Topic Model for Pure Membership Data with No Anchor Words

Zihao Wang

7/5/2018

```
library(CountClust)
```

```
## Loading required package: ggplot2
```

```
library(maptpx)
```

```
## Loading required package: slam
```

```
library(classtpx)
```

```
##
## Attaching package: 'classtpx'
```

```
## The following objects are masked from 'package:maptpx':
##
##     expit, logit, rdir, stm_tfidf
```

```
library(radmixture)
set.seed(12345)
```

```
X_K5 = read.csv("../Top_data/countX_K5.csv", header = FALSE, sep = ",")
Pi_K5 = read.csv("../Top_data/Pi_K5.csv", header = FALSE, sep = ",")
A_K5 = read.csv("../Top_data/topA_K5.csv", header = FALSE, sep = ",")
X_K5 = as.matrix(X_K5)
Pi_K5 = as.matrix(Pi_K5)
A_K5 = as.matrix(A_K5)
zero_row = which(rowSums(X_K5) == 0)
```

run the data on topics() and compute loglikelihood

```
fit.map = topics(X_K5,K = 5)
```

```
## Removed 4 blank documents.
##
## Estimating on a 2996 document collection.
## Fitting the 5 topic model.
## log posterior increase: 3142.5, done.
```

```
theta1 = t(fit.map$theta)
omega1 = fit.map$omega
fitted_probs1 <- omega1%*%theta1
loglik1 <- sum(Pi_K5[-zero_row,]*log(fitted_probs1))
loglik1
```

```
## [1] -164.9887
```

```
square_loss = norm((Pi_K5[-zero_row,] - fitted_probs1),"F")
square_loss
```

```
## [1] 11.67833
```

Continue Topic algo:Use A and X to compute W