Santhanakrishnan Ramani
SID: 105720585

Assignment 6

Part 1

Sample Document

'PNRRPPRRPPBPPPPNNPPNPRRNPDNPPRBPPNPPB PPPRPPPPDNPPP'

Code

```
ALPHA = 0.1;
BETA = 0.01;
M = 200; % no of documents
N = 50;  % no of words per document
V = 20;  % no of words in the vocabulary
K = 3;   % no of topics
Vocab = char(65:84); % Vocabulary with
words from A to T

DS = zeros(1,M*N);
for i= 1:M
    DS( ((N*i)-(N-1)) : (N*i)) =
(ones(1,N)+i-1);
end

Theta = drchrnd(ones(1,K)*ALPHA, M);
Phi = drchrnd(ones(1,V)*BETA, K);

z= zeros(M,N);
WS = zeros(1,M*N);
Doc = cell(M,1);

for doc = 1:M
    for word=1:N
        z(doc,word) =
find(mnrnd(1,Theta(doc,:),1));
        WS((N*doc)-(N-word)) =
find(mnrnd(1,Phi(z(doc,word),:),1));
    end
    Doc{doc} = Vocab(WS(((N*doc)-(N-1)) :
(N*doc)));
end

WO = cellstr(Vocab');
```

Sample Topic Distribution

Word	P(Word|Topic = Topic1)
'A'	2.33E-36
'B'	1.31E-51
'C'	1.90E-67
'D'	2.08E-56
'E'	**0.01699877**
'F'	1.48E-73
'G'	3.97E-12
'H'	4.63E-86
'I'	**0.108610873**
'J'	3.13E-172
'K'	6.14E-44
'L'	3.64E-18
'M'	2.85E-35
'N'	**0.874390356**
'O'	3.13E-110
'P'	3.26E-57
'Q'	3.35E-10
'R'	2.08E-79
'S'	1.15E-67
'T'	2.00E-14

Part 2

Code

```
E = 1;
M = 3;
T = 3;
K = 20;
N = 500;
SEED = 3;
OUTPUT = 0;

[ WP,DP,Z ] = GibbsSamplerLDA( WS,DS,T,N,ALPHA,BETA,SEED,OUTPUT );
[S] = WriteTopics( WP , BETA , WO , K, E, M, 'topics.txt' );
```

True Vs Recovered topics Comparison

	P(Word|Topic = Topic1)	
Word	'true topics'	'recovered topics'
'A'	2.33E-36	0
'B'	1.31E-51	0
'C'	1.90E-67	0
'D'	2.08E-56	0
'E'	**0.01699877**	**0.01691**
'F'	1.48E-73	0
'G'	3.97E-12	0
'H'	4.63E-86	0
'I'	**0.108610873**	**0.10356**
'J'	3.13E-172	0
'K'	6.14E-44	0
'L'	3.64E-18	0
'M'	2.85E-35	0
'N'	**0.874390356**	**0.87948**
'O'	3.13E-110	0
'P'	3.26E-57	0
'Q'	3.35E-10	0
'R'	2.08E-79	0
'S'	1.15E-67	0
'T'	2.00E-14	0

Part 3

Code

```
DP = full(DP);
ent = 0;
for i = 1:size(DP,1)
    dr = sum(DP(i,:));
    if dr ~= 0
        DP(i,:) = DP(i,:)/dr;
        temp = DP(i,DP(i,:) > 0);
        ent  = ent + sum(temp .*
log(temp));
    end
end
fprintf('entropy of topic distribution:
%f \n', (-1 * ent)/200);
```

α	entropy of topic distributions
	(Kept β constant $= 0.01$)
0.1	0.2456
1	0.8162
10	1.036
100	1.074

Code

```
WP = full(WP);
WP = [WP;zeros(V-size(WP,1),3)];
ent = 0;
for i = 1:size(WP,2)
    dr = sum(WP(:,i));
    if dr ~= 0
        WP(:,i) = WP(:,i)/dr;
        temp = WP(WP(:,i) > 0,i);
        ent  = ent + sum(temp .*
log(temp));
    end
end
fprintf('entropy of words distribution:
%f \n',(-1 * ent)/3);
```

β	entropy of word distributions
	(Kept α constant $= 0.1$)
0.01	0.2780
0.1	1.4319
1	2.5401
10	2.9591

Part 4

Used Psych **Review Abstracts (bag of words)**

bagofwords_psychreview (doc word counts)
words_psychreview (vocabulary)

Topic: *Recall (memory)*

Original List		Revised List	
Word	**P(W|T)**	**Word**	**P(W|T)*P(T|W)**
memory	0.13064	memory	0.104273993
recall	0.03764	recall	0.038451722
recognition	0.0367	retrieval	0.035613399
retrieval	0.03638	items	0.028089572
items	0.03163	list	0.023460411
item	0.02531	item	0.023431122
list	0.02278	recognition	0.02058446
associative	0.0174	associative	0.012476748
information	0.01614	storage	0.010773939
effects	0.01424	trace	0.01010101
serial	0.01392	cue	0.010010884
number	0.01297	traces	0.009785354
process	0.01234	familiarity	0.007820137
storage	0.01234	serial	0.007168459
cue	0.01202	stored	0.006674763
order	0.01107	amnesia	0.005556413
context	0.01076	episodic	0.005539264
study	0.00981	forgetting	0.005213425
trace	0.00981	cued	0.004906753
traces	0.00981	interference	0.004788024

Topic: *Scientific Method*

Original List		Revised List	
Word	**P(W|T)**	**Word**	**P(W|T)*P(T|W)**
theory	0.09097	theory	0.023507273
predictions	0.0292	predictions	0.015318532
experiments	0.02471	assumptions	0.009574608
results	0.02359	achievement	0.009305933
theories	0.02209	difficulty	0.008078066
assumptions	0.02097	support	0.006440102
presented	0.01797	outcomes	0.006377421
support	0.0176	experiments	0.00627891
proposed	0.01722	results	0.005785596
general	0.01647	theories	0.005719271
level	0.0146	belief	0.00533614
studies	0.01423	case	0.005047638
experimental	0.01311	hypotheses	0.004963164
discussed	0.01273	treatment	0.004726823
empirical	0.01273	general	0.004691741
derived	0.01236	derived	0.004540395
related	0.01123	predict	0.004466848
basic	0.01086	independence	0.004274913
case	0.01011	level	0.004273648
predicted	0.00974	randomness	0.004265219