

vocab size 1000; avg doc length 100

