Scraping Text Data from a Newspaper

# 기사 리스트의 url 가져오기

library(stringr)

## Warning: package 'stringr' was built under R version 3.3.3

baseurl = "http://news.donga.com/List/Economy"   
  
###1 기사 목록 페이지에서 특정 기사의 페이지를 불러와 내용 수집하기 ####  
  
# 기사 목록 페이지 불러오기  
read.baseurl = readLines(baseurl, warn = FALSE, encoding = "UTF-8")  
  
# 기사들의 URL 추출  
rawlinks = read.baseurl[str\_detect(read.baseurl, "<div class='articleList'>")]  
  
links = unique(unlist(str\_extract\_all(rawlinks, "http://news.donga.com/List/Economy/[:digit:]/[:digit:]{2}/[:digit:]{8}/[:digit:]{8}/[:digit:]")))  
  
head(links)

## [1] "http://news.donga.com/List/Economy/3/01/20180409/89527854/2"  
## [2] "http://news.donga.com/List/Economy/3/01/20180409/89527585/1"  
## [3] "http://news.donga.com/List/Economy/3/01/20180409/89527457/1"  
## [4] "http://news.donga.com/List/Economy/3/01/20180409/89527447/1"  
## [5] "http://news.donga.com/List/Economy/3/01/20180409/89527278/2"  
## [6] "http://news.donga.com/List/Economy/3/01/20180409/89527040/1"

# 첫번째 기사의 내용만 가져오기

# 첫번째 기사 읽어오기  
ex.article = links[1]  
read.article = readLines(ex.article, warn = FALSE, encoding = "UTF-8")  
head(read.article)

## [1] "<!doctype html>"   
## [2] "<html lang=\"ko\">"   
## [3] "<head>"   
## [4] "<title>원스톱 오피스텔 ‘대농지구 롯데캐슬 시티’</title>"   
## [5] "<meta http-equiv='Content-Type' content='text/html; charset=utf-8'>"  
## [6] "<meta http-equiv=\"X-UA-Compatible\" content=\"IE=edge\">"

# 특정 기사의 소스에서 기사 내용 부분만 수집  
raw.article = c()  
j = 1  
for(i in 1:length(read.article)) {  
 first = str\_extract(read.article[i], "<div class=\"article\_txt\">")  
 if (!is.na(first) == TRUE) {  
 while(TRUE) {  
 raw.article[j] = read.article[i]  
 j = j + 1  
 i = i + 1  
 if (str\_detect(read.article[i], "<div class='recommend'>")) break  
 }  
 }  
}  
  
raw.article = raw.article[str\_detect(raw.article, "[가-힣]")]  
head(raw.article)

## character(0)

# html 태그 등 제거   
if (is.na(raw.article[1]) == TRUE) { article = NA   
} else {  
 article = raw.article[1]  
 i=2  
 while(TRUE) {  
 if (is.na(raw.article[2]) == TRUE) break  
 article = str\_c(article, raw.article[i])  
 i = i + 1  
 if (is.na(raw.article[i]) == TRUE) break  
 }  
 article = str\_replace\_all(article, "<.+?>", " ")  
}  
  
article

## [1] NA

# 모든 기사의 내용만 자동적으로 가져오기

# 경제 관련 기사 목록 페이지 1쪽부터 10쪽까지의 URL을 저장  
url.lists = list()  
for(i in 1:10) {  
 url.lists[[i]] = paste("http://news.donga.com/List/Economy?p=", (i-1)\*16+1,   
 sep = "")  
}  
head(url.lists)

## [[1]]  
## [1] "http://news.donga.com/List/Economy?p=1"  
##   
## [[2]]  
## [1] "http://news.donga.com/List/Economy?p=17"  
##   
## [[3]]  
## [1] "http://news.donga.com/List/Economy?p=33"  
##   
## [[4]]  
## [1] "http://news.donga.com/List/Economy?p=49"  
##   
## [[5]]  
## [1] "http://news.donga.com/List/Economy?p=65"  
##   
## [[6]]  
## [1] "http://news.donga.com/List/Economy?p=81"

# 저장된 10개의 URL 읽어들이기  
read.url.lists = lapply(url.lists, readLines,  
 warn = FALSE, encoding = "UTF-8")  
  
# extract.Links 함수: 저장된 10개의 URL로부터 각각의 기사들의 URL을 불러옴  
# 뉴스 기사 URL 형식: http://news.donga.com/List/Economy/숫자1자리/숫자2리/숫자8자리/숫자8자리/숫자1자리  
extract.Links = function(read.url) {  
 a.list = read.url[str\_detect(read.url, "<div class='articleList'>")]  
   
 article.links = unique(unlist(str\_extract\_all(a.list,   
 "http://news.donga.com/List/Economy/[:digit:]/[:digit:]{2}/[:digit:]{8}/[:digit:]{8}/[:digit:]")))  
 return(article.links)  
}  
  
articlelinks = unlist(lapply(read.url.lists, extract.Links))  
head(articlelinks)

## [1] "http://news.donga.com/List/Economy/3/01/20180409/89527854/2"  
## [2] "http://news.donga.com/List/Economy/3/01/20180409/89527585/1"  
## [3] "http://news.donga.com/List/Economy/3/01/20180409/89527457/1"  
## [4] "http://news.donga.com/List/Economy/3/01/20180409/89527447/1"  
## [5] "http://news.donga.com/List/Economy/3/01/20180409/89527278/2"  
## [6] "http://news.donga.com/List/Economy/3/01/20180409/89527040/1"

# 각각의 기사 URL에함수 extract.Links 적용  
read.articlelinks = lapply(as.list(articlelinks), readLines, warn = FALSE,  
 encoding = "UTF-8")  
  
  
# extract.Articles 함수: 읽어들인 기사 소스로부터 기사 내용에 해당하는 부분만 추출   
extract.Articles = function(read.url) {  
 raw.articles = c()  
 j = 1  
 for(i in 1:length(read.url)) {  
 first = str\_extract(read.url[i], "<div class=\"article\_txt\">")  
 if (!is.na(first) == TRUE) {  
 while(TRUE) {  
 raw.articles[j] = read.url[i]  
 j = j + 1  
 i = i + 1  
 if (str\_detect(read.url[i], "<div class='recommend'>")) break  
 }  
 }  
 }  
 raw.articles = raw.articles[str\_detect(raw.articles, "[가-힣]")]  
   
 return(raw.articles)  
}  
  
# 읽어들인 기사 자료에 함수 extract.Articles 적용  
rawarticle.lists = lapply(read.articlelinks, extract.Articles)  
rawarticle.lists[[1]]

## character(0)

unlist(lapply(rawarticle.lists, length))

## [1] 0 2 2 0 4 2 2 4 4 0 0 0 0 0 0 0 0 0 2 0 0 0 2 0 1 1 1 4 0 0 0 0 2 0 0  
## [36] 0 0 2 1 2 0 2 1 2 0 0 0 0 4 2 2 0 1 0 1 1 1 0 0 2 1 0 0 2 2 4 0 4 4 4  
## [71] 4 4 4 4 2 6 6 5 6 4 6 5 6 4 8 2 1 1 3 2 2 0 0 0 5 1 0 0 3 1 0 0 3 1 1  
## [106] 0 0 0 1 1 2 0 0 0 3 1 1 3 3 0 1 3 3 0 0 1 1 1 0 0 0 4 1 0 0 0 0 1 4 4  
## [141] 0 1 4 4 4 1 1 2 1 0 0 1 1 0 0 0 1 0 1 0 1 0 1 0 2 1 0 1 0 0 2 0 1 1 2  
## [176] 0 1 2 2 0 1 2 2 0 0 0 2 1 0 1 2 1 1 2 2 2 1 6 4 5

# filter.Article 함수: 추출한 기사 내용에서 html 태그 등 제외   
filter.Article = function(x) {  
 if (is.na(x[1]) == TRUE) { Articles = NA }  
 else {  
 Articles = x[1]  
 i=2  
 while(TRUE) {  
 if (is.na(x[2]) == TRUE) break  
 Articles = str\_c(Articles, x[i])  
 i = i + 1  
 if (is.na(x[i]) == TRUE) break  
 }  
 Articles = str\_replace\_all(Articles, "<.+?>", " ")  
 }  
 return(Articles)  
}  
  
# 함수 filter.Article 적용   
article.lists = lapply(rawarticle.lists, filter.Article)  
str(article.lists)

## List of 200  
## $ : logi NA  
## $ : chr " <U+FEFF>최대쟁점은 유령주식 ‘재물’ 인정 여부일 듯 삼성증권 “대내외 수단 고려해 엄중 문책할 것” ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>조합측 “조합원 의견 수렴 후 수의계약 방법 결정” 문정동 136번지·대치쌍용2차 시공사 선정에도 영향 ""| \_\_truncated\_\_  
## $ : logi NA  
## $ : chr " 숙성 커피열매를 사용한 ‘조지아 고티카 빈티지’ 신제품 3종 출시 사진제공=조지아 고티카 코카-콜라사의"| \_\_truncated\_\_  
## $ : chr " <U+FEFF>野 “金, 적폐의 전형…靑 ‘김기식 지키기’ 눈물겹다” 與 “무리한 정치공세 묵과하지 않겠다” ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>배당 사고 파장 업계로 확대…종합검사도 부활 과잉 규제시 시장 위축…‘제2 벤처붐’에도 영향 ""| \_\_truncated\_\_  
## $ : chr " GS건설이 올해 1분기(1~3월) 흑자전환에 성공했다. GS건설은 공시를 통해 올해 1분기 법인세비용 차감 전 사업"| \_\_truncated\_\_  
## $ : chr " ㈜트라이크람스코리아 강찬고 대표(왼쪽)와 덕태상무비서유한공사 하진 대표가 협약서를 들고 기념촬영을 하고 있다. "| \_\_truncated\_\_  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : chr " <U+FEFF>농협·국민·하나은행…“가이드라인 준수 여부 확인” “법인계좌 자금 이동 적정성 집중 모니터링” ""| \_\_truncated\_\_  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : chr " <U+FEFF>농협·국민·하나은행…“가이드라인 준수 여부 확인” “법인계좌 자금 이동 적정성 집중 모니터링” ""| \_\_truncated\_\_  
## $ : logi NA  
## $ : chr " <U+FEFF>2005년 이후 13년만…국내 첫 타이틀 도전 ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>배우자 함께…文대통령 꽃다발 증정 ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>더스틴 존슨 59주째 1위…스피스 5위·파울러 6위 ""| \_\_truncated\_\_  
## $ : chr " ▲ 팩토리얼 ‘이지더블유’ 계절 관계없이 연이은 미세먼지 주의보에 실외 뿐 아니라 실내로 유입되는 유해물"| \_\_truncated\_\_  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : chr " <U+FEFF>6일 사고 이후 7·8일 회사 홈피에 사과문만 올려 직원 일탈·시스템 미흡…靑 국민청원 17만 넘어 ""| \_\_truncated\_\_  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : chr " <U+FEFF>이동걸 “제시한 자구안이 최선…미제출시 법정관리” 채권단 관계자들 창원 상주하며 노사확약서 제출 독려 ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>영남 지역에서는 소주 16.9도 일반적…“전 세계적 저도주 열풍” ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>민생사법경찰단, 불법중개업소 통신·금융 내역 등 확보 “시장교란 행위, 형사처벌·행정처분 이중처벌로 엄벌할"| \_\_truncated\_\_  
## $ : logi NA  
## $ : chr " <U+FEFF>이동걸 “제시한 자구안이 최선…미제출시 법정관리” 채권단 관계자들 창원 상주하며 노사확약서 제출 독려 ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>영남 지역에서는 소주 16.9도 일반적…“전 세계적 저도주 열풍” ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>민생사법경찰단, 불법중개업소 통신·금융 내역 등 확보 “시장교란 행위, 형사처벌·행정처분 이중처벌로 엄벌할"| \_\_truncated\_\_  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : chr " 사진제공=뽕뜨락피자 피자는 납작한 빵을 의미한 ‘삐따(Pitta)에 어원을 두고 있고, 이탈리아 남부 도시들이 "| \_\_truncated\_\_  
## $ : chr " <U+FEFF>9~19일 특별점검·현장검사…구성훈 대표 면담 “다른 상장 증권사 사고 우려 有…시스템 개선” ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>주식 착오 입고 1분 뒤인 오전 9시31분에 인지 오전 10시8분에야 실제 주문차단 조치 이뤄져 ""| \_\_truncated\_\_  
## $ : logi NA  
## $ : chr " <U+FEFF>교통부, 자동차관리법 시행규칙 개정안 입법예고 ""| \_\_truncated\_\_  
## $ : logi NA  
## $ : chr " <U+FEFF>“무차입 공매도 있었는지 점검할 것” ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>마지막 날 3언더파 추가, 최종합계 1오버파 공동 32위 ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>“소송충당금, 과징금 등 간접 비용 발생할 듯” ""| \_\_truncated\_\_  
## $ : logi NA  
## $ : logi NA  
## $ : chr " <U+FEFF>“국민 눈높이에 부적절 비판 수용…죄송” 거듭 사과 “삼성선 직원 실수라지만 그보다 심각” 대응案 발표 ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>“소송충당금, 과징금 등 간접 비용 발생할 듯” ""| \_\_truncated\_\_  
## $ : logi NA  
## $ : logi NA  
## $ : chr " <U+FEFF>“국민 눈높이에 부적절 비판 수용…죄송” 거듭 사과 “삼성선 직원 실수라지만 그보다 심각” 대응案 발표 ""| \_\_truncated\_\_  
## $ : chr " <U+FEFF>파울러-스피스 마지막 날 맹추격 펼치며 2-3위…매킬로이는 공동 5위 김시우 공동 24위 선전…‘3년만의 출전’ "| \_\_truncated\_\_  
## $ : chr " 인터파크와 롯데카드가 컬처 프로젝트 ‘May I Love You?’를 진행한다. 22일까지 총 500명을 선정해 서울 "| \_\_truncated\_\_  
## $ : logi NA  
## $ : chr " 요금제에 따라 600∼7300원 보상 총보상액 200억∼300억원 추정 SK텔레콤이 6일 발생한 LTE 서비스 장애로 피"| \_\_truncated\_\_  
## $ : chr " 삼성증권 사태 관련 청와대 게시판에 올라온 청원글. 사진출처｜청와대 국민청원 게시판 캡처 구성훈 대표 “도덕"| \_\_truncated\_\_  
## $ : chr " BMW M3는 세단의 모습을 지녔지만 내부에는 6기통 터보차저 엔진으로 대표되는 첨단 고성능차의 특성을 두루 갖추고 "| \_\_truncated\_\_  
## $ : chr " 인피니티의 프리미엄 스포츠 쿠페 ‘뉴 Q60’. 사진제공｜인피니티 인피니티 코리아가 프리미엄 스포츠 쿠페 인"| \_\_truncated\_\_  
## $ : chr " ▲ 오늘의 운세 도표. 스포츠동아 ■ 오늘의 운세｜박경호의 운수 좋은 날 도원역리학회 도원학당 원장 [스"| \_\_truncated\_\_  
## $ : chr " 14일 개그맨 윤성호 김진철 등 출연 ‘서울 광진구와 함께하는 개그테크노 파티(힘내라! 소상공인 개그콘서트"| \_\_truncated\_\_  
## $ : chr " 사진제공｜한화 갤러리아 서울 압구정 갤러리아명품관은 19일까지 이탈리아 패션과 라이프스타일을 소개하는 ‘"| \_\_truncated\_\_  
## $ : chr " 관영 환추시보 “美 쳐부술것” 美-中 항모 남중국해 맞불 훈련 "  
## $ : chr " 한국경제硏 5개 업종 고용 분석 화장품업 허가제→등록제 완화… 업체수 5배, 일자리 2배로 늘어 “재정 부담 주지않는 최"| \_\_truncated\_\_  
## $ : chr " [규제 완화가 일자리 해법]2015년 5위로 떨어져… 中은 3위 기업규제-노동시장 경직성 원인… 자동차-철강 등 주력산업 위"| \_\_truncated\_\_  
## $ : chr " [규제 완화가 일자리 해법]한경연, 5개 업종 효과 분석 크게보기 한국경제연구원이 8일 ‘규제 "| \_\_truncated\_\_  
## $ : chr " [여권, 워싱턴 한미연구소 압박 논란]김기식 보좌관 시절 정무위 활동 현재 장하성 정책실장 밀착수행 靑 “KIEP서 먼저 "| \_\_truncated\_\_  
## $ : chr " [여권, 워싱턴 한미연구소 압박 논란]한미硏 예산지원 중단 배경은 청와대가 미국 존스홉킨스대 한"| \_\_truncated\_\_  
## $ : chr " [규제 완화가 일자리 해법]2015년 5위로 떨어져… 中은 3위 기업규제-노동시장 경직성 원인… 자동차-철강 등 주력산업 위"| \_\_truncated\_\_  
## $ : chr " [규제 완화가 일자리 해법]한경연, 5개 업종 효과 분석 크게보기 한국경제연구원이 8일 ‘규제 "| \_\_truncated\_\_  
## $ : chr " [여권, 워싱턴 한미연구소 압박 논란]김기식 보좌관 시절 정무위 활동 현재 장하성 정책실장 밀착수행 靑 “KIEP서 먼저 "| \_\_truncated\_\_  
## $ : chr " [여권, 워싱턴 한미연구소 압박 논란]한미硏 예산지원 중단 배경은 청와대가 미국 존스홉킨스대 한"| \_\_truncated\_\_  
## $ : chr " 김기식, 외유성 해외출장 논란 사과 “해당기관에 혜택 준적 없어” 해명… 女비서 동행 논란엔 “정책 담당” 靑 “임명"| \_\_truncated\_\_  
## $ : chr " 교촌치킨 배달료 명목 2000원↑… 도미노피자-CGV 1000원 올려 택시비 등 공공요금도 ‘들썩’ "  
## $ : chr " 주식 배당오류 사태 후폭풍 "  
## $ : chr " 8조원 투입했는데 생사 기로에 "  
## $ : chr " 빅데이터 시대 변화하는 의료계 대형병원 중심 ‘의료인 창업’ 급증… VR 게임으로 시야장애 치료, 인공지능 활용 신약 "| \_\_truncated\_\_  
## $ : chr " 36홀로 축소된 롯데렌터카오픈… 4언더 추가, 9언더 시즌 첫 승 LPGA 2연속 컷탈락 부진 씻어 "  
## $ : chr " 3R 4오버 공동40위, 미흡해도 4R 티켓 최고 235만원 치솟아 김시우도 공동21위까지 순위 올려 "  
## $ : logi NA  
## $ : logi NA  
## $ : logi NA  
## $ : chr " 4월 중순 美서 정상회담때 납치자-철강관세 해결 주목 北은 日에 “납치문제 끝났다” 도널드 트럼프 미"| \_\_truncated\_\_  
## $ : chr " 옛 공산국가 ‘新독재 포퓰리즘’ "  
## $ : logi NA  
## $ : logi NA  
## $ : chr " 3월 한국에 ‘포르토피노’ 내놓은 페라리 넥텔 지사장 오스트리아 출신의 디터 넥텔 페라리 극동 및 중"| \_\_truncated\_\_  
## [list output truncated]

print(article.lists[[1]])

## [1] NA

# 연습문제

<http://news.donga.com/에서> 정치와 경제에서 최근 기사 100개씩 내려받아 기사의 url, 제목, 내용, 그리고 기사의 분류(정치, 경제) 코드를 저장하시오. 데이터를 랜덤하게 7:3으로 훈련 및 검증 데이터로 나누고 단순베이즈 분류와 k-NN을 이용하여 텍스트 내의 단어의 빈도를 이용하여 분류해보고 정확도와 수행시간등 종합적으로 성능을 평가하시오.