GetProbableTerms.R

ttran

Wed Apr 29 14:25:40 2015

Get cluster labels using a "more probable" method of terms

@description Function extracts probable terms from a set of documents. Probable here implies more probable than in a corpus overall @param docnames = character vector of rownames of dtm for set of documents @param dtm = a document term matrix, preferably a Matrix package sparse matrix @param p.terms = if not NULL, a numeric vector representing the probabiliy of each term in the corpus whose names correspond to colnames(dtm). @details A numeric vector of the format p.terms. The entries of the vectors correspond to the difference in the probability of drawing a term from the set of documents given by docnames and the probability of drawing that term from the corpus overall (p.terms) @export @examples mydocs <- rownames(ldatheta[ , 5 ] >= 0.25 ] # documents with a topic proportion of .25 or higher for topic 5

term.probs <- Matrix::colSums(dtm) / sum(Matrix::colSums(dtm))

GetProbableTerms(docnames=mydocs, dtm=dtm, p.terms=term.probs)

GetProbableTerms <- function(docnames, dtm, p.terms=NULL){  
   
   
 # if p.terms is NULL, then create p.terms  
 if( is.null(p.terms) ){  
 p.terms <- Matrix::colSums(dtm) / sum(Matrix::colSums(dtm))  
 }  
   
 # get probability of terms given docnames  
 if( length(docnames) == 1 ){   
 p.terms.given.docs <- dtm[ docnames , ]   
 }else{  
 p.terms.given.docs <- Matrix::colSums(dtm[ docnames , ])  
 }  
   
 p.terms.given.docs <- p.terms.given.docs/sum(p.terms.given.docs)  
   
 # get our result, the difference  
 result <- p.terms.given.docs - p.terms  
   
 names(result) <- colnames(dtm)  
   
 return(result)  
   
}