Feedback — Week 2 Quiz

Help Center

You submitted this quiz on **Sat 11 Jul 2015 12:30 AM PDT**. You got a score of **10.00** out of **10.00**.

Question 1

You are given a unigram language model θ distributed over a vocabulary set V composed of **only** 4 words: “the”, “global”, “warming”, and “effects”. The distribution of θ is given in the table below:

w	$P(w\|\theta)$
the	0.3
global	0.2
warming	0.2
effects	X

What is X, i.e. $P(\text{“effects”}|\theta)$?

Your Answer		Score	Explanation
0.3	✔	1.00	
0.2			
0.1			
0			
Total		1.00 / 1.00	

Question 2

Assume you are given the same unigram language model as in Question 1. Which of the following is **not** true?

Your Answer		Score	Explanation
$P(\text{“the global warming effects”}\|\theta) < P(\text{“global warming effects”}\|\theta)$			
$P(\text{“text mining”}\|\theta) = 0$			
$P(\text{“global warming”}\|\theta) > P(\text{“warming global”}\|\theta)$	✔	1.00	
$P(\text{“global warming”}\|\theta) = 0.04$			
Total		1.00 / 1.00	

Question 3

Assume that words are being generated by a mixture of two unigram language models, θ_1 and θ_2, where $P(\theta_1) = 0.5$ and $P(\theta_2) = 0.5$. The distributions of the two models are given in the table below:

w	$P(w\|\theta_1)$	$P(w\|\theta_2)$
sports	0.35	0.05
basketball	0.2	0.05
fast	0.3	0.3
computer	0.1	0.4
smartphone	0.05	0.2

Then the probability of observing “computer” from this mixture model is: $P(\text{“computer”}) =$

Your Answer		Score	Explanation
0.4			
0.05			
0.25	✔	1.00	
0.45			
Total		1.00 / 1.00	

Question 4

Assume the same given as in Question 3. We now want to infer which of the two word distributions, θ_1 and θ_2, has been used to generate "computer", and would thus like to compute the probability that it has been generated using θ_1 and θ_2, i.e., $P(\theta_1|\text{“computer”})$, and $P(\theta_2|\text{“computer”})$, respectively, then the values of $P(\theta_1|\text{“computer”})$ and $P(\theta_2|\text{“computer”})$ are:

Hint: Apply Bayes Rule (This is similar to what we need to compute in the E-step of the EM algorithm)

Your Answer		Score	Explanation
0.9 and 0.1			
0.2 and 0.8	✔	1.00	
0.8 and 0.2			
0.1 and 0.9			
Total		1.00 / 1.00	

Question 5

Suppose words are being generated using a mixture of two unigram language models θ_1 and θ_2. Let $P(w)$ denote the probability of generating a word w from this mixture model.

If $P(\theta_1) = 1$ then which of the following statements is true?

Your Answer		Score	Explanation
$P(w|\theta_2) = 0$, for any word w			
$P(w|\theta_1) = 0$, for any word w			
$P(w) = P(w|\theta_1)$, for any word w	✔	1.00	
Total		1.00 / 1.00	

Question 6

Let $\theta_1, \ldots, \theta_k$ be the k unigram language models output by PLSA. Then, for a specific word w the following relation always holds: $\sum_{i=1}^{k} P(w|\theta_i) = 1$.

Your Answer		Score	Explanation
False	✔	1.00	
True			
Total		1.00 / 1.00	

Question 7

You are given a document d that contains only two words: “the” and “machine”. Assume that this document was generated from a mixture of two unigram language models: a known background language model θ_B and an unknown topic language model θ_d. Let $P(\theta_B) = \lambda$ and $P(\theta_d) = 1 - \lambda$ and assume that $P(\text{“the”}|\theta_B) = 0.9$ and $P(\text{“machine”}|\theta_B) = 0.1$. We want to estimate θ_d using maximum likelihood. Then, as λ increases, $P(\text{“machine”}|\theta_d)$ will:

Hint: First get the maximum likelihood estimates of the two words in θ_d (refer to the lecture on "Probabilistic Topic Models: Mixture Model Estimation"). Then, write $P(\text{“machine”}|\theta_d)$ as a function of λ and study the behavior of the function.

Your Answer		Score	Explanation
Increase	✔	1.00	
Decrease			
Remain the same			
Total		1.00 / 1.00	

Question 8

When using PLSA to mine topics from a text collection, the number of parameters of the PLSA model stays the same as we keep adding new documents into the text collection assuming that the new documents do not introduce new words that have not occurred in the current text collection.

Your Answer		Score	Explanation
True			
False	✔	1.00	
Total		1.00 / 1.00	

Question 9

Suppose we have the following word counts for two documents d_1 and d_2.

Table 1: Counts for words in document set

Vocabulary Words	$c(w, d_1)$	$c(w, d_2)$	$P(w\|\theta_B)$
text	5	0	0.15
mining	4	0	0.05
the	4	4	0.50
fifa	0	5	0.10
football	0	2	0.20

We are interested in applying topic modeling to discover two topics, θ_0 and θ_1, in our corpus of two documents. Suppose that we run PLSA with the number of topics set to 2 (i.e. $k = 2$) while using an additional known (fixed) background word distribution θ_B as shown in Table 1. Using the EM algorithm, and after n iterations, the E-step gives the following estimates:

Table 2: Output of E-step after n iterations.

Documents	**Words**	$P(z_{w,d} = 0)$	$P(z_{w,d} = 1)$	$P(z_{w,d} = B)$
d_1	text	1.00	0.00	0.20
	mining	1.00	0.00	0.10

	the	0.60	0.40	0.90
d_2	the	0.40	0.60	0.90
	fifa	0.00	1.00	0.10
	football	0.00	1.00	0.10

Assume $\lambda_B = P(\theta_B) = 0.20$ and recall that $P(z_{w,d} = 0) + P(z_{w,d} = 1) = 1$ as discussed in the lectures. After completing the M-step, $P(\text{the}|\theta_0) =$?

Your Answer		Score	Explanation
0.16			
0.4			
0.05	✔	1.00	
0.24			
Total		1.00 / 1.00	

Question 10

Assume the same given as in Question (9) and recall that $\pi_{d_1,0} + \pi_{d_1,1} = 1$. What is $\pi_{d_1,0}$?

Your Answer		Score	Explanation
0.98	✔	1.00	
1.00			
0.99			
0.84			
Total		1.00 / 1.00	