Text Documentation Classification

Using Latent Semantic Analysis and Singular Value Decomposition

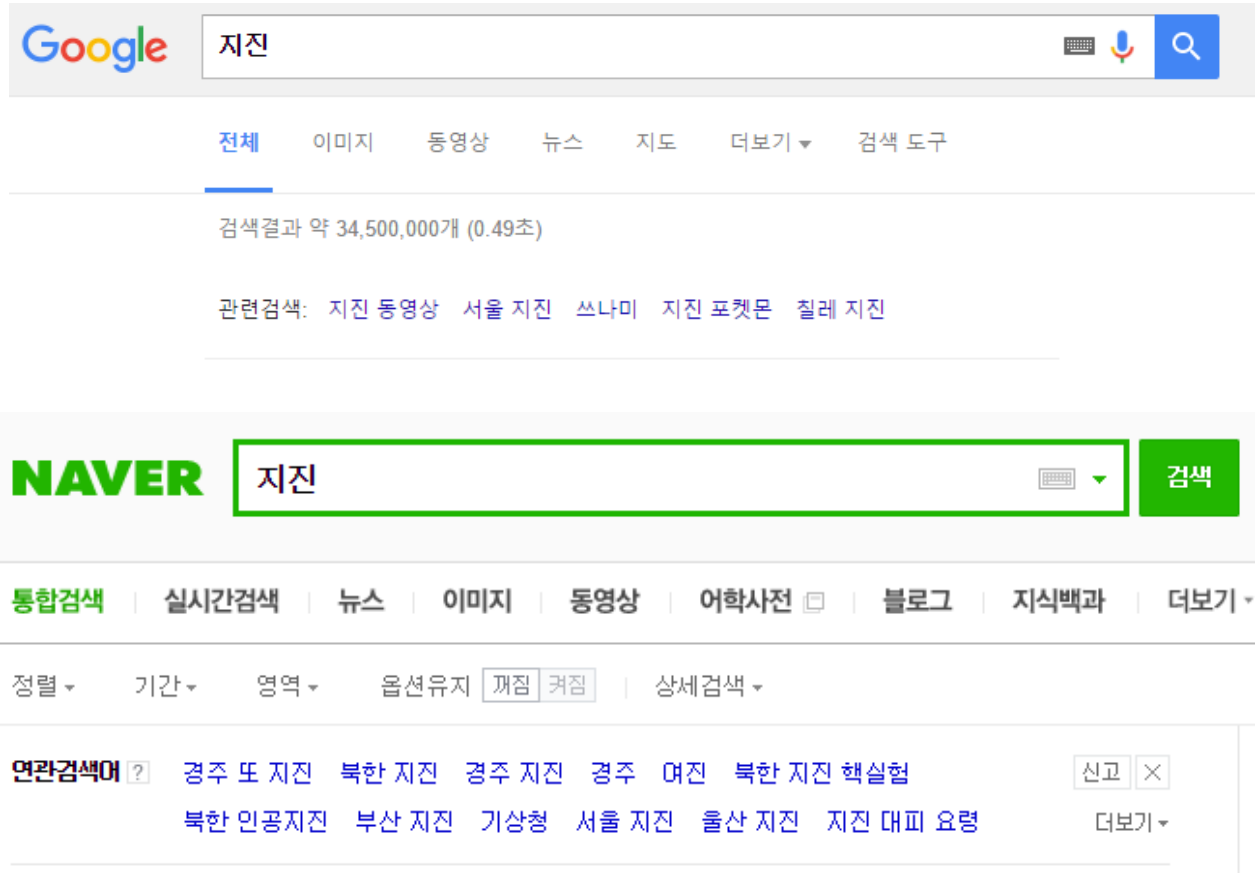
Junyoung Jang

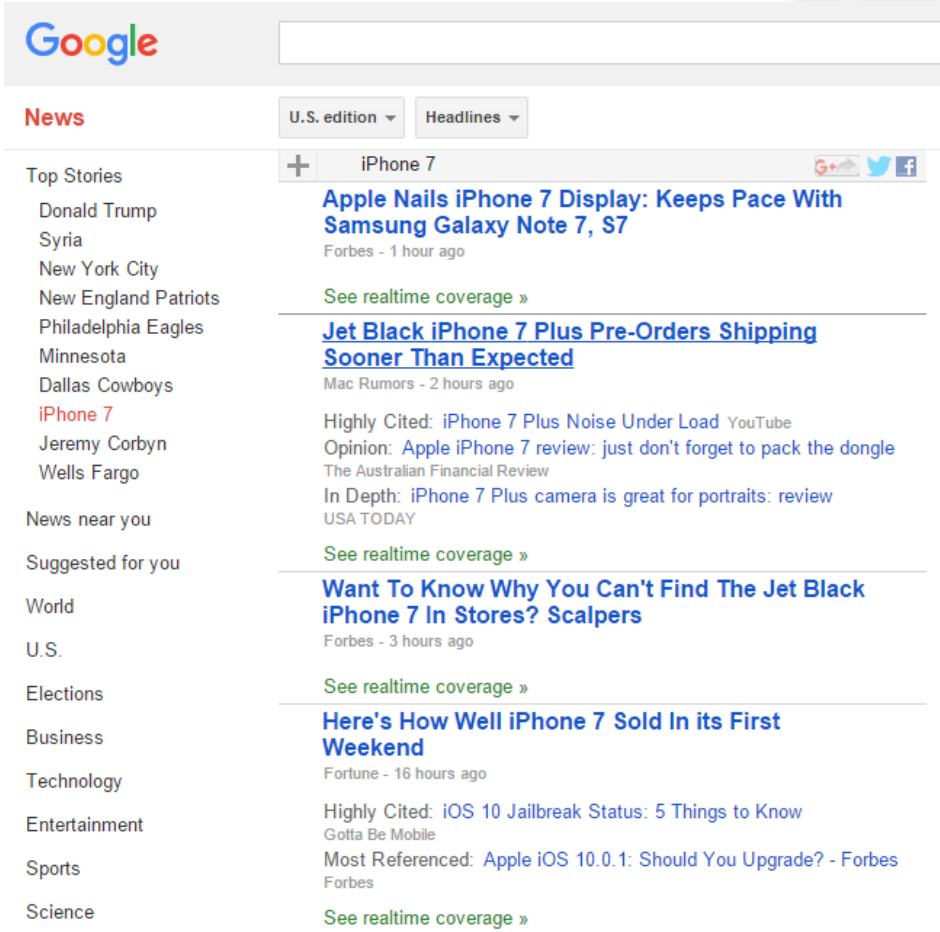
lakino@yonsei.ac.kr
http://cse.yonsei.ac.kr

Department of Computational Science and Engineering
Yonsei University

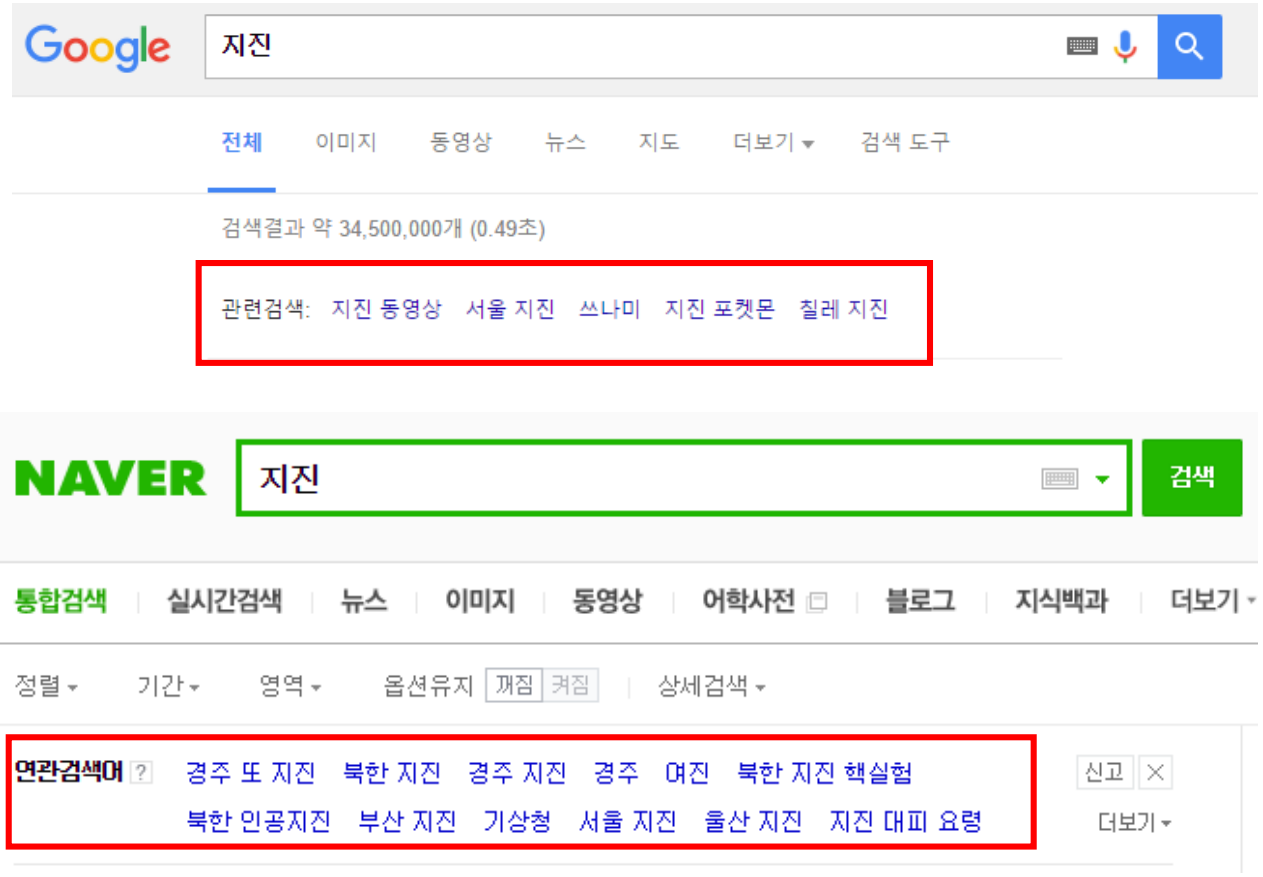
- Motivation
- Latent Semantic Analysis (LSA)
- Singular Value Decomposition (SVD)
- Simulation

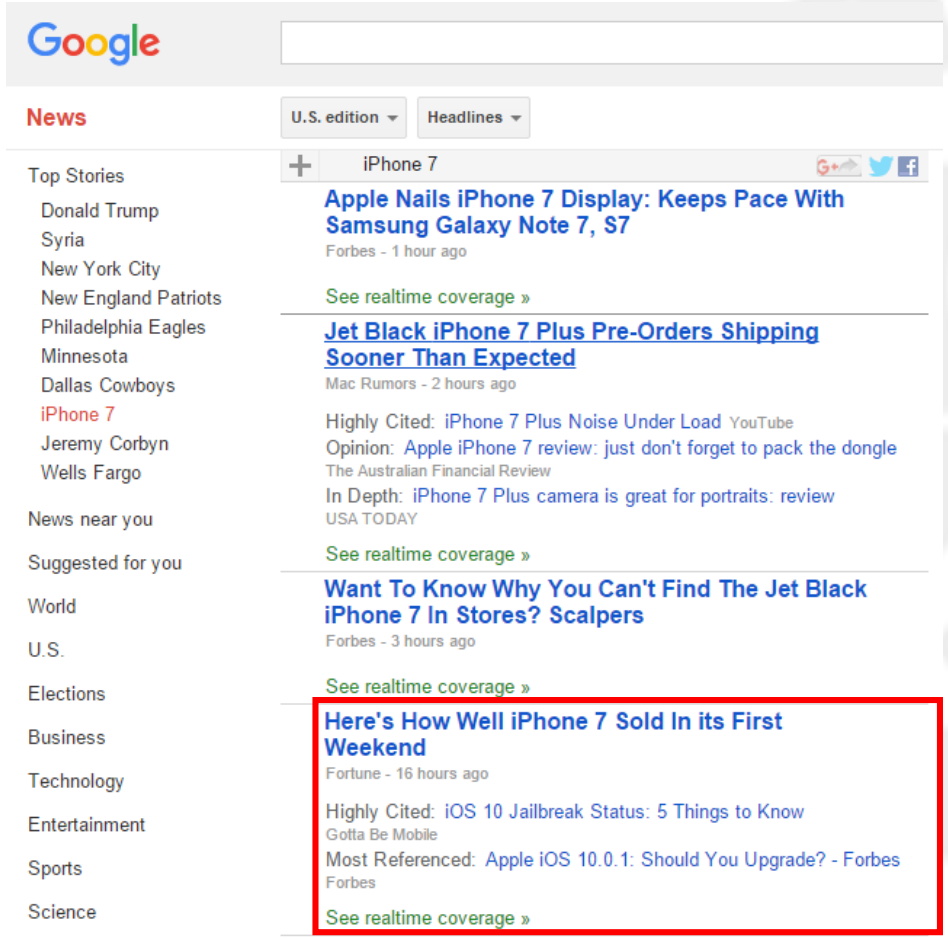
– Motivation


Google
지진
전체
이미지
동영상
뉴스
지도
더보기
검색 도구
검색결과 약 34,500,000개 (0.49초)
관련검색: 지진 동영상 서울 지진 쓰나미 지진 포켓몬 칠레 지진
NAVER
지진
검색
통합검색
실시간검색
뉴스
이미지
동영상
어학사전
블로그
지식백과
더보기
정렬
기간
영역
옵션유지
꺼짐
켜짐
상세검색
연관검색어
경주 또 지진
북한 지진
경주 지진
경주
여진
북한 지진 핵실험
북한 인공지진
부산 지진
기상청
서울 지진
울산 지진
지진 대피 요령
신고
더보기


Google
News
U.S. edition
Headlines
Top Stories
Donald Trump
Syria
New York City
New England Patriots
Philadelphia Eagles
Minnesota
Dallas Cowboys
iPhone 7
Jeremy Corbyn
Wells Fargo
News near you
Suggested for you
World
U.S.
Elections
Business
Technology
Entertainment
Sports
Science
iPhone 7
Apple Nails iPhone 7 Display: Keeps Pace With Samsung Galaxy Note 7, S7
Forbes - 1 hour ago
See realtime coverage »
Jet Black iPhone 7 Plus Pre-Orders Shipping Sooner Than Expected
Mac Rumors - 2 hours ago
Highly Cited: iPhone 7 Plus Noise Under Load YouTube
Opinion: Apple iPhone 7 review: just don't forget to pack the dongle
The Australian Financial Review
In Depth: iPhone 7 Plus camera is great for portraits: review
USA TODAY
See realtime coverage »
Want To Know Why You Can't Find The Jet Black iPhone 7 In Stores? Scalpers
Forbes - 3 hours ago
See realtime coverage »
Here's How Well iPhone 7 Sold In its First Weekend
Fortune - 16 hours ago
Highly Cited: iOS 10 Jailbreak Status: 5 Things to Know
Gotta Be Mobile
Most Referenced: Apple iOS 10.0.1: Should You Upgrade? - Forbes
Forbes
See realtime coverage »

– Motivation


Google
지진
전체 이미지 동영상 뉴스 지도 더보기 검색 도구
검색결과 약 34,500,000개 (0.49초)
관련검색: 지진 동영상 서울 지진 쓰나미 지진 포켓몬 칠레 지진
NAVER
지진
검색
통합검색 실시간검색 뉴스 이미지 동영상 어학사전 블로그 지식백과 더보기
정렬 기간 영역 옵션유지 꺼짐 켜짐 상세검색
연관검색어 경주 또 지진 북한 지진 경주 지진 경주 여진 북한 지진 핵실험
북한 인공지진 부산 지진 기상청 서울 지진 울산 지진 지진 대피 요령
신고
더보기


Google
News
U.S. edition
Headlines
Top Stories
Donald Trump
Syria
New York City
New England Patriots
Philadelphia Eagles
Minnesota
Dallas Cowboys
iPhone 7
Jeremy Corbyn
Wells Fargo
News near you
Suggested for you
World
U.S.
Elections
Business
Technology
Entertainment
Sports
Science
iPhone 7
Apple Nails iPhone 7 Display: Keeps Pace With Samsung Galaxy Note 7, S7
Forbes - 1 hour ago
See realtime coverage »
Jet Black iPhone 7 Plus Pre-Orders Shipping Sooner Than Expected
Mac Rumors - 2 hours ago
Highly Cited: iPhone 7 Plus Noise Under Load YouTube
Opinion: Apple iPhone 7 review: just don't forget to pack the dongle
The Australian Financial Review
In Depth: iPhone 7 Plus camera is great for portraits: review
USA TODAY
See realtime coverage »
Want To Know Why You Can't Find The Jet Black iPhone 7 In Stores? Scalpers
Forbes - 3 hours ago
See realtime coverage »
Here's How Well iPhone 7 Sold In its First Weekend
Fortune - 16 hours ago
Highly Cited: iOS 10 Jailbreak Status: 5 Things to Know
Gotta Be Mobile
Most Referenced: Apple iOS 10.0.1: Should You Upgrade? - Forbes
Forbes
See realtime coverage »

– Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a method for discovering hidden concepts in document(sentence) and term(word).

: An Example

d_1 : Romeo and Juliet.

d_2 : Juliet: O happy dagger!

d_3 : Romeo died by dagger.

d_4 : "Live free or die", that's the New-Hampshire's motto

d_5 : Did you know, New-Hampshire is in New England.

– Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a method for discovering hidden concepts in document(sentence) and term(word).

: An Example

d_1 : Romeo and Juliet.

d_2 : Juliet: O happy dagger!

d_3 : Romeo died by dagger.

d_4 : "Live free or die", that's the New-Hampshire's motto

d_5 : Did you know, New-Hampshire is in New England.

$\Rightarrow$ $Ranked\ Top$: $d_3 > d_2 > d_4 > d_1 > d_5$

– Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a method for discovering hidden concepts in document(sentence) and term(word).

: An Example

d_1 : Romeo ~~and~~ Juliet.

d_2 : Juliet: ~~O~~ happy dagger!

d_3 : Romeo died ~~by~~ dagger.

d_4 : "Live free ~~or~~ die", ~~that's the~~ New-Hampshire's ~~motto~~

d_5 : ~~Did you know,~~ New-Hampshire ~~is in New England.~~

$$A = \begin{array}{r} \\ \text{romeo} \\ \text{juliet} \\ \text{happy} \\ \text{dagger} \\ \text{live} \\ \text{die} \\ \text{free} \\ \text{new-hampshire} \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & d_4 & d_5 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

where i=1,2, … , #rank of matrix

– Singular Value Decomposition (SVD)

$$A = \begin{array}{r} \\ \text{romeo} \\ \text{juliet} \\ \text{happy} \\ \text{dagger} \\ \text{live} \\ \text{die} \\ \text{free} \\ \text{new-hampshire} \end{array} \begin{array}{c} \begin{array}{ccccc} d_1 & d_2 & d_3 & d_4 & d_5 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

: Doc-doc Matrix $B = A^T A$
: Term-term Matrix $C = AA^T$

using SVD, $A = U\Sigma V^T$

, $BV = (A^T A)V = \sigma_i^2 V$

, $CU = (AA^T)U = \sigma_i^2 U$

where i=1,2, … , #rank of matrix.

V and U is eigenvector of B, C.

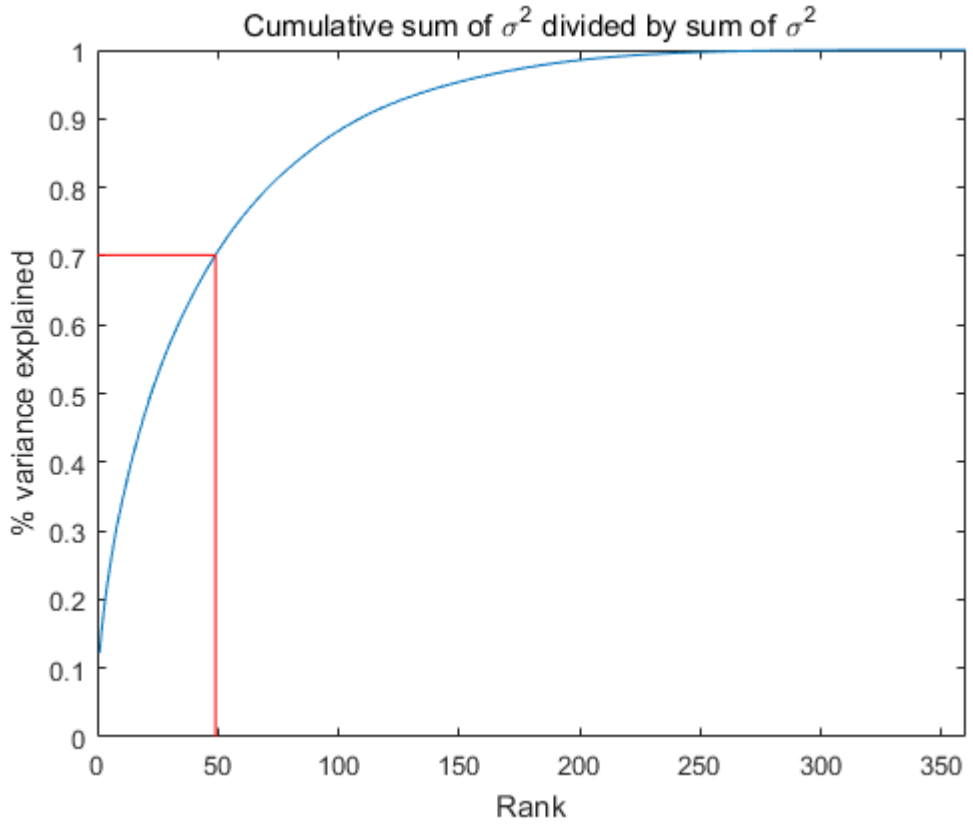
Namely, the terms(word) are represented by the row vectors : $\boldsymbol{U\Sigma}$

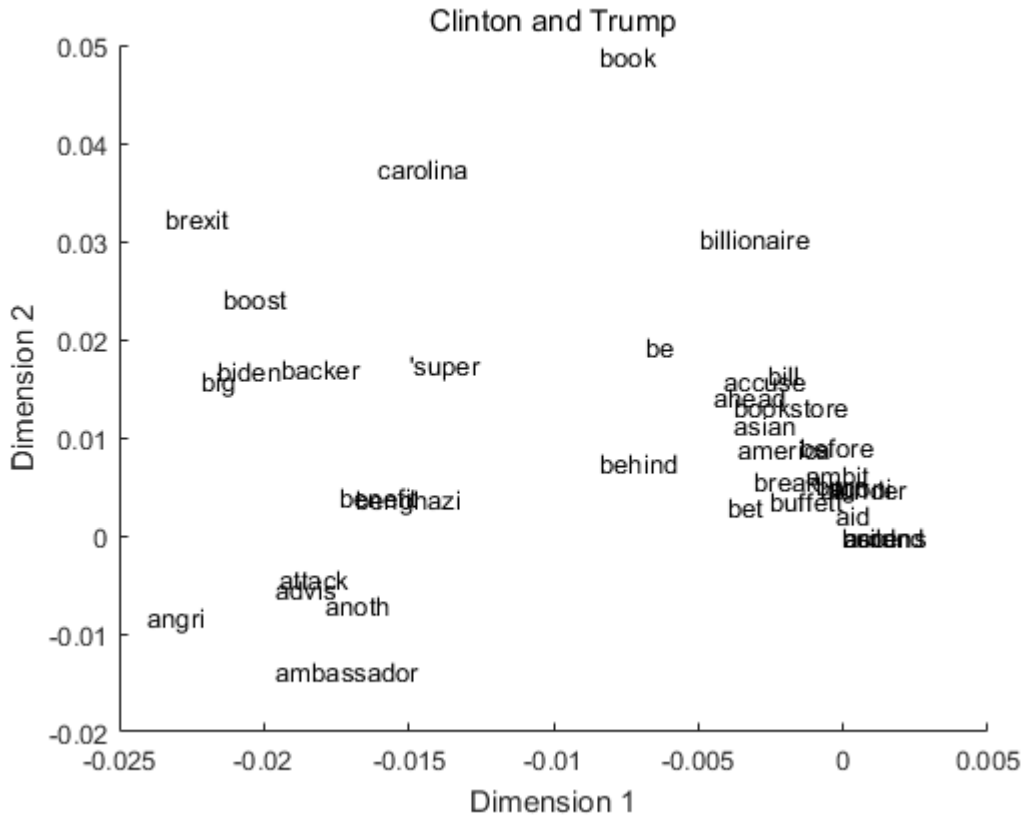
Whereas the documents(sentence) by the column vectors : $\boldsymbol{\Sigma V^T}$

– Simulation

: Web Crawler = The Korea Herald (http://www.koreaherald.com/)

: **#article** = 500





– Simulation

: Web Crawler = The Korea Herald (http://www.koreaherald.com/)
: **#article** = 500

'뱅가지 청문회'로 회생하는 힐러리, 온라인 후원금도 쇄도

[헤럴드경제=김태열 기자]미국 하원의 22일(현지시간) **뱅가지** 특위 청문회는 결과적으로 민주당 유력 대선주자인 힐러리 클린턴 전 국무장관에게 여러모로 도움이 된 자리였다는 분석이...큰 걸림돌 하나를 제거했을 뿐 아니라 청문회 당일 온라인 후원금이 쇄도했기 때문이다. **뱅가지** 청문회는 외견상 2012년 9월 리비아 **뱅가지**에서 발생한 미 영사관 습격사건을 다루기... [2015/10/24 09:01]

[사우스캐롤라이나 민주 경선] 클린턴, 사우스캐롤라이나 압승…'슈퍼 화요일'...

뉴시스 | 2016.02.28. | 네이버뉴스

도착한 민주당 경선주자인 버니 샌더스 상원의원이 **사우스캐롤라이나** 경선에 대한 입장을 표명하고 있다. 2016.02.28 **클린턴** 후보는 **사우스캐롤라이나** 출구조사에서 압승했다는 결과가 발표된 승리연설을 했다. ksk@newsis.com

美 CEO, 억만장자들이 트럼프를 대하는 태도는…경멸 · 무시 · 두려움

[헤럴드경제=신수정 기자] 미국 내 다수의 최고경영자(CEO)들은 사석에서 도널드 **트럼프** 공화당 대선후보를 경멸하고 있다. 하지만 보복을 당할지도 모른다는 두려움때문에 공개적으로...도)'라는 점을 강조했다. 이유는 '두려움'이 대부분이다. IT 기업들은 '애플'처럼 **트럼프**로부터 공개적인 공격을 당할까봐 두려워한다. 은행들은 **트럼프**가 뭘가를 무너뜨릴까봐 두려... [2016/09/21 10:50]

트럼프 신간이 美서점 유머 섹션에…"당신을 웃길 책"

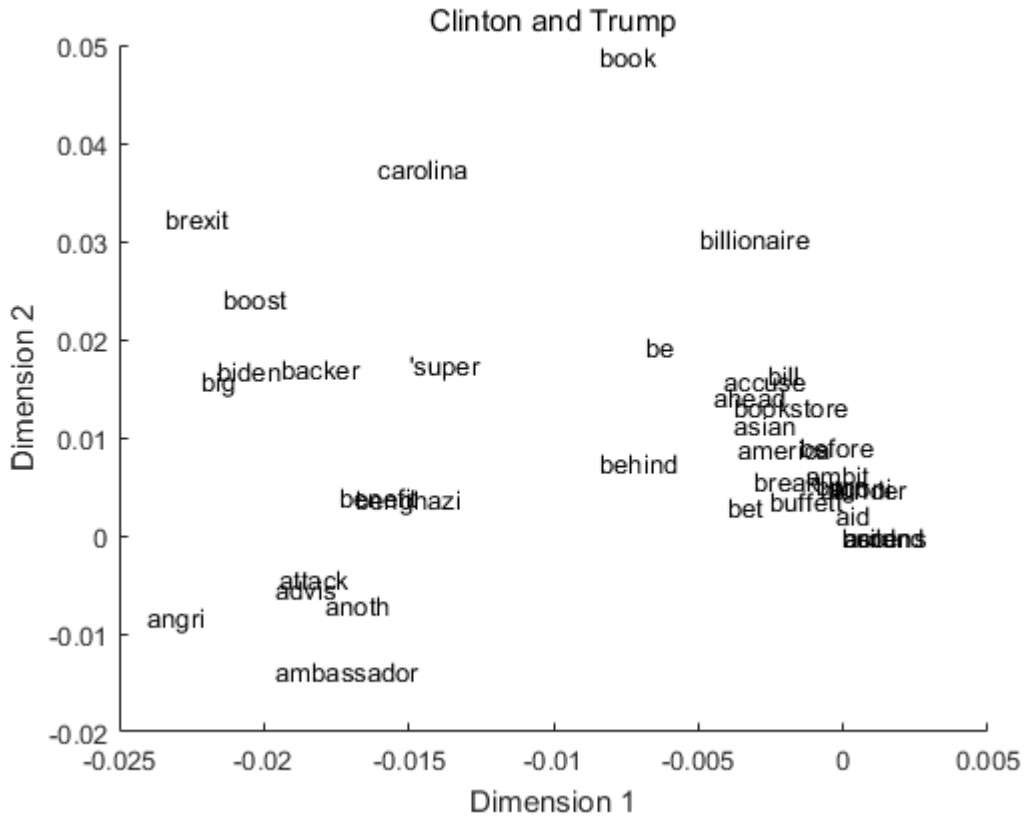
뉴시스 | 2015.12.30. | 네이버뉴스

29일(현지시간) 보수 온라인매체 '데일리 콜러'에 따르면 마이애미주의 한 대형**서점** 체인 반스 앤 노블은 **트럼프** 후보의 저서 '불구가 된 미국: 어떻게 미국을 다시 위대하게 만들 것인가(Crippled America: How to Make...

꿩 먹고 알 먹은 트럼프…선거 자금으로 자기 책 대량 구매

중앙일보 | 2016.08.25. | 네이버뉴스

미 온라인매체 데일리비스트는 연방선거위원회(FEC)를 인용, 지난 5월 10일 **트럼프** 측이 반스앤노블 **서점**에 5만 5055달러(약 6100만원)을 지급했다고 24일(현지시간) 보도했다. 지난해 **트럼프**가 펴낸 자서전...



– Simulation

: Web Crawler = The Korea Herald (http://www.koreaherald.com/)
: **#article** = 500

'벵가지 청문회'로 회생하는 힐러리, 온라인 후원금도 쇄도

[헤럴드경제=김태열 기자]미국 하원의 22일(현지시간) **벵가지** 특위 청문회는 결과적으로 민주당 유력 대선주자인 힐러리 클린턴 전 국무장관에게 여러모로 도움이 된 자리였다는 분석이...큰 걸림돌 하나를 제거했을 뿐 아니라 청문회 당일 온라인 후원금이 쇄도했기 때문이다. **벵가지** 청문회는 외견상 2012년 9월 리비아 **벵가지**에서 발생한 미 영사관 습격사건을 다루기... [2015/10/24 09:01]

[사우스캐롤라이나 민주 경선] 클린턴, 사우스캐롤라이나 압승…'슈퍼 화요일'...

뉴시스 | 2016.02.28. | 네이버뉴스

도착한 민주당 경선주자인 버니 샌더스 상원의원이 **사우스캐롤라이나** 경선에 대한 입장을 표명하고 있다. 2016.02.28 **클린턴** 후보는 **사우스캐롤라이나** 출구조사에서 압승했다는 결과가 발표된 승리연설을 했다. ksk@newsis.com

美 CEO, 억만장자들이 트럼프를 대하는 태도는…경멸 · 무시 · 두려움

[헤럴드경제=신수정 기자] 미국 내 다수의 최고경영자(CEO)들은 사석에서 도널드 **트럼프** 공화당 대선후보를 경멸하고 있다. 하지만 보복을 당할지도 모른다는 두려움때문에 공개적으로...도)'라는 점을 강조했다. 이유는 '두려움'이 대부분이다. IT 기업들은 '애플'처럼 **트럼프**로부터 공개적인 공격을 당할까봐 두려워한다. 은행들은 **트럼프**가 뭘가를 무너뜨릴까봐 두려... [2016/09/21 10:50]

트럼프 신간이 **美서점** 유머 섹션에…"당신을 웃길 책"

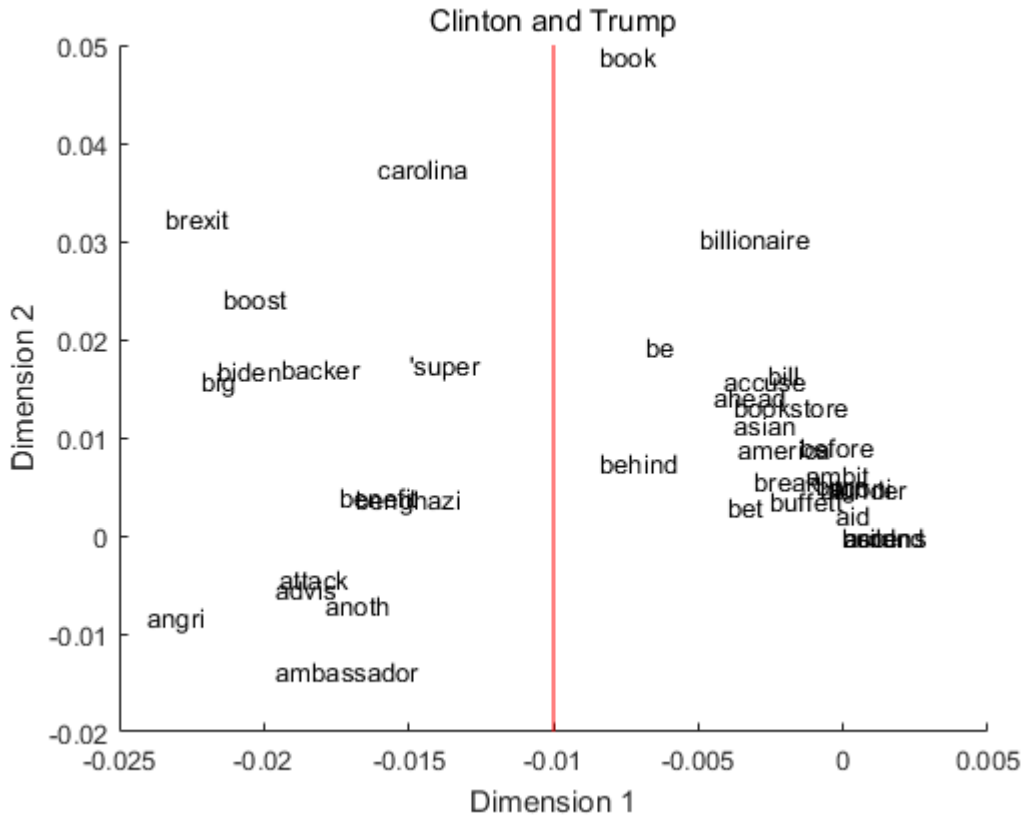
뉴시스 | 2015.12.30. | 네이버뉴스

29일(현지시간) 보수 온라인매체 '데일리 콜러'에 따르면 마이애미주의 한 대형**서점** 체인 반스 앤 노블은 **트럼프** 후보의 저서 '불구가 된 미국: 어떻게 미국을 다시 위대하게 만들 것인가(Crippled America: How to Make...

꿩 먹고 알 먹은 트럼프…선거 자금으로 자기 책 대량 구매

중앙일보 | 2016.08.25. | 네이버뉴스

미 온라인매체 데일리비스트는 연방선거위원회(FEC)를 인용, 지난 5월 10일 **트럼프** 측이 반스앤노블 **서점**에 5만 5055달러(약 6100만원)을 지급했다고 24일(현지시간) 보도했다. 지난해 **트럼프**가 펴낸 자서전...



Reference

1. Thomas K Landauer , Peter W. Foltz & Darrell Laham (1998). "An introduction to latent semantic analysis". Discourse Processes, Vol. 25.
2. Dongho Shin (2000). "A Study on Content-Based Information Retrieval System using LSA". Thesis.
3. Rafael E. Banchs (2012). "Text Mining with MATLAB". Springer