

# Tutorial 11:

## Natural Language Processing

### Question 1

a)

Frequencies of unigrams:

a=5                      b=3                      c=3                      d=1

Frequencies of bigrams (note that there is also a stop character):

aa=2	ba=1	ca=0	da=1
ab=1	bb=1	cb=1	db=0
ac=2	bc=0	cc=1	dc=0
ad=0	bd=1	cd=0	dd=0

Although not needed for this question, a more accurate model will also include a stop character:

astop = 0                      bstop=0                      cstop=1                      dstop=0

b)

Probabilities for unigram model:

P(a)=5/12                      P(b)=3/12                      P(c)=3/12                      P(d)=1/12

Conditional probabilities for bigram model:

P(a a)=2/5	P(a b)=1/3	P(a c)=0/3	P(a d)=1/1
P(b a)=1/5	P(b b)=1/3	P(b c)=1/3	P(b d)=0/1
P(c a)=2/5	P(c b)=0/3	P(c c)=1/3	P(c d)=0/1
P(d a)=0/5	P(d b)=1/3	P(d c)=0/3	P(d d)=0/1

And for the stop characters:

P(stop|a) = 0/5                      P(stop|b) = 0/3                      P(stop|c) = 1/3                      P(stop|d) = 0/1

c)

Probability of the sequence 'abba' according to the unigram model:

$P(\text{abba}) = P(a) \cdot P(b) \cdot P(b) \cdot P(a) = 5/12 \cdot 3/12 \cdot 3/12 \cdot 5/12 = 0.01085$

Probability of the sequence 'abba' according to the bigram model:

$P(\text{abba}) = P(a) \cdot P(b|a) \cdot P(b|b) \cdot P(a|b) = 5/12 \cdot 1/5 \cdot 1/3 \cdot 1/3 = 0.009259$

### Question 2

Q = "garbage recycle"

a)

**Unigram model:**

$P(Q|D1,r) = P(\text{word}=\text{"garbage"}|D1,r) \cdot P(\text{word}=\text{"recycle"}|D1,r)$

$$= (35/398) * (13/398)$$

$$\approx 0.00287$$

$$P(Q|D2,r) = P(\text{word}=\text{"garbage"}|D2,r) * P(\text{word}=\text{"recycle"}|D2,r)$$

$$= (15/491) * (14/491)$$

$$\approx 0.000871$$

### Bi-gram model:

There are two ways of working out the probability of the query phrase for the bi-gram model.

The first is the probability of the bigram given all of the possible bigrams in the document:

$$P(Q|D1,r) = P(\text{word pair}=\text{"garbage recycle"}|D1,r)$$

$$= (3/397)$$

$$\approx 0.00756$$

$$P(Q|D2,r) = P(\text{word pair}=\text{"garbage recycle"}|D2,r)$$

$$= (4/490)$$

$$\approx 0.00816$$

The second is using the probabilities and conditional probabilities as in the previous question. That is, the probability of the first word multiplied by the probability of the second word given the first word:

$$P(Q|D1,r) = P(\text{"garbage"}|D1,r) * P(\text{"recycle"}|\text{"garbage"},D1,r)$$

$$= 35/398 * 3/35$$

$$= 3/398$$

$$\approx 0.00754$$

$$P(Q|D2,r) = P(\text{"garbage"}|D2,r) * P(\text{"recycle"}|\text{"garbage"},D2,r)$$

$$= 15/491 * 4/15$$

$$= 4/491$$

$$\approx 0.00815$$

Either can be used, and end up giving very similar values. The first version is used for part b.

b)

### Unigram model:

$$P(r|D,Q)/P(-r|D,Q) = P(Q|D,r)P(r|D) / [\alpha (1-P(r|D))]$$

(for some constant alpha).

$$P(r|D1) = 4/200000 = 1/50000$$

$$P(r|D2) = 8/200000 = 1/25000$$

So

$$P(r|D1,Q)/P(-r|D1,Q) = P(Q|D1,r)P(r|D1) / [\alpha (1-P(r|D1))]$$

$$= (.00287/\alpha) * (1/49999) = 5.74 * 10^{(-8)}/\alpha$$

$$P(r|D2,Q)/P(-r|D2,Q) = P(Q|D2,r)P(r|D2) / [\alpha (1-P(r|D2))]$$

$$= (.000871/\alpha) * (1/24999) = 3.48 * 10^{(-8)}/\alpha$$

We don't need to know alpha to work out that D1 (document 1) is more relevant under the uni-gram model because the ratio

$$[P(r|D1,Q)/P(-r|D1,Q)] / [P(r|D2,Q)/P(-r|D2,Q)] > 1.$$

### Bi-gram model:

$$P(r|D,Q)/P(-r|D,Q) = P(Q|D,r)P(r|D) / [\alpha (1-P(r|D))]$$

(note that this constant alpha will be different to that in the unigram model)

$$P(r|D1) = 4/200000 = 1/50000$$

$$P(r|D2) = 8/200000 = 1/25000$$

So

$$P(r|D1,Q)/P(-r|D1,Q) = P(Q|D1,r)P(r|D1) / [\alpha (1-P(r|D1))] \\ = (.00756/\alpha)*(1/49999) = 1.51*10^{(-7)}/\alpha$$

$$P(r|D2,Q)/P(-r|D2,Q) = P(Q|D2,r)P(r|D2) / [\alpha (1-P(r|D2))] \\ = (.00816/\alpha)*(1/24999) = 3.26*10^{(-7)}/\alpha$$

Since the value for D2 is larger, under the bi-gram model, we would choose D2 (document 2) as the most relevant to this query.

The uni-gram and bi-gram models do not have to agree. Provided we have enough data (words) to fit a bigram model fairly accurately, it should be better than the unigram model, provided people do mean sentence-like associations between adjacent words in a query.

## Question 3

Unigram model only.

So we just need to find out the relative frequencies of the query words "train" and "classifier". So we need to work out which of the words count as either of these (after stemming), count how many there are in each document and count the total number of words in each document. Depending on the stemming rules you use and your treatment of numbers, you may end up with slightly different values for the total number of words and the count of each of the query words in each document.

D1 contains 56 words

D2 contains 61 words

Stemming rules used:

training -> train

overtraining -> train

classifiers -> classifier

Treatment of numbers:

- starts with a numeral, continues if following character is a numeral or a '.' or a ','. If have a '.' or ',', next character must be a numeral, or the number word ends.

-> this makes 18.10 a number

- but 748-752 is two numbers.

Counts of words:

	D1	D2
train	2	5
classifier	3	2

Assuming that both paragraphs are equally relevant to general queries,  
ie:  $P(r|D1) = P(r|D2)$ .

Which document is most relevant to the query?

Check if

$$\frac{P(r|D1,Q)/P(-r|D1,Q)}{P(r|D2,Q)/P(-r|D2,Q)} > 1 ?$$

If so, D1 will be deemed more relevant. Otherwise we choose D2.

This ratio is:

$$\frac{P(Q|D1,r)P(r|D1)}{[\alpha (1-P(r|D1))]}$$

\*

$$\frac{[\alpha (1-P(r|D2))]}{P(Q|D2,r)P(r|D2)}$$

We know that  $\frac{P(r|D1)}{[\alpha (1-P(r|D1))]}$

$= \frac{P(r|D2)}{[\alpha (1-P(r|D2))]}$ , so that cancels, leaving us  
with the ratio  $\frac{P(Q|D1,r)}{P(Q|D2,r)}$ .

$$\frac{P(Q|D1,r)}{P(Q|D2,r)} = \frac{(2/56)*(3/56)}{((5/61)*(2/61))}$$
$$\approx 0.6536$$

This is  $< 1$ , so D2 (Paragraph 2) is more relevant to this query under the unigram model.