





R을 이용한 텍스트 감정분석

여론과 감성 발견하기

김형준

Data Analyst / (주) 퀸트랩 / 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
- 교통안전공단
- 삼성전자
- 이지웰페어
- 웅진씽크빅
- 중소기업진흥공단
- 한화
- 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
- 본 발표자료는 Interactive Plots이 포함되어 있습니다. [클릭하세요](#)

텍스트 분석

텍스트 분석 목적

: 사람들은 생각과 감정을 언어로 표현합니다. 뉴스 댓글, 상품평, 커뮤니티, SNS 등에 사람들이 남기는 텍스트를 모아서 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

Choose a dataset:

2013y

Numbers of Words to view:
50

Minimum Frequency of Words to view:
1 3 15

Update View



대통령 취임사 Shiny

박근혜 - 노무현



노무현 - 박근혜



영화 이미테이션 게임 & 베네딕트



★★★★★ 10 베스트 베네딕트 영화라면 언제나 기대기대!!

아크(ark0****) | 2013.09.24 17:27 | 신고

👍 공감 300 👎 비공감 30

★★★★★ 10 베스트 베니 기대된다

ㅇㅇ(pe1r****) | 2013.09.15 00:04 | 신고

👍 공감 201 👎 비공감 24

★★★★★ 10 베스트 기대기대^^ 베니 홍해랏

올라올라(yunb****) | 2013.09.22 12:59 | 신고

👍 공감 155 👎 비공감 21

★★★★★ 10 베스트 베니 ㅠㅠㅠㅠㅠㅠㅠㅠㅠㅠ

빅더(raw****) | 2013.09.16 19:50 | 신고

👍 공감 139 👎 비공감 20

★★★★★ 10 베스트 베니!!!! 짱 기대된다

천류한(knhj****) | 2013.10.13 16:26 | 신고

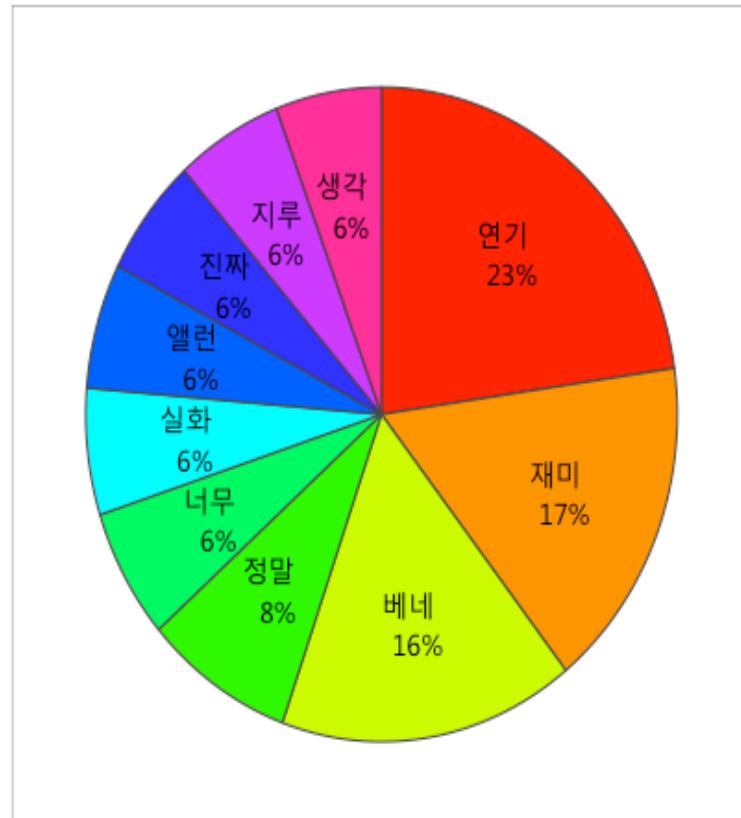
👍 공감 130 👎 비공감 19

분석 예시 - Text

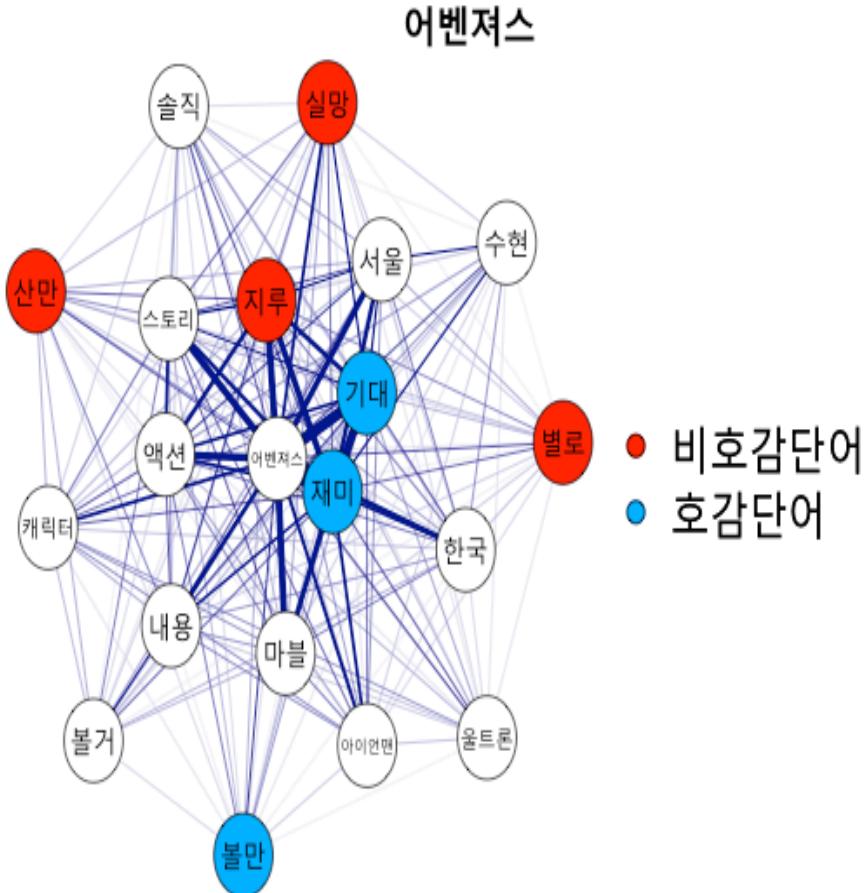
이미테이션 게임 개봉 전



이미테이션 게임 개봉 후



텍스트와 감정



library(KoNLP)

library(tm)

library(qgraph)

한국어 감정사전

불필요(stopwords) 단어사전

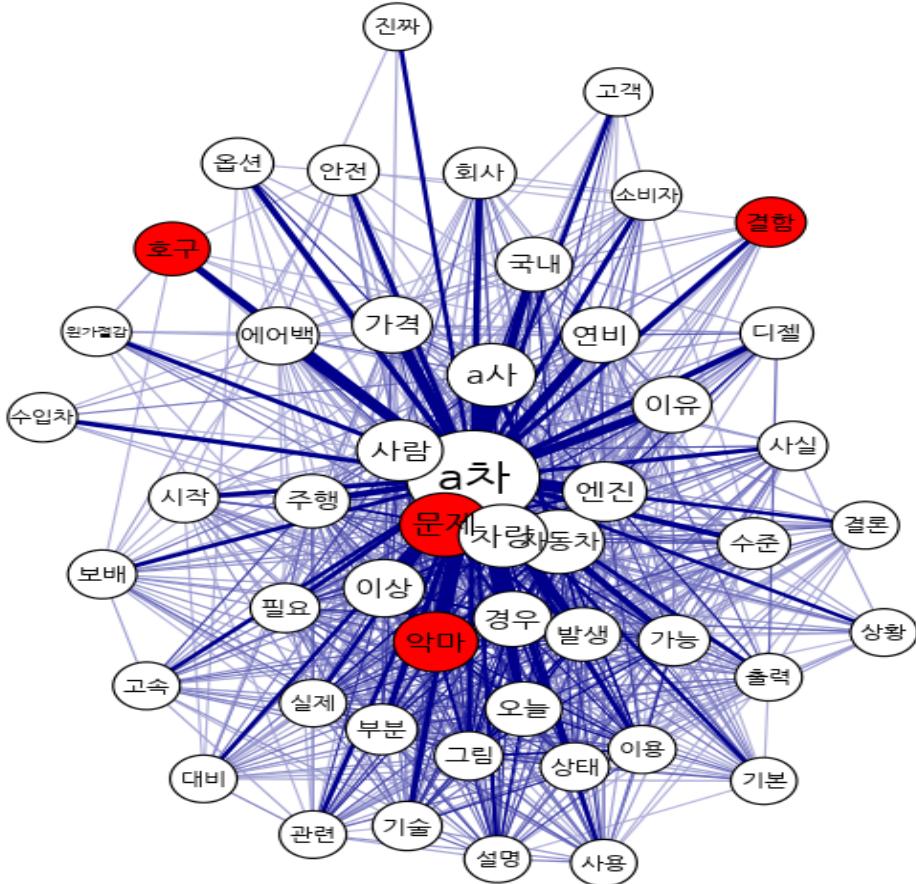
실망

[1] "허접" "3d" "4d" "cg빨" "기대"

지루

[1] "초반" "산만" "전개" "감정" "전편"

텍스트와 감정



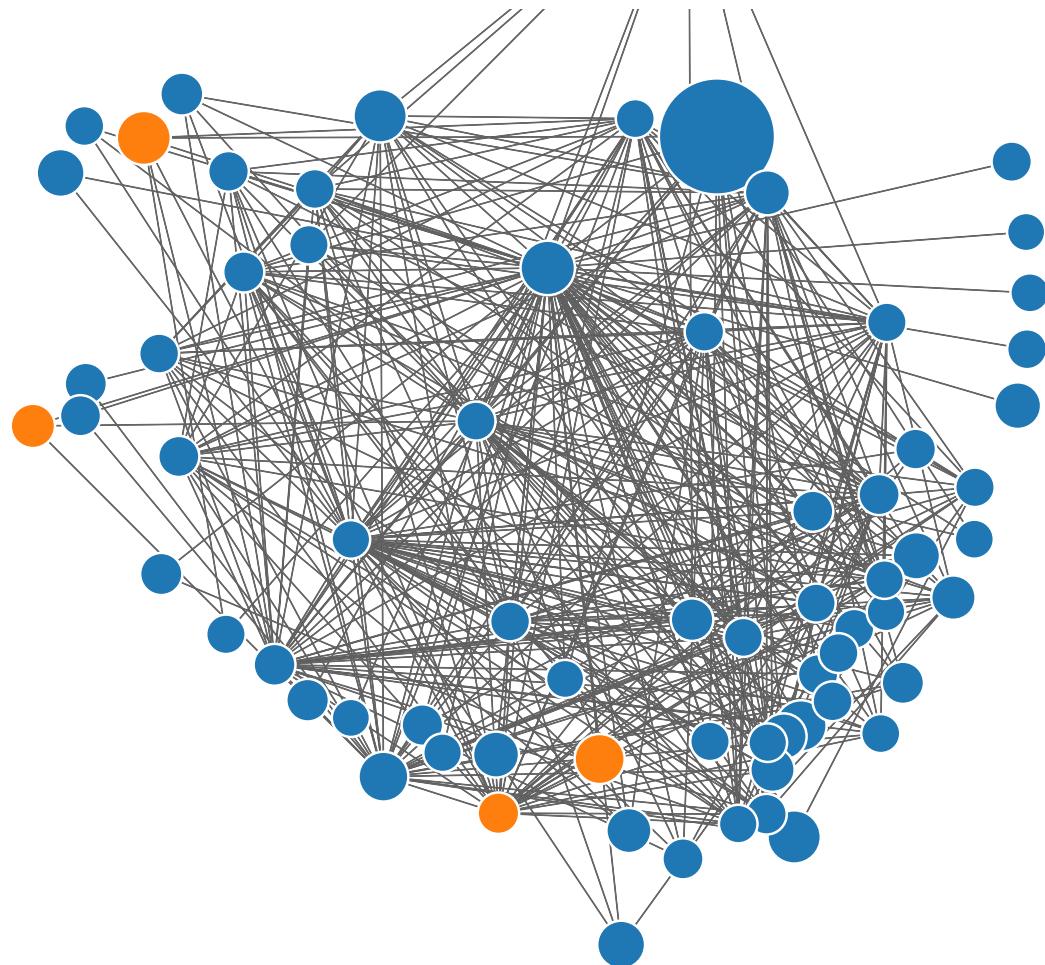
문제

- ```
[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://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](http://clab.snu.ac.kr/arssa/doku.php?id=app_dict_1.0)
- [www.openhangul.com](http://www.openhangul.com)

# 사전 만드는 법

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

Construction of a sentimental word dictionary.

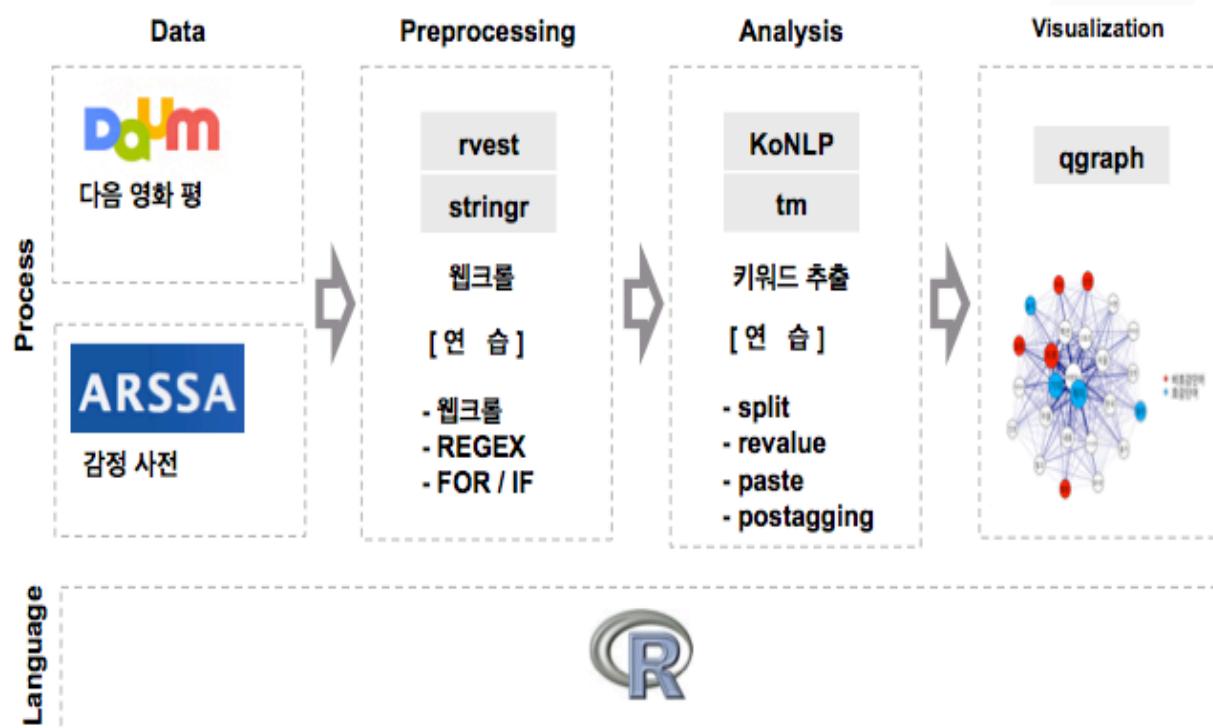
*Paper presented at the Proceedings of the 19th ACM international conference on Information and knowledge management.*

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

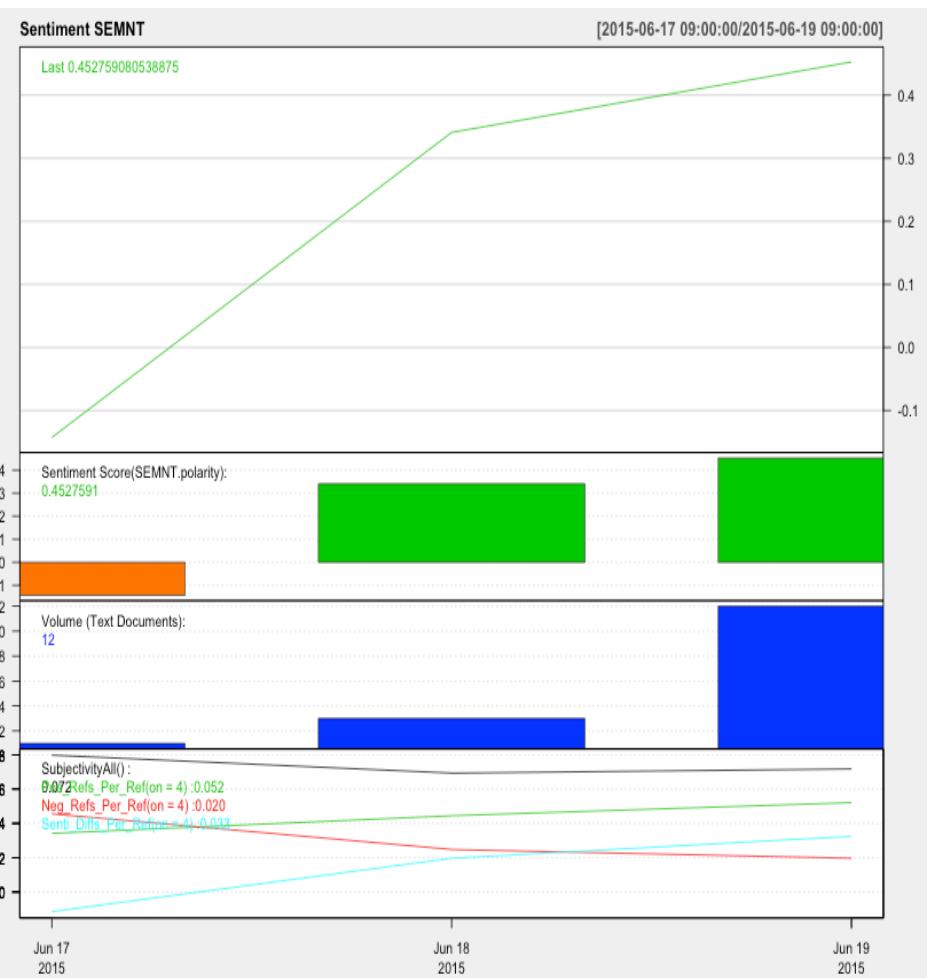
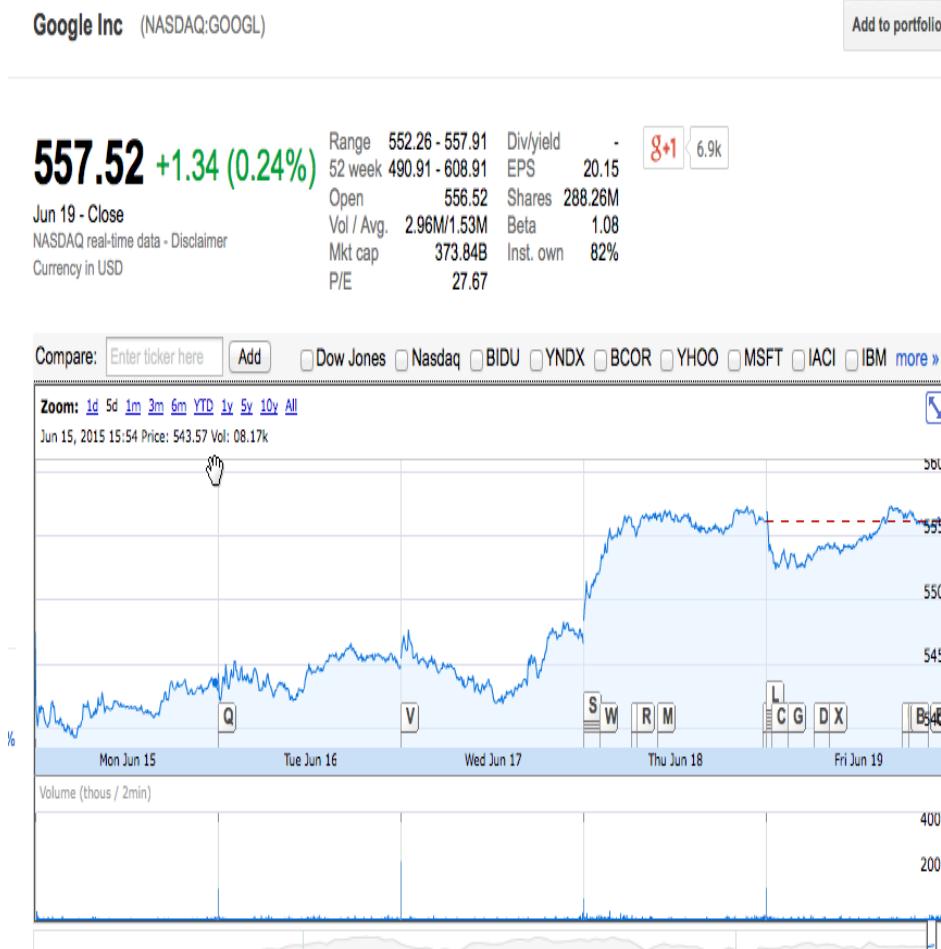
Building emotional dictionary for sentiment analysis of online news.

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

# Workflow



# 감정 점수 (tm.plugin.sentiment)



# 감정 점수

## Sentiment Indicators<sup>2</sup>

$$polarity = \frac{p - n}{p + n} \quad (1)$$

$$subjectivity = \frac{n + p}{N} \quad (2)$$

$$pos\_refs\_per\_ref = \frac{p}{N} \quad (3)$$

$$neg\_refs\_per\_ref = \frac{n}{N} \quad (4)$$

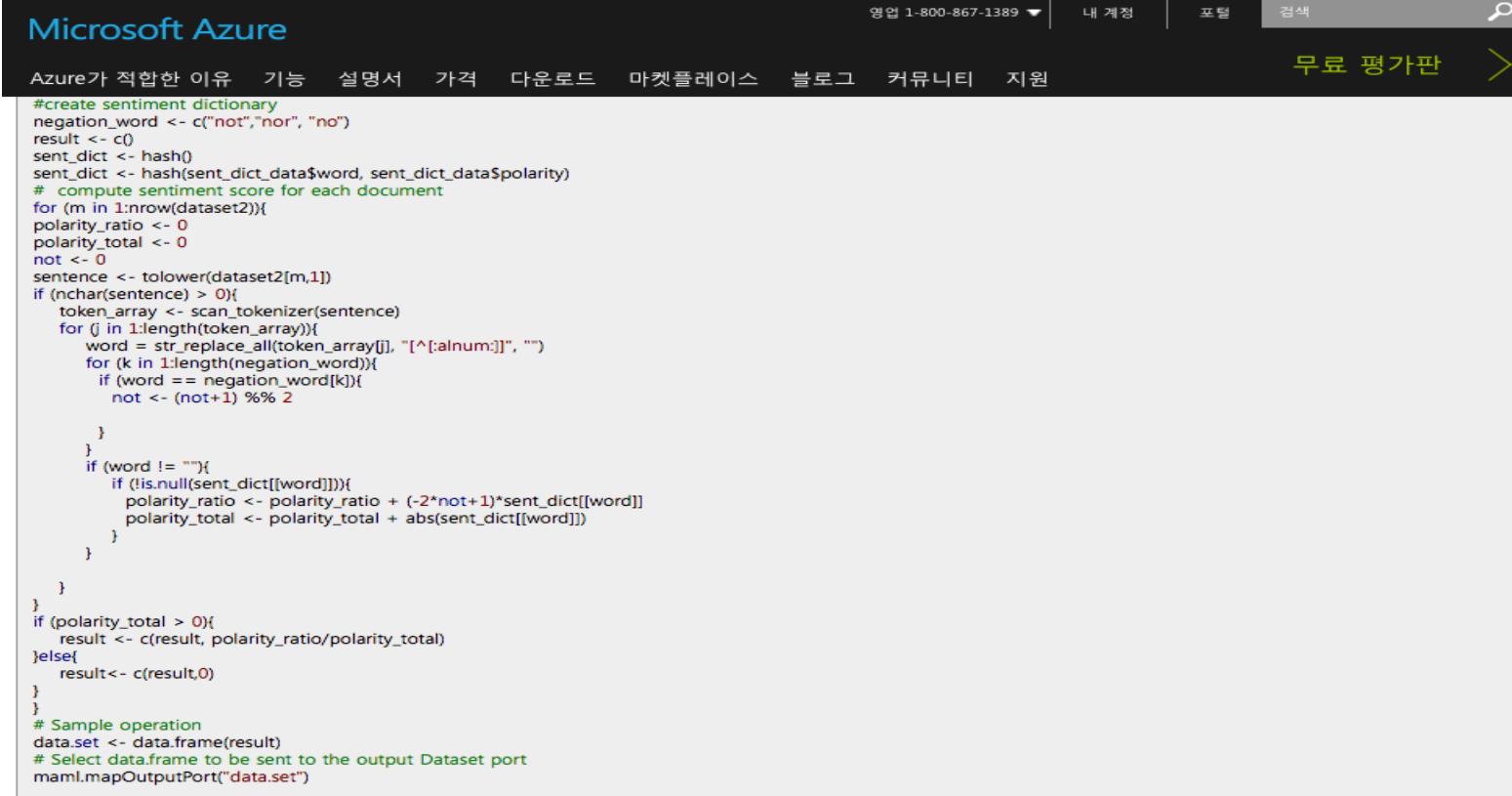
$$senti\_diffs\_per\_ref = \frac{p - n}{N} \quad (5)$$

---

<sup>2</sup>taken from the Lydia/Textmap project

- 출처 : Mario Annau(2010)

# 부정어 처리 (안, 않, 못)



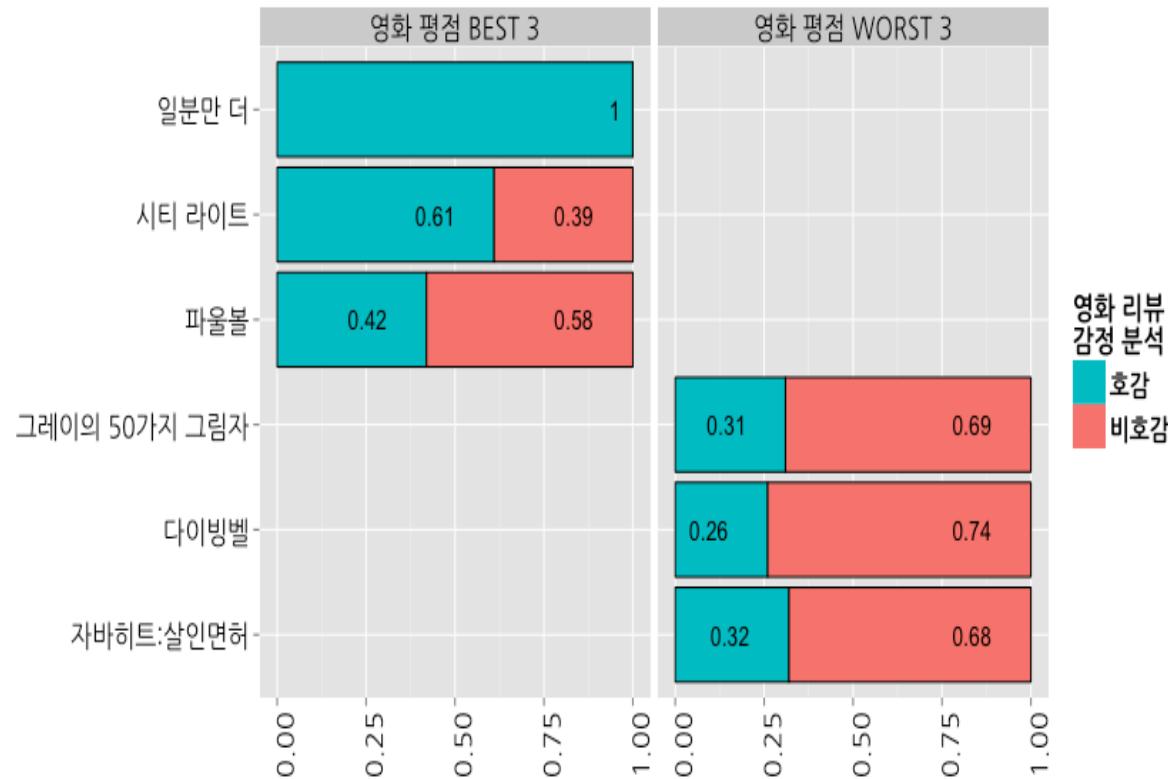
The screenshot shows the Microsoft Azure portal interface. At the top, there's a navigation bar with links for 'Azure가 적합한 이유', '기능', '설명서', '가격', '다운로드', '마켓플레이스', '블로그', '커뮤니티', and '지원'. On the right side of the bar, there are buttons for '영업 1-800-867-1389', '내 계정', '포털', '검색', ' 무료 평가판 >', and a magnifying glass icon.

```
#create sentiment dictionary
negation_word <- c("not", "nor", "no")
result <- c()
sent_dict <- hash()
sent_dict <- hash(sent_dict$data$word, sent_dict$data$polarity)
compute sentiment score for each document
for (m in 1:nrow(dataset2)){
 polarity_ratio <- 0
 polarity_total <- 0
 not <- 0
 sentence <- tolower(dataset2[m,1])
 if (nchar(sentence) > 0){
 token_array <- scan_tokenizer(sentence)
 for (j in 1:length(token_array)){
 word = str_replace_all(token_array[j], "[^[:alnum:]]", "")
 for (k in 1:length(negation_word)){
 if (word == negation_word[k]){
 not <- (not+1) %% 2
 }
 }
 if (word != ""){
 if (!is.null(sent_dict[[word]])){
 polarity_ratio <- polarity_ratio + (-2*not+1)*sent_dict[[word]]
 polarity_total <- polarity_total + abs(sent_dict[[word]])
 }
 }
 }
 }
 if (polarity_total > 0){
 result <- c(result, polarity_ratio/polarity_total)
 }else{
 result<- c(result,0)
 }
}
Sample operation
data.set <- data.frame(result)
Select data.frame to be sent to the output Dataset port
maml.mapOutputPort("data.set")
```

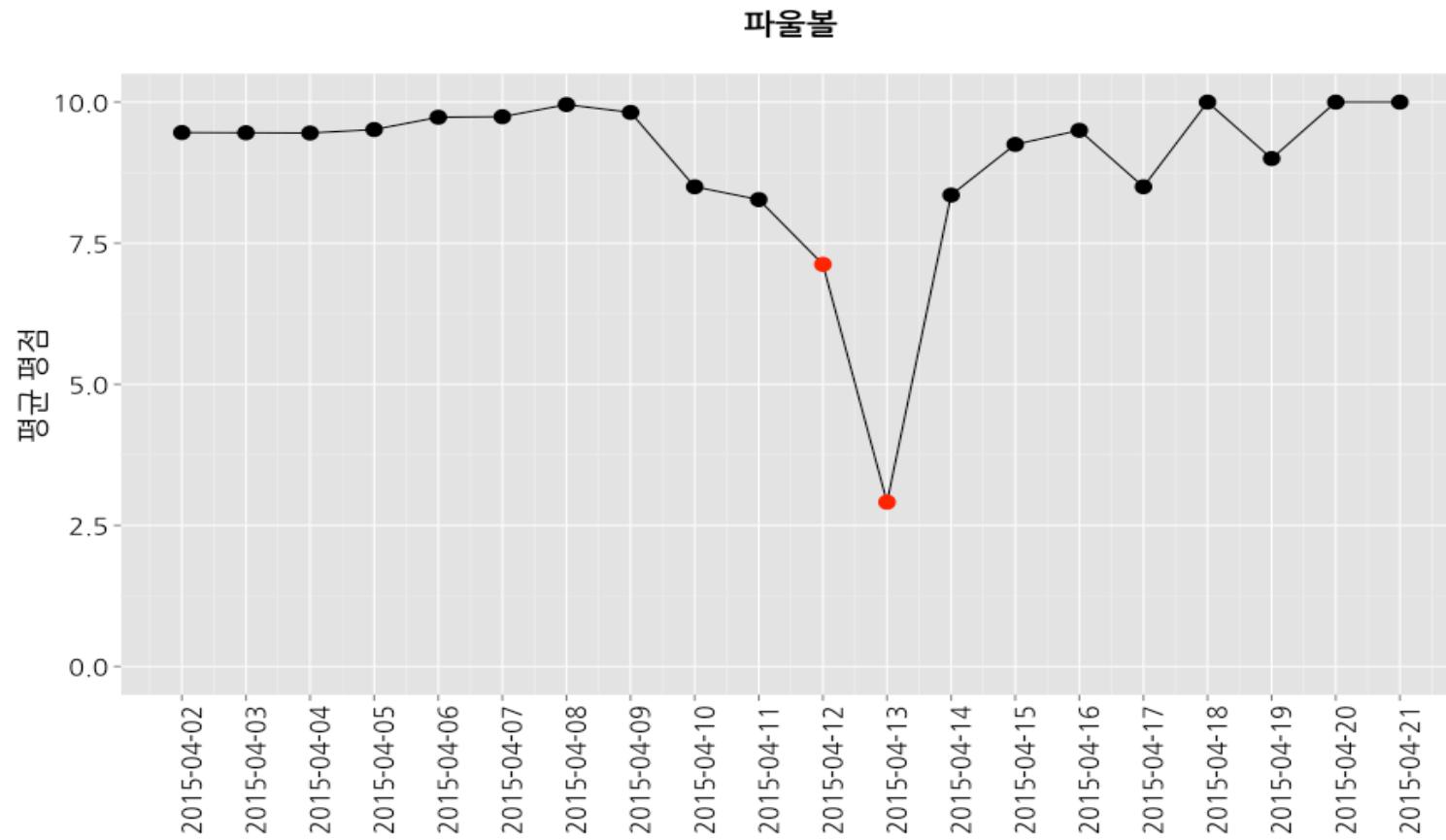
## 제한 사항

알고리즘 관점에서 봤을 때 어휘집 기반 감정 분석은 특정 필드에 대해 분류 방법보다 성능이 좋지 않을 수도 있는 일반 감정 분석 도구입니다. 부정 문제 가 잘 처리되지 않습니다. 여기서는 프로그램에 몇 가지 부정 단어를 하드 코드하지만, 더 좋은 방법은 부정 사전을 사용하고 몇 가지 규칙을 작성하는 것입니다. 이 웹 서비스는 Amazon 리뷰와 같은 길고 복잡한 문장보다 트윗, Facebook 게시물과 같은 짧고 간단한 문장에 대해 더 잘 수행됩니다.

# 감정 점수



# 감정 점수



# 감정 점수

**스포츠** ‘빈볼’ 이동걸 퇴장, 황재균 분노…한화–롯데 벤치클리어링  
이동걸, 황재균 항해 연달아 몸쪽 위협구  
몸에 맞자 양 팀 선수들 쏟아져 나와 신경전

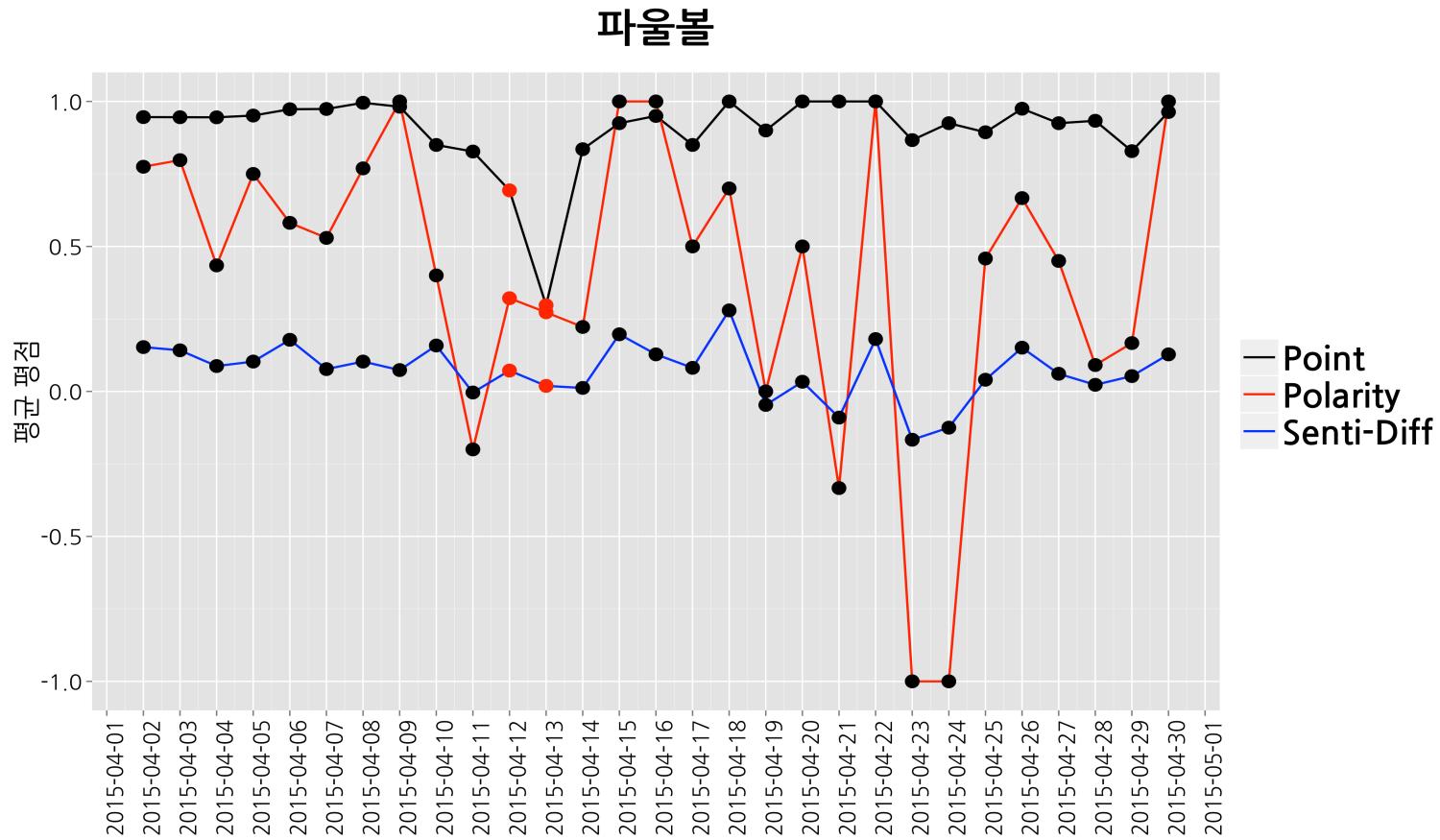
기사본문 댓글 바로가기 등록 : 2015-04-12 22:28 +가 →가 인쇄하기

스포츠 = 김도엽 객원기자 [기사 더보기](#) +



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

# 감정 점수



# WHY?

- ## [1] "한국야구의 진정한 발전을 위해서 이런 셋같은 영감탱이의 우상화, 신격화는 막아야.....영감님 훑는 작자들은 내가 롯데팬이라고 뒤집어씌울듯...ㅉㅉ"
- ## [2] "성큰옹..선수는 그냥 소모품임? 실망..."
- ## [3] "야신은 무슨 잘못도 인정 안하는 노망난 할배지"
- ## [4] "동걸이 인생은 내 알바아니지"
- ## [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 아니라야비의 신이게 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 빗나가고 세번째 공을 던지려 들때 이동걸의 비참한 표정이 뇌리에 깊이 박혀 지워지지가 않는다. 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 김성큰"

# Probabilistic Topic Models

LDA

Blei, David M. and Ng, Andrew and Jordan, Michael. (2003).

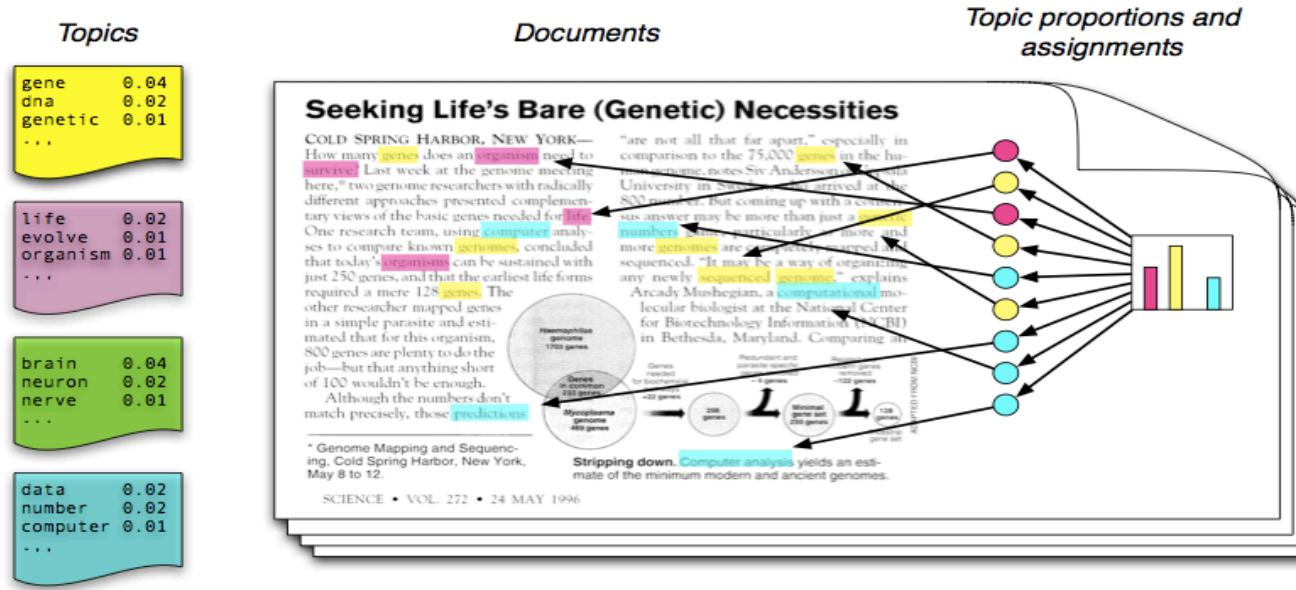
Latent Dirichlet allocation.

*Journal of Machine Learning Research*

참고자료

# LDA

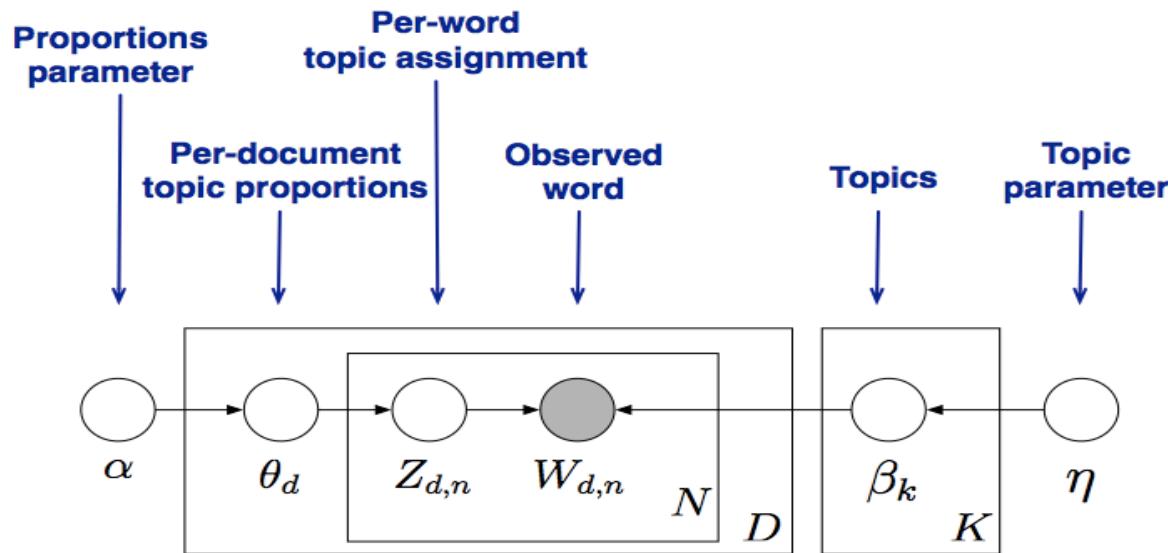
## Latent Dirichlet allocation (LDA)



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

# LDA

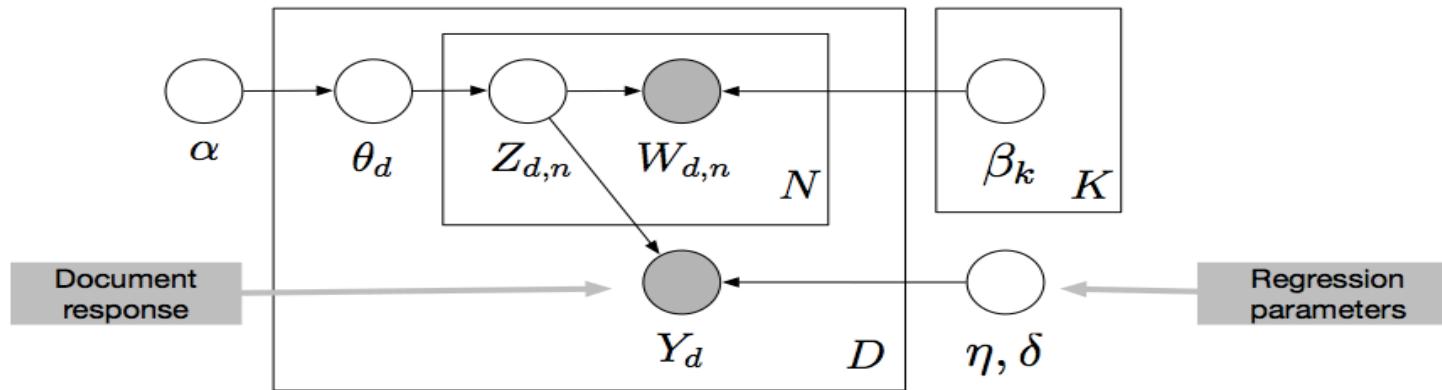
## LDA as a graphical model



$$p(\beta, \theta, \mathbf{z}, \mathbf{w}) = \left( \prod_{i=1}^K p(\beta_i | \eta) \right) \left( \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

# SLDA

## Supervised LDA



- ① Draw topic proportions  $\theta | \alpha \sim \text{Dir}(\alpha)$ .
- ② For each word
  - Draw topic assignment  $z_n | \theta \sim \text{Mult}(\theta)$ .
  - Draw word  $w_n | z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$ .
- ③ Draw response variable  $y | z_{1:N}, \eta, \sigma^2 \sim N(\eta^\top \bar{z}, \sigma^2)$ , where

$$\bar{z} = (1/N) \sum_{n=1}^N z_n.$$

# 대안

## SLDA

Blei and McAuliffe, (2008).

Supervised topic models.

*Advances in Neural Information Processing Systems*, pages 121–128. MIT Press.

## Cross-Validation

- Training Set과 Test Set을 7:3으로 분할

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  |

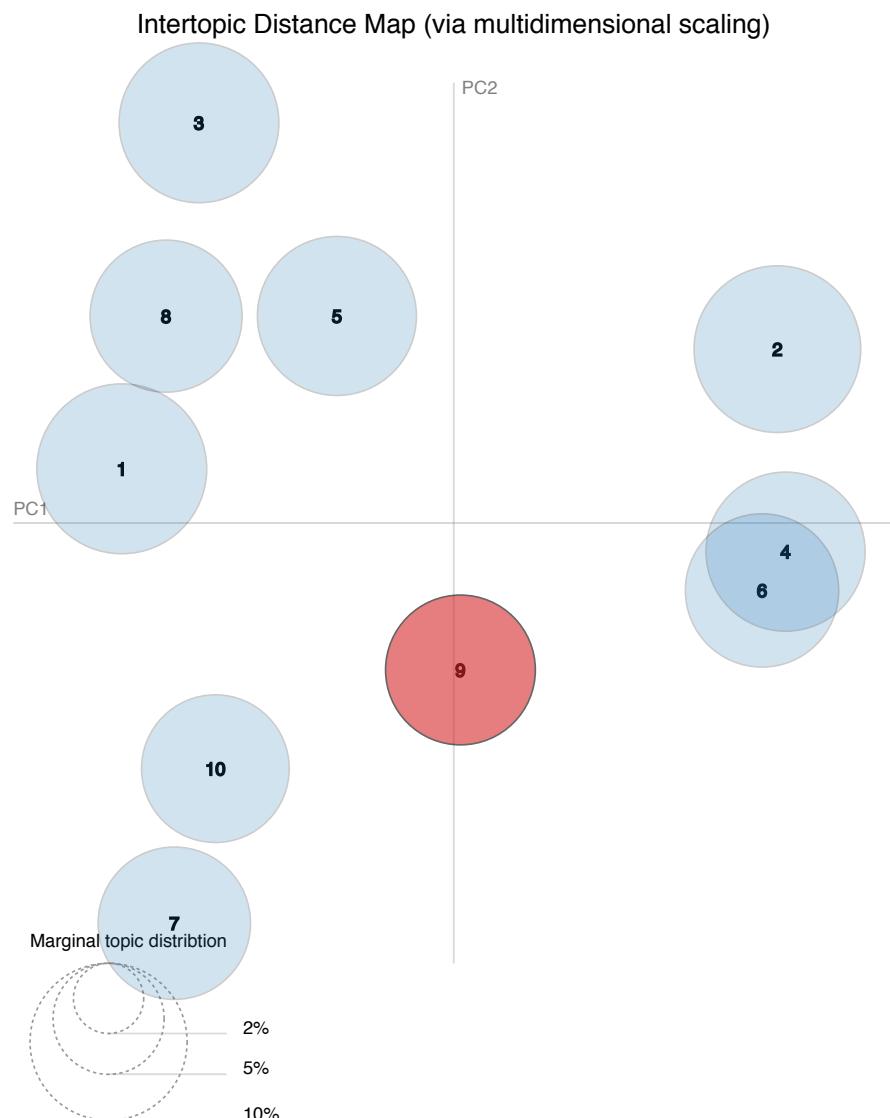
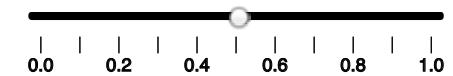
```
library(lda)
library(topicmodels)
library(LDAvis)
library(servr)
```

# Graph

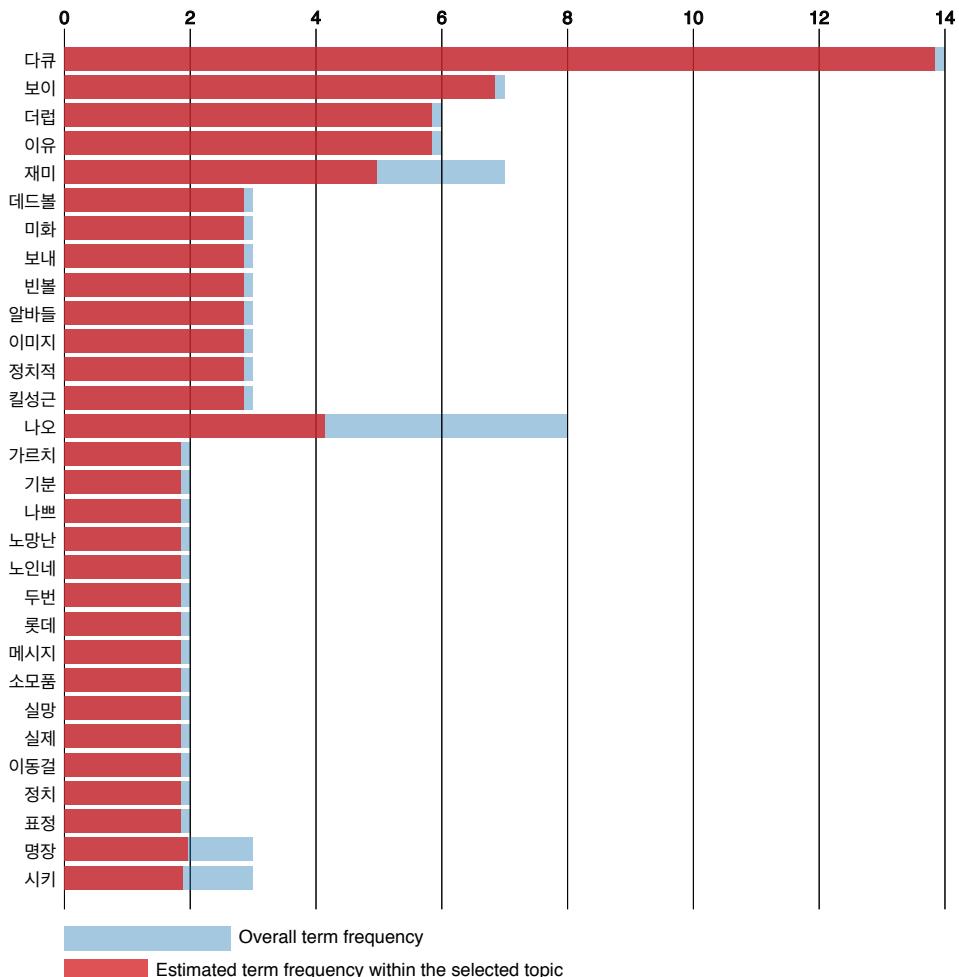
Selected Topic: 9

Slide to adjust relevance metric:<sup>(2)</sup>

$\lambda = 0.51$



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



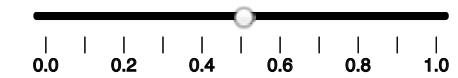
1. saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

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

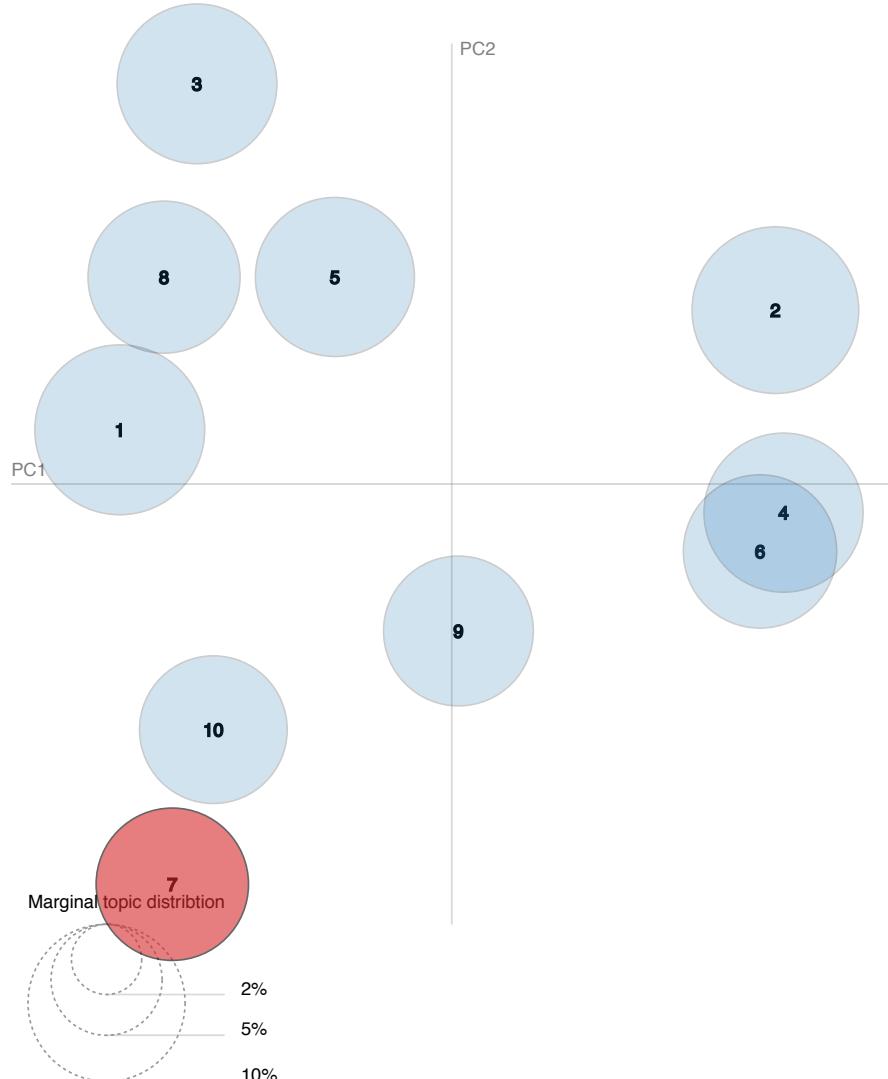
Selected Topic: 7

Slide to adjust relevance metric:<sup>(2)</sup>

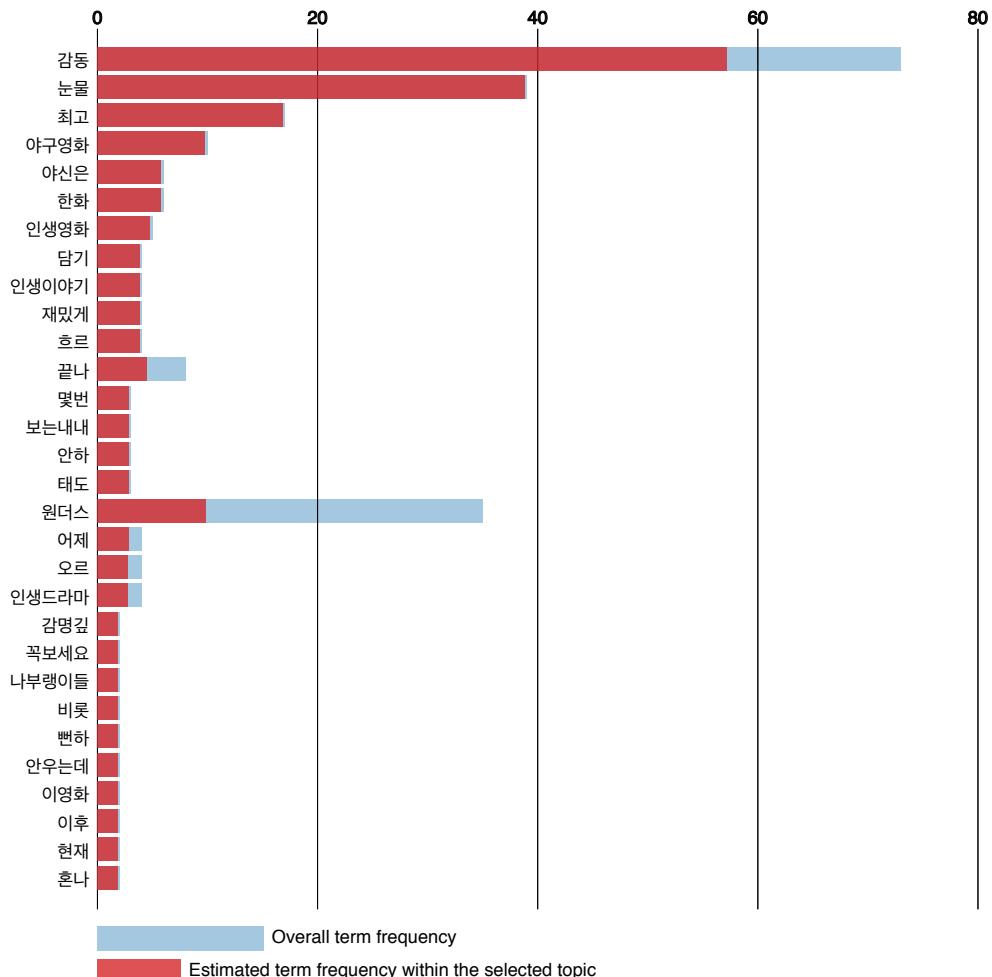
$\lambda = 0.51$



Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 7 (9.4% of tokens)

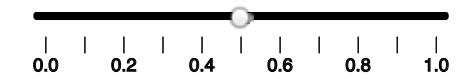


1. saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)  
2. relevance(term w | topic t) =  $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Sievert & Shirley (2014)

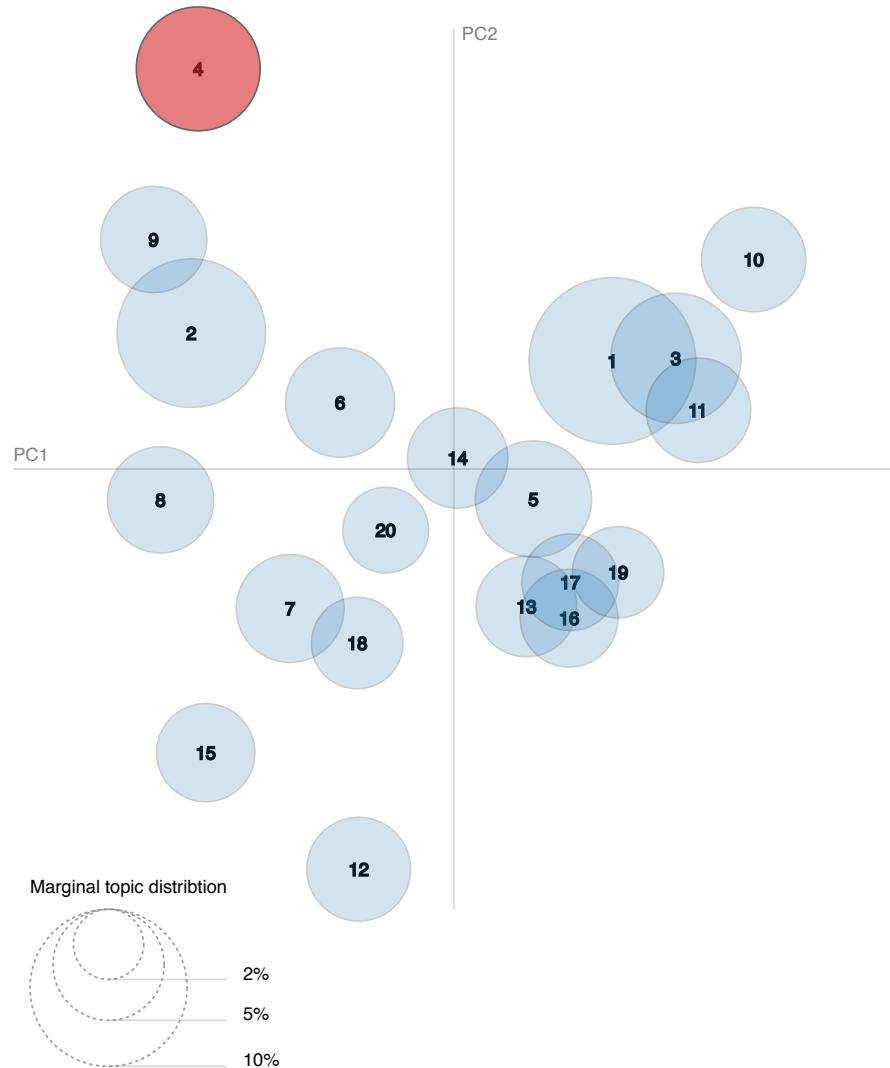
Selected Topic: 4

Slide to adjust relevance metric:<sup>(2)</sup>

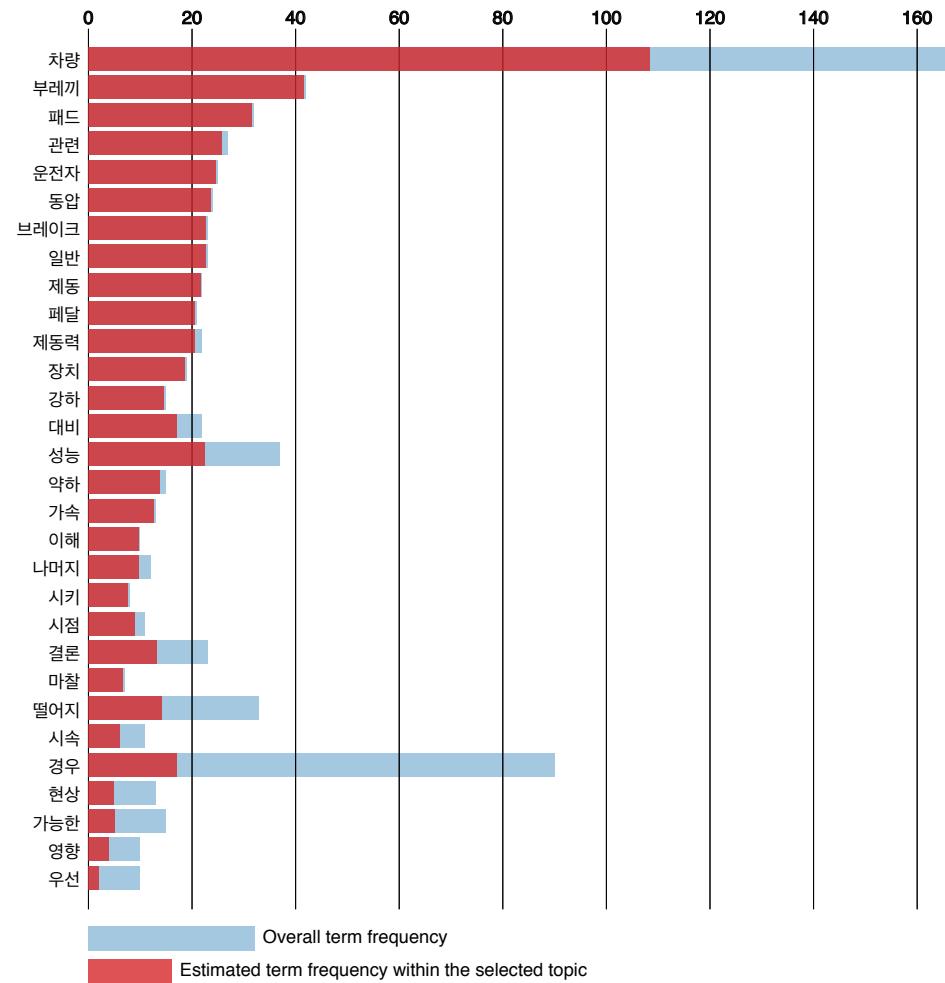
$\lambda = 0.5$



Intertopic Distance Map (via multidimensional scaling)



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



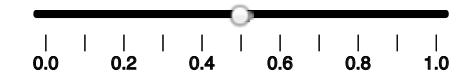
1. saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

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

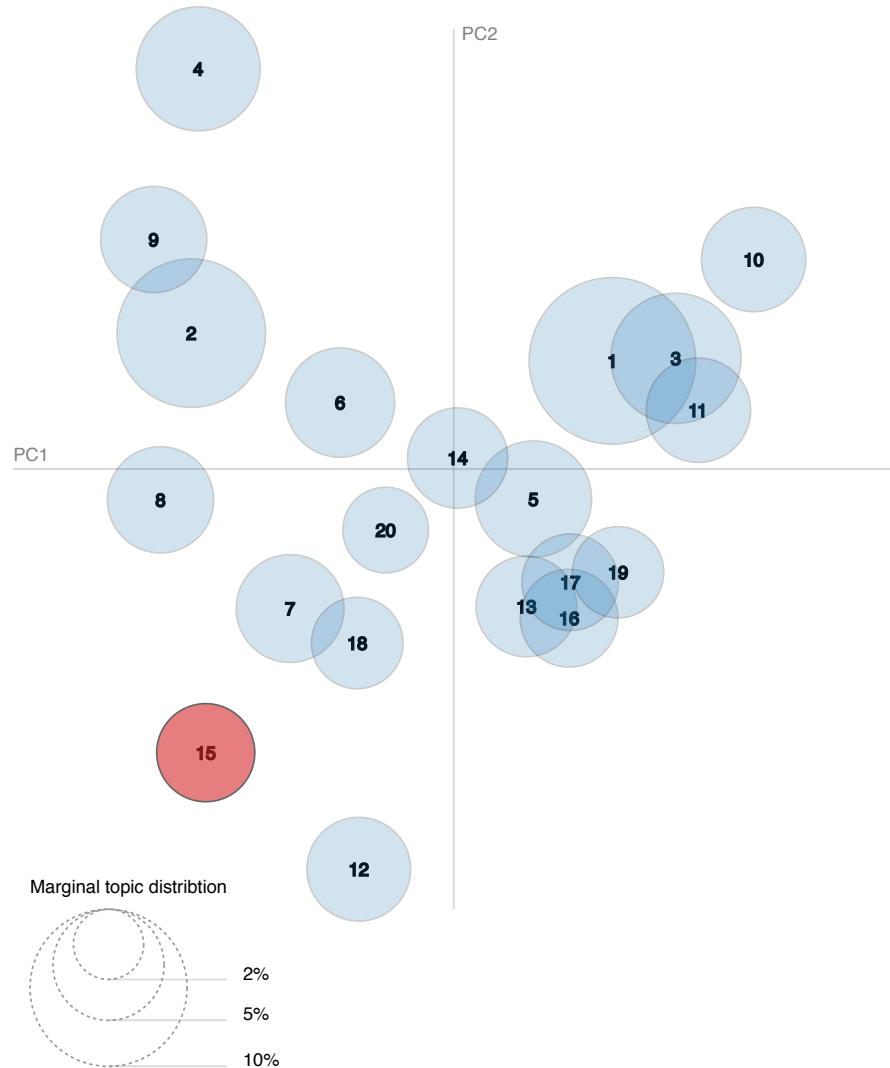
Selected Topic: 15

Slide to adjust relevance metric:<sup>(2)</sup>

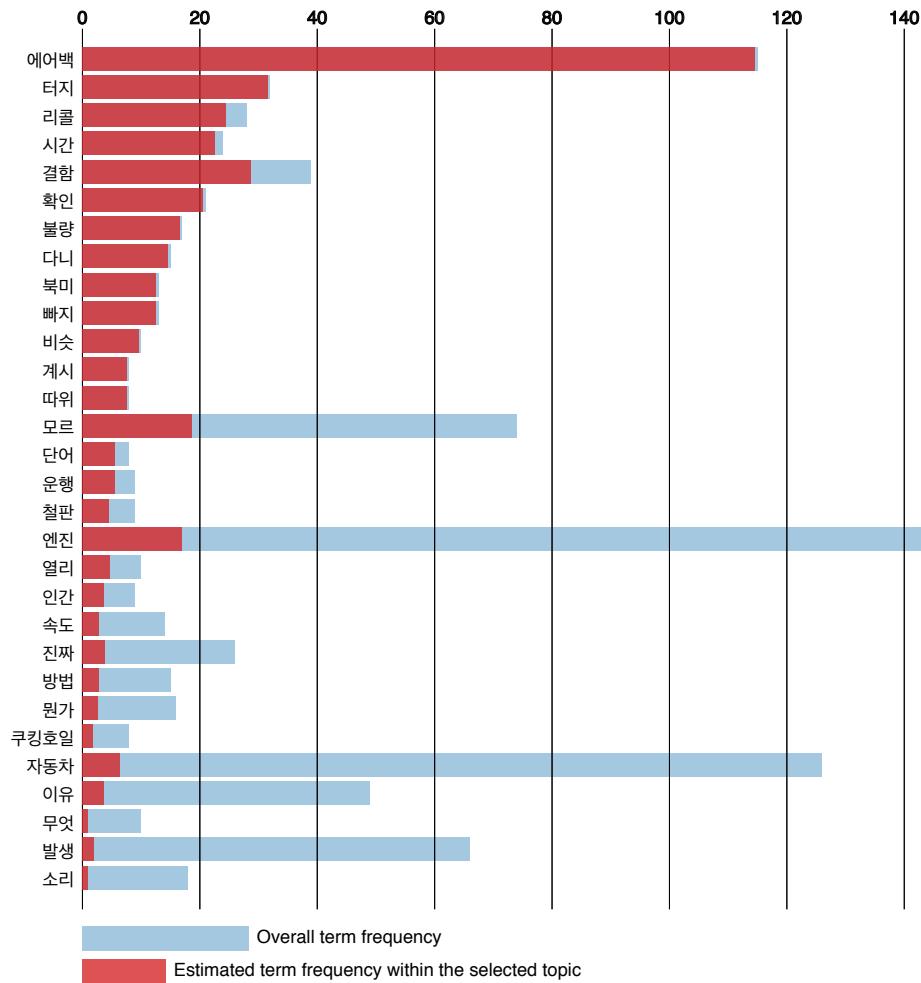
$\lambda = 0.5$



Intertopic Distance Map (via multidimensional scaling)



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



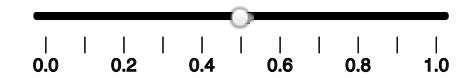
1. saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

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

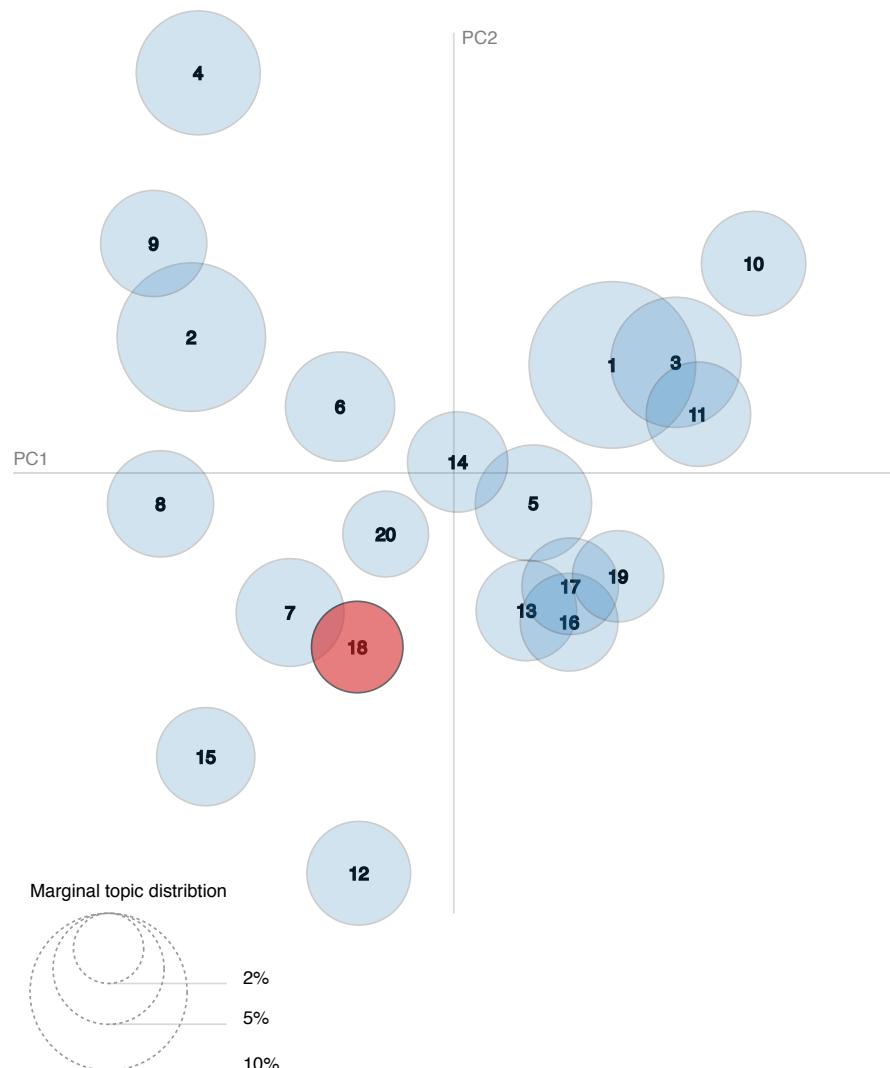
Selected Topic: 18

Slide to adjust relevance metric:<sup>(2)</sup>

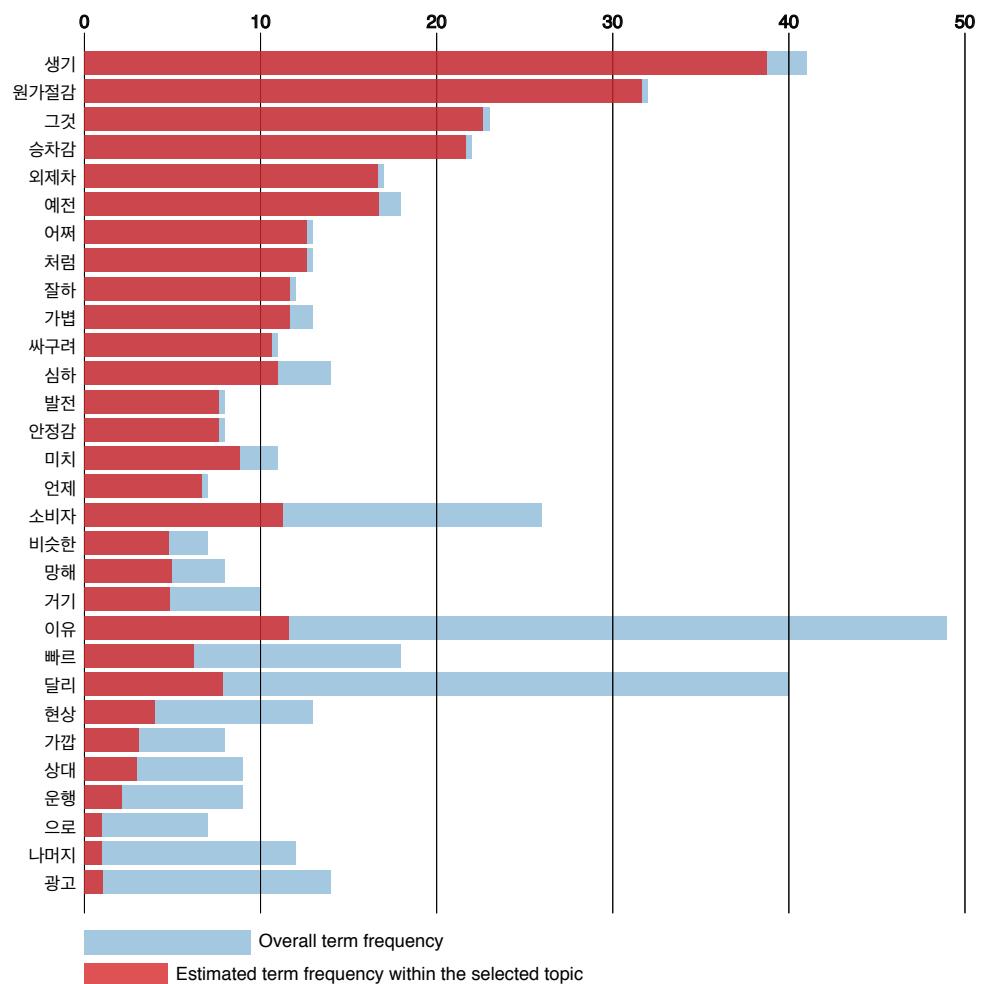
$\lambda = 0.5$



Intertopic Distance Map (via multidimensional scaling)



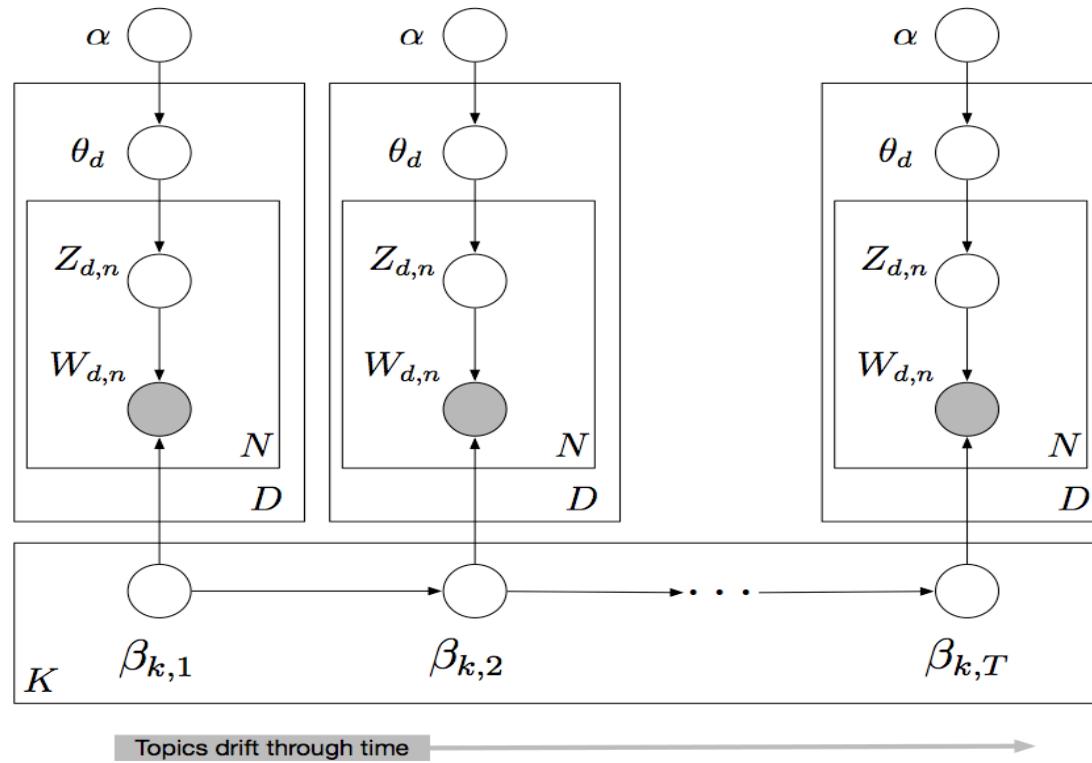
Top-30 Most Relevant Terms for Topic 18 (3.4% of tokens)



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)

# Dynamic Topic Model

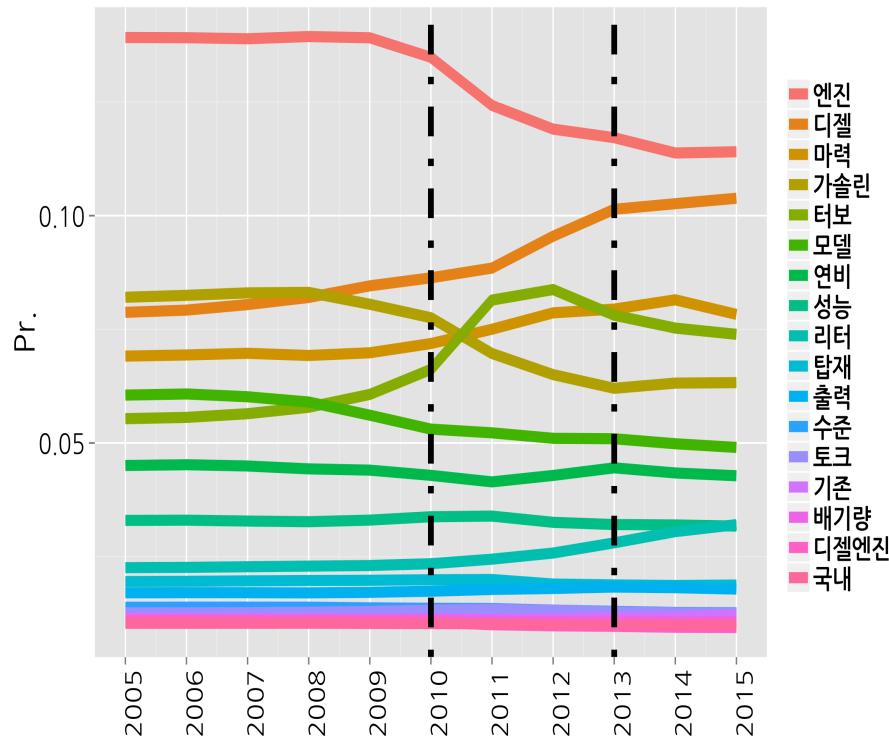


Blei, D. M., & Lafferty, J. D. (2006)

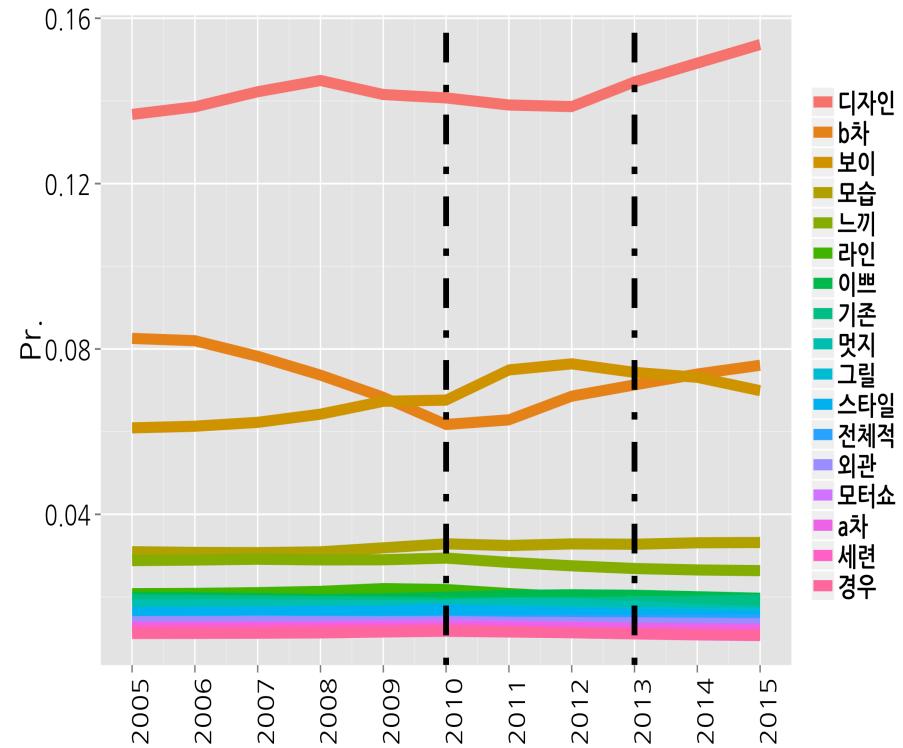
Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM.

# A차종의 장점 요인

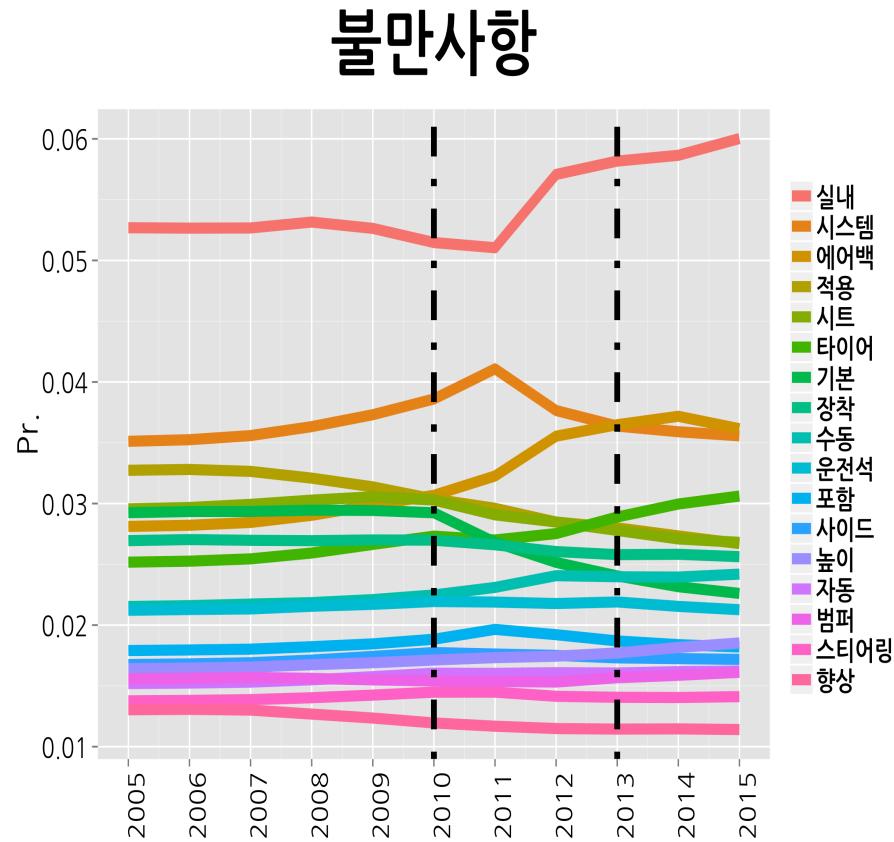
## 마력 & 연비



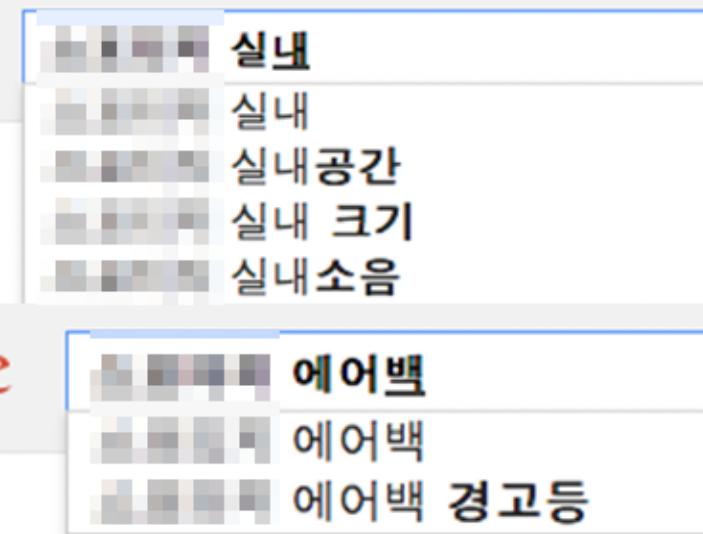
## 디자인 & 세련



# A차종의 불만 요인



Google

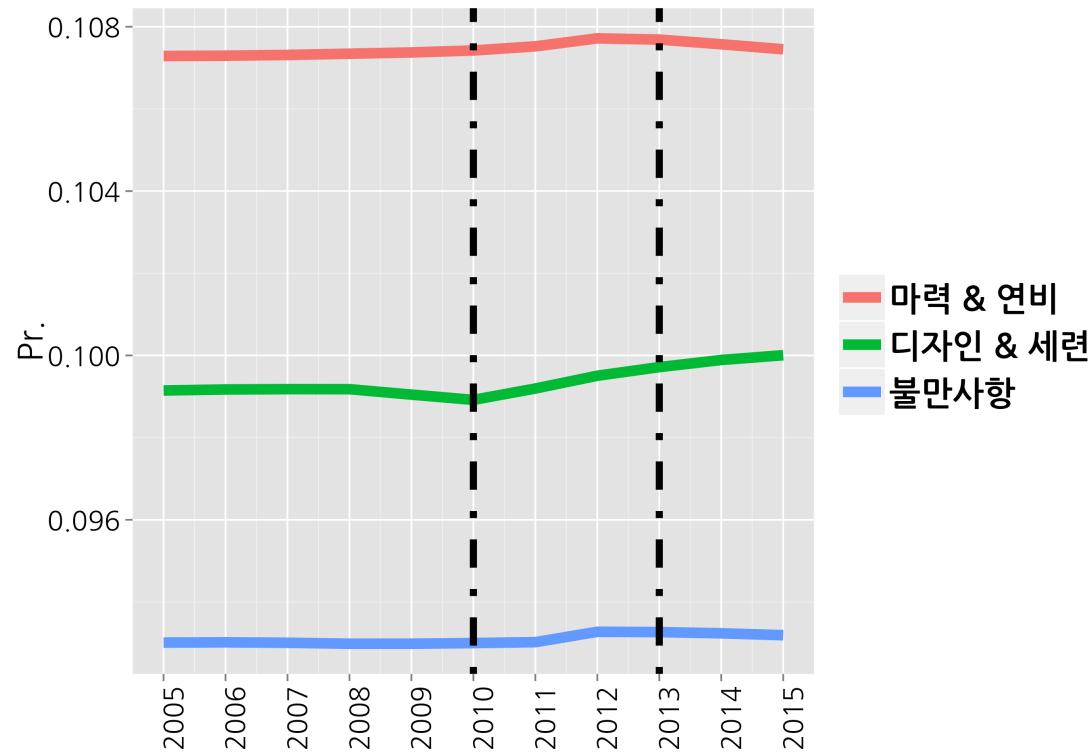


<크지않은 수납공간.....!>

<너무 큰 바램을 가져서는 안되는 시트부분>

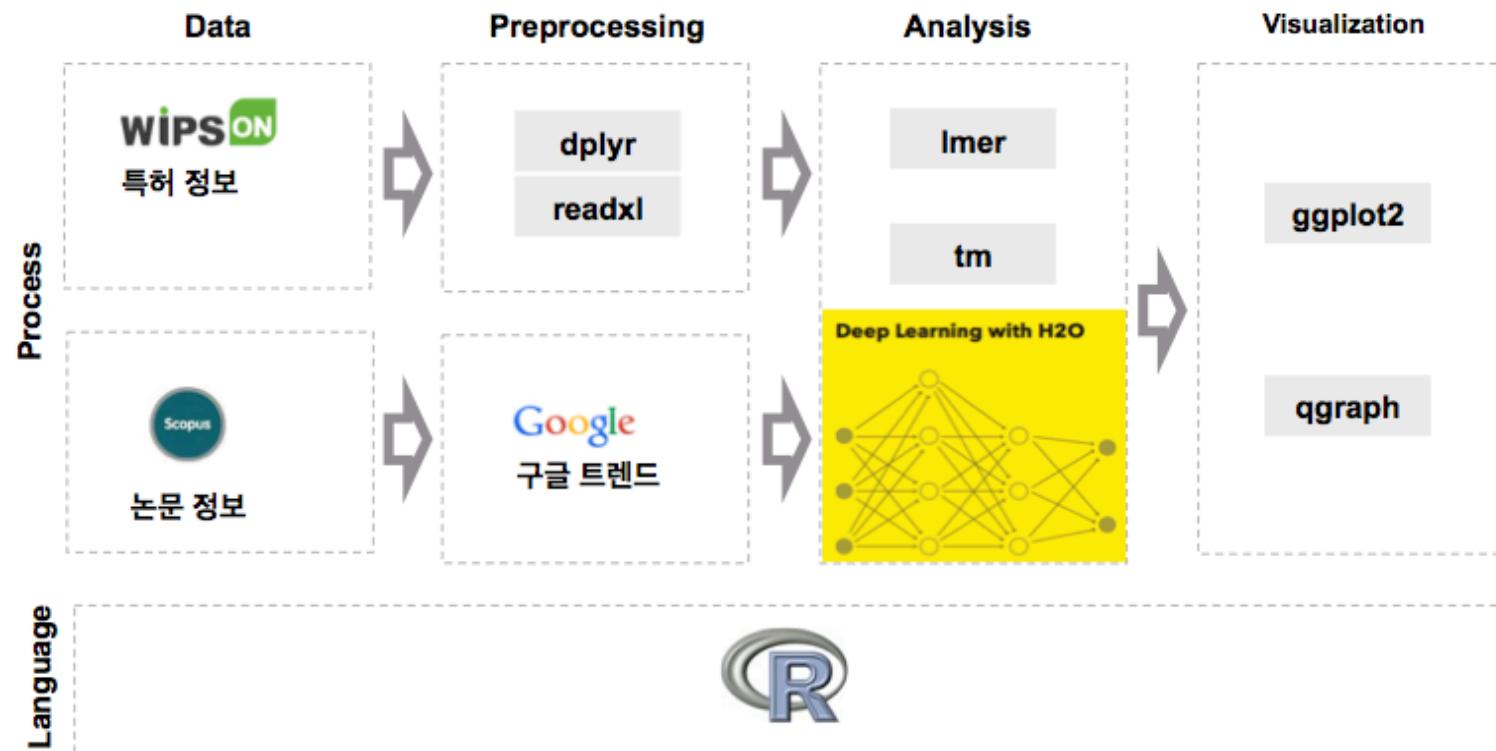
# Marginal Topic Distribution

## Marginal Topic Probability



# Deep-Learning

## 분석과정 (SYSTEM)

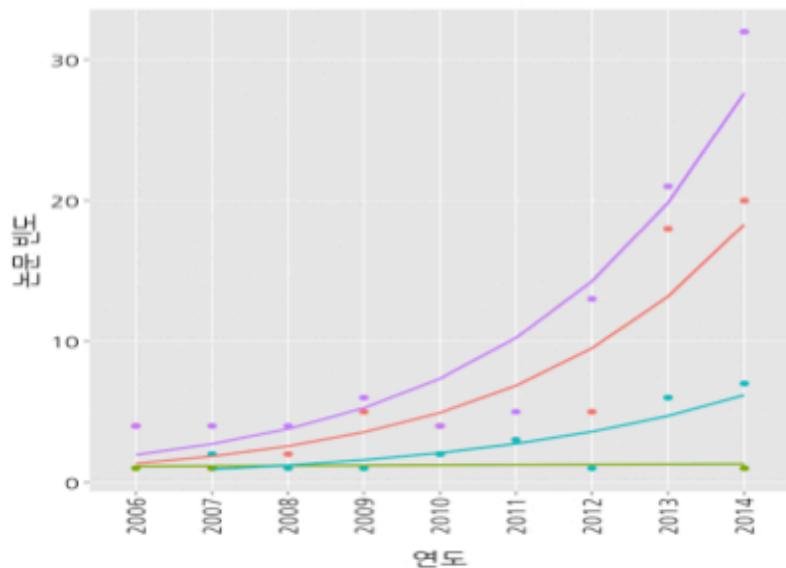


# Deep-Learning

논문 기준 핵심기술 기술통계

|                   | A   | B  | C  | D  | E  | F   | G   | H  |
|-------------------|-----|----|----|----|----|-----|-----|----|
| 비율                | 32% | 0% | 2% | 3% | 0% | 11% | 51% | 0% |
| 피인용<br>지수<br>(평균) | 4.5 | 0  | 0  | 1  | 0  | 3   | 2   | 0  |

연도 별 핵심기술 논문



논문의 경우

IPC 코드 부재로 핵심 기술별 분류가 특허보다  
상대적으로 어려움

→ 미국 특히 Abstract를 *Deep Learning*으로 학습 후  
→ 논문 Abstract를 이용하여 핵심 기술 분류

분류 결과

미국 특허와 마찬가지로  
(G)와 (A), 그리고 (F) 기술개발이 활발함

기술

- 차량 IT 통합 인터페이스 시스템(G)
- 안전 맵 연동 능동 안전시스템(A)
- V2I 기반 횡방향 통합제어 시스템(F)
- 주행지역 상황인지 센서 퓨전 시스템(D)

# 정리

감정 분석

사전 기반

- Wordnet / Sentiwordnet

주제에 따라 달라질 수 있는 감정 단어

- 우리 어머니를 생각하면 **안타까운** 마음이 크다
- 분석을 망쳐서 **안타깝다**

기계 학습 및 통계 모형 기반

- 데이터가 많이 필요
- 복잡한 모형은 구현이 어려움

그 외 이슈들

한국어 형태소 분석

부정어 처리

- 이중부정

단어 순서 및 위치 처리

- N-gram + LDA
- Conditional Random Fields
- Recursive Neural Network
- Recurrent Neural Network
- Convolution Neural Network

# 워크숍 관련 사이트

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

## 코스

### 텍스트 데이터 분석: 키워드를 넘어 토픽으로

현재 수강 중인 코스입니다.

#### 온오프믹스 링크

| 제목                                             | 수강 시작      | 수강 끝       |                      |
|------------------------------------------------|------------|------------|----------------------|
| 텍스트에서 여론과 감정을 발견하기 : R을 이용한 텍스트 데이터 분석         | 2015-05-10 | 2015-06-30 | <a href="#">강의실로</a> |
| 텍스트에서 여론과 감정을 발견하기 : R을 이용한 텍스트 데이터 분석 (05/30) | 2015-05-26 | 2100-01-01 | <a href="#">강의실로</a> |
| R을 이용한 웹 크롤링                                   | 2015-06-10 | 2100-01-01 | <a href="#">강의실로</a> |

감사합니다