Xiaowen Mao
PID: A53220159
x9mao@eng.ucsd.edu

Oct 25

ASSIGNMENT 4.3

(a)

tokens	unigram probability
A	0.018407
AND	0.017863
AT	0.004313
AS	0.003992
AN	0.002999
ARE	0.002990
ABOUT	0.001926
AFTER	0.001347
ALSO	0.001310
ALL	0.001182
A.	0.001026
ANY	0.000632
AMERICAN	0.000612
AGAINST	0.000596
ANOTHER	0.000428
AMONG	0.000374
AGO	0.000357
ACCORDING	0.000348
AIR	0.000311
ADMINISTRATION	0.000292
AGENCY	0.000280
AROUND	0.000277
AGREEMENT	0.000263
AVERAGE	0.000259
ASKED	0.000258
ALREADY	0.000249
AREA	0.000231
ANALYSTS	0.000226
ANNOUNCED	0.000227
ADDED	0.000221
ALTHOUGH	0.000214
AGREED	0.000212
APRIL	0.000207
AWAY	0.000202

(b)

Most likely words	bigram probability
<UNK>	0.615020
U.	0.013372
FIRST	0.011720
COMPANY	0.011659
NEW	0.009451

(c)

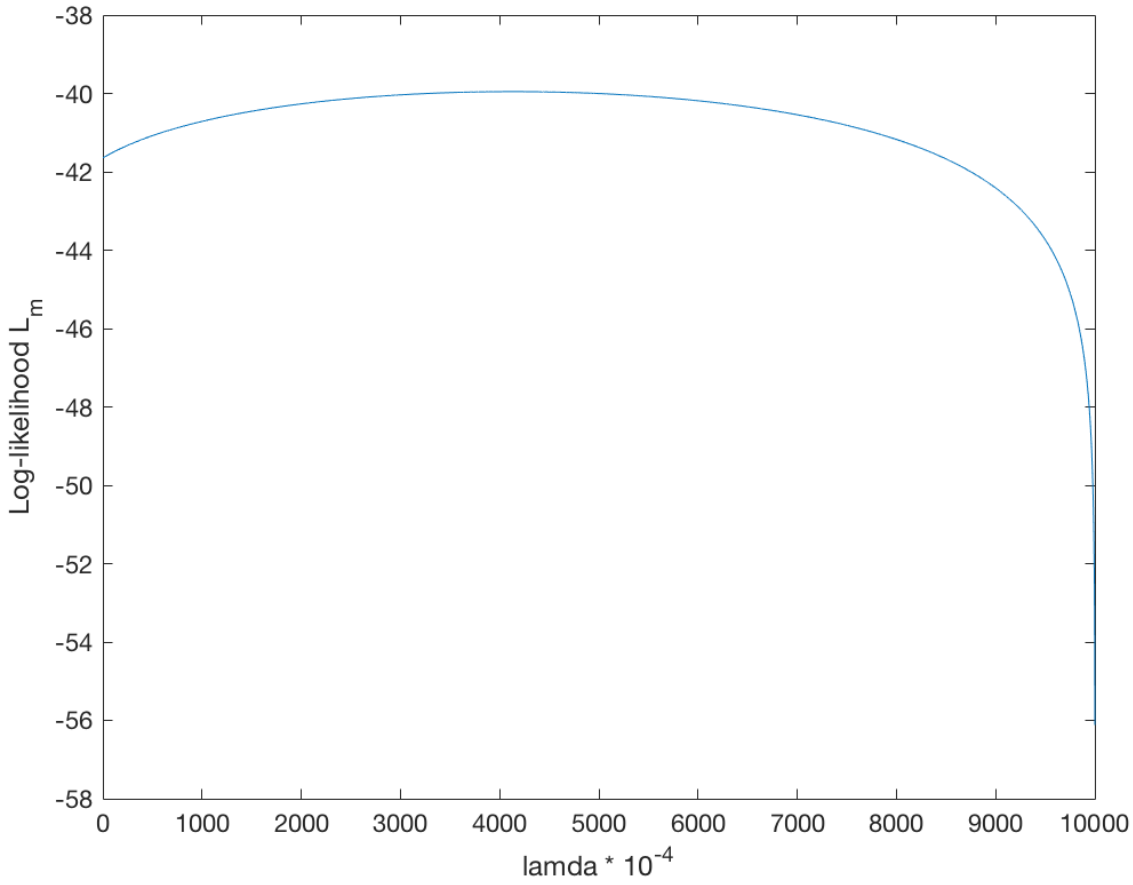
$L_u : -64.509440$
$L_b : -44.740469$
The bigram model has a higher log-likelihood.

(d)

$L_u : -41.643460$
$L_b : -\infty$
"NINETEEN OFFICIALS" and "SOLD FIRE" are not observed in the training corpus.
The unobserved pairs have caused some likelihoods calculated in the bigram model turned to 0. Thus, the total log-likelihood of bigram model has been $log(0)$, which is invalid. So we defined the result as $-\infty$ in this case.

(e)

The optimal value: $\lambda = 0.41$, where $L_m = -39.95$



1 Source Code

Listing 1: hw4_3.m

```
clear all

[words] = textread('vocab.txt', '%s');
[counts] = textread('unigram.txt', '%d');
[w1,w2,count] = textread('bigram.txt', '%d %d %d');

%% ------ (a) ------

P_unigram = zeros(size(counts));
total = sum(counts);
for i = 1 : length(counts)
    P_unigram(i) = counts(i)/total;
end
% sorted by frequency in acsending order
i = 1;
fprintf('----- (a) -----\n');
fprintf('tokens:\t \t unigram probability: \n');
while(i < 500)
    if(words{i}(1) == 'A')
        fprintf('%s \t \t %f \n', words{i}, P_unigram(i));
    end
    i = i+1;
end

%% ------ (b) ------

P_bigram = zeros(size(w1));
total = sum(count);

for i = 1 : length(w1)
    P_bigram(i) = count(i)/counts(w1(i));
end
Bigram = [w1 w2 count P_bigram];
% extract tuples where w1 == THE
tmp_the = [];
for i = 1 : length(w1)
    if(Bigram(i,1) == 4)
        tmp_the = [tmp_the; Bigram(i,:)];
    end
end
% sorting
[sorted_p, index] = sort(tmp_the(:,4),'descend');
```

```
fprintf('----- (b) -----\n');
fprintf('Most likely: \t \t bigram probability:\n');
for i = 1 : 5
    w2_index = tmp_the(index(i), 2);
    fprintf('%s \t \t \t %f\n', words{w2_index}, tmp_the(index(i),4));
end

%% ------ (c) ------

sentence = {'<s>', 'LAST', 'WEEK', 'THE', 'STOCK', ...
            'MARKET', 'FELL', 'BY', 'ONE', 'HUNDRED', 'POINTS'};
indexes = zeros(size(sentence));
P_u = zeros(size(sentence));
P_b = zeros(size(sentence));
indexes(1) = 2;
for i = 2 : length(sentence)
    for j = 1 : length(words)
        if(strcmpi(words{j}, sentence(i)))
            P_u(i) = P_unigram(j);
            indexes(i) = j;
        end
    end
end

for i = 2 : length(sentence)
    for j = 1 : length(w1)
        if(w1(j) == indexes(i-1) && w2(j) == indexes(i))
            P_b(i) = P_bigram(j);
        end
    end
end
L_u = 0;
L_b = 0;
for i = 2 : length(sentence)
     L_u = L_u + log(P_u(i));
     L_b = L_b + log(P_b(i));
end
fprintf('----- (c) -----\n');
fprintf('L_u: %f \nL_b: %f\n', L_u,L_b);

%% ------ (d) ------

sentence_2 = {'<s>', 'THE', 'NINETEEN', 'OFFICIALS', ...
              'SOLD', 'FIRE', 'INSURANCE'};
indexes_2 = zeros(size(sentence_2));
P_u_2 = zeros(size(sentence_2));
```

```
P_b_2 = zeros(size(sentence_2));
indexes_2(1) = 2;
for i = 2 : length(sentence_2)
    for j = 1 : length(words)
        if(strcmpi(words{j}, sentence_2(i)))
            P_u_2(i) = P_unigram(j);
            indexes_2(i) = j;
        end
    end
end
fprintf('----- (d) -----\n');
for i = 2 : length(sentence_2)
    for j = 1 : length(w1)
        if(w1(j) == indexes_2(i-1) && w2(j) == indexes_2(i))
            P_b_2(i) = P_bigram(j);

        end
    end
end
fprintf('two pairs:\n');
for i = 2 : length(P_b_2)
    if(P_b_2(i) == 0)
       fprintf('w1 : %s \t w2 : %s\n', words{indexes_2(i-1)},words{indexes_2(i)});
    end
end
L_u_2 = 0;

for i = 2 : length(sentence_2)
     L_u_2 = L_u_2 + log(P_u_2(i));
end

fprintf('L_u: %f \n', L_u_2);

%% ------ (e) ------

lamda = 0 : 0.0001 :1;
L_m = zeros(size(lamda));
for i = 1 : length(lamda)
    P_m = (1-lamda(i))*P_u_2 + lamda(i)*P_b_2;
    for j = 2 : length(sentence_2)
     L_m(i) = L_m(i) + log(P_m(j));
    end
end
plot(L_m);
xlabel('lamda * 10^{-4}');
ylabel('Log-likelihood L_m');
```

```
135 hold on
136 [ymax indmax]=max(L_m);
137 xmax=lamda(indmax);
138 fprintf('------(e)------\n');
139 fprintf('max lamda=%4.2f, L_m =%4.2f\n',xmax,ymax);
```

Submitted by Xiaowen Mao on Oct 25.