Seperation Posts by Document

Step 1: Exclude univs that are hard to seperate posts

library(ggplot2)  
load("data\\fbdata.RData")  
  
temp <- file[c(file$from\_name!="전대숲-전국대학생 통합 대나무숲"),]  
temp <- temp[c(temp$from\_name!="단국대학교 대나무숲"),]  
temp <- temp[c(temp$from\_name!="서울과학기술대학교 대나무숲"),]  
temp <- temp[c(temp$from\_name!="홍익대학교 대나무숲"),]  
temp <- temp[c(temp$from\_name!="서울시립대학교 대나무숲"),]  
temp <- temp[c(temp$from\_name!="건국대학교 대나무숲"),]  
temp <- temp[c(temp$from\_name!="한국외국어대학교 글로벌캠퍼스 대나무숲"),]  
  
file <- temp; rm(temp)

Step2: set Seperate Rules

# Whole university names  
univs.name = levels(factor(file$from\_name))  
  
# Univs with posts that can be seperated by #...(hashtag)  
univs.hashsep.name <- c("인천대학교 대나무숲", "의학과, 의예과 대나무숲",  
 "국립 경상대학교 대나무숲", "동덕여자대학교 대나무숲",  
 "충북대학교 대나무숲", "강남대학교 대나무숲",  
 "한국항공대학교 대나무숲", "부산대학교 대나무숲",  
 "세종대학교 대나무숲", "아주대학교 대나무숲" ,  
 "성균관대학교 대나무숲", "한양대학교 대나무숲",  
 "CAU Bamboo Grove", "서강대학교 대나무숲",  
 "한성대학교 대나무숲", "숙명여대 대나무숲",  
 "숭실대학교 대나무숲")  
  
univs.hashsep <- subset(file, from\_name %in% univs.hashsep.name)  
levels(factor(univs.hashsep$from\_name)) %in% univs.hashsep.name

## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [15] TRUE TRUE TRUE

# Univs with posts that can be seperated by (1)(2)...numbering  
univs.numbered.name <- c("연세대학교 대나무숲", "SNU Bamboo Grove")  
  
univs.numbered <- subset(file, from\_name %in% univs.numbered.name)  
levels(factor(univs.numbered$from\_name)) %in% univs.numbered.name

## [1] TRUE TRUE

# Univs don't need seperation  
univs.nosep.name <- setdiff(univs.name, union(univs.hashsep.name, univs.numbered.name))  
  
univs.nosep <- subset(file, from\_name %in% univs.nosep.name)  
levels(factor(univs.nosep$from\_name)) %in% univs.nosep.name

## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

#final check  
nrow(file) == nrow(univs.hashsep) + nrow(univs.nosep) + nrow(univs.numbered)

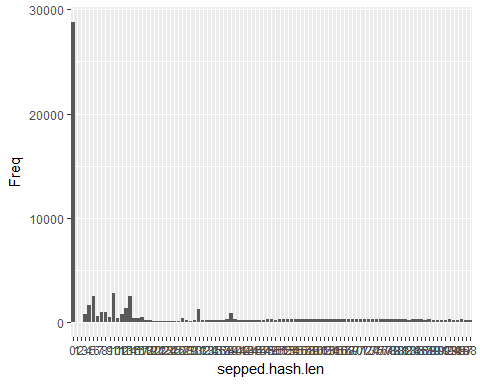
## [1] TRUE

Step3: Seperating

Step3-1: Split hashed Unviersiies such that #….#….

#leave only messages   
univs.hashsep <- univs.hashsep$message  
univs.numbered <- univs.numbered$message  
univs.nosep <- univs.nosep$message  
  
#split  
sepped.hash <-unlist(sapply(X = univs.hashsep ,FUN = function(x) strsplit(x,"#"),USE.NAMES = FALSE))  
  
  
#check  
sepped.hash.len <- sapply(sepped.hash, nchar)  
sepped.hash.len <- unname(sepped.hash.len)  
  
#when nchar <= 14, we can see dramatic increase of Frequency   
tab <- table(sepped.hash.len)  
tab <- as.data.frame(tab)  
tab <- subset(tab, as.numeric(sepped.hash.len) < 100)  
ggplot(data=tab, aes(x = sepped.hash.len, y = Freq)) + geom\_histogram(stat="identity")

## Warning: Ignoring unknown parameters: binwidth, bins, pad



# We can see that seperated messages with nchar <= 14 are meaningless.  
sepped.hash[sepped.hash.len==17]

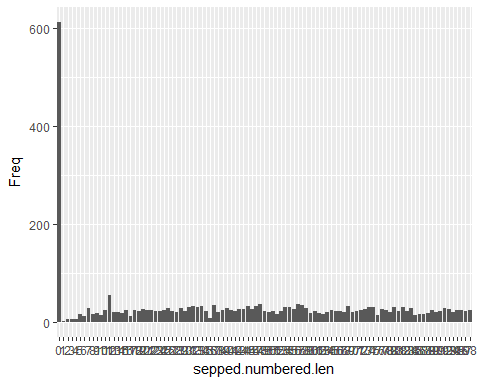
## [1] "87\n\n중대 카우버거 맛있어요?"   
## [2] "숲숲이의\_이야기☆\n\n안녕하세요 "   
## [3] "상쾌한\_새벽에는\_세개씩묶어야지\n"   
## [4] "사랑으로\_보듬어주실분을\_찾아요\n"   
## [5] "대나무25166\n사랑은 타이밍."   
## [6] "학교>\n과장이없지않아있지만 현실"   
## [7] "야옹이도\nㅠㅠㅠㅠㅠㅠㅠㅠㅠㅠㅠㅠ"  
## [8] "고민>\n이걸 기뻐해야 하나..."   
## [9] "연애>\n서강대 학우분들 모두 다"   
## [10] "기타>\n엉엉엉 교수님 미워이이잉"   
## [11] "학교>\n저희 종강은 언제인가요?"   
## [12] "가난충\n\n가난충의한끼식사.jpg"   
## [13] "미필들희소식JPG\n\n아주 축하해"   
## [14] "좋아하는거 아니면 끼부리지마ㅡㅡ"   
## [15] "792\n내일 중-중간고사 봅니다"   
## [16] "한양대분실물\n백번 분실물입니다!"   
## [17] "7190번\n사진 주인을 찾습니다"   
## [18] "교육공학과\_유다경씨\_감사합니다<U+2764>"  
## [19] "아름사자 )들 잊지 않기...★"   
## [20] "23150번째사자후\n1학년 꿀잼"   
## [21] "도와줘요선배님들뿌잉뿌잉뿌뿌잉\n\n"  
## [22] "22297번째사자후\n반성문입니다"   
## [23] "\*\*\*학과\_성추행\_사건\_제보\n\n"   
## [24] "95년생 신검나왔어요 군대가욧!"   
## [25] "740\n썸 좀 타 고 싶 다\n\n"   
## [26] "혼잣말\_제보\_모음\_끝\n\n아대숲 "   
## [27] "혼잣말\_제보\_모음\_3\n\n아대숲 "   
## [28] "미안해요\_사랑해요\_♡\n\n아대숲 "   
## [29] "자정의\_흥신소\_두번째\n\n아대숲 "   
## [30] "분실물!\_주인에게로!\n\n아대숲 "   
## [31] "이런거\_하지맠ㅋㅋㅋㅋ\n\n아대숲 "   
## [32] "분실물과\_찾아요\_추억\n\n아대숲 "   
## [33] "9279번째 선구자의 외침:\n\n"   
## [34] "한국사\_국정화\_반대의견\n아대숲 "   
## [35] "한국사\_국정화\_찬성의견\n아대숲 "   
## [36] "8682제보에대한\_의견\n아대숲 "   
## [37] "대나무숲은총학이아니라규\n아대숲 "   
## [38] "늦어서 미안해요 :(\n\n아대숲 "   
## [39] "제보가 없더어....\n\n아대숲 "   
## [40] "▼아주대학교대나무숲 직접작성\n\n"   
## [41] "아대숲\_266번째\_선구자외침:\n"   
## [42] "아대숲\_263번째\_선구자외침:\n"   
## [43] "아대숲\_255번째\_선구자외침:\n"   
## [44] "아대숲\_254번째\_선구자외침:\n"   
## [45] "아대숲\_253번째\_선구자외침:\n"   
## [46] "아대숲\_252번째\_선구자외침:\n"   
## [47] "아대숲\_251번째\_선구자외침:\n"   
## [48] "아대숲\_250번째\_선구자외침:\n"   
## [49] "아대숲\_249번째\_선구자외침:\n"   
## [50] "아대숲\_248번째\_선구자외침:\n"   
## [51] "아대숲\_247번째\_선구자외침:\n"   
## [52] "아대숲\_246번째\_선구자외침:\n"   
## [53] "아대숲\_245번째\_선구자외침:\n"   
## [54] "아대숲\_244번째\_선구자외침:\n"   
## [55] "아대숲\_240번째\_선구자외침:\n"   
## [56] "아대숲\_239번째\_선구자외침:\n"   
## [57] "아대숲\_238번째\_선구자외침:\n"   
## [58] "아대숲\_237번째\_선구자외침:\n"   
## [59] "아대숲\_236번째\_선구자외침:\n"   
## [60] "아대숲\_235번째\_선구자외침:\n"   
## [61] "아대숲\_234번째\_선구자외침:\n"   
## [62] "아대숲\_233번째\_선구자외침:\n"   
## [63] "아대숲\_232번째\_선구자외침:\n"   
## [64] "아대숲\_231번째\_선구자외침:\n"   
## [65] "아대숲\_230번째\_선구자외침:\n"   
## [66] "아대숲\_229번째\_선구자외침:\n"   
## [67] "아대숲\_228번째\_선구자외침:\n"   
## [68] "아대숲\_227번째\_선구자외침:\n"   
## [69] "아대숲\_226번째\_선구자외침;\n"   
## [70] "아대숲\_224번째\_선구자외침:\n"   
## [71] "아대숲\_223번째\_선구자외침:\n"   
## [72] "아대숲\_222번째\_선구자외침:\n"   
## [73] "아대숲\_221번째\_선구자외침:\n"   
## [74] "아대숲\_220번째\_선구자외침:\n"   
## [75] "아대숲\_218번째\_선구자외침:\n"   
## [76] "아대숲\_217번째\_선구자외침:\n"   
## [77] "아대숲\_216번째\_선구자외침:\n"   
## [78] "아대숲\_215번째\_선구자외침:\n"   
## [79] "아대숲\_214번째\_선구자외침:\n"   
## [80] "아대숲\_213번째\_선구자외침:\n"   
## [81] "아대숲\_212번째\_선구자외침:\n"   
## [82] "아대숲\_211번째\_선구자외침:\n"   
## [83] "아대숲\_210번째\_선구자외침:\n"   
## [84] "아대숲\_209번째\_선구자외침:\n"   
## [85] "아대숲\_208번째\_선구자외침:\n"   
## [86] "아대숲\_207번째\_선구자외침:\n"   
## [87] "아대숲\_206번째\_선구자외침:\n"   
## [88] "아대숲\_205번째\_선구자외침:\n"   
## [89] "아대숲\_204번째\_선구자외침:\n"   
## [90] "아대숲\_203번째\_선구자외침:\n"   
## [91] "아대숲\_202번째\_선구자외침:\n"   
## [92] "아대숲\_201번째\_선구자외침:\n"   
## [93] "아대숲\_200번째\_선구자외침:\n"   
## [94] "아대숲\_199번째\_선구자외침:\n"   
## [95] "아대숲\_198번째\_선구자외침:\n"   
## [96] "아대숲\_197번째\_선구자외침:\n"   
## [97] "아대숲\_196번째\_선구자외침:\n"   
## [98] "아대숲\_195번째\_선구자외침:\n"   
## [99] "아대숲\_194번째\_선구자외침:\n"   
## [100] "아대숲\_193번째\_선구자외침:\n"   
## [101] "아대숲\_192번째\_선구자외침:\n"   
## [102] "아대숲\_191번째\_선구자외침:\n"   
## [103] "아대숲\_190번째\_선구자외침:\n"   
## [104] "아대숲\_189번째\_선구자외침:\n"   
## [105] "아대숲\_188번째\_선구자외침:\n"   
## [106] "아대숲\_187번째\_선구자외침:\n"   
## [107] "아대숲\_186번째\_선구자외침:\n"   
## [108] "아대숲\_185번째\_선구자외침:\n"   
## [109] "아대숲\_184번째\_선구자외침:\n"   
## [110] "아대숲\_183번째\_선구자외침:\n"   
## [111] "아대숲\_182번째\_선구자외침:\n"   
## [112] "아대숲\_181번째\_선구자외침:\n"   
## [113] "아대숲\_180번째\_선구자외침:\n"   
## [114] "아대숲\_179번째\_선구자외침:\n"   
## [115] "아대숲\_178번째\_선구자외침:\n"   
## [116] "아대숲\_177번째\_선구자외침:\n"   
## [117] "아대숲\_176번째\_선구자외침:\n"   
## [118] "아대숲\_175번째\_선구자외침:\n"   
## [119] "아대숲\_174번째\_선구자외침:\n"   
## [120] "아대숲\_173번째\_선구자외침:\n"   
## [121] "아대숲\_172번째\_선구자외침:\n"   
## [122] "아대숲\_171번째\_선구자외침:\n"   
## [123] "아대숲\_170번째\_선구자외침:\n"   
## [124] "아대숲\_169번째\_선구자외침:\n"   
## [125] "아대숲\_168번째\_선구자외침:\n"   
## [126] "아대숲\_167번째\_선구자외침:\n"   
## [127] "아대숲\_166번째\_선구자외침:\n"   
## [128] "아대숲\_165번째\_선구자외침:\n"   
## [129] "아대숲\_164번째\_선구자외침:\n"   
## [130] "아대숲\_163번째\_선구자외침:\n"   
## [131] "아대숲\_162번째\_선구자외침:\n"   
## [132] "아대숲\_161번째\_선구자외침:\n"   
## [133] "아대숲\_160번째\_선구자외침:\n"   
## [134] "아대숲\_159번째\_선구자외침:\n"   
## [135] "아대숲\_158번째\_선구자외침:\n"   
## [136] "아대숲\_157번째\_선구자외침:\n"   
## [137] "아대숲\_156번째\_선구자외침:\n"   
## [138] "아대숲\_155번째\_선구자외침:\n"   
## [139] "아대숲\_154번째\_선구자외침:\n"   
## [140] "아대숲\_153번째\_선구자외침:\n"   
## [141] "아대숲\_152번째\_선구자외침:\n"   
## [142] "아대숲\_151번째\_선구자외침:\n"   
## [143] "아대숲\_150번째\_선구자외침:\n"   
## [144] "아대숲\_149번째\_선구자외침:\n"   
## [145] "아대숲\_148번째\_선구자외침:\n"   
## [146] "아대숲\_147번째\_선구자외침:\n"   
## [147] "아대숲\_146번째\_선구자외침:\n"   
## [148] "아대숲\_145번째\_선구자외침:\n"   
## [149] "851번째샛벌\n\n 안녕하세요. "   
## [150] "부산대대나무숲118번\n\n본글은 "   
## [151] "이것은 네 별, 저것은 내 별\""   
## [152] "5636번째울림\n보고싶다 ㅎㅇㅅ"   
## [153] "보고 싶다. 썼다 지웠다.\"\n\n"   
## [154] "당신 없이 남은 모든 나는\n\n\""   
## [155] "네 볼이 붉은 이유와 같을꺼야."   
## [156] "303번째울림\n화공 화이팅!!!"   
## [157] "155\n해군1차 붙으신분들\n띱."   
## [158] "일상\n\n말 무시하는거 개싫다;;"   
## [159] "학교\n\n방학가장 늦은 학교는…?"   
## [160] "엠제이), 그리고 예과 2학년("   
## [161] "공부\n전체 등수 언제 마오나요?"   
## [162] "분실물\n내 영혼을 돌려도...."   
## [163] "31816번\n아이고! 종강이야!"   
## [164] "26952번\n제 205 학군단\n"   
## [165] "14986번\n이혜민 보고싶다\n\n"   
## [166] "항공대동아리 시험관련 글이면 "   
## [167] "67번\n\n스킨쉽 팁좀 주세여ㅠㅠ"   
## [168] "사연은 오는데\n댓글은 없네ㅠㅠ\n"   
## [169] "453번\n두근두근 설렌다. \n\n"   
## [170] "5129번째\_메아리 \n도와주세요"   
## [171] "6635번째\_메아리\n종강이네요."   
## [172] "<휴학 및 복학 관련 질문>\n\n"   
## [173] "3호) ☆<U+301C>（<U+309D>。∂）차쟈땅!\n\n"  
## [174] "세종대\_대나무숲\_나고민있어요\n\n"   
## [175] "세종대\_대나무숲\_우리토론해요\n\n"   
## [176] "세종대\_대나무숲\_나할말있어요\n\n"   
## [177] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [178] "세종대\_대나무숲\_나할말있어요\n\n"   
## [179] "세종대\_대나무숲\_나할말있어요\n\n"   
## [180] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [181] "세종대\_대나무숲\_나고민잇어오\n\n"   
## [182] "세종대\_대나무숲\_나고민있어요\n\n"   
## [183] "세종대\_대나무숲\_나고민있어요\n\n"   
## [184] "세종대\_대나무숲\_나고민있어요\n\n"   
## [185] "세종대\_대나무숲\_나고민있어요\n\n"   
## [186] "세종대\_대나무숲\_나할말있어요\n\n"   
## [187] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [188] "세종대\_대나무숲\_연애는어려워\n\n"   
## [189] "세종대\_대나무숲\_나고민있어요\n\n"   
## [190] "세종대\_대나무숲\_나할말있어요\n\n"   
## [191] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [192] "세종대\_대나무숲\_나할말있어요\n\n"   
## [193] "세종대\_대나무숲\_나고민있어요\n\n"   
## [194] "세종대\_대나무숲\_나할말있어요\n\n"   
## [195] "세종대\_대나무숲\_나할말있어요\n\n"   
## [196] "세종대\_대나무숲\_나고민있어요\n\n"   
## [197] "세종대\_대나무숲\_나할말있어요\n\n"   
## [198] "세종대\_대나무숲\_나고민있어요\n\n"   
## [199] "세종대\_대나무숲\_우리토론해요\n\n"   
## [200] "세종대\_대나무숲\_나할말있어요\n\n"   
## [201] "세종대\_대나무숲\_나고민있어요\n\n"   
## [202] "세종대\_대나무숲\_나고민있어요\n\n"   
## [203] "세종대\_대나무숲\_우리토론해요\n\n"   
## [204] "세종대\_대나무숲\_나할말있어요\n\n"   
## [205] "세종대\_대나무숲\_나고민있어요\n\n"   
## [206] "세종대\_대나무숲\_나할말있어요\n\n"   
## [207] "세종대\_대나무숲\_나고민있어요\n\n"   
## [208] "세종대\_대나무숲\_나할말있어요\n\n"   
## [209] "세종대\_대나무숲\_나할말있어요\n\n"   
## [210] "세종대\_대나무숲\_나할말있어요\n\n"   
## [211] "세종대\_대나무숲\_나할말있어요\n\n"   
## [212] "세종대\_대나무숲\_나할말있어요\n\n"   
## [213] "세종대\_대나무숲\_나고민있어요\n\n"   
## [214] "세종대\_대나무숲\_나할말있어요\n\n"   
## [215] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [216] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [217] "세종대\_대나무숲\_나할말있어요\n\n"   
## [218] "세종대\_대나무숲\_나할말있어요\n\n"   
## [219] "세종대\_대나무숲\_나할말있어요\n\n"   
## [220] "세종대\_대나무숲\_나할말있어요\n\n"   
## [221] "세종대\_대나무숲\_나할말있어요\n\n"   
## [222] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [223] "세종대\_대나무숲\_나할말있어요\n\n"   
## [224] "세종대\_대나무숲\_나고민있어요\n\n"   
## [225] "세종대\_대나무숲\_나고민있어요\n\n"   
## [226] "세종대\_대나무숲\_나할말있어요\n\n"   
## [227] "세종대\_대나무숲\_나할말있어요\n\n"   
## [228] "세종대\_대나무숲\_나할말있어요\n\n"   
## [229] "세종대\_대나무숲\_나할말있어요\n\n"   
## [230] "세종대\_대나무숲\_나고민있어요\n\n"   
## [231] "세종대\_대나무숲\_나할말있어요\n\n"   
## [232] "세종대\_대나무숲\_나고민있어요\n\n"   
## [233] "세종대\_대나무숲\_나할말있어요\n\n"   
## [234] "세종대\_대나무숲\_나할말있어요\n\n"   
## [235] "세종대\_대나무숲\_나할말있어요\n\n"   
## [236] "세종대\_대나무숲\_나고민있어요\n\n"   
## [237] "세종대\_대나무숲\_우리토론해요\n\n"   
## [238] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [239] "세종대\_대나무숲\_나할말있어요\n\n"   
## [240] "세종대\_대나무숲\_나할말있어요\n\n"   
## [241] "세종대\_대나무숲\_나할말있어요\n\n"   
## [242] "세종대\_대나무숲\_나할말있어요\n\n"   
## [243] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [244] "세종대\_대나무숲\_나고민있어요\n\n"   
## [245] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [246] "세종대\_대나무숲\_나할말있어요\n\n"   
## [247] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [248] "세종대\_대나무숲\_나고민있어요\n\n"   
## [249] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [250] "세종대\_대나무숲\_나고민있어요\n\n"   
## [251] "세종대\_대나무숲\_나할말있어요\n\n"   
## [252] "세종대\_대나무숲\_나고민이써요\n\n"   
## [253] "세종대\_대나무숲\_나할말있어요\n\n"   
## [254] "세종대\_대나무숲\_나할말있어요\n\n"   
## [255] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [256] "세종대\_대나무숲\_나할말있어요\n\n"   
## [257] "세종대\_대나무숲\_나할말있어요\n\n"   
## [258] "세종대\_대나무숲\_나고민있어요\n\n"   
## [259] "세종대\_대나무숲\_나할말있어요\n\n"   
## [260] "세종대\_대나무숲\_나할말있어요\n\n"   
## [261] "세종대\_대나무숲\_나할말있어요\n\n"   
## [262] "세종대\_대나무숲\_나할말있어요\n\n"   
## [263] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [264] "세종대\_대나무숲\_우리토론해요\n\n"   
## [265] "세종대\_대나무숲\_나할말있어요\n\n"   
## [266] "세종대\_대나무숲\_나할말있어요\n\n"   
## [267] "세종대\_대나무숲\_나고민있어요\n\n"   
## [268] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [269] "세종대\_대나무숲\_나할말있어요\n\n"   
## [270] "세종대\_대나무숲\_나할말있어요\n\n"   
## [271] "세종대\_대나무숲\_나할말있어요\n\n"   
## [272] "세종대\_대나무숲\_나할말있어요\n\n"   
## [273] "세종대\_대나무숲\_나할말있어요\n\n"   
## [274] "세종대\_대나무숲\_나할말있어요\n\n"   
## [275] "세종대\_대나무숲\_나할말있어요\n\n"   
## [276] "세종대\_대나무숲\_나고민있어요\n\n"   
## [277] "세종대\_대나무숲\_나고민있어요\n\n"   
## [278] "세종대\_대나무숲\_나할말있어요\n\n"   
## [279] "세종대\_대나무숲\_우리토론해요\n\n"   
## [280] "세종대\_대나무숲\_나할말있어요\n\n"   
## [281] "세종대\_대나무숲\_나할말있어요\n\n"   
## [282] "세종대\_대나무숲\_나할말있어요\n\n"   
## [283] "세종대\_대나무숲\_나할말있어요\n\n"   
## [284] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [285] "세종대\_대나무숲\_나할말있어요\n\n"   
## [286] "세종대\_대나무숲\_나할말있어요\n\n"   
## [287] "세종대\_대나무숲\_나할말있어요\n\n"   
## [288] "세종대\_대나무숲\_나할말있어요\n\n"   
## [289] "세종대\_대나무숲\_나고민있어요\n\n"   
## [290] "세종대\_대나무숲\_나할말있어요\n\n"   
## [291] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [292] "세종대\_대나무숲\_나할말있어요\n\n"   
## [293] "세종대\_대나무숲\_우리토론해요\n\n"   
## [294] "세종대\_대나무숲\_나고민있어요\n\n"   
## [295] "세종대\_대나무숲\_나할말있어요\n\n"   
## [296] "세종대\_대나무숲\_나할말있어요\n\n"   
## [297] "세종대\_대나무숲\_나고민있어요\n\n"   
## [298] "세종대\_대나무숲\_나고민있어요\n\n"   
## [299] "세종대\_대나무숲\_나할말있어요\n\n"   
## [300] "세종대\_대나무숲\_나할말있어요\n\n"   
## [301] "세종대\_대나무숲\_나고민있어요\n\n"   
## [302] "세종대\_대나무숲\_나고민있어요\n\n"   
## [303] "세종대\_대나무숲\_나할말있어요\n\n"   
## [304] "세종대\_대나무숲\_나고민있어요\n\n"   
## [305] "세종대\_대나무숲\_나할말있어요\n\n"   
## [306] "세종대\_대나무숲\_나할말있어요\n\n"   
## [307] "세종대\_대나무숲\_나할말있어요\n\n"   
## [308] "세종대\_대나무숲\_나고민있어요\n\n"   
## [309] "세종대\_대나무숲\_나고민있어요\n\n"   
## [310] "세종대\_대나무숲\_나할말있어요\n\n"   
## [311] "세종대학교\_대나무숲\_나화나요\n\n"   
## [312] "세종대\_대나무숲\_나연애해요 \n\n"   
## [313] "세종대\_대나무숲\_나할말있어요\n\n"   
## [314] "세종대\_대나무숲\_나고민있어요\n\n"   
## [315] "세종대\_대나무숲\_나고민있어요\n\n"   
## [316] "세종대\_대나무숲\_나할말있어요\n\n"   
## [317] "세종대\_대나무숲\_나할말있어요\n\n"   
## [318] "세종대\_대나무숲\_나할말있어요\n\n"   
## [319] "세종대\_대나무숲\_나할말있어요\n\n"   
## [320] "세종대\_대나무숲\_나할말있어요\n\n"   
## [321] "세종대학교\_대나무숲\_나화나요\n\n"   
## [322] "세종대학교\_대나무숲\_널찾아요\n\n"   
## [323] "세종대\_대나무숲\_나고민있어요\n\n"   
## [324] "세종대\_대나무숲\_나고민있어요\n\n"   
## [325] "세종대\_대나무숲\_나고민있어요\n\n"   
## [326] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [327] "세종대\_대나무숲\_나할말있어요\n\n"   
## [328] "세종대\_대나무숲\_나할말있어요\n\n"   
## [329] "세종대\_대나무숲\_나고민있어요\n\n"   
## [330] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [331] "세종대\_대나무숲\_나할말있어요\n\n"   
## [332] "세종대\_대나무숲\_나불만있어요\n\n"   
## [333] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [334] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [335] "세종대\_대나무숲\_나할말있어요\n\n"   
## [336] "세종대\_대나무숲\_나할말있어요\n\n"   
## [337] "세종대\_대나무숲\_나할말있어요\n\n"   
## [338] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [339] "세종대\_대나무숲\_나할말있어요\n\n"   
## [340] "세종대\_대나무숲\_화장실화나요\n\n"   
## [341] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [342] "세종대\_대나무숲\_나고민있어요\n\n"   
## [343] "세종대\_대나무숲\_나할말있어요\n\n"   
## [344] "세종대\_대나무숲\_나고민있어요\n\n"   
## [345] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [346] "세종대\_대나무숲\_나할말있어요\n\n"   
## [347] "세종대\_대나무숲\_우리토론해요\n\n"   
## [348] "세종대\_대나무숲\_나할말있어요\n\n"   
## [349] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [350] "세종대\_대나무숲\_나고민있어요\n\n"   
## [351] "세종대\_대나무숲\_나할말있어요\n\n"   
## [352] "세종대\_대나무숲\_나고민있어요\n\n"   
## [353] "세종대\_대나무숲\_나할말있어요\n\n"   
## [354] "세종대\_대나무숲\_나할말있어요\n\n"   
## [355] "세종대\_대나무숲\_나할말있어요\n\n"   
## [356] "세종대\_대나무숲\_나고민있어요\n\n"   
## [357] "세종대\_대나무숲\_나할말있어요\n\n"   
## [358] "세종대\_대나무숲\_나할말있어요\n\n"   
## [359] "세종대\_대나무숲\_나할말있어요\n\n"   
## [360] "세종대\_대나무숲\_나고민있어요\n\n"   
## [361] "세종대\_대나무숲\_나할말있어요\n\n"   
## [362] "세종대\_대나무숲\_나할말있어요\n\n"   
## [363] "세종대\_대나무숲\_나할말있어요\n\n"   
## [364] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [365] "세종대\_대나무숲\_난군인이에요\n\n"   
## [366] "세종대\_대나무숲\_나할말있어요\n\n"   
## [367] "세종대\_대나무숲\_나할말있어요\n\n"   
## [368] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [369] "세종대\_대나무숲\_나고민있어요\n\n"   
## [370] "세종대\_대나무숲\_나헤어졌어요\n\n"   
## [371] "세종대\_대나무숲\_나할말있어요\n\n"   
## [372] "세종대\_대나무숲\_나할말있어요\n\n"   
## [373] "세종대\_대나무숲\_나할말있어요\n\n"   
## [374] "세종대\_대나무숲\_나할말있어요\n\n"   
## [375] "세종대\_대나무숲\_나할말있어요\n\n"   
## [376] "세종대\_대나무숲\_학과와사람들\n\n"   
## [377] "세종대\_대나무숲\_학과와사람들\n\n"   
## [378] "세종대\_대나무숲\_학과와사람들\n\n"   
## [379] "세종대\_대나무숲\_학과와사람들\n\n"   
## [380] "세종대\_대나무숲\_학과와사람들\n\n"   
## [381] "세종대\_대나무숲\_학과와사람들\n\n"   
## [382] "세종대\_대나무숲\_학과와사람들\n\n"   
## [383] "세종대\_대나무숲\_학과와사람들\n\n"   
## [384] "세종대\_대나무숲\_학과와사람들\n\n"   
## [385] "세종대\_대나무숲\_학과와사람들\n\n"   
## [386] "세종대\_대나무숲\_학과와사람들\n\n"   
## [387] "세종대\_대나무숲\_학과와사람들\n\n"   
## [388] "세종대\_대나무숲\_학과와사람들\n\n"   
## [389] "세종대\_대나무숲\_학과와사람들\n\n"   
## [390] "세종대\_대나무숲\_학과와사람들\n\n"   
## [391] "세종대\_대나무숲\_학과와사람들\n\n"   
## [392] "세종대\_대나무숲\_학과와사람들\n\n"   
## [393] "세종대\_대나무숲\_학과와사람들\n\n"   
## [394] "세종대\_대나무숲\_학과와사람들\n\n"   
## [395] "세종대\_대나무숲\_학과와사람들\n\n"   
## [396] "세종대\_대나무숲\_우리는세종인\n\n"   
## [397] "세종대\_대나무숲\_학과와사람들\n\n"   
## [398] "세종대\_대나무숲\_학과와사람들\n\n"   
## [399] "세종대\_대나무숲\_학과와사람들\n\n"   
## [400] "세종대\_대나무숲\_학과와사람들\n\n"   
## [401] "세종대\_대나무숲\_좋아하는사람\n\n"   
## [402] "세종대\_대나무숲\_학과와사람들\n\n"   
## [403] "세종대\_대나무숲\_학과와사람들\n\n"   
## [404] "세종대\_대나무숲\_학과와사람들\n\n"   
## [405] "세종대\_대나무숲\_학과와사람들\n\n"   
## [406] "세종대\_대나무숲\_학과와사람들\n\n"   
## [407] "세종대\_대나무숲\_학과와사람들\n\n"   
## [408] "세종대\_대나무숲\_그것이알고싶\n\n"   
## [409] "세종대\_대나무숲\_넌이별난아직\n\n"   
## [410] "세종대\_대나무숲\_학과와사람들\n\n"   
## [411] "세종대\_대나무숲\_학과와사람들\n\n"   
## [412] "세종대\_대나무숲\_헌내기의걱정\n\n"   
## [413] "세종대\_대나무숲\_새내기의걱정\n\n"   
## [414] "세종대\_대나무숲\_학과와사람들\n\n"   
## [415] "세종대\_대나무숲\_학과와사람들\n\n"   
## [416] "세종대\_대나무숲\_학과와사람들\n\n"   
## [417] "세종대\_대나무숲\_학과와사람들\n\n"   
## [418] "세종대\_대나무숲\_학과와사람들\n\n"   
## [419] "세종대\_대나무숲\_학과와사람들\n\n"   
## [420] "세종대\_대나무숲\_학과와사람들\n\n"   
## [421] "세종대\_대나무숲\_너말하는거야\n\n"   
## [422] "세종대\_대나무숲\_학과와사람들\n\n"   
## [423] "세종대\_대나무숲\_학과와사람들\n\n"   
## [424] "세종대\_대나무숲\_학과와사람들\n\n"   
## [425] "세종대\_대나무숲\_학과와사람들\n\n"   
## [426] "세종대\_대나무숲\_학과와사람들\n\n"   
## [427] "세종대\_대나무숲\_통학버스운행\n\n"   
## [428] "세종대\_대나무숲\_수강신청특집\n\n"   
## [429] "세종대\_대나무숲\_거짓말탐지기\n\n"   
## [430] "세종대\_대나무숲\_수강신청특집\n\n"   
## [431] "세종대\_대나무숲\_수강신청특집\n\n"   
## [432] "세종대\_대나무숲\_학과와사람들\n\n"   
## [433] "세종대\_대나무숲\_학과와사람들\n\n"   
## [434] "세종대학교\_대나무숲\_dash\n\n"   
## [435] "세종대\_대나무숲\_질투는나의힘\n\n"   
## [436] "세종대\_대나무숲\_그만좀주무셔\n\n"   
## [437] "세종대\_대나무숲\_학과와사람들\n\n"   
## [438] "세종대\_대나무숲\_내생각좀해줘\n\n"   
## [439] "세종대\_대나무숲\_학과와사람들\n\n"   
## [440] "세종대\_대나무숲\_오빠오빠오빠\n\n"   
## [441] "세종대\_대나무숲\_학과와사람들\n\n"   
## [442] "세종대\_대나무숲\_학과와사람들\n\n"   
## [443] "세종대\_대나무숲\_학과와사람들\n\n"   
## [444] "세종대\_대나무숲\_모두들외롭다\n\n"   
## [445] "세종대\_대나무숲\_그러지마세요\n\n"   
## [446] "세종대\_대나무숲\_오빠대란\_3\n\n"   
## [447] "세종대\_대나무숲\_오빠대란\_1\n\n"   
## [448] "세종대\_대나무숲\_연애의후유증\n\n"   
## [449] "세종대\_대나무숲\_족보브레이커\n\n"   
## [450] "세종대\_대나무숲\_수강신청특집\n\n"   
## [451] "세종대\_대나무숲\_학과와사람들\n\n"   
## [452] "세종대\_대나무숲\_학과와사람들\n\n"   
## [453] "세종대\_대나무숲\_학과와사람들\n\n"   
## [454] "세종대\_대나무숲\_너자신을알라\n\n"   
## [455] "세종대\_대나무숲\_발표는힘들어\n\n"   
## [456] "세종대\_대나무숲\_은근슬쩍슬픈\n\n"   
## [457] "세종대\_대나무숲\_학과와사람들\n\n"   
## [458] "세종대\_대나무숲\_흔한연애패턴\n\n"   
## [459] "세종대\_대나무숲\_학과와사람들\n\n"   
## [460] "세종대\_대나무숲\_학과와사람들\n\n"   
## [461] "세종대\_대나무숲\_학과와사람들\n\n"   
## [462] "세종대\_대나무숲\_학과와사람들\n\n"   
## [463] "세종대\_대나무숲\_학과와사람들\n\n"   
## [464] "세종대\_대나무숲\_학과와사람들\n\n"   
## [465] "세종대\_대나무숲\_훈남사냥꾼들\n\n"   
## [466] "세종대\_대나무숲\_학과와사람들\n\n"   
## [467] "세종대\_대나무숲\_학과와사람들\n\n"   
## [468] "세종대\_대나무숲\_학과와사람들\n\n"   
## [469] "세종대\_대나무숲\_학과와사람들\n\n"   
## [470] "세종대\_대나무숲\_학과와사람들\n\n"   
## [471] "세종대\_대나무숲\_학과와사람들\n\n"   
## [472] "세종대\_대나무숲\_수강정정특집\n\n"   
## [473] "세종대\_대나무숲\_학과와사람들\n\n"   
## [474] "세종대\_대나무숲\_수강신청특집\n\n"   
## [475] "세종대학교\_대나무숲\_고발폭로\n\n"   
## [476] "세종대\_대나무숲\_수강신청특집\n\n"   
## [477] "세종대\_대나무숲\_학과와사람들\n\n"   
## [478] "세종대\_대나무숲\_학과와사람들\n\n"   
## [479] "세종대\_대나무숲\_학과와사람들\n\n"   
## [480] "세종대\_대나무숲\_학과와사람들\n\n"   
## [481] "세종대\_대나무숲\_학과와사람들\n\n"   
## [482] "세종대\_대나무숲\_학과와사람들\n\n"   
## [483] "세종대\_대나무숲\_학과와사람들\n\n"   
## [484] "세종대\_대나무숲\_학과와사람들\n\n"   
## [485] "세종대\_대나무숲\_학과와사람들\n\n"   
## [486] "세종대\_대나무숲\_학과와사람들\n\n"   
## [487] "세종대\_대나무숲\_학과와사람들\n\n"   
## [488] "세종대\_대나무숲\_학과와사람들\n\n"   
## [489] "세종대\_대나무숲\_학과와사람들\n\n"   
## [490] "세종대\_대나무숲\_학과와사람들\n\n"   
## [491] "세종대\_대나무숲\_학과와사람들\n\n"   
## [492] "세종대\_대나무숲\_학과와사람들\n\n"   
## [493] "세종대\_대나무숲\_학과와사람들\n\n"   
## [494] "세종대\_대나무숲\_학과와사람들\n\n"   
## [495] "세종대\_대나무숲\_학과와사람들\n\n"   
## [496] "세종대\_대나무숲\_학과와사람들\n\n"   
## [497] "1011번\n익명제보해요 게시물 "   
## [498] "세종대\_대나무숲\_학과와사람들\n\n"   
## [499] "세종대\_대나무숲\_학과와사람들\n\n"   
## [500] "세종대\_대나무숲\_학과와사람들\n\n"   
## [501] "세종대\_대나무숲\_학과와사람들\n\n"   
## [502] "세종대\_대나무숲\_학과와사람들\n\n"   
## [503] "세종대\_대나무숲\_학과와사람들\n\n"   
## [504] "세종대\_대나무숲\_학과와사람들\n\n"   
## [505] "539번\n어장인거아는데 좋다\n\n"   
## [506] "7693번\n통학시간 배틀 갑시다"   
## [507] "사람을\_찾는\_글은\_끝이\_없다\n\n"   
## [508] "9459번\ndkdldjsaos!"   
## [509] "셤기간\_업로드\_늦어\_미안합니다\n"   
## [510] "연락한번\_해줍시다\_사람죽겠네.\n"   
## [511] "21424번\n짜증난다 헤어질까!"   
## [512] "27725번\n던.파.좋.아.\n\n"   
## [513] "29098번\n씨씨 어때요? \n\n"   
## [514] "33236번\n보고싶다 ㅎㄱ아\n\n"   
## [515] "청파제 후기 모음글 1 ~ \n\n"   
## [516] "진리의 사람 바이 사람...\n\n"   
## [517] "이번엔 꼭 찾으시길...^^\n\n"   
## [518] "새내기질문 학교질문글 모음~\n\n"

# Therefore, remove them  
sepped.hash14 <- sepped.hash[sepped.hash.len > 14] # strict version  
sepped.hash17 <- sepped.hash[sepped.hash.len > 17] # naive version (more loss)

Step3-2: Split numbered Univs such that   (1)…(2)…

#Test the seperation rule  
check <- gregexpr("\\(\\d\\)", univs.numbered)  
target.index <- unlist(lapply(check, function(x) x[1] != -1 ))  
  
#Working good  
target <- univs.numbered[target.index]  
test <- strsplit(target[1:10],split = "\\(\\d\\)")  
  
# SPlit!  
sepped.numbered <- unlist(sapply(X = univs.numbered, FUN = function(x) strsplit(x, "\\(\\d\\)"), USE.NAMES = FALSE))  
  
#check  
sepped.numbered.len <- sapply(sepped.numbered, nchar)  
sepped.numbered.len <- unname(sepped.numbered.len)  
  
#Here we can see that nchar == 0 is outstanding  
tab <- table(sepped.numbered.len)  
tab <- as.data.frame(tab)  
tab <- subset(tab, as.numeric(sepped.numbered.len) < 100)  
ggplot(data=tab, aes(x = sepped.numbered.len, y = Freq)) + geom\_histogram(stat="identity")

## Warning: Ignoring unknown parameters: binwidth, bins, pad



# And nchar == 0 are meaningless  
sepped.numbered[sepped.numbered.len==1]

## [1] "ㅎ" "!" "~"

# Therefore, remove them  
sepped.numbered <- sepped.numbered[sepped.numbered.len != 0]

Step4: Merge

#version1 : Strict (less loss, but a lot of noise)  
Sepped.v1 <- c(univs.nosep, sepped.hash14, sepped.numbered)  
  
#Version2: Naive (more loss, but less noise)  
Sepped.v2 <- c(univs.nosep, sepped.hash17, sepped.numbered)  
setwd("C:\\Users\\kyuchul\\Documents\\Carrer\\Rzip\\data")  
save(Sepped.v1, Sepped.v2, file="Bamboo\_Seperated.RData")