

여백지스 오락 10점 100점
할크 **꿀잼** 액션 서울
할인 아이언맨 헐리웃
영화 킹스맨 히어로 마블 감독

  QuantLab
주식회사 퀀트랩



R을 이용한 텍스트 감정분석

여론과 감성 발견하기

김형준

Data Analyst / (주) 퀀트랩 / kim@mindscale.kr



발표자 소개

- 김형준 (kim@mindscale.kr)
- 서울대학교 인류학 / 심리학 학사
- 서울대학교 인지과학 석사

(현)

- (주)퀀트랩 Analytic Director
- [온오프라인 R 교육](#)
- 기업 데이터 분석 및 컨설팅

(전)

- 품질 / 클레임 / 인사 데이터 분석
- 홈페이지 및 서버 관리

회사 소개

퀀트랩 소개

- 2011년 설립
- 데이터 분석, 직무역량평가, 전문성 개발 전문 컨설팅 기업

members



유재명

서울대학교 산업공학과
서울대학교 인지과학 박사(수료)



황창주

서울대학교 심리학과
서울대학교 심리학 박사(수료)



김형준

서울대학교 인류학과 / 심리학과
서울대학교 인지과학 석사

clients

- LG생활건강
- LG U+
- NC소프트
- SK플래닛
- 중소기업진흥공단
- 이지웰페어
- 현대자동차

나에게 R이란?

1. 통계 프로그램 : 모형화 / 예측
2. 시각화 도구 : ggplot2 / Web과 연동
3. 발표 자료 도구 : slidify
4. 언어 처리 도구 : 텍스트 분석
5. Matlab / Python -> R

텍스트 분석

텍스트 분석 목적

: 사람들은 생각과 감정을 언어로 표현합니다. 뉴스 댓글, 상품평, 커뮤니티, SNS 등에 사람들이 남기는 텍스트를 모아 분석해보면 기존의 방법론으로 알기 어려웠던 여러 가지 정보를 얻을 수 있습니다.

감정 분석 목적

: 특정 키워드(이미지, 제품 등)에 대한 감정을 점수화하여 별도의 여론 조사 없이 감정의 정도를 예측할 수 있습니다. 또한, 감정의 이유를 분석하여 부정적인 요소를 개선할 수 있습니다.

통계 분석 목적

: 주어진 데이터를 통해 미래를 예측 + 통계 모형을 통해 현상을 설명

분석 예시 - Text

최초의 텍스트 분석

형태소 분석기

- 형태소 분석기 KLT2000 (강승식)

R

- wordcloud
- shiny

결과

- 신축 기숙사 공용 공간 확대
- 기존 기숙사 흡연 구역 재배정

사생들의 건의 사항 분석 (2013)

Dormitory Issue

Chose a dataset:

2013y

Numbers of Words to view:

50

Minimum Frequency of Words to view:

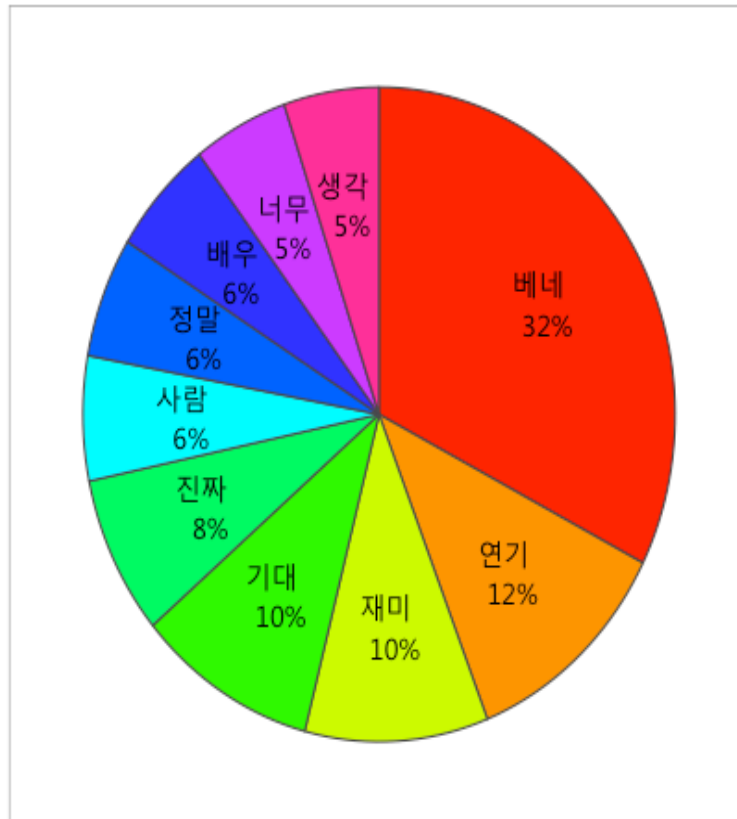
1 3 15

Update View

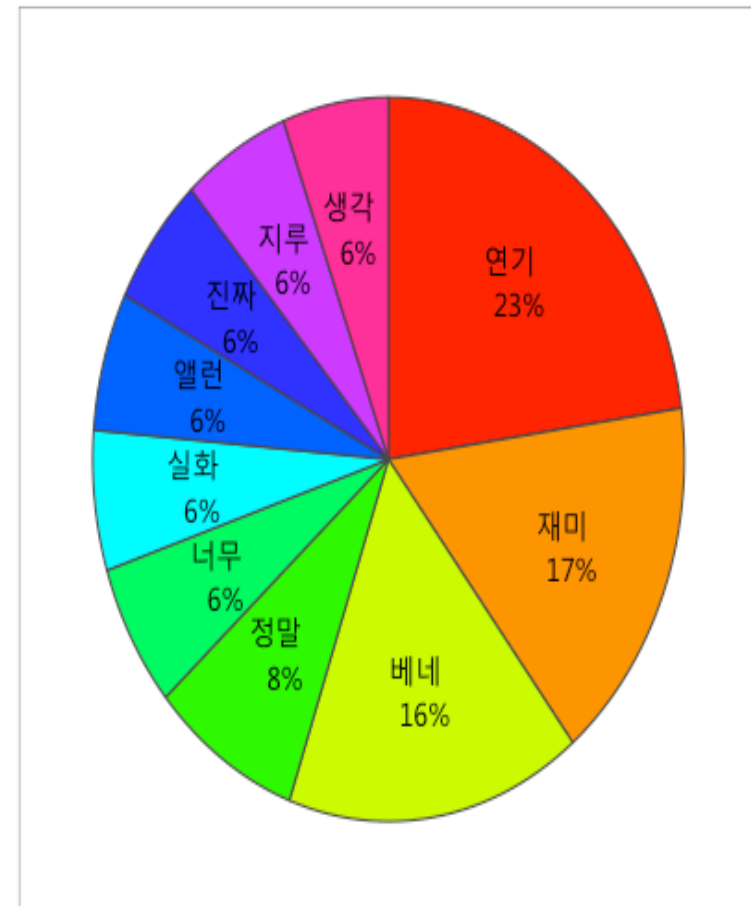


분석 예시 - Text

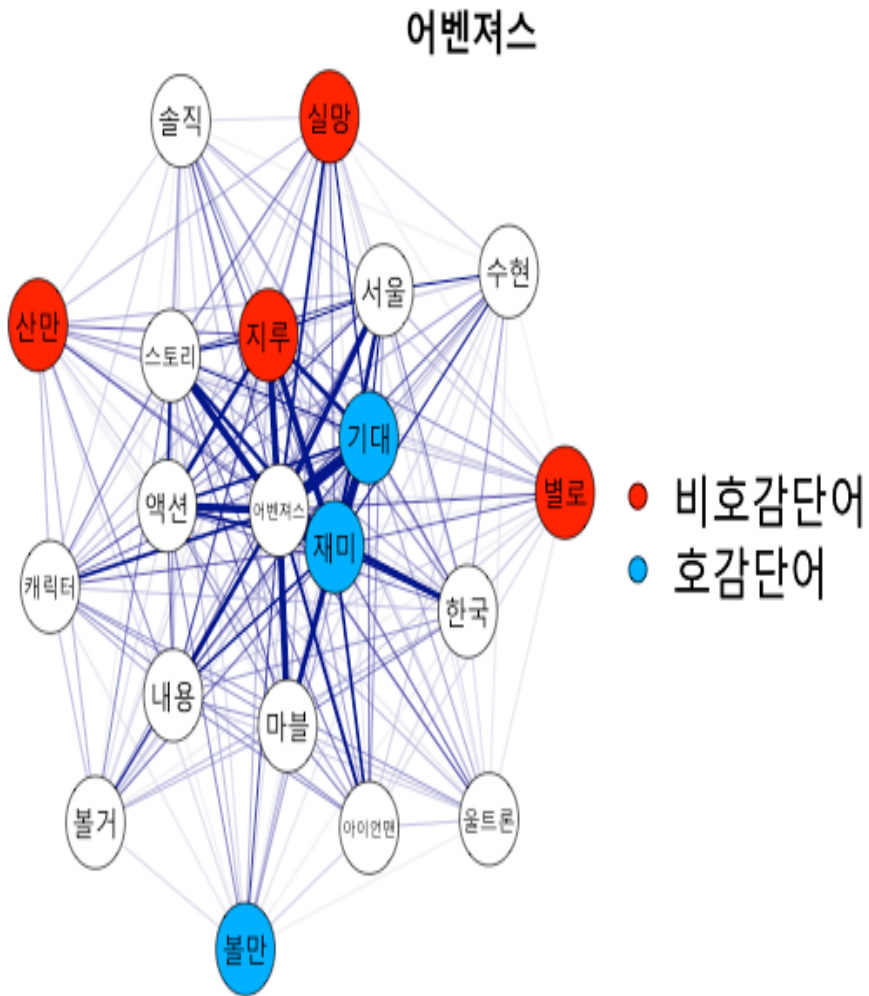
이미테이션 게임 개봉 전



이미테이션 게임 개봉 후



텍스트와 감정



```
library(KoNLP)
library(tm)
library(qgraph)
```

한국어 감정사전

불필요(stopwords) 단어사전

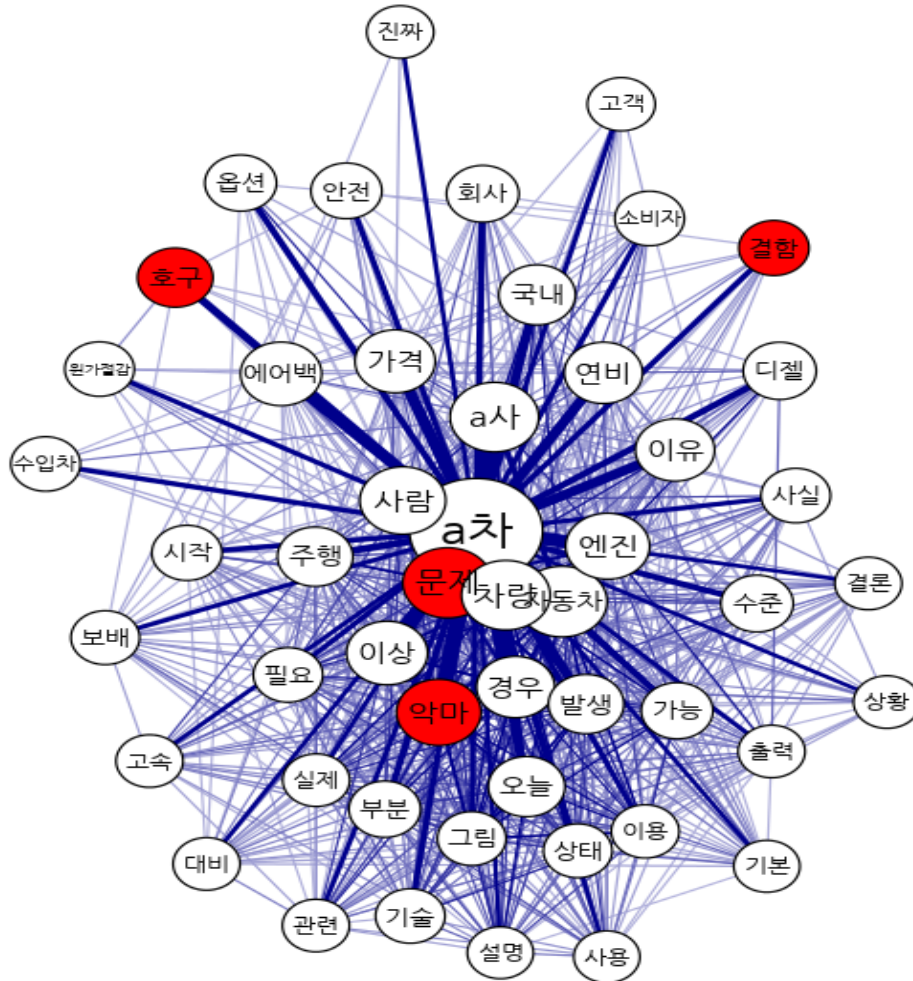
실망

```
## [1] "허접" "3d" "4d" "cg빨" "기대"
```

지루

```
## [1] "초반" "산만" "전개" "감정" "전편"
```

텍스트와 감정



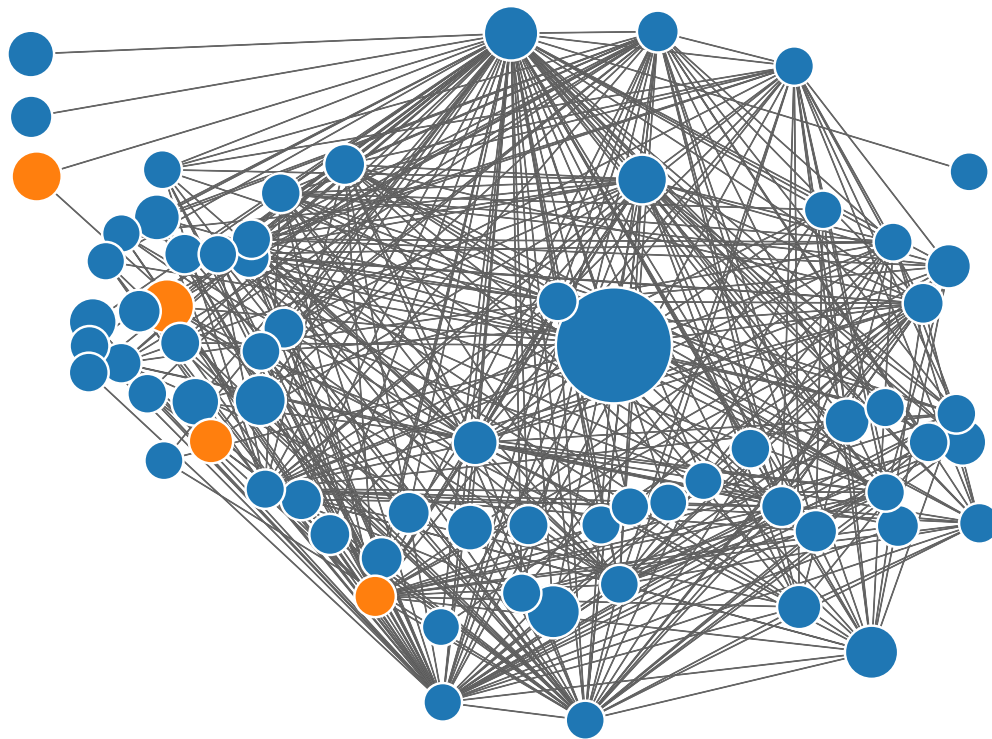
문제

[1] "발생" "차량" "해결" "방향" "무관" "상태" "판단"
"동일" "소음" "엔진"

결함

[1] "심각" "리콜" "기미" "대형사고" "머플러" "목숨"
[7] "앞바퀴" "직관" "확인" "국토"

텍스트와 감정



```
library(networkD3)
```

How?

필요한 것

형태소 분석 및 단어 파싱

- tm / tau / NLP / openNLP
- KoNLP

감정사전

- tm.plugin.sentiment
- http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/
- <http://word.snu.ac.kr/kosac/>
- http://clab.snu.ac.kr/arssa/doku.php?id=app_dict_1.0
- www.openhangul.com

사전 만드는 법

Dragut, E. C., Yu, C., Sistla, P., & Meng, W. (2010).

Construction of a sentimental word dictionary.

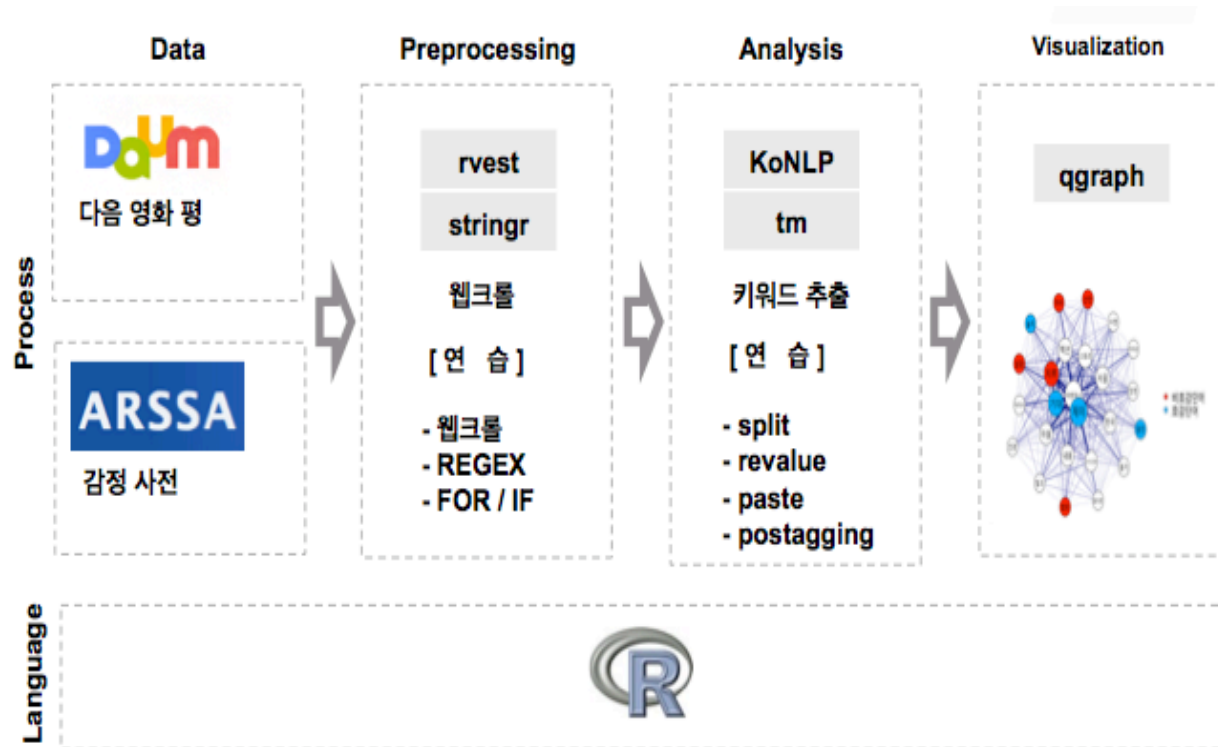
Paper presented at the Proceedings of the 19th ACM international conference on Information and knowledge management.

Rao, Y., Lei, J., Wenyin, L., Li, Q., & Chen, M. (2014).

Building emotional dictionary for sentiment analysis of online news.

World Wide Web, 17(4), 723-742.

Workflow



Scoring

감정 점수

Google Inc (NASDAQ:GOOGL)

Add to portfolio

557.52 +1.34 (0.24%)

Jun 19 - Close

NASDAQ real-time data - Disclaimer

Currency in USD

Range 552.26 - 557.91 Div/yield -
52 week 490.91 - 608.91 EPS 20.15
Open 556.52 Shares 288.26M
Vol / Avg. 2.96M/1.53M Beta 1.08
Mkt cap 373.84B Inst. own 82%
P/E 27.67

8+1 6.9k

Compare: ☐ Dow Jones ☐ Nasdaq ☐ BIDU ☐ YNDX ☐ BCOR ☐ YHOO ☐ MSFT ☐ IACI ☐ IBM

Zoom: 1d 5d 1m 3m 6m YTD 1y 5y 10y All

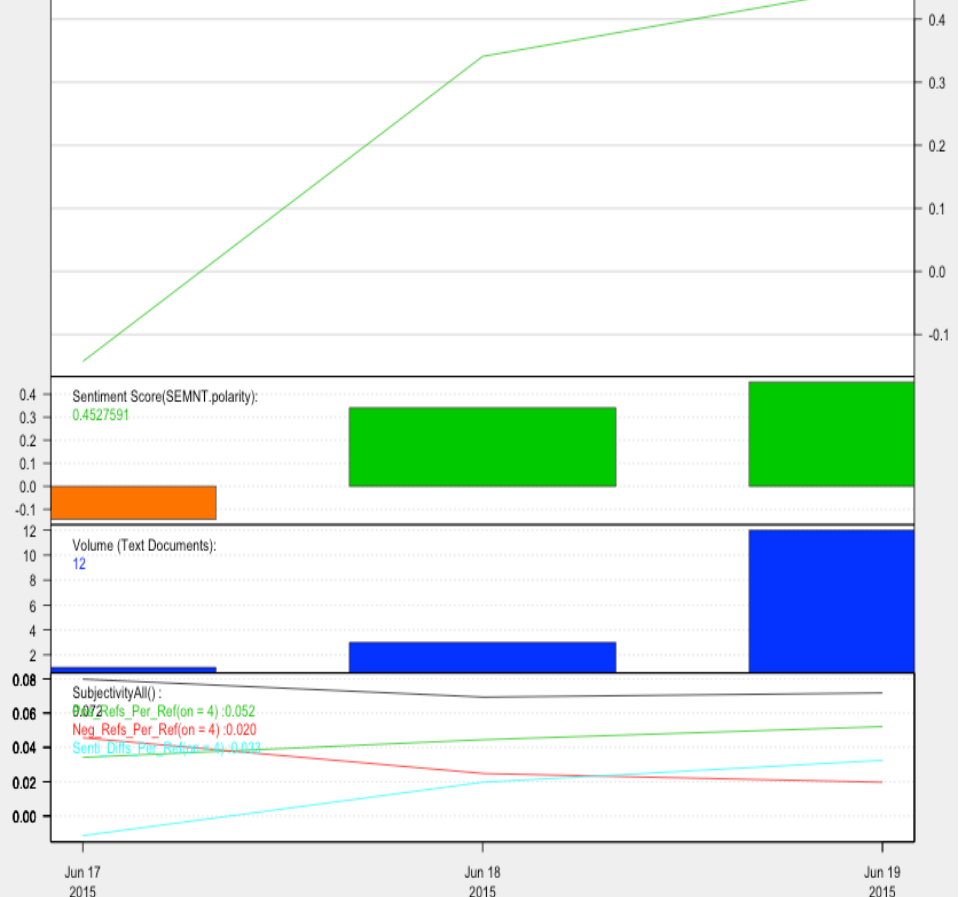
Jun 15, 2015 15:54 Price: 543.57 Vol: 08.17k



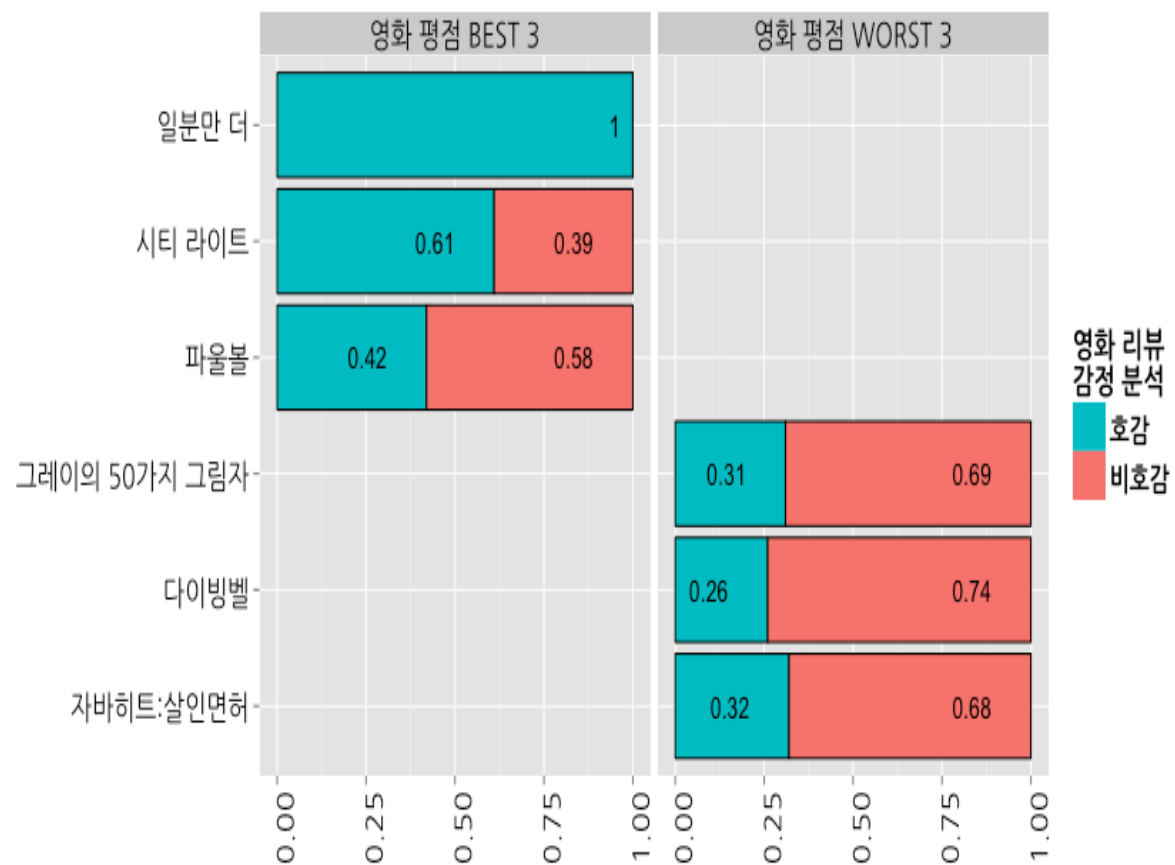
Sentiment SEMNT

[2015-06-17 09:00:00/2015-06-19 09:00:00]

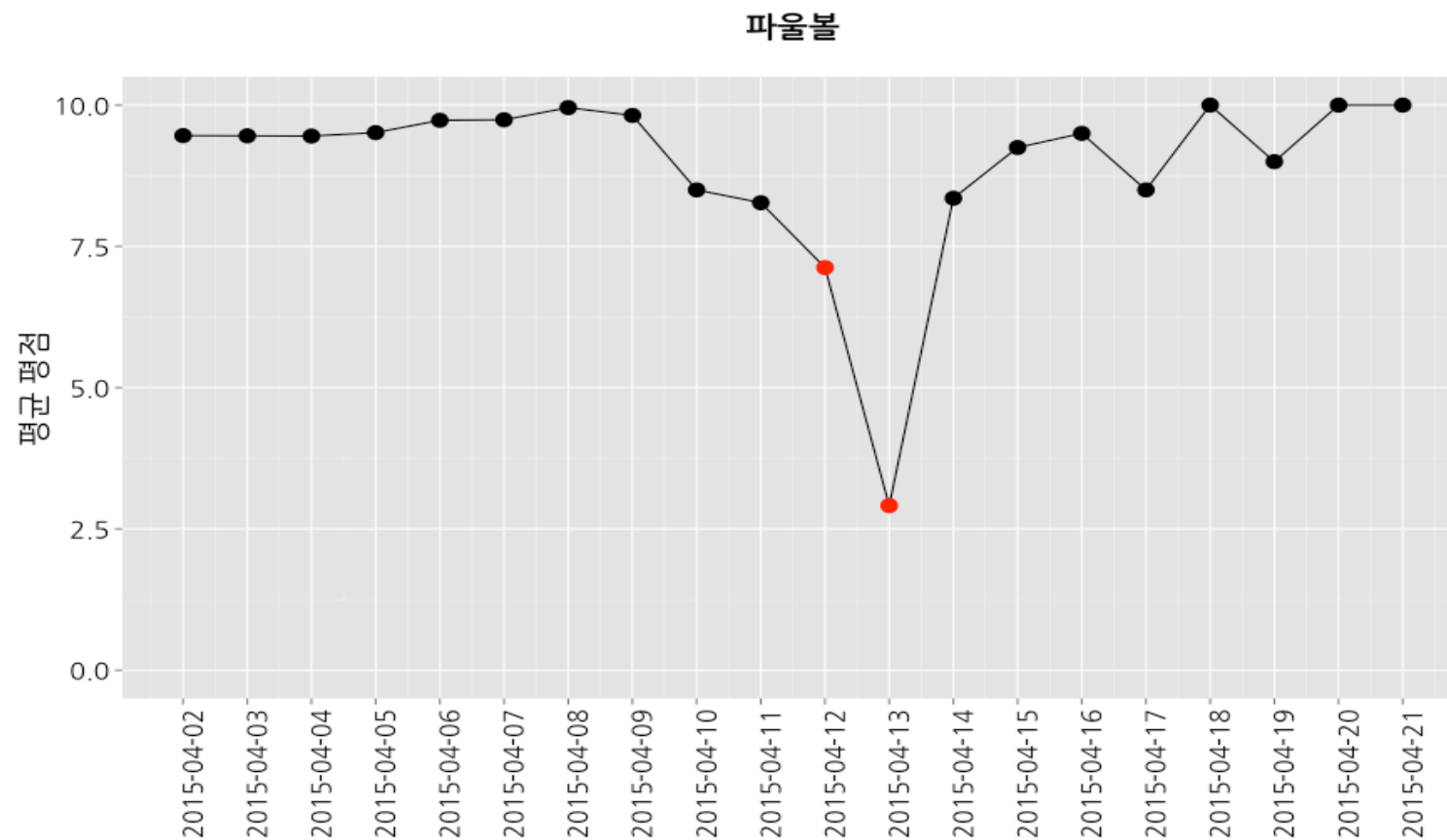
Last 0.452759080538875



감정 점수 & 해석



감정 점수 실패 사례




감정 점수 실패 사례

스포츠

‘빈볼’ 이동걸 퇴장, 황재균 분노…한화-롯데 벤치클리어링
이동걸, 황재균 향해 연달아 몸쪽 위협구
몸에 맞자 양 팀 선수들 쏟아져 나와 신경전

기사본문 댓글 바로가기 등록 : 2015-04-12 22:28 [+가](#) [-가](#) [인쇄하기](#)

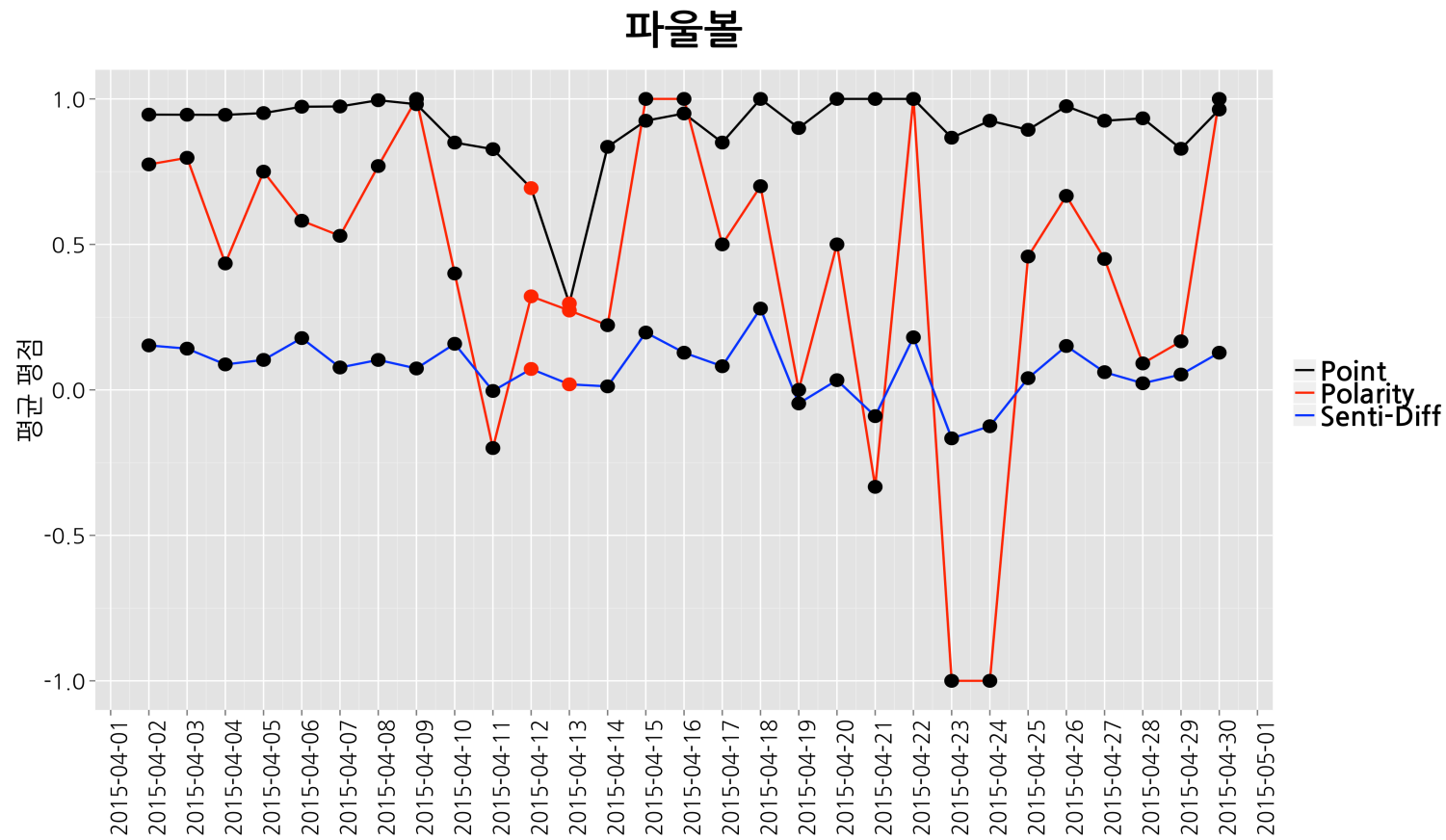
스포츠 = 김도엽 객원기자 [기사더보기 +](#)





▲ 한화 이동걸이 롯데 황재균에게 빈볼을 던져 퇴장 당했다. (MBC 스포츠 방송 캡처)

감정 점수 실패 사례



WHY?

- ## [1] "한국야구의 진정한 발전을 위해서 이런 췌같은 영감탱이의 우상화, 신격화는 막아야....영감님 닮는 작자들은 내가 롯데팬이라고 뒤집어씌울듯...ㅉㅉ"
- ## [2] "성큰옹..선수는 그냥 소모품임? 실망..."
- ## [3] "야신은 무슨 잘못도 인정 안하는 노망난 할배지"
- ## [4] "동걸이 인생은 내 알아야니지"
- ## [5] "이동현 전병두 이승호 정대현 김성길 신윤호 김현욱 박정현 고효준 장문석 : 감독님 팔이 안올라가요 ㅠㅠ"
- ## [6] "제목 틀렸습니다. 데드볼이라고 해야지 않나 시포요."
- ## [7] "독립구단에서도 연봉은 역대로 받으셨죠"
- ## [8] "빈볼시키고 선수를 소모품처럼 버리고..."
- ## [9] "빈볼이라쓰고실투라부른다"
- ## [10] "이동걸만 불쌍...."
- ## [11] "빈볼왕101010010101"
- ## [12] "영화가 얼마나 사람의 시야를 흐리게 만드는지 분명히 보여준다. 감성팔이를 하려면 최소한 감성이 있는 사람이 해야하지 않을까? 선수들을 인간적으로"
- ## [13] "이만수 종신갓동니뮤ㅠ"
- ## [14] "0점 왜 못주는거죠? 꼭 주고 싶습니다ㅠ"
- ## [15] "빈볼 던지라고 시켜놓고 자기는 안시켰다고 그 투수만 제구안되는 병신으로 만들어버리네."
- ## [16] "인간의 탈을 쓴 더러운 양아치.야구의 신이 아니라아버의 신에게 딱 킬성근의 본모습.킬성근의 가식에 치가 떨린다."
- ## [17] "미화 하나는 잘 시키는 역겨운 한국."
- ## [18] "파울볼? 김성근하면 역시 빈볼이지"
- ## [19] "추잡한 늙은이 야구계를 떠나라"
- ## [20] "야구계에서 사라지십쇼. 언제까지 그렇게 더러운 플레이로 팬들 눈살을 찌푸리게 하실 겁니까? 이게 한 두번이어야 그러려니 하지..sk 때부터 악질입니"
- ## [21] "황재균을 향한 공이 두번 빛나가고 세번째 공을 던지려 들때 이동걸의 비참한 표정이 뇌리에 깊이 박혀 지워지지않는다. 33살의 무명이 4살어린 유"
- ## [22] "당신이 추구하는 야구 어제 아주 잘 보았습니다 ^^ 남 가르치기전에 자신부터 돌아보시길 ㅎㅎ"
- ## [23] "빈볼 더티야구 노답..."
- ## [24] "이미지 세탁왕 제일교포 김성근"

대안

- Training Set과 Test Set을 7:3으로 분할
- SLDA 사용
- Blei, David M. and Ng, Andrew and Jordan, Michael. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003.

```
library(lda)
```

예측한 점수와 실제 점수간 상관관계

	X	TEST.POINT	POLARITY	SENTI.DIFF	SLDA
1	test.point	1.00	0.01	0.07	0.66
2	Polarity	0.01	1.00	0.75	-0.01
3	Senti-Diff	0.07	0.75	1.00	0.05
4	slda	0.66	-0.01	0.05	1.00

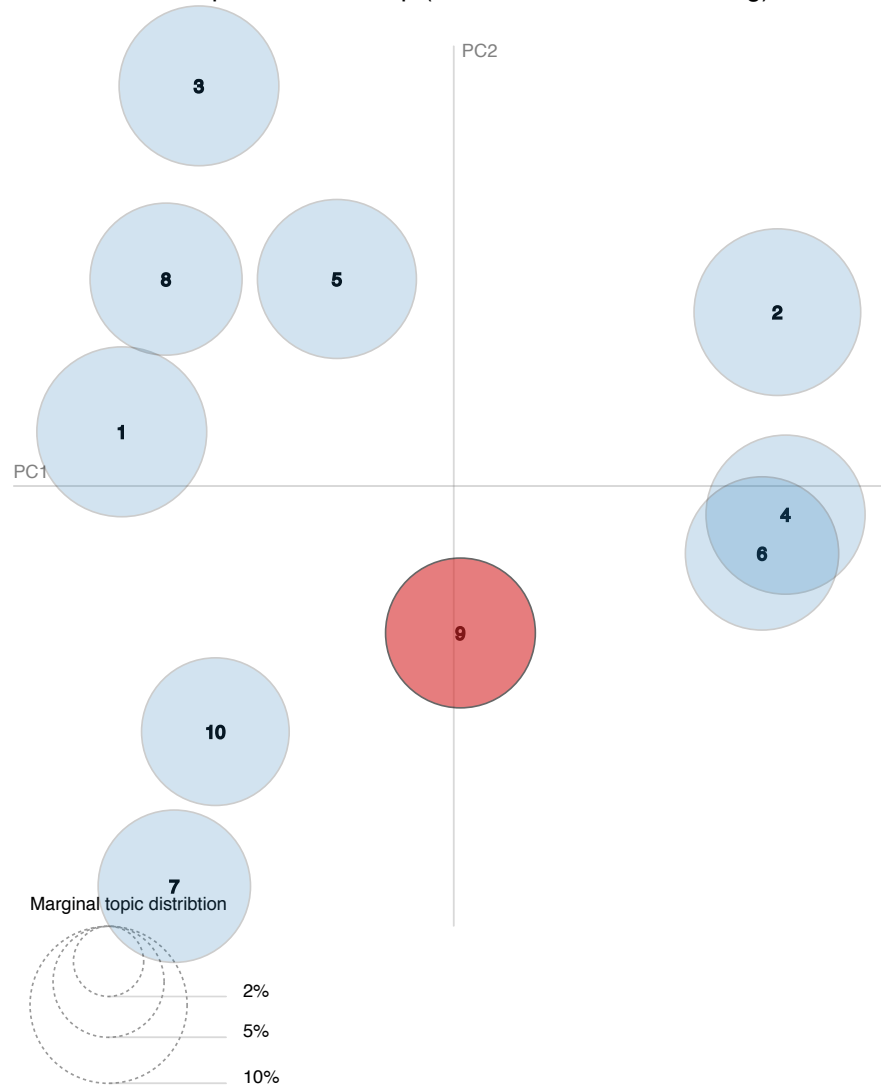
Selected Topic:

Slide to adjust relevance metric:⁽²⁾

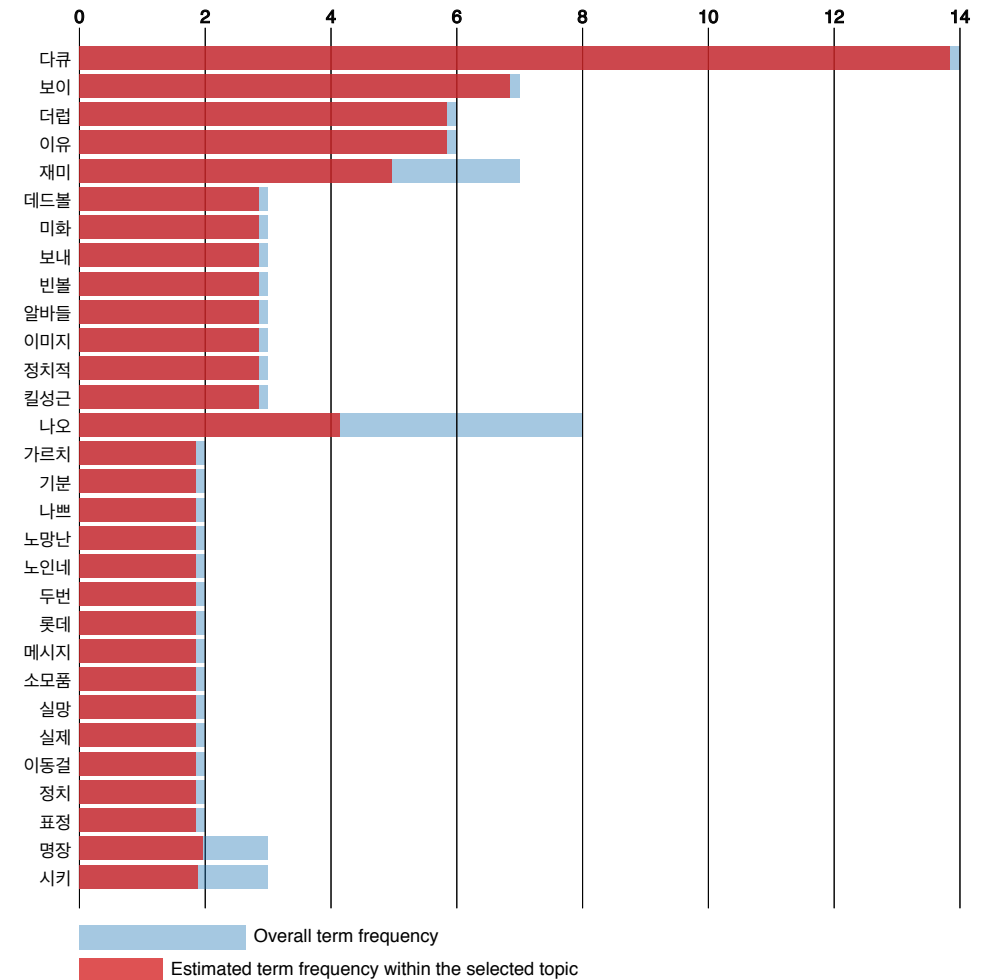
$\lambda = 0.51$



Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 9 (9.1% of tokens)

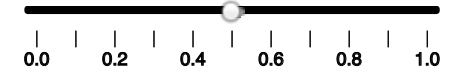


1. saliency(term w) = frequency(w) * $[\sum_t p(t | w) * \log(p(t | w)/p(t))]$ for topics t ; see Chuang et. al (2012)
2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

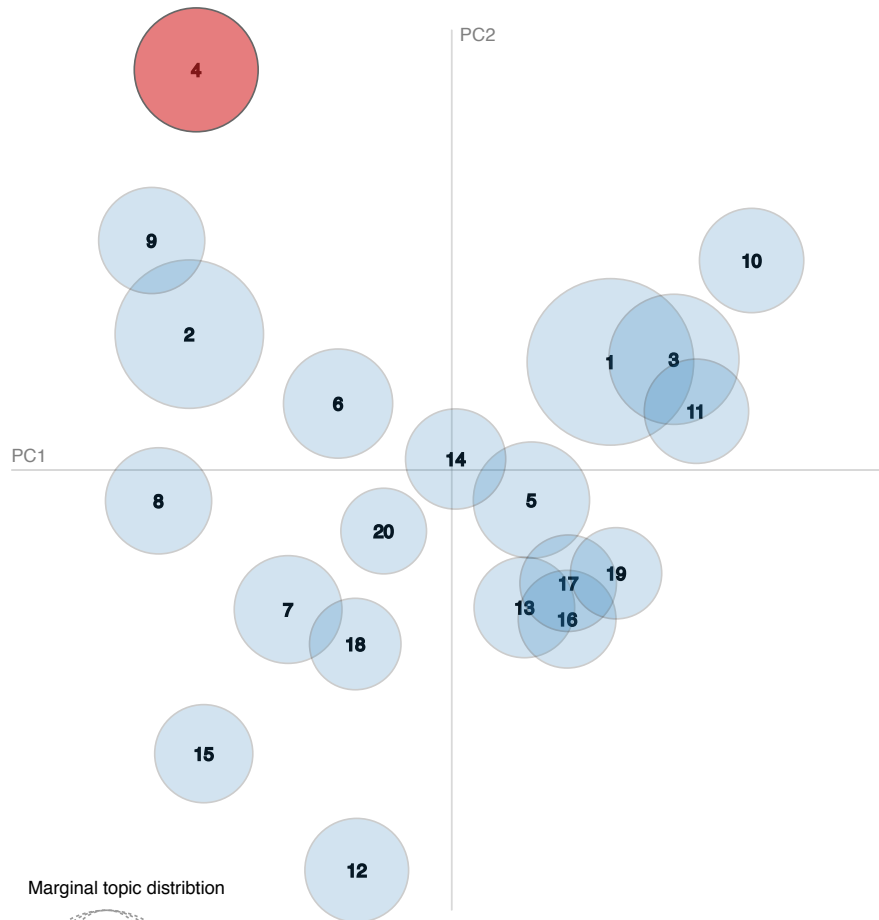
Selected Topic:

Slide to adjust relevance metric:⁽²⁾

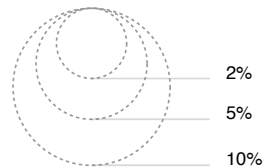
$\lambda = 0.5$



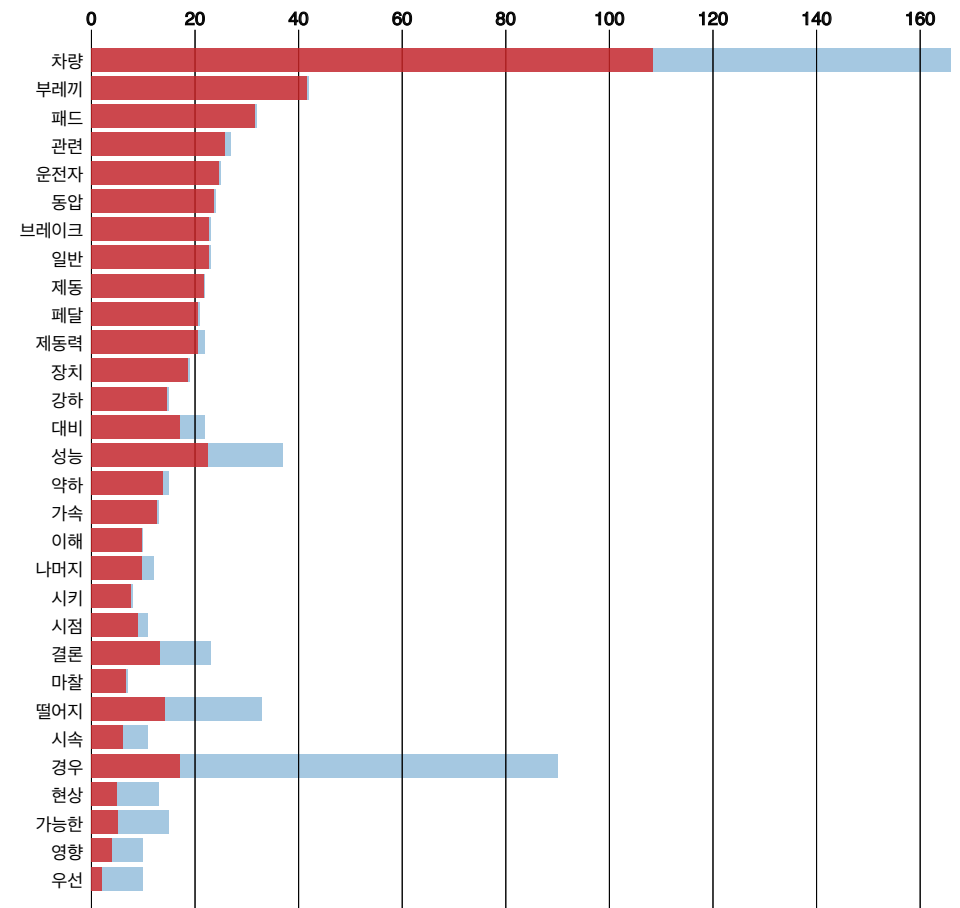
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 4 (6.2% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

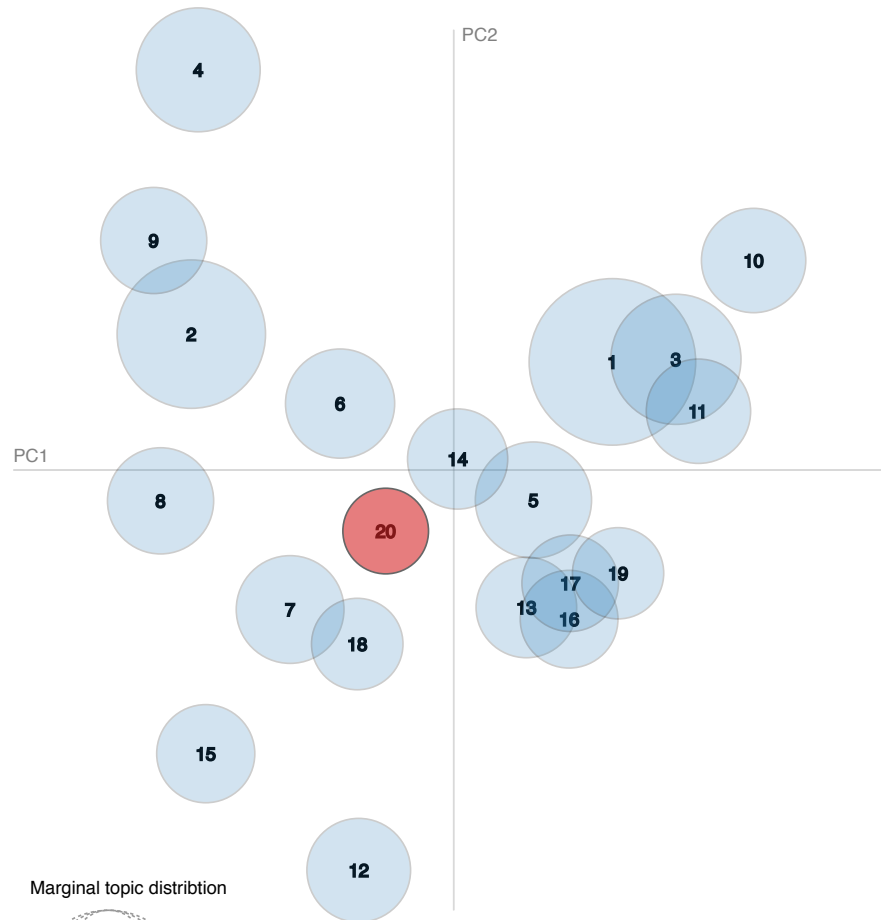
Selected Topic: 20 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:⁽²⁾

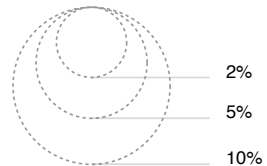
$\lambda = 0.5$



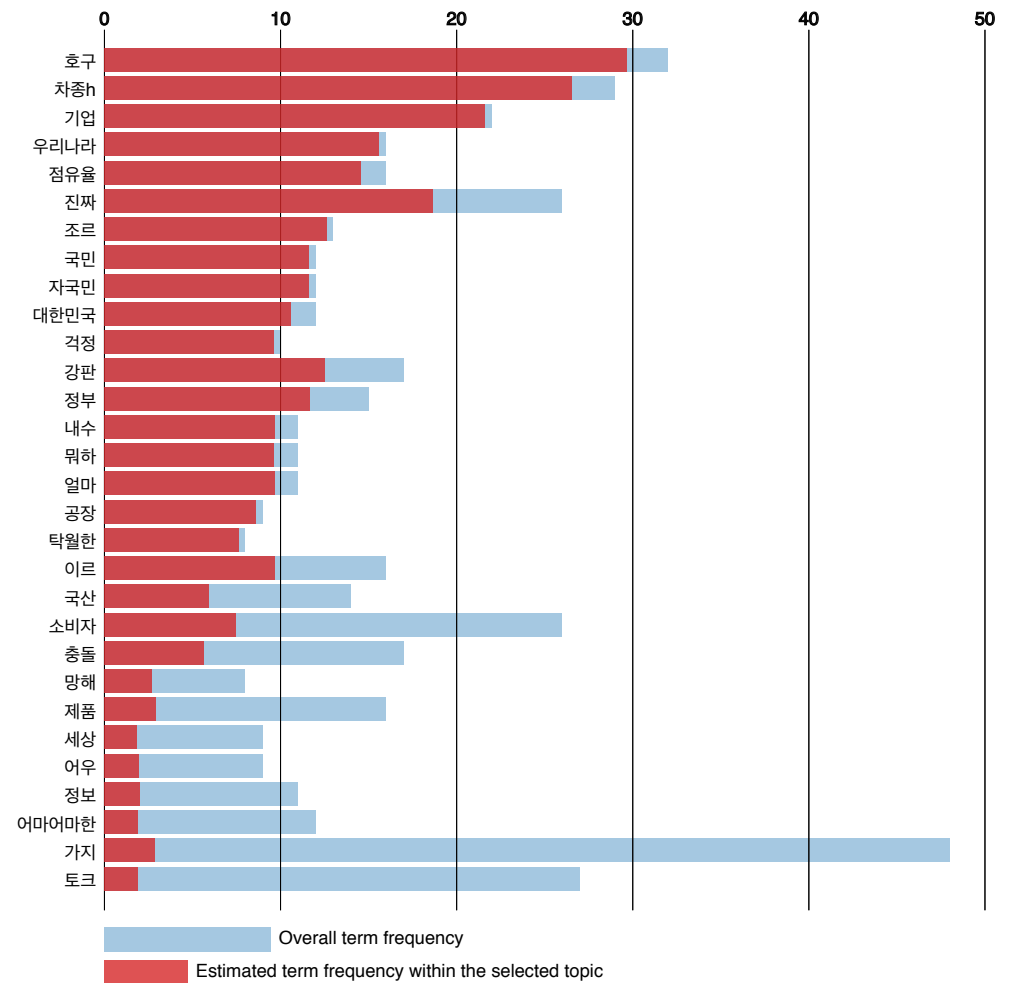
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 20 (3% of tokens)

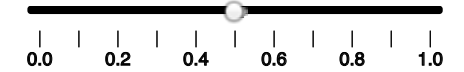


1. saliency(term w) = frequency(w) * $[\sum_t p(t | w) * \log(p(t | w)/p(t))]$ for topics t ; see Chuang et. al (2012)
2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

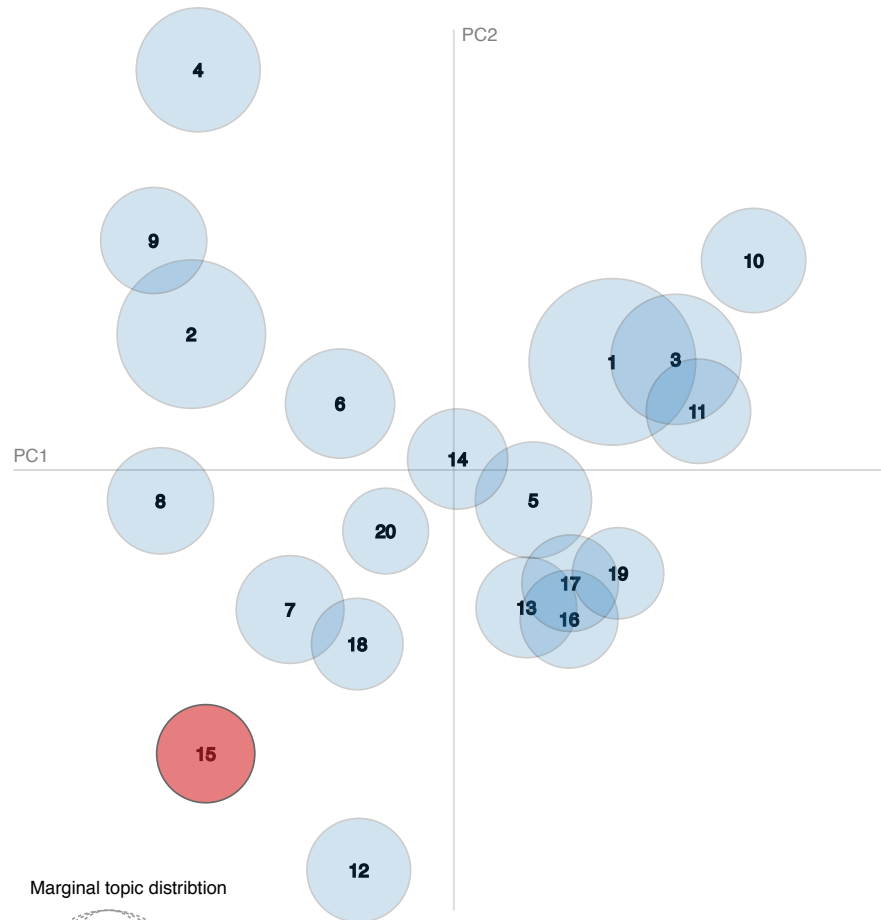
Selected Topic: 15 Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric:⁽²⁾

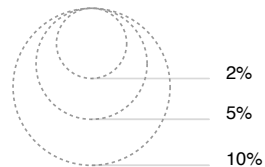
$\lambda = 0.5$



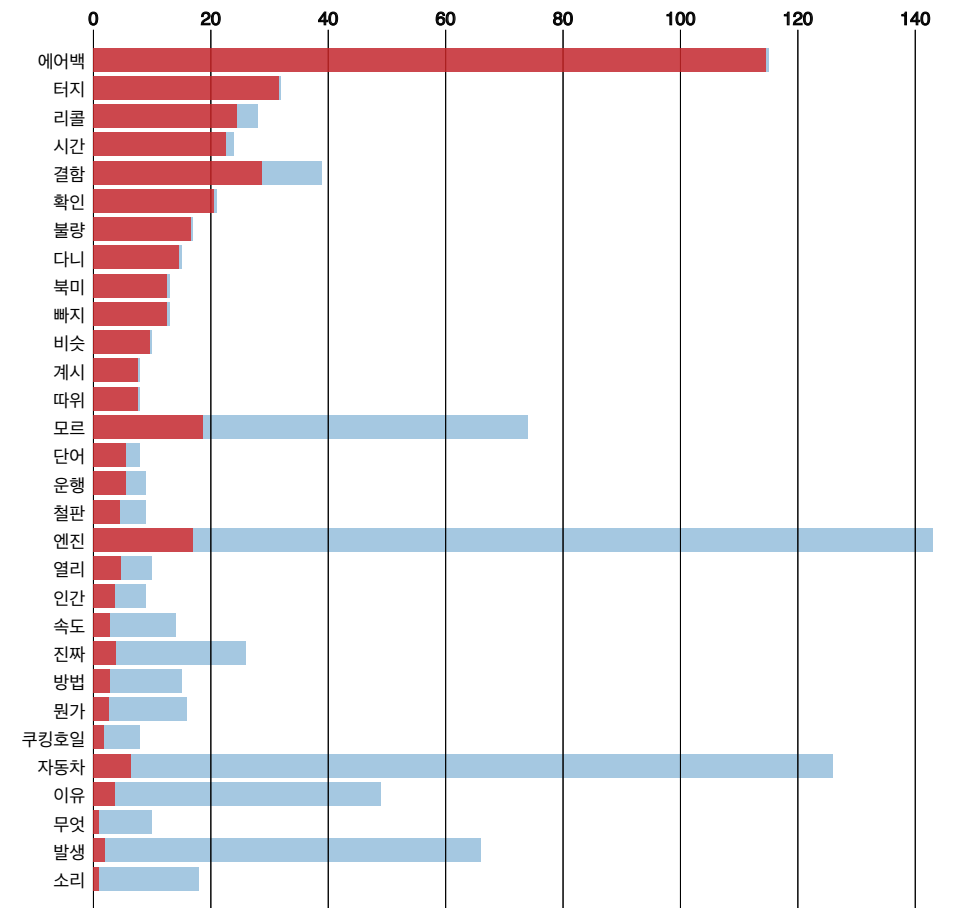
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 15 (3.9% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t ; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

워크숍 관련 온라인 사이트

<http://course.mindscale.kr/course/text-analysis>

코스

현재 수강 중인 코스입니다.

제목	수강 시작	수강 끝	
텍스트에서 여론과 감정을 발견하기 : R을 이용한 텍스트 데이터 분석	2015-05-10	2015-06-30	강의실로
텍스트에서 여론과 감정을 발견하기 : R을 이용한 텍스트 데이터 분석 (05/30)	2015-05-26	2100-01-01	강의실로
R을 이용한 웹 크롤링	2015-06-10	2100-01-01	강의실로