Journal of the Korean Data Analysis Society (April 2015)
Vol. 17, No. 2 (B), pp. 719-728

Stylometry를 이용한 영화 흥행 예측*

백광일[1], 김규곤[2], 최승배[3], 강창완[4]

요 약

최근 스마트폰 등 모바일 장치의 대중화로 인하여 시간과 장소에 구애받지 않고 인터넷에 접속할 수 있는 소셜 미디어 시대로 들어섰으며 이러한 환경에서 기업이나 국가는 소셜 미디어 분석을 통한 가치 있는 정보 획득을 최우선 과제로 설정하고 있다. 더욱이 요즘 들어 다양한 분야에서 빅데이터가 발생하고 있고 이러한 빅데이터를 활용한 성공사례가 증가하고 있는 상황에서 음악이나 영화 컨텐츠 산업과 같은 문화 컨텐츠 분야에서 빅데이터 분석 역시 흥미있는 주제로 자리잡고 있다. 본 연구에서는 소셜 미디어 분석에서 핵심적인 분석방법인 텍스트 마이닝을 소개하고 텍스트 마이닝 분석 방법 중 Stylometry 방법에 대한 고찰과 이를 이용한 영화 흥행 예측을 시도하고자 한다. 분석 도구로는 R 프로그램을 이용하였고 온라인 상의 영화 댓글 데이터 수집부터 흥행 예측을 위한 소셜네트워크 분석까지 실제 사용한 R 프로그램을 제시하였다. 사례분석에서는 Stylometry 기법을 이용하여 흥행영화인 “명량”을 기본 데이터로 하여 같은 시점에 동시 개봉한 두 개의 후보 영화(제보자, 슬로우 비디오) 중 상대적으로 더 흥행을 할 영화를 예측하였고 예측 결과는 “제보자”로 나타났다.

주요용어 : 소셜미디어, 텍스트 마이닝, Stylometry, R 프로그램.

1. 서론

신호감지 센서와 모바일 장치의 급성장으로 말미암아 빅데이터 기술이 산업 전반에 확산되면서 빅데이터 기술이 요즘의 최대 이슈가 되고 있다. 빅데이터의 의미가 기존의 처리 범위를 넘어서고 비정형화된 데이터까지 다루는 특성으로 인하여 숫자데이터 뿐만 아니라 문자데이터를 포함하고 있고 이로 인해 자연스럽게 문자데이터 분석 기술이 주목을 받게 되었다. 특히 트위터, 페이스북 등 온라인상의 소셜네트워크서비스(SNS)는 폭발적인 비정형데이터의 발생을 유도하였고 문자 데이터 분석을 통해 감추어진 그리고 가치 있는 정보를 추출하는 노력이 필요하게 되면서 텍스트 마이닝 기법이 대두되게 되었다. 텍스트 마이닝은 비/반정형 텍스트 데이터에서 자연어 처리 기술을 기반으로 유용한 정보를 추출, 가공하는 것을 목적으로 하는 기술이다. 이러한 텍스트 마이닝의 소개와 활용 연구는 Oh, Cho, Kang, Lim(2010), Oh, Jin(2012), Kim, Jin(2013), Kim(2012) 등을 참고할 수 있다. 본 연구에서는 텍스트 마이닝 활용 사례 중 흥행 영화 예측을 위한 텍스트 마이닝을 소개하고 이를 위한 분석을 무료로 제공되는 R 프로그램을 이용하여 분석하는 과정을 소개함으로써 텍

*이 논문은 2014학년도 동의대학교 교내연구비에 의해 연구되었음(2014AA483).
[1]614-714 부산광역시 부산진구 엄광로 176, 동의대학교 데이터정보학과 박사과정.
E-mail : hjang@koreascoring.com
[2]614-714 부산광역시 부산진구 엄광로 176, 동의대학교 데이터정보학과 교수. E-mail : kkkim@deu.ac.kr
[3]614-714 부산광역시 부산진구 엄광로 176, 동의대학교 데이터정보학과 교수. E-mail : csb4851@deu.ac.k
[4](교신저자) 614-714 부산광역시 부산진구 엄광로 176, 동의대학교 데이터정보학과 교수.
E-mail : cwkang@deu.ac.kr
[접수 2015년 1월 19일; 수정 2015년 4월 7일; 게재확정 2015년 4월 10일]

스트 자료 분석가들에 도움이 되고자 한다. 활용한 사례는 한국 영화사상 가장 큰 흥행을 기록한 '명량'의 댓글을 이용하여 10월 2일 동시 개봉한 두 개의 영화 중 좀 더 많은 흥행을 할 영화를 Stylometry라는 텍스트 마이닝 기법으로 예측하고자 한다.

2. 텍스트 마이닝과 Stylometry

2.1. Stylometry

Stylometry는 원작자를 찾기 위하여 문서를 분석하는 통계적 방법을 말하는데 역사적으로는 19세기 중엽 영국의 논리학자인 Augustus de Morgan이 단어 길이가 저자의 지표가 될 수 있다는 것을 주장한 것으로 기원을 찾을 수 있다(Ramyaa, Khaled, 2004). 그러나 실질적 관심은 1964년 미국의 통계학자인 Mosteller와 Wallace가 'The Federalist Papers'의 원작자 찾기 문제를 사용 단어의 빈도수로 해결하면서 비롯되었다. 현대의 Stylometry는 컴퓨터 계산능력을 이용한 서체(style)의 측도를 의미하며 장르의 분류, 언어 변화의 역사 연구, 문체 분석, 원저자 확인, 법언어학(forensic linguistics)에 활용되고 있다. 이러한 Stylometry는 기본적으로 단어나 혹은 문장의 길이를 대상으로 하며 문자 쌍의 빈도, 단어들의 분포 등을 분석하며 문서와 단어로 구성된 2차원 빈도 행렬에 대한 대응분석, 주성분 분석, 문서들 간의 군집분석 등을 이용하기도 한다(Holmes, Kardos, 2003).

2.2. R을 이용한 텍스트 마이닝

텍스트 마이닝은 비정형화된 문자데이터를 구조화된 데이터로 변환시키는 알고리즘을 적용하고 이를 분석하여 정보를 추출하는 전 과정을 의미한다. 전통적으로 데이터 마이닝과 유사하게 문서 군집화(document clustering)와 문서 분류(document classification)가 텍스트 마이닝의 주요 분석이라 할 수 있다(Kim, 2009) . 이러한 텍스트 마이닝 분석은 상용프로그램으로 SAS Text Miner와 SPSS Clementine 등이 알려져 있고 오픈 소스 통계분석 프로그램 R에서 패키지 <tm>과 <RcmdrPlugin.temis>을 이용할 수 있다(http://CRAN.R-project.org). 이에 대한 자세한 내용은 Feiner, Hornik, Meyer (2008)를 참고할 수 있다.

3. 사례분석

3.1. 데이터

본 연구에서 이용한 데이터는 포털사이트 다음(movie.daum.net)에서 2014. 7. 30일 개봉한 "명량", 그리고 10월 2일 동시 개봉한 "제보자"와 "슬로우 비디오"의 영화 댓글 데이터이다. 다음은 영화댓글을 수집하는 R 프로그램 명령문이고 실제 분석에 사용한 데이터는 해당 영화 댓글만을 발췌, 정제하였고 그 중 stylometry 분석을 하기 위한 기본 영화인 "명량"은 7월 30일을 기준으로 개봉 전 댓글문서(MR_b.txt), 개봉 당일 문서(MR_0730.txt) 이후 8월 4일까지 날짜별 댓글문서로 구분하여 분석하였다.

```
library(XML)
library(xtable)
library(tm)
GetDaumMovieData <- function(num){ url = gsub(" ","", paste("http ://movie.daum.net /review
  /netizen_point/movieNetizenPoint.do?type=after&page=",num))
```

```
  doc = htmlTreeParse(url, useInternalNodes = T, encoding ="UTF-8")
  xpathSApply(doc, "//div[@class='commentList']", xmlValue)
  movie_nm <- xpathSApply(doc, "//span[@class='movieTitle fs11']", xmlValue)
  score<-gsub("네티즌별점","",xpathSApply(doc, "//span[@class='star_small']", xmlValue))
  reg_date<-xpathSApply(doc, "//span[@class='date']", xmlValue)
  contents<-gsub("네티즌별점|\r|\t|\n","",xpathSApply(doc, "//span[@class='comment article']",
  xmlValue))
  contents<-removeNumbers(contents)
  contents<-removePunctuation(contents)
  contents<-stripWhitespace(contents)
  senti_data <-cbind(movie_nm,score,reg_date,contents)
  return(senti_data) }
DATA<-NULL
StartPage <- 1
EndPage <- 10
for (i in StartPage:EndPage) { getData<-GetDaumMovieData(i)
  DATA<-rbind(DATA,getData) }
write.csv(DATA,"MOVIE_DATA_00001_10000.csv")
```

R 프로그램 수행으로부터 수집된 데이터는 아래 Figure 1과 같은 형태이며 변수는 4개 변수로 일련번호, 영화평점, 날짜, 내용 순으로 이루어져 있다. 영화 "명량"인 경우 문서는 날짜별로 구분하여 총 7개 문서이고 총 댓글 수는 3084개로 분석대상으로 삼았다.

A1 fx 6822

	A	B	C	D
1	6822	9	2014.07.21	출정 직전의 고뇌하는 이순신과 목숨을 잃을 수도 있다는 사실에 주저하는 민초들의 심정이 후반부 열심
2	6821	10	2014.07.21	진짜 꼭 봐야할 영화에요 전투신도 스케일이 어마어마 해요
3	6820	9	2014.07.22	믿고 보십시요
4	6819	6	2014.07.22	류승룡 씨는 왜 메인포스터에 나왔는지 잘 모르겠을 정도였고 최민식 씨가 연기하신 이순신 장군님은 가
5	6818	10	2014.07.22	정말 최고 주연 조연 배우들 모두다 최고였다조반은 다소 지루한감이 없지않아 있었지만 끝까지 박진감
6	6817	10	2014.07.22	시사회가서 봤는데 보고나서 소름이 돋았습니다 완전 짱입니다 이순신장군의 지략이 정말 최고였고 싸
7	6816	7	2014.07.22	중후반부의 전투신을 위해 만들어진 영화 이순신과 주변 인물들간의 갈등을 보여주는 초반은 어수선하기
8	6815	8	2014.07.22	명량해전의 통쾌함보다 이순신 장군과 아들 이회의 대화장면은 영화의 핵심이라 할만해 보였습니다 드라
9	6814	9	2014.07.22	절체절명의 위기 주위의 근심과 두려움 만류를 뿌리치고 고뇌의 결단과 죽음을 무릅쓴 전략 수행과정을
10	6813	10	2014.07.22	명량을 만날수 있었던 건 천행입니다그 분들이 목숨 걸고 지켰던 이 나라를 제대로 지키지 못한 지금의
11	6812	7	2014.07.23	대체왜 이 영화엔 꼬랑지를 내기는가 전문가들ㅋㅋㅋ
12	6811	9	2014.07.24	시사 봤는데 개봉하면 또 볼려구 기다림명량
13	6810	10	2014.07.24	그냥 우라나라국민이면 꼭보세요 재미을 떠나서 관람하세요 영화보고 재미없다고 하는사람 봐서요 일본
14	6809	9	2014.07.24	재미있어요 구루지만 나을때 진짜 무서움 이순신 정말 대단 마지막에 노꾼들 하는말이 압권 우리 후세들
15	6808	8	2014.07.25	이순신 장군 시사회 보고 왔습니다 드라마 책속에 그려진 인물을 이 영화 한편에 다 담아내기에는 역시
16	6807	8	2014.07.25	성웅 이순신이 아닌 인간 이순신의 고뇌와 갈등 그리고 해상 액션이 꽤 볼만한 영화

Figure 1. Reviews of movie, Myeongryang

3.2. 명량 댓글 분석

이 절에서는 "명량" 댓글에 대한 텍스트 마이닝을 R 패키지 <RcmdrPlugin.temis>를 병행하여 분석하고 영화 댓글에서 어떠한 정보를 획득할 수 있는지 살펴보기로 한다. 먼저 "명량" 데이터는 문서-단어 빈도 분석을 기본으로 하되 가장 빈도가 많은 단어들의 시각적 분석, 댓글 문서들의 군집분석, 문서와 단어들의 대응분석을 시도하였다.

Figure 2는 가장 많이 댓글에 나타난 단어들의 빈도수에 비례하여 글자크기를 시각화한 워드클라우드(word cloud) 그림이다. 한편 날짜별 댓글 7개 문서들의 계보적 군집분석 결과는 Figure 3에 나타나 있다.

Figure 3은 Ward 방법을 이용하여 날짜별 댓글 간 계층적 군집분석 결과인데, 군집을 보면 유사하게 묶인 댓글은 개봉 다음날과 그 다음날로 가장 먼저 유사하였고 개봉 전 시사회 이후의 댓글은 가장 나중에 묶이는 것을 볼 수 있다.

Figure 2. Wordcloud of Myeongryang

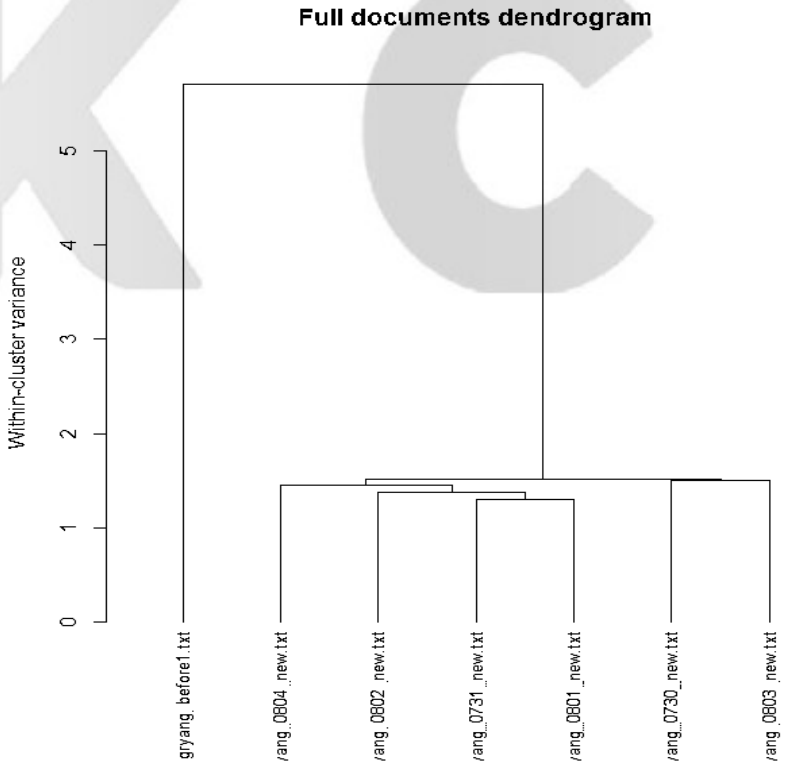

Figure 3. Hierarchical clustering

Figure 4는 댓글 문서-단어 행렬에 대하여 대응분석을 한 결과로서 '이순신', '감동', '인간' 이라는 단어와 개봉 전 댓글문서(MR_b1.txt)가 대응되는 것을 볼 수 있다.

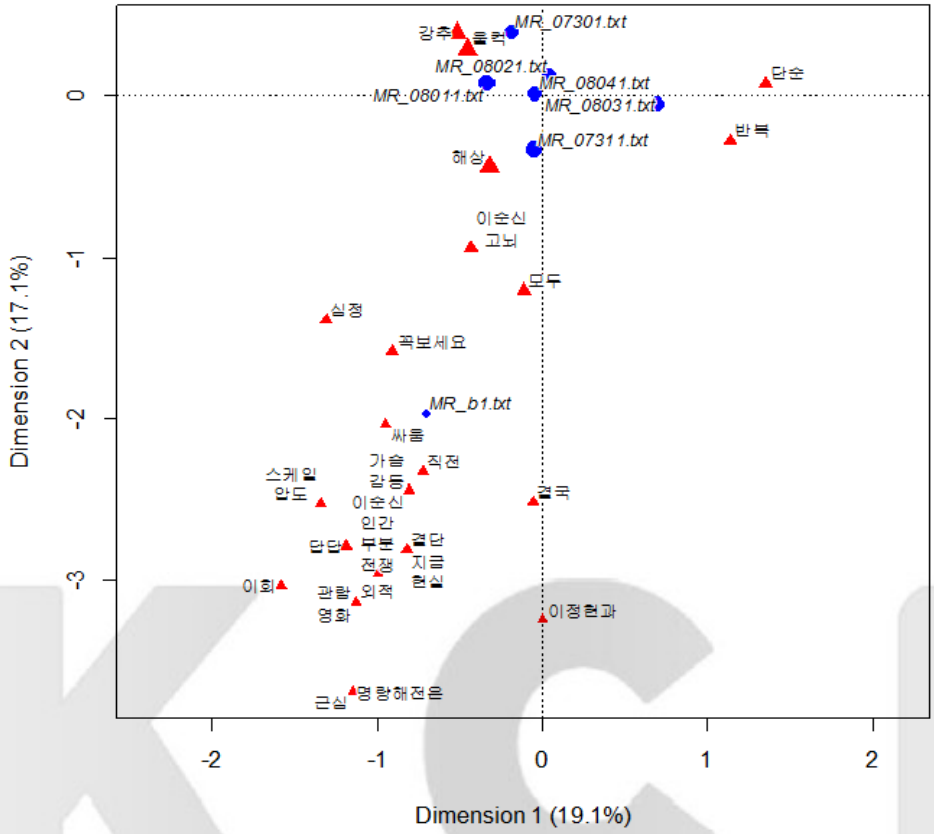

Figure 4. Correspondence analysis of Myeongryang review

3.3. 흥행 영화 예측

이 절에서는 흥행한 영화 "명량" 댓글을 기반으로 10월 2일 동시 개봉한 두 영화 "제보자"와 "슬로우 비디오" 중 어느 영화가 상대적으로 더 흥행할 것인지 stylometry 기법을 이용하여 예측하기로 한다. 분석과정과 R 스크립트는 다음과 같다.

단계 1) 먼저 "명량" 영화의 댓글과 "제보자" 및 "슬로우 비디오" 댓글의 단어 빈도행렬을 구한다.
단계 2) 세 영화의 빈도 행렬에서 일정 기준을 만족하는 빈번한 단어들을 구한다.
단계 3) 흥행한 기준 영화 "명량"의 빈번한 단어수와 일치하는 단어 빈도수를 구하고 비교한다.

아래는 위 분석과정에 해당하는 R 스크립트이다. 그리고 텍스트 마이닝에서의 대응분석과 군집분석을 위한 패키지 'RcmdrPlugin.temis'의 실행 화면은 Figure 5와 같다.

이에 대한 결과를 보면 "명량"과 "제보자"의 일치 단어 빈도 수는 66회, "명량"과 "슬로우 비디오"와의 일치 단어 빈도 수는 46회로 stylometry 방법에 의한 흥행 예측은 "제보자"로 나타났다. 한편 텍스트 마이닝을 위한 R commander 패키지를 이용하여 다양하게 단어들의 분석을 할 수 있는

```
library(tm)
movie <- Corpus(DirSource("c:/movie/"))
meta(movie, "Heading", "local") <- c("MR", "JBJ", "SV")
movieMR <- DocumentTermMatrix(movie[1])
movieJBJ <-DocumentTermMatrix(movie[2])
movieSV <-DocumentTermMatrix(movie[3])
MR <-findFreqTerms(movieMR, 1)
JBJ <- findFreqTerms(movieJBJ, 1)
SV <- findFreqTerms(movieSV, 1)
length(intersect(MR, JBJ))
length(intersect(MR, SV))
```

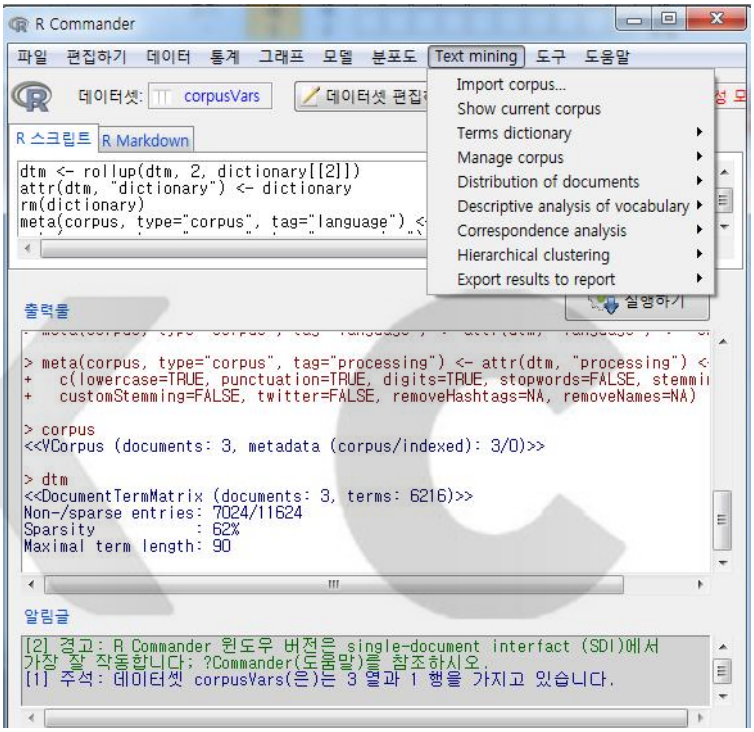

Figure 5. R commander for Text mining

데 가장 빈번한 주요 단어들의 분포의 동일성 검정을 시행한 결과가 Table 1에 나타나있다. 먼저 '명량' 댓글에서 '재미'라는 단어의 빈도수는 전체 댓글에서의 빈도수에 비해 유의수준 5%에서 더 극단적이지 못하다고 할 수 있으며, 반면 '슬로우 비디오' 댓글에서는 '재미'라는 단어가 유의수준 5%에서 더 극단적으로 발생하였음을 알 수 있다.

Table 2에서 가장 빈번한 단어들의 분포가 세 영화 간에 차이가 있다는 것이 통계적으로 유의하며 앞서 결과 마찬가지로 "명량"과 "제보자" 댓글이 유사함을 알 수 있다.

부가적으로 세 영화의 문서-단어간 대응분석 결과(Figure 6)를 보면 '몰입' 단어와 "명량"과 "제보자"가 대응되고 있음을 알 수 있고 또한, 세 영화 댓글간의 계층적 군집분석결과(Figure 7)도 "명량" 댓글과 "제보자" 댓글이 같은 군집으로 묶이는 것을 볼 수 있다.

Table 1. Comparison of level's term percent with global's term percent

		% term/Level	Global %	Level	Global	t value	p-value
MR	감동(impression)	1.83	1.83	100	213	-0.060	0.525
	연기(acting)	1.37	1.52	75	177	-1.150	0.125
	재미(fun)	0.66	0.80	36	93	-1.500	0.067
	몰입(immersion)	0.62	0.56	34	65	0.747	0.228
	기대(expectation)	0.53	0.66	29	77	-1.526	0.067
JBJ	감동(impression)	0.42	1.83	15	213	-8.520	<0.001
	연기(acting)	1.86	1.52	67	177	1.890	0.030
	재미(fun)	0.75	0.80	27	93	-0.280	0.389
	몰입(immersion)	0.83	0.56	30	65	2.450	0.007
	기대(expectation)	0.67	0.66	24	77	-0.070	0.529
SV	감동(impression)	3.82	1.83	98	213	7.740	<0.001
	연기(acting)	1.36	1.52	35	177	-0.650	0.259
	재미(fun)	1.17	0.80	30	93	2.170	0.015
	몰입(immersion)	-	0.56	1	65	-	-
	기대(expectation)	0.93	0.66	24	77	1.750	0.041

Number of terms : MR(5458), JBJ(3603), SV(3876), Corpus total(11629)
%Term/Level : the percent of the term's occurrences in all terms occurrences in the level.
Global % : the percent of the term's occurrences in all terms occurrences in the corpus.
Level : the number of occurrences of the term in the level ("internal").
Global : the number of occurrences of the term in the corpus.

Table 2. Summary of most frequent terms

	감동(impression)	연기(acting)	재미(fun)	몰입(immersion)	기대(expectation)
MR	100	75	36	34	29
JBJ	15	67	27	30	24
SV	98	35	30	1	24
Total	213	177	93	65	77
chi-square test : 94.97 (p < 0.01)					

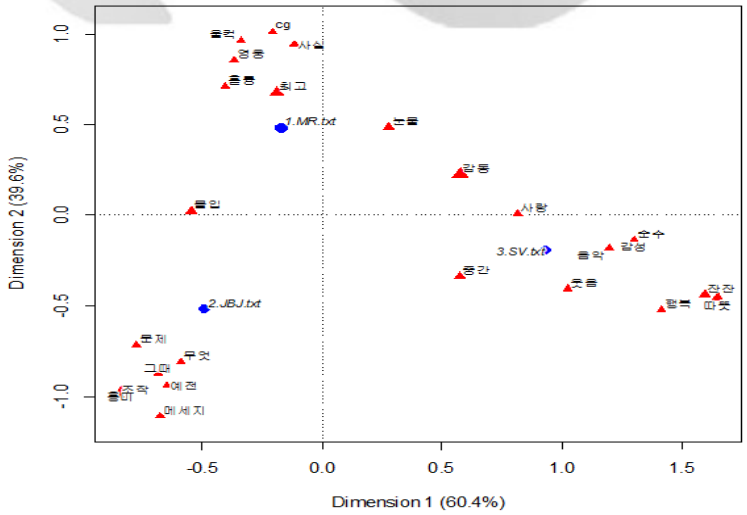

Figure 6. Correspondence analysis of 3 movie reviews

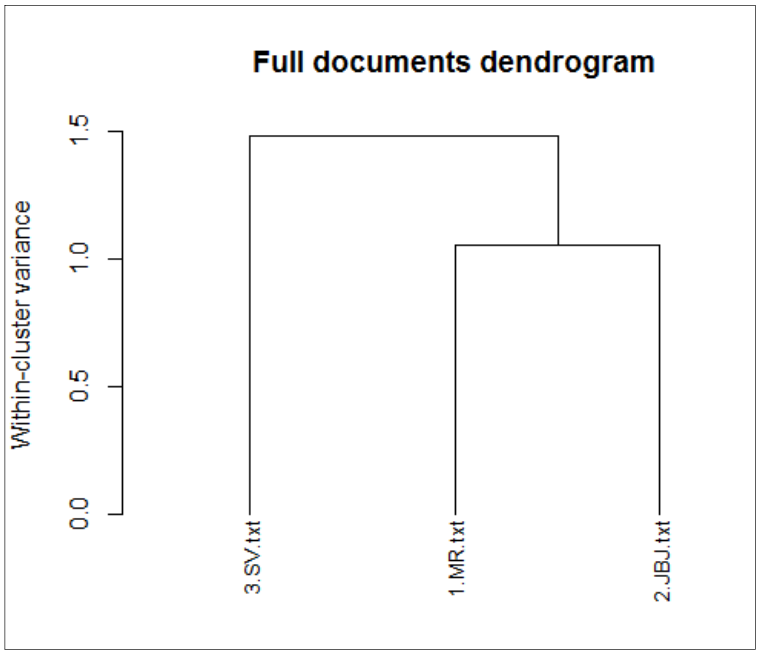

Figure 7. Clustering of 3 movie reviews

4. 결론

최근 빅데이터 시대에 들어서 수많은 데이터들이 만들어 졌다가 의미 없이 버려지는 와중에 가치 있는 정보를 추출할 수는 없는지가 주요 관심사가 되고 있다. 그 중 비정형화된 문서 데이터에 대한 분석방법으로 텍스트 마이닝은 빅데이터 분석의 중요한 도구로 인식되고 있고 이에 대한 연구가 점차 확대되고 있는 실정이다. 본 연구에서는 텍스트 마이닝 응용 기법 중 Stylometry라는 빈도 기반 분석 기법을 소개하고 최근 동시 개봉한 두 영화 “제보자”와 “슬로우 비디오”에 대하여 어느 영화가 더 흥행할 지를 이미 흥행한 영화 “명량” 댓글 문서를 기반으로 예측하였다. 예측 결과는 “제보자”가 “슬로우 비디오”보다 상대적으로 더 흥행할 것으로 예측되었고 최종 관객 수는 제보자가 1,755,181명이었고 “슬로우 비디오”는 1,169,546명으로 나타났다.

References

Bouchet-Valat, M., Bastin, G. (2013). *RcmdrPlugin.temis : Graphical integrated text mining solution,* URL http://CRAN.R-project.org/package=RcmdrPlugin.temis.

Feinerer, I., Hornik, K., Meyer, D. (2008). Text mining infrastructure in R, *Journal of Statistical Software,* 25, 5.

He, R. C., Rasheed, K. (2004). Using machine learning techniques for stylometry, *Proceedings of the International Conference on Artificial Intelligence,* 897-903.

Holmes, D., Kardos, J. (2003). Who was the author? An introduction to Stylometry, *Chance,* 16(2), 5-8.

Korea Creative Content Agency (2014). A case study of big data for contents area, *Culture and Technology*, 34, 50-63. (in Korean).

Kim, D. (2009). On the silhouette plot in cluster analysis, *Journal of the Korean Data Analysis Society*, 11(6), 2955-2964. (in Korean).

Kim, J., Jin, S. (2012). A study on application of opinion mining based on big data, *Journal of the Korean Data*

Analysis Society, 15(1), 101-113. (in Korean).

Oh, H., Cho, S., Kang, C., Lim, D. (2010). Fashion company's claim data analysis using text mining, *Journal of the Korean Data Analysis Society,* 12(1), 297-306. (in Korean).

Oh, S., Jin, S. (2012). A study on analysis of internet shopping mall customers' reviews by text mining, *Journal of the Korean Data Analysis Society,* 14(1), 125-138. (in Korean).

Prediction for the Films Success using Stylometry*

Gwangil Baek[1], *Kyu Kon Kim*[2], *Seung Bae Choi*[3], *Changwan Kang*[4]

Abstract

In recent years, another interesting application field of text mining techniques has been stylometry research. Textual stylometry deals with identifying the linguistic style of text documents. Typical research topics are the authorship identification problem, i.e., who wrote a specific text passage, or linguistic forensic tests. The advance of text mining techniques and computing power has led to a steady rise in usage of text mining for stylometry. Classical textual stylometry mainly deals with historical documents subject to unclear author-document correspondence. Examples are poems of Shakespeare, books of the Wizard of Oz, or the Federalist Papers. In this paper we introduce text mining in Stylometry using R and predict the movie success. In case study, we predicted that 'Jeboja' would be more successful than 'Slow Video' using Stylometry method.

Keywords : social media, text mining, Stylometry, R program.

*This work was supported by Dongeui University Grant(2014AA483).

[1]Doctoral Course, Department of Information Statistics, Dongeui University, Busan 614-714, Korea. E-mail : kibak@deu.ac.kr

[2]Professor, Department of Data Information Science, Dong-Eui University, Busan 614-714, Korea. E-mail : kkkim@deu.ac.kr

[3]Professor, Department of Data Information Science, Dong-Eui University, Busan 614-714, Korea. E-mail : csb4851@deu.ac.kr

[4](Corresponding Author) Professor, Department of Data Information Science, Dong-Eui University, Busan 614-714, Korea. E-mail : cwkang@deu.ac.kr

[Received 19 January 2015; Revised 7 April 2015; Accepted 10 April 2015]