BST249 HW5

Sijia Huo

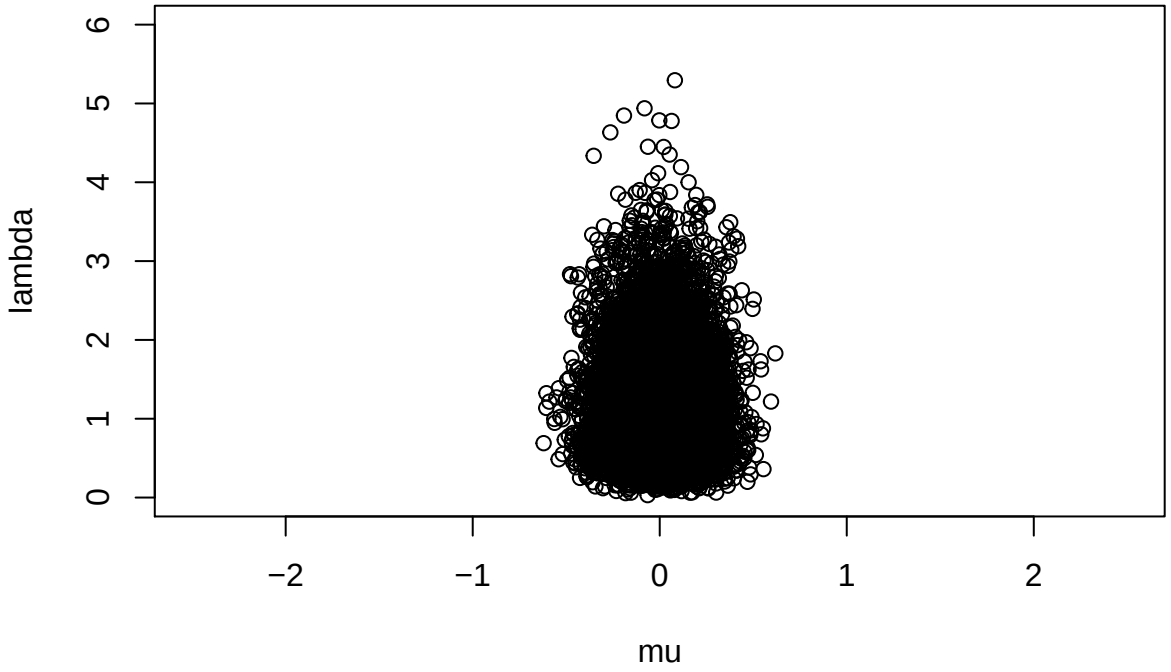
4/8/2022

1.

(a)

Based on the lecture slides, we have $q^{\text{new}}(\mu) = \mathcal{N}\left(\mu \mid \bar{x}, (n\mathrm{E}(\lambda))^{-1}\right) = \mathcal{N}\left(0, 1/(5 \times (5+1)/5)\right) = \mathcal{N}\left(0, \frac{1}{6}\right)$ and $q^{\text{new}}(\lambda) = \text{Gamma}\left(\lambda \mid n/2+1, \frac{1}{2}\left(n\hat{\sigma}^2 + 1/\mathrm{E}(\lambda)\right)\right) = \text{Gamma}\left(\frac{7}{2}, \frac{1}{2} \times (5 + \frac{5}{6}\right) = \text{Gamma}\left(\frac{7}{2}, \frac{35}{12}\right)$.

```
mu = rnorm(10000,0,1/6)
lambda = rgamma(10000,7/2,35/12)
plot(mu,lambda,xlim = c(-2.5,2.5), ylim = c(0,6))
```



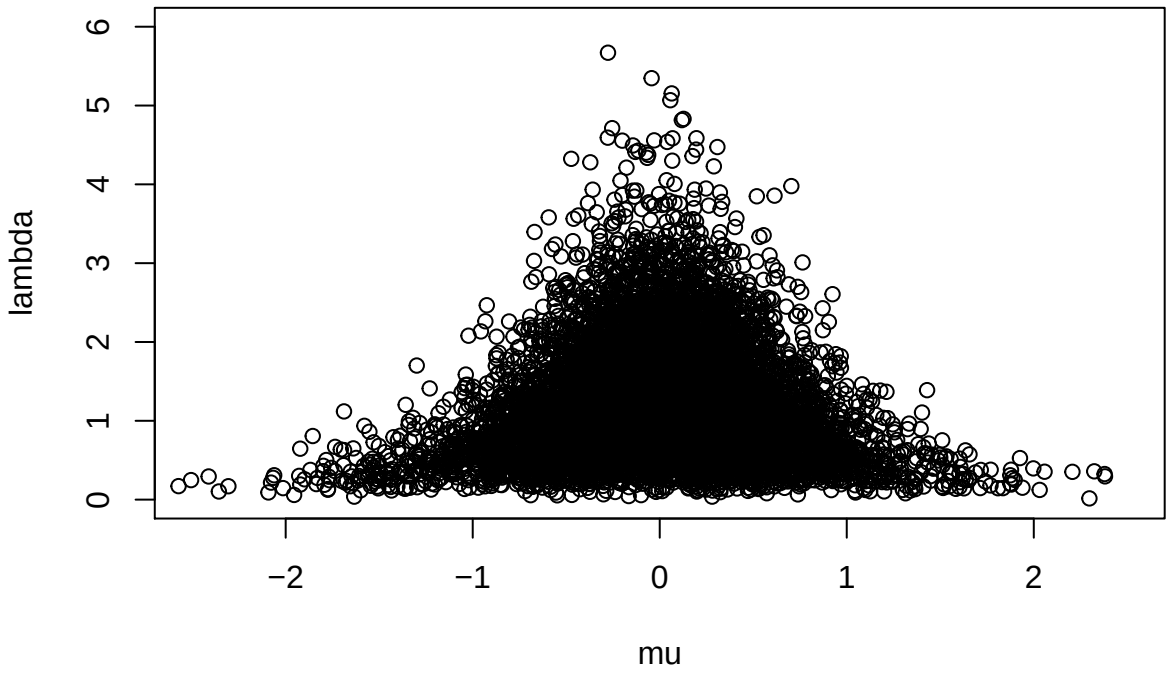
(b)

Based on the true posterior provided, we have $p(\mu, \lambda \mid x_{1:n}) = \text{NormalGamma}(\mu, \lambda \mid m, c, a, b) = \text{NormalGamma}(0, 5, 3, \frac{5}{2})$. Compared to the plot in part (a), this one is more spread out with a larger variance, which is as expected.

```
# function to sample n observations from normal-gamma distribution
rnormalgamma = function(n,m,c,a,b){
lambda_sample = rgamma(n,a,b)
mu_sample = m + rnorm(n)*(c*lambda_sample)^(-1/2)
return(cbind(mu_sample,lambda_sample))
```

```
}

#sample & plot
samples_1b = rnormalgamma(10000,0,5,3,5/2)
plot(samples_1b[,1],samples_1b[,2],xlim = c(-2.5,2.5), ylim = c(0,6), xlab = "mu",ylab = "lambda")
```



2.

When computing the expectation of $h(\theta)$ with respect to $\pi(\theta)$, the importance sampling approximation follows the formula of $\mathrm{E}h(\theta) \approx \frac{1}{N}\sum_{i=1}^{N} h\left(\theta_i'\right)\frac{\pi\left(\theta_i'\right)}{q\left(\theta_i'\right)}$. That's to say, we can sample θ_i' from the proposal distribution q and calculate the mean of $h(\theta')$ with the importance weights of $\frac{\pi\left(\theta_i'\right)}{q\left(\theta_i'\right)}$. Based on the law of large number (LLN), we have

$$\frac{1}{N}\sum_{i=1}^{N} h\left(\theta_i'\right)\frac{\pi\left(\theta_i'\right)}{q\left(\theta_i'\right)} \xrightarrow{P} E_q\left(h\left(\theta'\right)\frac{\pi\left(\theta'\right)}{q\left(\theta'\right)}\right) = \int h\left(\theta'\right)\frac{\pi\left(\theta'\right)}{q\left(\theta'\right)}q(\theta')d\theta' = E_\pi\left(h(\theta')\right) = E_\pi\left(h(\theta)\right)$$

Therefore, as the number of importance samples grows, the asymptotic guarantees of correctness can be satisified.

When choosing proposal distribution for importance sampling, we prefer those $q(\theta)$ that are not too small in areas where $\pi(\theta)$ is large and are a little more spread out than $\pi(\theta)$. By doing this, we can avoid overweighting some samples and avoid a large RMSE. In our case, since 1) the classic VI approach doesn't blow up when $q(\theta) \ll \pi(\theta)$ 2) $\pi(\theta)$ is actually less spread out than $\pi(\theta)$, the IS method may not work very well in general if the sampling size is small. However, as sample size grows, the performance using this proposal distribution will improve. And since $q\left(\theta\right)$ is in general a good approximation of $\pi(\theta)$, the performance won't be too bad anyway.

3.

We start our derivatoin from $\pi(z, w, \beta)$. All the distributions here are directly taken from the course slides.

$$
\begin{aligned}
\pi(z, w, \beta) &= p(z, w, \beta \mid x) \\
&\propto p(w) \cdot p(Z = z|w) \cdot p(\beta|z) \cdot p(X = x|w, z, \beta) \\
&\propto \prod_{i=1}^{N} \prod_{k=1}^{K} w_{ik}^{\alpha_k - 1} \cdot \prod_{i=1}^{N} \prod_{l=1}^{L} \prod_{k=1}^{K} w_i^{\mathrm{I}(z_{il}=k)} \cdot \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{\lambda_v - 1} \cdot \prod_{i=1}^{N} \prod_{l=1}^{L} \prod_{k=1}^{K} \prod_{v=1}^{V} \beta_{kv}^{\mathrm{I}(x_{il}=v)\mathrm{I}(z_{il}=k)}
\end{aligned}
$$

Therefore, we finally have

$$
\begin{aligned}
&\log \pi(z, w, \beta) \\
=&\sum_{i=1}^{N} \sum_{k=1}^{K} (\alpha_k - 1) \log(w_{ik}) + \sum_{i=1}^{N} \sum_{l=1}^{L} \sum_{k=1}^{K} \mathrm{I}(z_{il} = k) \log(w_{ik}) \\
&+ \sum_{k=1}^{K} \sum_{v=1}^{V} (\lambda_v - 1) \log(\beta_{kv}) + \sum_{i=1}^{N} \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{v=1}^{V} \mathrm{I}(x_{il} = v)\mathrm{I}(z_{il} = k) \log(\beta_{kv}) + \text{const} \\
=&\sum_{i,\ell,k,v} \mathrm{I}\left(x_{i\ell} = v\right) \mathrm{I}\left(z_{i\ell} = k\right) \log\left(\beta_{kv}\right) + \sum_{i,\ell,k} \mathrm{I}\left(z_{i\ell} = k\right) \log\left(w_{ik}\right) \\
&+ \sum_{i,k} \left(\alpha_k - 1\right) \log\left(w_{ik}\right) + \sum_{k,v} \left(\lambda_v - 1\right) \log\left(\beta_{kv}\right) + \text{const}
\end{aligned}
$$

4.

Suppose at step a, we have distributions of $q(w)$, $q(\beta)$, and $q(z)$ as what is given in the page 32 of lecture 12, we will derive $q^{\text{new}}(z), q^{\text{new}}(w),$ and $q^{\text{new}}(\beta)$ at step a+1.

$$
\begin{aligned}
h_1(w) &= \mathrm{E}_q\left(\log \pi(\theta) \mid w\right) \\
&= \mathrm{E}_q\left(\sum_{i,\ell,k,v} \mathrm{I}\left(x_{i\ell} = v\right) \mathrm{I}\left(z_{i\ell} = k\right) \log\left(\beta_{kv}\right) + \sum_{i,\ell,k} \mathrm{I}\left(z_{i\ell} = k\right) \log\left(w_{ik}\right)\right. \\
&\left. + \sum_{i,k} \left(\alpha_k - 1\right) \log\left(w_{ik}\right) + \sum_{k,v} \left(\lambda_v - 1\right) \log\left(\beta_{kv}\right) + \text{const} \mid w\right) \\
&= \mathrm{E}_q\left(\sum_{i,\ell,k} \mathrm{I}\left(z_{i\ell} = k\right) \log\left(w_{ik}\right) + \sum_{i,k} \left(\alpha_k - 1\right) \log\left(w_{ik}\right) \mid w\right) + \text{const} \\
&= \mathrm{E}_q\left(\sum_{i,k} \left(\sum_{\ell} \mathrm{I}(z_{il} = k) \log(w_{ik}) \mid w_{ik}\right)\right) + \sum_{i,k} \left(\alpha_k - 1\right) \log\left(w_{ik}\right) + \text{const} \\
&= \sum_{i,k} \sum_{\ell} \mathbb{P}(z_{il} = k) \log(w_{ik}) + \sum_{i,k} \left(\alpha_k - 1\right) \log\left(w_{ik}\right) + \text{const} \\
&= \sum_{i,k} \left(\sum_{\ell} t_{ilk} + \left(\alpha_k - 1\right)\right) \log(w_{ik}) + \text{const}
\end{aligned}
$$

Given that $q^{\text{new}}(w) \propto \exp\left(h_1(w)\right)$, we have

$$
\begin{aligned}
q^{\text{new}}(w) &\propto w^{\sum_{i,k}\left(\sum_{\ell} t_{ilk}+(\alpha_k-1)\right)} \\
&\propto \prod_{ik} w^{\sum_{\ell} t_{ilk}+\alpha_k-1} \\
&\propto \prod_{i=1}^{n} \text{Dirichlet}\left(w_i \mid r_{i1}^{new}, \ldots, r_{iK}^{new}\right)
\end{aligned}
$$

Where $r_{ik}^{new} = \sum_{\ell=1}^{L_i} t_{ilk} + \alpha_k$

$$
\begin{aligned}
h_2(\beta) &= \mathrm{E}_q\left(\log \pi(\theta) \mid \beta\right) \\
&= \mathrm{E}_q\left(\sum_{i,\ell,k,v} \mathrm{I}\left(x_{i\ell}=v\right)\mathrm{I}\left(z_{i\ell}=k\right)\log\left(\beta_{kv}\right) + \sum_{i,\ell,k}\mathrm{I}\left(z_{i\ell}=k\right)\log\left(w_{ik}\right)\right. \\
&\left. + \sum_{i,k}\left(\alpha_k-1\right)\log\left(w_{ik}\right) + \sum_{k,v}\left(\lambda_v-1\right)\log\left(\beta_{kv}\right) + \text{const} \mid \beta\right) \\
&= \mathrm{E}_q\left(\sum_{i,\ell,k,v} \mathrm{I}\left(x_{i\ell}=v\right)\mathrm{I}\left(z_{i\ell}=k\right)\log\left(\beta_{kv}\right) + \sum_{k,v}\left(\lambda_v-1\right)\log\left(\beta_{kv}\right) \mid \beta\right) + \text{const} \\
&= \mathrm{E}_q\left(\sum_{k,v}\left(\sum_{i\ell}\mathrm{I}\left(x_{i\ell}=v\right)\mathrm{I}\left(z_{i\ell}=k\right)\log(\beta_{kv})\right)\right) + \sum_{k,v}\left(\lambda_v-1\right)\log\left(\beta_{kv}\right) + \text{const} \\
&= \sum_{k,v}\sum_{i\ell}\mathbb{P}(z_{il}=k)\mathrm{I}\left(x_{i\ell}=v\right)\log(\beta_{kv}) + \sum_{k,v}\left(\lambda_v-1\right)\log\left(\beta_{kv}\right) + \text{const} \\
&= \sum_{k,v}\left(\sum_{i\ell} t_{ilk}\mathrm{I}\left(x_{i\ell}=v\right) + \left(\lambda_v-1\right)\right)\log(\beta_{kv}) + \text{const}
\end{aligned}
$$

Therefore, we further have

$$
\begin{aligned}
q^{\text{new}}(\beta) &\propto \exp\left(h_2(\beta)\right) \\
&\propto \prod_{kv} \beta_{kv}^{\sum_{i\ell} t_{ilk}\mathrm{I}(x_{i\ell}=v)+\lambda_v-1} \\
&\propto \prod_{k=1}^{K} \text{Dirichlet}\left(\beta_k \mid s_{k1}^{new} \ldots, s_{kV}^{new}\right)
\end{aligned}
$$

Where $s_{kv}^{new} = \sum_{i=1}^{n}\sum_{\ell=1}^{L_i}\mathrm{I}\left(x_{i\ell}=v\right)t_{ilk} + \lambda_v$

$$
\begin{aligned}
h_3(z) &= \mathrm{E}_q\left(\log \pi(\theta) \mid z\right) \\
&= \mathrm{E}_q\left(\sum_{i,\ell,k,v} \mathrm{I}\left(x_{i\ell}=v\right)\mathrm{I}\left(z_{i\ell}=k\right)\log\left(\beta_{kv}\right) + \sum_{i,\ell,k}\mathrm{I}\left(z_{i\ell}=k\right)\log\left(w_{ik}\right) \mid z\right) + \text{const} \\
&= \sum_{i,\ell,k,v}\mathrm{I}\left(x_{i\ell}=v\right)\mathrm{I}\left(z_{i\ell}=k\right)\mathrm{E}_q\left(\log(\beta_{kv})\right) + \sum_{i,\ell,k}\mathrm{I}\left(z_{i\ell}=k\right)\mathrm{E}_q\left(\log\left(w_{ik}\right)\right) + \text{const} \\
&= \sum_{i,\ell,k,v}\mathrm{I}\left(x_{i\ell}=v\right)\mathrm{I}\left(z_{i\ell}=k\right)\left(\psi\left(s_{kv}\right) - \psi\left(\sum_{v'=1}^{V} s_{kv'}\right)\right) + \sum_{i,\ell,k}\mathrm{I}\left(z_{i\ell}=k\right)\left(\psi\left(r_{ik}\right) - \psi\left(\sum_{k'}^{K} r_{ik'}\right)\right) + \text{const} \\
&= \sum_{i,\ell,k}\mathrm{I}\left(z_{i\ell}=k\right)\left\{\sum_{v}\mathrm{I}\left(x_{i\ell}=v\right)\left(\psi\left(s_{kv}\right) - \psi\left(\sum_{v'=1}^{V} s_{kv'}\right)\right) + \left(\psi\left(r_{ik}\right) - \psi\left(\sum_{k'=1}^{K} r_{ik'}\right)\right)\right\} + \text{const} \\
&= \sum_{i,\ell}\sum_{k}\mathrm{I}\left(z_{i\ell}=k\right)u_{i\ell k} \\
&\text{where } u_{i\ell k} = \psi\left(r_{ik}\right) - \psi\left(\sum_{k'} r_{ik'}\right) + \sum_{v=1}^{V}\mathrm{I}\left(x_{i\ell}=v\right)\left(\psi\left(s_{kv}\right) - \psi\left(\sum_{v'} s_{kv'}\right)\right)
\end{aligned}
$$

Therefore, we further have

$$
\begin{aligned}
q^{\text{new}}\left(z\right) &\propto \exp\left(h_3(z)\right) \\
&\propto \prod_{i\ell}\prod_{k}\exp\left(u_{i\ell k}\right)^{\mathrm{I}(z_{i\ell}=k)} \\
&\propto \prod_{i\ell}\prod_{k}\left(\frac{\exp\left(u_{i\ell k}\right)}{\sum_{k'=1}^{K}\exp\left(u_{i\ell k'}\right)}\right)^{\mathrm{I}(z_{i\ell}=k)} \\
&\propto \prod_{i=1}^{n}\prod_{\ell=1}^{L_i}\text{Categorical}\ \left(z_{i\ell} \mid t_{i\ell}^{new}\right)
\end{aligned}
$$

Where $t_{i\ell k}^{new} = \frac{\exp\left(u_{i\ell k}\right)}{\sum_{k'=1}^{K}\exp\left(u_{i\ell k'}\right)}$

5.

```
library(gtools)

lda = function(x,V,K,alpha,lambda,tau){

  n = dim(x)[1] # number of documents
  L = rowSums(!is.na(x)) # number of words per document
  wc = table(factor(x[!is.na(x)], levels = 1:V))  # count frq of words across files
  t_k = rdirichlet(1,rep(1,K)) # random assign t, but same across words and files initially
  r_new = matrix(rep(alpha,n),nrow = n,byrow = TRUE) + matrix(L) %*% t_k # initially t are equal across
  s_new =  matrix(rep(lambda,K),nrow =K ,byrow = TRUE) + t_k %*% t(wc)
  converge = FALSE # flag indicating whether to converge

  while(!converge){
    r_prev = r_new
```

```
    s_prev = s_new
    r_new = matrix(rep(alpha,n),nrow = n,byrow = TRUE) # new r_ik
    s_new = matrix(rep(lambda,K),nrow =K ,byrow = TRUE) # new s_kv
    psi_r = digamma(r_prev) - digamma(rowSums(r_prev))
    pri_s = digamma(s_prev) - digamma(rowSums(s_prev))
    for (i in 1:n){
      T = matrix(0,nrow = L[i], ncol=K) # construct matrix T
      for (l in 1:L[i]){
        word = x[i,l]
        u_l = psi_r[i,] + pri_s[,word]
        T[l,] = exp(u_l)/sum(exp(u_l))
        s_new[,word] = s_new[,word] + T[l,]
      }
      r_new[i,] = r_new[i,]+colSums(T)
    }

    converge = (sqrt(mean((r_new - r_prev)^2)) < tau) & (sqrt(mean((s_new - s_prev)^2)) < tau)
  }
  return(list(r_new,s_new))
}
```

6.

(a)

```
data = read.csv("~/Documents/Spring2022/BST249/Homework/HW5/homework-5-data/ap.csv", header=FALSE)
K = 25
V = 10473
alpha = rep(1/K,K)
lambda = rep(1/V,V)
tau = 0.001
result = lda(data,V,K,alpha,lambda,tau)

# save the result to files, avoid running again.
write.table(result[[1]],"/Users/scarletthuo/Documents/Spring2022/BST249/Homework/HW5/r_matrix_v1.csv", i
write.table(result[[2]],"/Users/scarletthuo/Documents/Spring2022/BST249/Homework/HW5/s_matrix_v1.csv", i
```

(b)

```
library(knitr)
# load data
r_matrix = as.matrix(read.table("/Users/scarletthuo/Documents/Spring2022/BST249/Homework/HW5/r_matrix_v
s_matrix = as.matrix(read.table("/Users/scarletthuo/Documents/Spring2022/BST249/Homework/HW5/s_matrix_v
vocab = read.table("~/Documents/Spring2022/BST249/Homework/HW5/homework-5-data/vocab.txt", quote="\"", c

pop_topic = colSums(r_matrix/rowSums(r_matrix))
frq_word = s_matrix/rowSums(s_matrix)
topic_rank =  order(-pop_topic)
word_rank = t(apply(frq_word,1,function(x) order(-x)))

# construct dataframe
table = sapply(1:8,function(x) vocab[word_rank[topic_rank[x],1:20]])
table = as.data.frame(table)
```

```
colnames(table)<-c('Daily Life','Local Issues','Federal Assistance','Business Policy','Judiciary','Crim
kable(table)
```

Daily Life	Local Issues	Federal Assistance	Business Policy	Judiciary	Criminal	Cold War	Politics
space	i	percent	new	court	police	soviet	government
church	people	states	president	federal	people	air	party
hit	years	new	plan	case	government	military	south
summer	new	texas	million	attorney	two	officials	political
look	two	rep	workers	judge	killed	two	minister
care	time	spending	union	drug	city	united	economic
car	year	democrats	last	trial	today	government	president
income	just	nation	company	law	president	troops	africa
recently	dont	aids	agreement	charges	monday	official	opposition
wanted	first	jackson	year	prison	authorities	force	meeting
calls	state	legislation	billion	department	group	war	million
fall	get	bill	united	state	political	spokesman	national
main	day	southern	program	office	three	agency	african
old	like	election	government	justice	fire	news	prime
research	say	fair	officials	district	told	plane	economy
young	last	high	offer	filed	man	communist	mrs
brought	back	measure	group	investigation	state	american	year
large	school	sen	states	jury	black	forces	first
los	three	state	work	years	officials	army	leader
film	going	plant	news	guilty	night	navy	parties

(c)

The result does make sense, to a certain extent. Some topics are indeed much more coherent than others. In my result, columns labeled “judiciary”“,”criminal“,”cold war“,and”politics" are much easier to be interperrated than others, specifically because the key words are correctly ranked at the top of the lists. Columns for “federal assicetance” and “business policy” are ok to interpret. First two topics are the hardest ones, for most of the vocabularies are just the commonly used words for any sentences under any topics. However, it still makes sense that they are ranked as the hottest topics because of the high frequencies of these words. The words that I wish to remove from the analysis include “i”, “look”,“just”, “say”, “two”, and “three”.