حانیه ناصری 810101286 تمرین 6 بازیابی هوشمند اطلاعات

سوال 1 عملی

در فایل نوت بوک قرار داده شده، توضیحات هر بخش کامل نوشته شده است.

سوالات تئوری

1)

الف)

We need to compute p(text | category) * p(category) for each category and then compare them to find most proper category.

P(‘a very close race’ | politics) * p(politics) =

p(‘a’ | politics) * p(‘very’ | politics) * p(‘close’ | politics) * p(‘race’ | politics) * p(politics)

P(‘a very close race’ | not politics) = ?

p(‘a’ | not politics) * p(‘very’ | not politics) * p(‘close’ | not politics) * p(‘race’ | not politics) * p(not politics)

p(politics) = 3/5

p(not politics) = 2/5

p(‘a’| politics) = 1/12

p(‘very’ | politics) = 1/12

p(‘close’ | politics) = 0

p(‘race’ | politics) = 1/12

p(‘a’| not politics) = 2/12

p(‘very’ | not politics) = 0

p(‘close’ | not politics) = 1/12

p(‘race’ | not politics) = 1/12

p(‘a very close race’ | politics) * p(politics) = 1/12 * 1/12 * 0 * 1/12 * 3/5 = 0

p(‘a very close race’ | not politics) * p(not politics) = 2/12 * 0 * 1/12 * 1/12 * 2/5 = 0

we can not decide about category !! we need laplace smoothing

ب)

We add a value of 0.5 to each word count in order to smooth!

P(politics) = 3/5

P(not politics) = 2/5

|V| = 17

P('a' | politics) = (1 + 0.5) / (12 + 0.5 * 17) = 1.5 / 20.5 = 0.0731707317

P('very' | politics) = (1 + 0.5) / (12 + 0.5 * 17) = 1.5 / 20.5 = 0.0731707317

P('close' | politics) = 0.5 / 20.5 = 0.0243902439

P('race' | politics) = (1 + 0.5) / (12 + 0.5 * 17) = 1.5 / 20.5 = 0.0731707317

P('a' | not politics) = (2 + 0.5) / (12 + 0.5 * 17) = 2.5 / 20.5 = 0.1219512195

P('very' | not politics) = 0.5 / (12 + 0.5 * 17) = 0.5 / 20.5 = 0. 0243902439

P('close' | not politics) = (1 + 0.5) / (12 + 0.5 * 17) = 1.5 / 20.5 = 0.0731707317

P('race' | not politics) = (1 + 0.5) / (12 + 0.5 * 17) = 1.5 / 20.5 = 0.0731707317

0.0731707317 * 0.0731707317 * 0. 0243902439 * 0.1219512195 * (2/5) >

0.0731707317 * 0.0731707317 * 0. 0243902439 * 0.0731707317 * (3/5)

As shown, the probability to be in 'non politics' category is more than 'politics' category!

Yes, as shown in this part Laplace smoothing helps us add some small possibility to words not seen in train data of a category.

سوال 2)

الف)

Precision = 70 / 70 + 30 = 0.7

Recall = 70 / 70 + 70 = 0.5

ب)

کاربر اول نمیخواهد لزوما همه ایمیل های درست را بگیرد. او بیشتر ترجیح می دهد که از بین ایمیل های دریافت شده اکثریت درست باشند نه اینکه همه لزوما دریافت شده باشند. به عبارتی دقت (precision) برایش مهم تر است. در مقابل، کاربر دوم ترجیح می دهد تا جایی که میشود همه درست ها را ببیند حتی شده به قیمت دیدن تعداد زیادی اسپم میان پیشبینی شده ها. پس recall بیشتر برایش مهم است.