텍스트 마이닝을 이용한 지능적 워드클라우드

김연창 · 지상수 · 박동서 · 이충호

한밭대학교

Intelligent Wordcloud Using Text Mining

Yeongchang Kim · Sangsu Ji · Dongseo Park · Choong Ho Lee

Hanbat National University

E-mail : wcfull@naver.com

요 약

본 논문은 텍스트 마이닝 기법으로 명사의 빈도수를 조사하여 워드클라우드를 나타내는 기존의 방법을 개선하여 지능적 워드클라우드를 구현하는 방법을 제안한다. 텍스트 마이닝 시에 명사 단어를 추출하는 사전에 누락된 신조어 등의 단어를 효과적으로 추가하고, 동사 등 다른 품사위주의 워드클라우드를 시각적으로 보여주는 방법을 제안한다. 실험에서 기존 명사의 빈도수 추출에는 KoNLP 패키지를 사용하였고, 지원되지 않는 신조어 80개를 추가하였고 빈도수를 수동으로 조사하여 추가하였다.

ABSTRACT

This paper proposes an intelligent word cloud by improving the existing method of representing word cloud by examining the frequency of nouns with text mining technique. In this paper, we propose a method to visually show word clouds focused on other parts, such as verbs, by effectively adding newly-coined words and the like to a dictionary that extracts noun words in text mining. In the experiment, the KoNLP package was used for extracting the frequency of existing nouns, and 80 new words that were not supported were added manually by examining frequency.

Keywords

Text Mining, Wordcloud, R Language

Ⅰ. 서 론

방대한 양의 텍스트를 분석하기 위하여 비구조적인 텍스트를 분석하는 방법으로 텍스트마이닝 기법이 각광을 받고 있다. 글자로 기록된 텍스트에서 중요한 단어는 빈도수에 비례한다는 가정 하에 빈도수를 조사하여 시각적인 형태로 나타내는 방법으로 워드클라우드가 최근 많이 사용되고 있다.

하지만 기존의 워드클라우드는 명사의 빈도수 위주로 되어 있어 중요한 단어가 동사나 형용사와 같은 단어에 포함되어 있는 경우에 그 중요도가 충분히 반영되지 않을 가능성이 있다. 또한 워드클라우드 생성 시에 참고하는 한글 패키지 KoNLP에 포함되어 있지 않은 신조어, 방언 등의 경우에는 워드클라우드에서 제외되어 버리는 단점이 있다.

본 논문에서는 기존 명사 위주의 워드클라우드를 동사나 다른 품사 위주로 보완하고 기존 패키지에 포함되어 있지 않은 신조어, 방언 등을 효과적으로 추가하는 방법을 제안한다.

Ⅱ. 텍스트마이닝과 기존 워드클라우드

텍스트 마이닝은 데이터 마이닝 방법과 정보 검색, 자연어 처리, 용어 및 정보추출과 같은 특징 추출, 문서 분류, 군집화, 연결 분석 등의 기법들이 결합된다. 단어 분류 또는 문장의 문법 구조 분석 등의 자연언어 처리 기술에 기반하고 있으며 문서 분류, 정보 추출, 문서 요약 등에 활용되고 있다.[1-2]

워드클라우드는 문서의 단어들을 분류하여 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심 단어를 시각적으로 돋보이게 하는 기법이다. 많이 언급될수록 단어를 크게 표현해 한눈에 들어올 수 있게 하는 기법 등이 있다. 주로 방대한 양의 정보를 다루는 빅데이터를 분석할 때 데이터의 특징을 도출하기 위해 활용한다.[1-2]

그림 1은 명사위주의 기존 워드클라우드를 보여준다. 기존의 방법에서 일반적으로 명사를 추출 시에 KoNLP 패키지를 사용하고 useSejongDic()이라는 명령어를 사용한다.[1]

그림 1. 기존의 워드 클라우드

텍스트에서 단어의 빈도수를 조사하고 워드클라우드로 표현하는 부분을 R언어로 구현하는 방법은 그림 2와 같다. [1]

```
> library(KoNLP)
> library(RColorBrewer)
> library(wordcloud)
> useSejongDic()
Backup was just finished!
370957 words dictionary was built.
> noun<- sapply(text,extractNoun,USE.NAMES = F)
> noun
> noun2 <- unlist(noun)
> noun2
> word_count <- table(noun2)
> word_count
noun2
                                14                              30
        98                       1                               1
         7                    가슴                            가지
         2                       2                               1
      감사                    감상                            감회
         2                       1                               1
      개발                    개방                            개척
         1                       1                               1
      거기                    거룩                              것
         1                       1                              26
```

그림 2. 한글 워드클라우드를 생성하기 위한 주요 소스코드

III. 신조어를 포함한 워드클라우드

현재까지 있는 KoNLP 한글 패키지에는 새롭게 추가된 신조어들이 거의 없다. 패키지에 포함되어 있지 않은 경우에는 조사가 붙어서 조사까지 포함되어 명사로 오분류된다. 이러한 경우를 개선하기 위하여 신조어를 수동으로 추가하여 워드클라우드로 나타낸 것은 그림 3과 같다. 이것은 신조어 80개를 추가하여 워드클라우드로 나타낸 것이다.

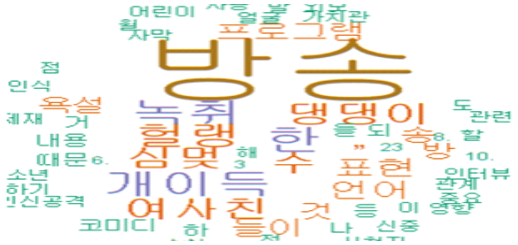

그림 3. 신조어를 포함한 워드클라우드

신조어를 수동으로 추가하는 방법[3]은 그림4와 같다.

```
> mergeUserDic(data.frame(c('여사친'),c('ncn')))
1 words were added to dic_user.txt.
```

```
찰스q.머피          ncn
티모시d.쿡          ncn
필립w.실러          ncn
여사친    ncn
```

그림 4. 신조어를 추가하는 소스코드 부분 및 KoNLP패키지 메모장 추가부분.

그러나 이것은 대량의 신조어나 방언을 추가하는 효과적인 방법이 아니므로 기존 KoNLP패키지에 수록되지 않은 단어들을 추출하여 .csv 파일[4]로 만들어서 한 번에 읽어 들이는 방법이 효과적이다. 따라서 다량의 한글 텍스트를 크롤링하여 기존 사전에 미수록된 단어를 추출하여 자동으로 .csv 파일을 생성하는 방법도 구현하여야 한다.

IV. 결론 및 향후계획

기존의 워드클라우드에 신조어를 추가하는 방법과 명사 위주의 기존의 워드클라우드를 개선하는 방법을 제안하였다. 현 단계에서 80개의 신조어를 추가하여 개선된 워드클라우드를 시험적으로 구현하였다. 향후 크롤링을 통하여 신조어를 효과적으로 추출하여 파일로 생성하는 방법을 구현할 계획이다. 또한 동사와 형용사 등 다른 품사 위주의 워드 클라우드도 구현할 계획이다.

References

[1] 장용식, 강희구, R로 배우는 코딩, 생능출판사, 2018. 01.
[2] 노규성, 김진화, 박성택, 김근원, 김도연, R과 Java로 크롤링하자, 생능출판사, 2017.02.
[3] 이영민, KoNLP를 이용한 한국어 형태소 분석, https://brunch.co.kr/, 2017.05.
[4] ___, https://ko.wikipedia.org/wiki/, 2019.05.