Assignment 7 CSCI 5822

Jensen Dempsey

April 12, 2018

1

A few examples of documents

```
GGCOGGCGGCGOOOHAGGGOHOGGGGGHGACLGAACHGGGGGAOOAOOGG
HOHCOHOOCCOOOOOOHOOOOOOOOHOOHOHOOOHOOOOHOHCOOOHHCO
JJJJJJJJJJJJJJJJJHJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```

Sample topic Distribution

```
O Topic Actual
A  1.0564535837178656e-133
B  5.736378787967005e-12
C  0.18818383042178446
D  5.8322098905640385e-09
E  4.298560712148518e-79
F  0.00011054741176300422
G  1.6781411749726992e-31
H  0.18265443132365894
I  1.0150161042532433e-10
J  0.0028117327270779963
K  6.462546618365091e-56
L  2.8690449671399296e-10
M  9.279743461686449e-15
N  4.94812644320681e-78
O  0.6262225197512795
P  4.627763221377831e-15
Q  1.3122530207720496e-05
R  4.252234260420952e-08
S  1.1372712686097287e-08
T  3.7557128068315915e-06
```

2

Recovered Topic Distribution, after running for many iterations.

```
O Topic Recovered
A  2.8393305391978237e-06
B  2.8393305391978237e-06
C  0.1854111235401568
D  2.8393305391978237e-06
E  2.8393305391978237e-06
F  0.00028677238445858956
G  2.8393305391978237e-06
H  0.18676178629349952
I  2.8393305391978237e-06
J  2.8393305391978237e-06
K  2.8393305391978237e-06
L  2.8393305391978237e-06
M  2.8393305391978237e-06
N  2.8393305391978237e-06
O  0.6274948884932576
P  2.8393305391978237e-06
Q  2.8393305391978237e-06
R  2.8393305391978237e-06
S  2.8393305391978237e-06
T  2.8393305391978237e-06
```

Compared to the true distribution it does capture the important values, C,H, and O. This is similar for the other topics as well, at least for the 1-3 most probable words. But then they are not on the same order of magnitude for the less probable words. That being said it does still have them has very unlikely.

I also compared the total squared error for each topic. Which just resulted in some small numbers 6.241e-05, 3.409e-05,1.416e-05 which does at least show that LDA can get somewhat close.

3

For this test I computed the entropy of topics with alphas of 0.1,0.2,0.5, and 1 giving me entropies of 0.3016, 0.3183, 0.3620, 0.4218 respectively. And as I increased alpha I did in fact get an increase in the entropy as expected.

For the word distribution I used betas of 0.01,0.02,0.05,0.1 giving me entropies of 0.50213, 1.5629, 0.5044, 1.5650. Which does not give me the response I was expecting. At least for two of the tests I did get a higher entropy with a higher beta. But the 0.05 test does not make sense.

4 Other Topics

G	Topic Actual	G	Topic Recovered
A	0.1260156134574155	A	0.1295469193091
B	1.4585736515067566e-136	B	3.2224800206238753e-06
C	1.8777690706865347e-22	C	3.2224800206238753e-06
D	7.129497119617989e-20	D	3.2224800206238753e-06
E	6.6714510415116445e-81	E	3.2224800206238753e-06
F	2.221027506323619e-75	F	3.2224800206238753e-06
G	0.8591290191642068	G	0.8526714359370976
H	1.114761857501676e-99	H	3.2224800206238753e-06
I	7.778210716762246e-24	I	3.2224800206238753e-06
J	6.591957349494043e-83	J	3.2224800206238753e-06
K	1.841484088473169e-32	K	3.2224800206238753e-06
L	0.014855367378124654	L	0.017726862593451465
M	5.19604756195978e-84	M	3.2224800206238753e-06
N	5.250335088226528e-21	N	3.2224800206238753e-06
O	7.135815958726154e-20	O	3.2224800206238753e-06
P	6.446576629565532e-133	P	3.2224800206238753e-06
Q	2.529810399446156e-13	Q	3.2224800206238753e-06
R	1.2167385747999263e-23	R	3.2224800206238753e-06
S	4.2295898086184235e-43	S	3.2224800206238753e-06
T	2.6783427092322356e-26	T	3.2224800206238753e-06

J	Topic Actual	J	Topic Recovered
A	2.1933987493537724e-26	A	2.962574076710448e-06
B	5.1225418518888516e-20	B	2.962574076710448e-06
C	7.628701521557582e-55	C	2.962574076710448e-06
D	1.2174288351105604e-07	D	2.962574076710448e-06
E	1.636280428027298e-70	E	2.962574076710448e-06
F	6.4679837952231e-14	F	2.962574076710448e-06
G	1.4997275756078248e-35	G	2.962574076710448e-06
H	0.019425859223397315	H	0.01666537503234461
I	3.288685821712107e-96	I	2.962574076710448e-06
J	0.9765844778521032	J	0.9791336949268801
K	2.1976738626135216e-15	K	2.962574076710448e-06
L	3.433255898521956e-13	L	2.962574076710448e-06
M	5.003319131763776e-05	M	2.962574076710448e-06
N	2.541678596550933e-07	N	2.962574076710448e-06
O	2.9161443687492997e-52	O	2.962574076710448e-06
P	0.003938726252135757	P	0.004150566281471006
Q	1.1584610115793119e-107	Q	2.962574076710448e-06
R	5.828518359254079e-115	R	2.962574076710448e-06
S	3.552767974236144e-22	S	2.962574076710448e-06
T	5.275698927178347e-07	T	2.962574076710448e-06