인문학 텍스트 마이닝

Crawling the text with R

- 네이버 고려사
  - http://terms.naver.com/list.nhn?cid=49629&categoryId=49629

국역 고려사

역사기록물 > 국역 고려사 > **국역 고려사: 세가** 584건

국역 고려사: 세가 | 2008. 8. 30. 책보러가기 >

동아대학교 석당학술원 | 경인문화사

국역 고려사 세가를 세트로 엮은『국역 고려사 세가』세트. 전12권으로 구성되어 있다. 자세히보기

더보기

- 태조 원년(918) 무인년

- http://terms.naver.com/entry.nhn?docId=1623779&categoryId=49629&cid=49629

국역 고려사

국역 고려사 : 세가

태조 원년(918) 무인년

• **원년 여름 6월**

병진일. 태조가 포정전(布政殿) [1]에서 즉위하여 국호를 고려(高麗) [2]라 하고 연호를 고쳐 천수(天授)라고 했다.
정사일. 다음과 같은 조서를 내렸다.

- 메모리정리하기

```
> gc()
          used (Mb) gc trigger  (Mb) max used  (Mb)
Ncells 1215028 64.9    2164898 115.7  2164898 115.7
Vcells 2414921 18.5    4701432  35.9  4648477  35.5
> rm(list=ls())
```

- 경로지정하기

```
> getwd()
[1] "/Users/Seongmin_M/Desktop/Class"
> setwd("/Users/Seongmin_M/Desktop/Class")
```

- rvest설치 및 불러오기

```
> install.packages("rvest")
trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.4/rvest_0.3.2.tgz'
Content type 'application/x-gzip' length 852813 bytes (832 KB)
==================================================
downloaded 832 KB

tar: Failed to set default locale

The downloaded binary packages are in
        /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//Rtmp5ZIez1/downloaded_packages
> library(rvest)
Loading required package: xml2
```

- url설정하기

```
> url <- "http://terms.naver.com/entry.nhn?docId=1623779&categoryId=49629&cid=49629"
```

- 사이트 html읽어오기

```
> Gorea <- read_html(url)
> Gorea
{xml_document}
<html lang="ko">
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=utf-8">\n<meta http-equiv="Content-Scri ...
[2] <body id="termBody" class="end">\n<script type="text/javascript">\n\tdocument.domain = "naver.com";\n</scri ...
```

- 사이트 title불러오기

```
> GoreaTitle <- html_nodes(Gorea, css='.headword_title')
>
> head(GoreaTitle)
{xml_nodeset (1)}
[1] <div class="headword_title">\n\t\t\t\t<p class="cite"><a href="list.nhn?categoryId=49629&amp;so=st4.asc" on ...
>
> str(GoreaTitle)
List of 1
 $ :List of 2
  ..$ node:<externalptr>
  ..$ doc :<externalptr>
  ..- attr(*, "class")= chr "xml_node"
 - attr(*, "class")= chr "xml_nodeset"
>
> GoreaTitle[1]
{xml_nodeset (1)}
[1] <div class="headword_title">\n\t\t\t\t<p class="cite"><a href="list.nhn?categoryId=49629&amp;so=st4.asc" on ...
```

- 추출된 Title 가져오기

```
> GoreaTitle[1] %>% html_nodes('h2') %>% html_text()
[1] "태조 원년(918) 무인년"
>
> pagetitle<-GoreaTitle[1] %>% html_nodes('h2') %>% html_text()
>
> pagetitle
[1] "태조 원년(918) 무인년"
```

- 텍스트 크롤링하기

```
> GoreaInfos <- html_nodes(Gorea, css='.size_ct_v2')
>
> head(GoreaInfos)
{xml_nodeset (1)}
[1] <div id="size_ct" class="size_ct_v2">\n\t\t\t\t\t<script type="text/javascript" src="https://audioapi.nmv.nav ...
>
> str(GoreaInfos)
List of 1
 $ :List of 2
  ..$ node:<externalptr>
  ..$ doc :<externalptr>
  ..- attr(*, "class")= chr "xml_node"
 - attr(*, "class")= chr "xml_nodeset"
>
> GoreaInfos[1]
{xml_nodeset (1)}
[1] <div id="size_ct" class="size_ct_v2">\n\t\t\t\t\t<script type="text/javascript" src="https://audioapi.nmv.nav ...
```

- 텍스트 크롤링하기

```
>
> pagetext<-GoreaInfos[1] %>% html_nodes('.txt') %>% html_text()
>
> pagetext
```

[1] "● 원년 여름 6월병진일. 태조가 포정전(布政殿)1)에서 즉위하여 국호를 고려(高麗)2)라 하고 연호를 고쳐 천수(天授)라고 했다.정사일. 다음과 같은 조서를 내렸다.“전 임금은 사군(四郡)3)이 흙 무너지듯 붕괴할 때에 도적의 무리들을 제거하고 점차로 영토를 넓혀갔다. 그러나 천하를 아우르기도 전에 갑자기 잔혹한 폭정으로 백성들을 다스렸으며 간사함을 가장 옳은 것으로 여기고 위협과 모욕을 가하는 것을 주된 통치수단으로 삼았다. 요역과 부세가 번거롭고 과중하여 인구는 줄어들고 농토는 텅 비게 되었다. 그런데도 오히려 궁실만은 크고 으리으리하며 옛 제도를 준수하지 않고 힘든 부역은 그칠 날이 없으니 결국 원망과 비난이 일어나게 된 것이다. 더군다나 함부로 연호를 정하고 황제를 칭했으며 처자를 살육한 죄는 천지간에 용납되지 못할 일이며 귀신과 사람이 함께 노할 일로서 왕업의 기반을 송두리째 추락시켰으니 어찌 경계하지 않으랴? 짐은 공들이 추대하는 마음에 힘입어 가장 높은 자리에 올랐으니 낡은 풍속을 고쳐 모든 것을 다함께 새롭게 만들려 한다. 마땅히 법도와 규범을 혁신하는 길4)을 쫓을 것이며 가까운 데서 얻는 원칙[伐柯之則5)]을 감계로 삼으리라. 임금과 신하는 물과 물고기처럼 서로 어울려 즐거움[魚水之歡6)]을 같이 할 것이며 온 천하는 태평시대의 경사[晏淸之慶7)]를 함께 누릴지니 나라의 모든 백성들은 다 짐의 뜻을 잘 알도록 하라.”이에 신하들이 절을 올리고 사례했다.“전 임금의 통치 기간에는 선량한 사람들이 악독한 피해를 입고 죄 없는 사람들이 잔혹한 학대를 받는 통에 남녀노소가 모두 불만에 싸여 원한을 품지 않은 이가 없었습니다. 이제 다행히 목숨을 보전하여 성스럽고 현명한 임금을 만날 수 있게 되었으니 어찌 힘을 다하여 성은에 보답하지 않겠습니까?”무오일. 왕이 한찬(韓粲) 총일(聰逸)에게 지시했다.“전 임금이 참소를 믿고 함부로 사람을 죽였는데, 경의 고향 청주(靑州)는 땅이 기름지고 호걸이 많았기 때문에 변란을 일으킬까 우려한 나머지 그 곳 사람들의 씨를 말리려 했다. 그리고는 군인 윤전(尹全)과 애견(愛堅) 등 80여 명을 소환했는데 이들은 아무 죄가 없는데도 형구를 찬 채 끌려오고 있으니 경은 빨리 가서 그들8)을 고향으로 돌려보내도록 하라.”경신일. 마군장군(馬軍將軍) 환선길(桓宣吉)이 역모를 꾀하다가 처형당했다.신유일. 다음과 같은 조서를 내렸다.“관직을 설치하고 직책을 분담하는 일에는 유능한 사람을 임명하는 것이 중요하며, 세상을 이롭게 하고 백성을 평안하게 하는 일에는 어진 이를 가려 뽑는 것이 우선이다. 관리들이 직무에 소홀하지만 않는다면 정치가 문란해질 까닭이 없는 것이다. 짐이 외람되게도 천명[景命9)]을 받아 큰 계획을 통어하려니, 높은 지위를 차지하게 되면 편안하기 어렵고 재능이 부족함을 두려워해야 한다는 말이 새삼 떠오른다. 오직 사람의 재능을 제대로 알지 못하고 관리를 선발함에 실수가 많아 어진 사람을 누락시켰다는 탄식을 야기시키고 선비를 얻는 도리에 어긋남이 있을까 걱정이다. 자나 깨나 머릿속을 떠나지 않는 것은 오로지 이것뿐이다. 조정 안팎의 여러 신료들이 모두 그 임무를 잘 감당할 수 있으면 현재 훌륭한 치적을 이룩할 수 있을 뿐 아니라 후대의 칭송까지 받을 수 있는 것이다. 마땅히 관리[列壁10)]를 등용하고 사람들을 시험함에 있어 반드시 힘써 잘 가려 뽑아 적재적소에 배치해야 할 것이다. 온 나라 사람들은 모두 짐의 뜻을 헤아릴지어다.”이에 따라 한찬(韓粲) 김행도(金行濤)11)를 광평시중(廣評侍中)으로, 한찬 금강(黔剛)을 내봉령(內奉令)으로, 한찬 임명필(林明弼)12)을 순군부령(徇軍部令)으로, 파진찬(波珍粲) 임희(林曦)13)를 병부령(兵部令)으로, 소판(蘇判) 진원(陳原)을 창부령(倉部令)으로, 한찬 염장(閻萇)을 의형대령(義形臺令)으로, 한찬 귀평(歸評)을 도항사령(都航司令)으로, 한찬 손형(孫逈)을 물장성령(物藏省令)으로, 소판 진경(秦勁)을 내천부령(內泉部令)으로, 파진찬 진정(秦靖)을 진각성령(珍閣省令)으로 각각 임명하였다. 이들은 모두가 품성이 단정하고 일을 공정하고 성실하게 처리했으며, 개국 초창기부터 왕을 잘 보좌하여 공훈을 세운 사람들이었다.알찬(閼粲) 임적여(林積璵)를 광평시랑(廣評侍郎)으로, 전 수순군부경(守徇軍部卿) 능준(能駿)과 창부경(倉部卿) 권식(權寔)을 함께 내봉경(內奉卿)으로, 알찬 김인(金堙)과 영준(英俊)을 함께 병부경(兵部卿)으로, 알찬 최문(崔汶)14)과 견술(堅術)15)을 창부경으로, 일길찬(一吉粲) 박인원(朴仁遠)16)과 김언규(金言規)17)를 백서성경(白書省卿)으로, 임상난(林湘煖)을 도항사경(都航司卿)으로,

- 강조문구 크롤링하기

```
> GoreaInfos[1] %>% html_nodes('strong') %>% html_text()
 [1] "● 원년 여름 6월"  "병진일."          "정사일."          "무오일."          "경신일."
 [6] "신유일."          "임술일."          "계해일."          "을축일."          "무진일."
[11] "기사일."          "● 가을 7월"       "임신일."          "계사일."          "병신일."
[16] "● 8월"            "기유일."          "경술일."          "신해일."          "갑인일."
[21] "계해일."          "병인일."          "● 9월"            "을유일."          "경인일."
[26] "계사일."          "갑오일."          "을미일."          "병신일."          "정유일."
[31] "● 겨울 10월"      "경신일."          "신유일."          "● 11월"           "元年 夏六月 丙辰"
[36] "丁巳"             "戊午"             "庚申"             "辛酉"             "壬戌"
[41] "癸亥"             "乙丑"             "戊辰"             "己巳"             "秋七月 壬申"
[46] "癸巳"             "丙申"             "八月 己酉"        "庚戌"             "辛亥"
[51] "甲寅"             "癸亥"             "丙寅"             "九月 乙酉"        "庚寅"
[56] "癸巳"             "甲午"             "乙未"             "丙申"             "丁酉"
[61] "冬十月 庚申"      "辛酉"             "十一月"
```

- 저장하기

```
> #저장하기
> Goreafilename<-paste("/Users/Seongmin_M/Desktop/Class/Gorea/",pagetitle,".txt", sep = "")
> write.table(pagetext,file=Goreafilename,sep="\n")
```