\cdot P(walks|Article)\cdot P(with|Article)\cdot P(her|Article)\cdot P(Article)$

$= 0.1 \times 0.1 \times 0.05 \times 0.05 \times \frac{1}{3}$

$= 0.00000833$

∴ D7 ควรถูกจัดใน Novel

(D8) $P(Novel \mid D8) = P(I \mid Novel) \cdot P(fight \mid Novel) \cdot P(a \mid Novel) \cdot P(dog \mid Novel) \cdot P(Novel)$

$= 0.05 \times 0.05 \times 0.05 \times 0.05 \times \frac{1}{3}$

$= 0.00000208$

$P(Diary \mid D8) = P(I \mid Diary) \cdot P(fight \mid Diary) \cdot P(a \mid Diary) \cdot P(dog \mid Diary) \cdot P(Diary)$

$= 0.15 \times 0.05 \times 0.15 \times 0.1 \times \frac{1}{3}$

$= 0.0000375$

$P(Article \mid D8) = P(I \mid Article) \cdot P(fight \mid Article) \cdot P(a \mid Article) \cdot P(dog \mid Article) \cdot P(Article)$

$= 0.1 \times 0.1 \times 0.2 \times 0.1 \times \frac{1}{3}$

$= 0.0000667$

∴ D8 ควรถูกจัดใน Article