Review Test Submission: Fall 2018 Assignment 3

User	Tianqi Wu
Course	Fall 2018-CS 510-Advanced Information Retrieval-Section A1
Test	Fall 2018 Assignment 3
Started	9/22/18 12:13 PM
Submitted	9/22/18 7:17 PM
Due Date	9/23/18 11:59 PM
Status	Completed
Attempt Score	100 out of 100 points
Time Elapsed	7 hours, 4 minutes
Results Displayed	All Answers, Submitted Answers, Correct Answers, Feedback, Incorrectly Answered Questions

Question 1

10 out of 10 points

Unigram Language Model is a special class of N-Gram Language Model where the next word in the document is assumed to be independent of the previous words generated by the model. Mathematically, this is written as $P(w_m|w_{m-1},...,w_1)$ = $P(w_m)$.

We can estimate the parameters of a Unigram Language Model through Maximum Likelihood Estimation. Given a particular document d and the vocabulary set V, the maximum likelihood estimator for a Unigram Language Model is given by the following formula:

$$P_{ML}(w|d) = \frac{c(w,d)}{\sum_{w' \in V} c(w',d)}$$

Now, let us assume that, we have seen the following document d: **"*science without religion is lame, religion without science is blind.*"** Assuming that this document was generated by a Unigram Language Model and words in the document d constitute the entire vocabulary, what are the parameters of this Unigram Language Model? Estimate the values of all these parameters using the maximum likelihood estimator ***(The symbol # indicates the 'beginning of sentence' marker)***.

Selected Answer:

p("science"|d) = 2/11

p("without"|d) = 2/11

p("religion"|d) = 2/11

p("is"|d) = 2/11

p("lame"|d) = 1/11

p("blind"|d) = 1/11

p(#|d) = 1/11

Answers:

p("science"|d) = 2/21

p("without"|d) = 2/21

p("religion"|d) = 2/21

p("is"|d) = 2/21

p("lame"|d) = 1/21

p("blind"|d) = 1/21

p(","|d) = 1/21

p(" "|d) = 9/21

p("."|d) = 1/21

p("science"|d) = 2/11

p("without"|d) = 2/11

p("religion"|d) = 2/11

p("is"|d) = 2/11

p("lame"|d) = 1/11

p("blind"|d) = 1/11

p(#|d) = 1/11

p("science"|d) = 2/10
p("without"|d) = 2/10
p("religion"|d) = 2/10
p("is"|d) = 2/10
p("lame"|d) = 1/10
p("blind"|d) = 1/10

p("science"|d) = 2/6
p("without"|d) = 2/6

p("religion"|d) = 2/6

p("is"|d) = 2/6

p("lame"|d) = 1/6

p("blind"|d) = 1/6

Question 2

10 out of 10 points

Continuing the setup of the previous question; now assume that, the entire vocabulary V consists of the set:

{science, religion, without, is, lame, blind, retrieval, model, BM25}

If we consider the same document d: **"*science without religion is lame, religion without science is blind.*"** and assume again that this document was generated by a Unigram Language Model, what are the parameters of this Unigram Language Model in this case? Estimate the values of all these parameters using the maximum likelihood estimator (The symbol # indicates a 'beginning of sentence' marker).

Selected Answer:

p("science"|d) = 2/10

p("without"|d) = 2/10

p("religion"|d) = 2/10

p("is"|d) = 2/10

p("lame"|d) = 1/10

p("blind"|d) = 1/10

p("retrieval"|d) = 0/10

p("model"|d) = 0/10

p("BM25"|d) = 0/10

Answers:

p("science"|d) = 2/10

p("without"|d) = 2/10

p("religion"|d) = 2/10

p("is"|d) = 2/10

p("lame"|d) = 1/10

p("blind"|d) = 1/10

p("retrieval"|d) = 0/10

p("model"|d) = 0/10

p("BM25"|d) = 0/10

p("science"|d) = 2/21

p("without"|d) = 2/21

p("religion"|d) = 2/21

p("is"|d) = 2/21

p("lame"|d) = 1/21

p("blind"|d) = 1/21

p(","|d) = 1/21

p(" "|d) = 9/21

p("."|d) = 1/21

p("retrieval"|d) = 0/21

p("model"|d) = 0/21

p("BM25"|d) = 0/21

p("science"|d) = 2/9

p("without"|d) = 2/9

p("religion"|d) = 2/9

p("is"|d) = 2/9

p("lame"|d) = 1/9

p("blind"|d) = 1/9

p("retrieval"|d) = 0/9

p("model"|d) = 0/9

p("BM25"|d) = 0/9

p("science"|d) = 2/11

p("without"|d) = 2/11

p("religion"|d) = 2/11

p("is"|d) = 2/11

p("lame"|d) = 1/11

p("blind"|d) = 1/11

p("retrieval"|d) = 0/11

p("model"|d) = 0/11

p("BM25"|d) = 0/11

p(#|d) = 1/11

Question 3

10 out of 10 points

Bigram Language Model is another special class of N-Gram Language Model where the next word in the document depends only on the immediate preceding word. Mathematically, this is written as the conditional probability, $P(w_m|w_{m-1},..,w_1) = P(w_m|w_{m-1})$.

Given the same document d: "science without religion is lame, religion without science is blind." and the vocabulary set from the previous question:

{science, religion, without, is, lame, blind, retrieval, model, BM25}

Assume that document d is now generated by a Bigram Language Model, how many parameters does this Bigram Language Model have? ***(Do not forget to consider the beginning of sentence marker # in your calculations)***

Selected Answer: 90
Correct Answer: 90
Answer range +/- 0 (90.0 - 90.0)

Response Feedback: How many P(A|B) do we need to specify? Since there are 9 unique words in the vocabulary, |A| = 9 , and |B| = 10 (since we need to count the beginning of the sentence marker # too). Thus there are: 9*10 = 90 parameters to specify.

Question 4

10 out of 10 points

Continued from Question 3:

Using the maximum likelihood estimator, estimate the values of the following parameters (assume # to be the start of the sentence marker):

P("without"|"science") = **[a]**

P("science"|"without") = **[b]**

P("religion"|"is") = **[c]**

P("is"|"religion") = **[d]**

P("is"|"blind") = **[e]**

P("lame"|"blind") = **[f]**

P("model"|"is") = **[g]**

P("model"|"retrieval") = **[h]**

P("science"|#) = **[i]**

P("BM25"|#) = **[j]**

(Round your answer to 1 decimal point. Example: 74 --> 74.0)

Specified Answer for: a 0.5
Specified Answer for: b 0.5
Specified Answer for: c 0.0
Specified Answer for: d 0.5
Specified Answer for: e 0.0
Specified Answer for: f 0.0

Specified Answer for: g 0.0

Specified Answer for: h 0.0

Specified Answer for: i 1.0

Specified Answer for: j 0.0

Correct Answers for: a		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.5	
Correct Answers for: b		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.5	
Correct Answers for: c		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0	
Correct Answers for: d		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.5	
Correct Answers for: e		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0	
Correct Answers for: f		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0	
Correct Answers for: g		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0	
Correct Answers for: h		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0	
Correct Answers for: i		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	1.0	
Correct Answers for: j		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0	

Question 5

20 out of 20 points

Let $p_{J-M}(w|d)$ represent a Unigram Language model smoothed using Jelinek-Mercer smoothing.

And let $p_{Dir}(w|d)$ represent a Unigram Language model smoothed using Dirichlet Prior smoothing.

Which of these statements are correct? (Choose all that apply)

Selected Answers:

- As $|d| \rightarrow \infty$, then $p_{J-M}(w|d)$ becomes proportional to the background language model.
- As $\lambda \rightarrow 1$, then $p_{J-M}(w|d)$ becomes equivalent to the background language model.
- As $\mu \rightarrow 0$, then $p_{Dir}(w|d)$ becomes equivalent to the maximum likelihood estimate of the unigram language model.
- As $\lambda \rightarrow 0$, then $p_{J-M}(w|d)$ becomes equivalent to the maximum likelihood estimate of the unigram language model.
- As $\mu \rightarrow \infty$, then $p_{Dir}(w|d)$ becomes equivalent to the background language model.

Answers:

- As $|d| \rightarrow \infty$, then $p_{J-M}(w|d)$ becomes proportional to the background language model.
- As $\lambda \rightarrow 0$, then $p_{J-M}(w|d)$ becomes equivalent to the background language model.
- As $|d| \rightarrow \infty$, then $p_{Dir}(w|d)$ becomes equivalent to the background language model.
- As $\mu \rightarrow \infty$, then $p_{Dir}(w|d)$ becomes equivalent to the maximum likelihood estimate of the unigram language model.
- As $|d| \rightarrow \infty$, then $p_{Dir}(w|d)$ becomes equivalent to the maximum likelihood estimate of the unigram language model.
- As $\lambda \rightarrow 1$, then $p_{J-M}(w|d)$ becomes equivalent to the background language model.
- As $\mu \rightarrow 0$, then $p_{Dir}(w|d)$ becomes equivalent to the background language model.
- As $\lambda \rightarrow 1$, then $p_{J-M}(w|d)$ becomes equivalent to the maximum likelihood estimate of the unigram language model.
- As $|d| \rightarrow \infty$, then $p_{J-M}(w|d)$ becomes proportional to the maximum likelihood estimate of the unigram language model.
- As $\mu \rightarrow 0$, then $p_{Dir}(w|d)$ becomes equivalent to the maximum likelihood estimate of the unigram language model.

As $\lambda \rightarrow 0$, then $p_{J-M}(w|d)$ becomes equivalent to the maximum likelihood estimate of the unigram language model.

As $\mu \rightarrow \infty$, then $p_{Dir}(w|d)$ becomes equivalent to the background language model.

Question 6

10 out of 10 points

Again, Consider the document d: **"*science without religion is lame, religion without science is blind.*"**

This time, assume that we have a background word distribution (pre-computed somehow) denoted by *REF* which is characterized as follows:

P_{REF}("science") = 0.18

P_{REF}("religion") = 0.17

P_{REF}("without") = 0.13

P_{REF}("is") = 0.02

P_{REF}("lame") = 0.05

P_{REF}("blind") = 0.04

P_{REF}("retrieval") = 0.16

P_{REF}("model") = 0.10

P_{REF}("BM25") = 0.15

Assume document d is generated by a Unigram Language Model. Estimate the parameters of the Unigram Language Model using Dirichlet Prior Smoothing assuming ***u*** = 4:

P("science"|*d*) = **[a]**

P("religion"|*d*) = **[b]**

P("without"|*d*) = **[c]**

P("is"|*d*) = **[d]**

P("lame"|*d*) = **[e]**

P("blind"|*d*) = **[f]**

P("retrieval"|*d*) = **[g]**

P("model"|*d*) = **[h]**

P("BM25"|*d*) = **[i]**

(Round your answers to 4 decimal places. Example: 0.12045 --> 0.1205)

Specified Answer for: a 0.1943

Specified Answer for: b 0.1914

Specified Answer for: c 0.1800

Specified Answer for: d 0.1486

Specified Answer for: e 0.0857

Specified Answer for: f 0.0829

Specified Answer for: g 0.0457

Specified Answer for: h 0.0286

Specified Answer for: i 0.0429

Correct Answers for: a		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1943	
Correct Answers for: b		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1914	
Correct Answers for: c		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1800	
Correct Answers for: d		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1486	
Correct Answers for: e		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0857	
Correct Answers for: f		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0829	
Correct Answers for: g		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0457	
Correct Answers for: h		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0286	
Correct Answers for: i		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0429	

Question 7

10 out of 10 points

Repeat the previous question assuming ***u*** = 100. Do the results match with your intuition and your formulas in Question 5?

P("science"|d) = **[a]**

P("religion"|d) = **[b]**

P("without"|d) = **[c]**

P("is"|d) = **[d]**

P("lame"|d) = **[e]**

P("blind"|d) = **[f]**

P("retrieval"|d) = **[g]**

P("model"|d) = **[h]**

P("BM25"|d) = **[i]**

(Round your answers to 4 decimal places. Example: 0.12045 --> 0.1205)

Specified Answer for: a 0.1818

Specified Answer for: b 0.1727

Specified Answer for: c 0.1364

Specified Answer for: d 0.0364

Specified Answer for: e 0.0545

Specified Answer for: f 0.0455

Specified Answer for: g 0.1455

Specified Answer for: h 0.0909

Specified Answer for: i 0.1364

Correct Answers for: a		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1818	
Correct Answers for: b		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1727	
Correct Answers for: c		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1364	
Correct Answers for: d		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0364	
Correct Answers for: e		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0545	
Correct Answers for: f		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0455	
Correct Answers for: g		
Evaluation Method	**Correct Answer**	**Case Sensitivity**

Exact Match	0.1455	
Correct Answers for: h		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0909	
Correct Answers for: i		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1364	

Question 8

20 out of 20 points

Repeat Question 6 with Jelinek-Mercer smoothing instead of Dirichlet Prior Smoothing assuming

$$\lambda = \{0.01, 0.9\}$$

Do the results match with your intuition about what happens when lambda approaches 0 or approaches 1?

$\lambda = 0.01$:

P("science"|d) = **[a]**

P("religion"|d) = **[b]**

P("without"|d) = **[c]**

P("is"|d) = **[d]**

P("lame"|d) = **[e]**

P("blind"|d) = **[f]**

P("retrieval"|d) = **[g]**

P("model"|d) = **[h]**

P("BM25"|d) = **[i]**

$\lambda = 0.9$:

P("science"|d) = **[j]**

P("religion"|d) = **[k]**

P("without"|d) = **[l]**

P("is"|d) = **[m]**

P("lame"|d) = **[n]**

P("blind"|d) = **[o]**

P("retrieval"|d) = **[p]**

P("model"|d) = **[q]**

P("BM25"|d) = **[r]**

(Round your answers to 4 decimal places. Examples: 0.12045 --> 0.1205 , and 0.012 --> 0.0120)

Specified Answer for: a	0.1998
Specified Answer for: b	0.1997
Specified Answer for: c	0.1993
Specified Answer for: d	0.1982
Specified Answer for: e	0.0995
Specified Answer for: f	0.0994
Specified Answer for: g	0.0016
Specified Answer for: h	0.0010
Specified Answer for: i	0.0015
Specified Answer for: j	0.1820
Specified Answer for: k	0.1730
Specified Answer for: l	0.1370
Specified Answer for: m	0.0380
Specified Answer for: n	0.0550
Specified Answer for: o	0.0460
Specified Answer for: p	0.1440
Specified Answer for: q	0.0900
Specified Answer for: r	0.1350

Correct Answers for: a		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1998	
Correct Answers for: b		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1997	
Correct Answers for: c		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1993	
Correct Answers for: d		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1982	
Correct Answers for: e		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0995	
Correct Answers for: f		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0994	
Correct Answers for: g		
Evaluation Method	**Correct Answer**	**Case Sensitivity**

Exact Match	0.0016	
Correct Answers for: h		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0010	
Correct Answers for: i		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0015	
Correct Answers for: j		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1820	
Correct Answers for: k		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1730	
Correct Answers for: l		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1370	
Correct Answers for: m		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0380	
Correct Answers for: n		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0550	
Correct Answers for: o		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0460	
Correct Answers for: p		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1440	
Correct Answers for: q		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.0900	
Correct Answers for: r		
Evaluation Method	**Correct Answer**	**Case Sensitivity**
Exact Match	0.1350	

Saturday, September 22, 2018 7:17:43 PM CDT

← OK