```
%---------------The load_data function ----------------------------------
function [a,b,c]=load_data()
%function [a,b,c]=load_data()
%a will get vocab, b will get unigram, c will get bigram

a = importdata('vocab.txt');%load the words data from txt file
b = importdata('unigram.txt');%load the unigram counts from txt file
c = importdata('bigram.txt');% load the bigram counts form txt file


%--------------The souce code for the problem 3.3(b)-----------------------
%Compute the maximum likehood estimate of the bigram distribution

[vocab_string,unigramdata,bigramdata] = load_data;%load the words data, unigram counts
%and bigram counts form txt file.
FID=fopen('outputtable_b', 'w+');

TF=strcmp('ONE',vocab_string);%find the index for 'ONE' in the vacabulary
r=find(TF == 1);
ONE_FIRST=find(bigramdata(:,1) == r);%find the postion(indices) of the word 'ONE' in
the bigramdata
sum_ONE = sum(bigramdata(ONE_FIRST,3));%the total counts that 'ONE' apprea followed by
any words

[sorted_wordnumber,IX] = sort(bigramdata(ONE_FIRST,3),'descend');%sort the words
following 'ONE' based on their counts
top_10_words = vocab_string(bigramdata(ONE_FIRST(1) - 1 + IX(1:10),2));%find the top
10 most
%like words w' to follow 'ONE'
top_10_probability = sorted_wordnumber(1:10,1)/sum_ONE;% calculate the maximum
likelihood estimate of the
%bigram distrituion for the 10 most like words w'

for i = 1:10
    fprintf(FID, '%-10s  %-1.9f \n', char(top_10_words(i)), top_10_probability(i));
end

fclose(FID);


%--------------The souce code for the problem 3.3(c)-----------------------
[vocab_string,unigramdata,bigramdata] = load_data;%load the words data, unigram counts
%and bigram counts form txt file.

totalwords = sum(unigramdata);%total words appears in the journal
sentence = {'<s>','THE','MARKET','FELL','BY','ONE','HUNDRED','POINTS','LAST','WEEK'};
n = size(sentence,2);
PU = 1;
PB = 1;

%compute Lu
for i = 2:n
    TF = strcmp(sentence(i),vocab_string);
    r = find(TF == 1);%find the index for ith word in the vacabulary
    count_i = unigramdata(r);%counts this word appreas
    pu(i) = count_i/totalwords;
    PU=PU*pu(i);
end
Lu = log(PU);
fprintf('Logrithm of unigram is Lu = %2.8f\n',Lu);

%compute Lb
for i = 1:n-1
    TF = strcmp(sentence(i), vocab_string);%find the ith word in the sentence in the
vocabulary
    ri = find(TF == 1);

    TF = strcmp(sentence(i+1), vocab_string);%find the i+1th word in the sentence in
the vocabulary
    rj = find(TF == 1);

    word_i = find(bigramdata(:,1) == ri);% the indices of ith word in the sentence in
```

```
bigram
    sum_i = sum(bigramdata(word_i,3));% the total counts that ith words followed by any word
    indiceij = find(bigramdata(word_i,2) == rj);%find the location where i+1th word follows ith word
    if(isempty(indiceij))
            fprintf('Pairs of adjacent words %10s and %10s are not observed\n', char(sentence(i)),char(sentence(i+1)));
    else
            wordij = bigramdata(word_i(1) - 1 + indiceij,3);%the counts that i+1 the word follows ith words;
            Pij=wordij/sum_i;
            %printf('P%d%d = %f',i,i+1,Pij);
            PB=PB*Pij;
    end

end
 Lb = log(PB);
 fprintf('logrithm of bigram is LB = %2.8f\n',Lb);



%--------------The souce code for the problem 3.3(d)-----------------------
[vocab_string,unigramdata,bigramdata] = load_data;%load the words data, unigram counts
%and bigram counts form txt file.

totalwords = sum(unigramdata);%total words appears in the journal
sentence = {'<s>','THE','FOURTEEN','OFFICIALS','SOLD','FIRE','INSURANCE'};
n = size(sentence,2);
PU = 1;
PB = 1;

%compute Lu
for i = 2:n
    TF = strcmp(sentence(i),vocab_string);
    r = find(TF == 1);%find the index for ith word in the vacabulary
    count_i = unigramdata(r);%counts this word appreas
    pui = count_i/totalwords;
    PU=PU*pui;
end
Lu = log(PU);
fprintf('Logrithm of unigram is Lu = %2.8f\n',Lu);

%compute Lb
for i = 1:n-1
    TF = strcmp(sentence(i), vocab_string);%find the ith word in the sentence in the vocabulary
    ri = find(TF == 1);

    TF = strcmp(sentence(i+1), vocab_string);%find the i+1th word in the sentence in the vocabulary
    rj = find(TF == 1);

    word_i = find(bigramdata(:,1) == ri);% the indices of ith word in the sentence in bigram
    sum_i = sum(bigramdata(word_i,3));% the total counts that ith words followed by any word
    indiceij = find(bigramdata(word_i,2) == rj);
    %find the location where i+1th word follows ith word
    if(isempty(indiceij))
            fprintf('Pairs of adjacent words %-9s follows %-9s are not observed\n', char(sentence(i+1)),char(sentence(i)));
            Pij=1;%Can't find ith word i followed  by (i+1)th word, so ignore this term in the log-likelihood
    else
            wordij = bigramdata(word_i(1) - 1 + indiceij,3);%the counts that i+1 the word follows ith words;
            Pij=wordij/sum_i;
            %size(Pij)
            PB=PB*Pij;
    end

end
 Lb = log(PB);
```

```
 fprintf('Logrithm of bigram is Lb = %2.8f\n',Lb);



 %--------------The souce code for the problem 3.3(e)-----------------------
N=1001;
lemda=linspace(0,1,N);
[vocab_string,unigramdata,bigramdata] = load_data;%load the words data, unigram counts
%and bigram counts form txt file.
totalwords = sum(unigramdata);%total words appears in the journal
sentence = {'<s>','THE','FOURTEEN','OFFICIALS','SOLD','FIRE','INSURANCE'};
n = size(sentence,2);
PM=1;
PMLEMDA=zeros(1,N-1);

for k=1:N

    PM=1;
    for i = 1:n-1
        %compute pu
        TF = strcmp(sentence(i+1),vocab_string);
        r = find(TF == 1);%find the index for i+1 th word in the vacabulary
        counti_next = unigramdata(r);%counts this word appreas
        pui_next = counti_next/totalwords;

        %compute pb
        TF = strcmp(sentence(i), vocab_string);%find the ith word in the sentence in
the vocabulary
        ri = find(TF == 1);

        TF = strcmp(sentence(i+1), vocab_string);%find the i+1th word in the sentence
in the vocabulary
        rj = find(TF == 1);
        word_i = find(bigramdata(:,1) == ri);% the indices of ith word in the sentence
in bigram
        sum_i = sum(bigramdata(word_i,3));% the total counts that ith words followed
by any word
        indiceij = find(bigramdata(word_i,2) == rj);
        %find the location where i+1th word follows ith word
        if(isempty(indiceij))%if can not find ith and i+1th words as adjacent words,
set the probability to zero.
            pbij=0;
        else
            wordij = bigramdata(word_i(1) - 1 + indiceij,3);%the counts that i+1 the
word follows ith words;
            pbij = wordij/sum_i;
        end

        Pmij=(1-lemda(k))*pui_next+lemda(k)*pbij;
        PM=PM*Pmij;
    end
    PMLEMDA(k)=log(PM);
end
plot(lemda(1:N-50),PMLEMDA(1:N-50));%lemda ranges from 0 to 0.95
[maxvalue,I] = max(PMLEMDA(1:N-50));
fprintf('maximum value is %2.8f with lemda %2.8f', maxvalue, lemda(I));
xlabel('Lemda');
ylabel('Log Likelihood');
```