Web Information Retrieval

Assignment 5

Team Name : Gamma

Members

Name	Matriculation Number
Mohammad Nizam Uddin	216101140
Md. Shohel Ahamad	216203438
MD Jakaria Nawaz	216203442
Shreya Chatterjee	216100848

1. Query Likelihood Model (14 Points) :

Doc 1: “teach education school university education”.
Doc 2: “education education campus teach teach”.
Doc 3: “university road school teach learning”.
Doc 4: “campus learning education learning”.

Query: “teach teach education campus”.

1.1

$$P_{M_{d_i}}(t_j)$$

Term(t)	D1	**D2**	**D3**	**D4**
campus	0	0.2	0	0
education	0.4	0.4	0	0.25
learning	0	0	0.2	0.50
road	0	0	0.2	0
school	0.2	0	0.2	0
teach	0.2	0.4	0.2	0
university	0.2	0	0.2	0

1.2.

Term(t)	$P_{M_c}(t_j)$
campus	0.105
education	0.263
learning	0.158
road	0.053
school	0.105
teach	0.210
university	0.105

2.

$$P_{\text{uni}}(d \mid q).$$

Term(t)	d1	**d2**	**d3**	**d4**	$P_{\text{uni}}(d \mid q).$
campus	0.2	0.4	0.2	0	0
education	0.4	0.4	0	0.25	0
teach	0.2	0.4	0.2	0	0

3.

$$P_{\text{interp-uni}}(d \mid q)$$

.

Query: “teach teach education campus”.

Term(t)	d1	**d2**	**d3**	**d4**	$P_{\text{interp-uni}}(d \mid q)$
campus	(0*.5+.105*.5) =0.525	(.2*.5+.105*.5) =0.625	(0*.5+.105*.5) =0.525	(0*.5+.105*.5) =0.525	0.09
education	(.4*.5+.263*.5)=0.3315	(.4*.5+.263*.5) =0.3315	(0*.5+.263*.5) =.1315	(.25*.5+.263*.5)=0.748	0.01
teach	(.2*.5+.21*.5) =0.205	(.4*.5+.21*.5)= 0.305	(.2*.5+.21*.5) =0.205	(0*.5+.21*.5)= 0.105	0.001

2. n-gram Models (10 Points):

1. Estimate the probability of a term sequence t1 t2 t3 t4 t5 appearing in a document.

Bigram: [{tf(t4 t5,d)} / {tf(t4,d)}] * [{tf(t2 t4,d)} / {tf(t4,d)}] * [{tf(t2 t3,d)} / {tf(t3,d)}] * [{tf(t1 t2,d)} / {tf(t1,d)}].
Trigram: [{tf(t3 t4 t5,d)} / {tf(t3 t4,d)}] * [{tf(t2 t3 t4,d)} / {tf(t2 t3,d)}] * [{tf(t1 t2 t3,d)} / {tf(t1 t2,d)}]

General Formula:

Bigram = $\prod_{n=2 \text{ to } n=i}$ [{tf(t_{n-1} t_n,d)} / {tf(t_{n-1},d)}]

Trigram = $\prod_{n=3 \text{ to } n=i}$ [{tf(t_{n-2} t_{n-1} t_n,d)} / {tf(t_{n-2} t_{n-1},d)}]

2.2

Doc 1: “rose is a rose is a rose is a rose”
Doc 2: “rose rose rose rose is is is a a a”
Doc 3: “rose is a rose”
Doc 4: “a rose is a”

	d1	d2	d3	d4
rose , t1	4/10=0.4	4/10=0.4	2/4=0.5	¼=0.25
is , t2	3/10=0.3	3/10=0.3	¼=0.25	¼=0.25
a , t3	3/10=0.3	3/10=0.3	¼=0.25	2/4=0.5

‘rose is a rose’

Probability according to unigram model:

Doc1 = tf(rose , d)/dl(d1) * tf(is , d)/dl(d1) * tf(a , d)/dl(d1) * tf(rose , d)/dl(d1)
=> 0.4 * 0.3 * 0.3 * 0.4
=> 0.0144
Doc2 = tf(rose , d)/dl(d2) * tf(is , d)/dl(d2) * tf(a , d)/dl(d2) * tf(rose , d)/dl(d2)
=> 0.4 * 0.3 * 0.3 * 0.4
=> 0.0144
Doc3 = tf(rose , d)/dl(d3) * tf(is , d)/dl(d3) * tf(a , d)/dl(d3) * tf(rose , d)/dl(d3)
=> 0.5 * 0.25 * 0.25 * 0.5
=> 0.0156
Doc4 = tf(rose , d)/dl(d4) * tf(is , d)/dl(d4) * tf(a , d)/dl(d4) * tf(rose , d)/dl(d4)
=> 0.25 * 0.25 * 0.5 * 0.25
=> 0.0078

Probability according to bigram model:

	d1	d2	d3	d4
rose is , t1	3/4=0.75	1/4=0.25	1/2=0.5	1/1=1
Is a , t2	3/3=1	1/3=0.33	1/1=1	1/1=1
a rose , t3	3/3=1	0/3=0	1/1=1	1/2=0.5

Doc1 = tf(rose is , d)/tf(rose , d) * tf(is a , d)/tf(is , d) * tf(a rose , d)/tf(a , d)
=> 0.75 * 1 * 1
=> 0.75
Doc2 = tf(rose is , d)/tf(rose , d) * tf(is a , d)/tf(is , d) * tf(a rose , d)/tf(a , d)
=> 0.25 * 0.33 * 0
=> 0
Doc3 = tf(rose is , d)/tf(rose , d) * tf(is a , d)/tf(is , d) * tf(a rose , d)/tf(a , d)
=> 0.5 * 1 * 1
=> 0.5
Doc4 = tf(rose is , d)/tf(rose , d) * tf(is a , d)/tf(is , d) * tf(a rose , d)/tf(a , d)
=> 1 * 1 * 0.5
=> 0.5

Probability according to trigram model:

	d1	d2	d3	d4
rose is a , t1	3/3=1	0/1=0	1/1=1	1/1=1
Is a rose , t2	3/3=1	0/1=0	1/1=1	0/1=0

Doc1 = tf(rose is a , d)/tf(rose is , d) * tf(is a rose , d)/tf(is a , d)
=> 1 * 1
=> 1
Doc2 = tf(rose is a , d)/tf(rose is , d) * tf(is a rose , d)/tf(is a , d)
=> 0 * 0

```
        => 0
Doc3    = tf(rose is a , d)/tf(rose is , d) * tf(is a rose , d)/tf(is a , d)
        => 1 * 1
        => 1
Doc4    = tf(rose is a , d)/tf(rose is , d) * tf(is a rose , d)/tf(is a , d)
        => 1 * 0
        => 0
```