

# Data 처리 & 토픽모델링

주재걸 교수님 연구실  
DAVIAN Lab.

2019. 07. 01

최성재

requirements.txt: <https://tinyurl.com/davian-require>

ipython notebook: <https://tinyurl.com/davian-ipynb>

anaconda: <https://www.anaconda.com/distribution/>

```
C:Users> pip install -r requirements.txt
```

```
C:Users> jupyter lab
```

# **Contents**

**1. Crawling**

**2. Data processing**

**3. Term-document matrix**

**4. Topic modeling (NMF)**

**5. Visualization**

## Library

- newspaper
  - 뉴스 페이지에서 자동으로 제목, 본문 사진, 영상 등을 추출하는 라이브러리
- konlpy
  - 한국어 형태소 분석기 (Hannanum, Kkma, Mecab, Twitter 등)
- bokeh
  - visualization tool

# Crawling

이러한 기사를 수천, 수만개 이상 수집해보자!

연관검색어 ? edg skt skt 고객센터 sk skt t1 tworld sk t월드 2017 롤드컵 skt 티월드  
skt 역전 입롤 롤드컵 skt edg skt vs edg skt 번호변경 skt 대리점 skt 멤버쉽 신고 X 더보기

뉴스 1-10 / 17건

✔ 관련도순 ✔ 최신순 ✔ 오래된순

뉴스검색 가이드

검색결과 자동고침 시작 ▶

**SKT, 장애청소년 ICT 메이커톤 대회 개최** 머니투데이 | 2017.10.01. | 네이버뉴스 |   
SK텔레콤은 한국장애인단체총연맹과 공동으로 지난달 28~29일 이틀간 '2017 SKT 장애청소년 ICT(정보통신기술) 메이커톤 대회'를 개최했다고 1일 밝혔다. 이번 대회는 장애청소년 112명으로 이뤄진 28개팀이 참가해...

↳ SKT, 장애 청소년과 메이커톤 대회 개최 ZDNet Korea | 2017.10.01. | 네이버뉴스  
↳ SKT 장애청소년 ICT메이커톤 대회 연합뉴스 | 2017.10.01. | 네이버뉴스  
↳ SKT, '2017 ICT메이커톤 대회' 개최 뉴스워커 | 2017.10.01.  
↳ SKT, '2017 ICT메이커톤 대회' 개최 아크로팬 | 2017.10.01.  
관련뉴스 24건 전체보기 >

**SKT, '티움'에 국내 거주 외국인 유학생 초청** ZDNet Korea | 2017.10.01. | 네이버뉴스 |   
외국인 유학생 20명에 티움 투어 기회 제공 (지디넷코리아=박수형 기자)SK텔레콤(대표 박정호)은 추석연

뉴스토픽

- 1 코스
- 2 에이
- 3 개성
- 4 학교
- 5 세탁
- 6 4차
- 7 박찬
- 8 경전
- 9 수원
- 10 홈앤

2017.10.11

뉴스검색 R

# Crawling

연관검색어 ? edg skt skt 고객센터 sk skt t1 tworld sk t월드 2017 롤드컵 skt 티월드  
skt 역전 입롤 롤드컵 skt edg skt vs edg skt 번호변경 skt 대리점 skt 멤버쉽 더보기

뉴스 1-10 / 17건 [뉴스검색가이드](#)

✔ 관련도순 ✔ 최신순 ✔ 오래된순 검색결과 자동고침 시작



[SKT, 장애청소년 ICT 메이커톤 대회 개최](#) 머니투데이 | 2017.10.01. | 네이버뉴스 | [🔗](#)  
SK텔레콤은 한국장애인단체총연맹과 공동으로 지난달 28~29일 이틀간 '2017 SKT 장애청소년 ICT(정보통신기술) 메이커톤 대회'를 개최했다고 1일 밝혔다. 이번 대회는 장애청소년 112명으로 이뤄진 28개팀이 참가해...  
↳ [SKT, 장애 청소년과 메이커톤 대회 개최](#) ZDNet Korea | 2017.10.01. | 네이버뉴스  
↳ [SKT 장애청소년 ICT메이커톤 대회](#) 연합뉴스 | 2017.10.01. | 네이버뉴스  
↳ [SKT, '2017 ICT메이커톤 대회' 개최](#) 뉴스워커 | 2017.10.01.  
↳ [SKT, '2017 ICT메이커톤 대회' 개최](#) 아크로팬 | 2017.10.01.  
[관련뉴스 24건 전체보기 >](#)



[SKT, '티움'에 국내 거주 외국인 유학생 초청](#) ZDNet Korea | 2017.10.01. | 네이버뉴스 | [🔗](#)  
외국인 유학생 20명에 티움 투어 기회 제공 (지디넷코리아=박수형 기자)SK텔레콤(대표 박정호)은 추석연

## 뉴스토픽

- 1 코스
- 2 에이
- 3 개성
- 4 학교
- 5 세탁
- 6 4차·
- 7 박찬
- 8 경전
- 9 수원
- 10 홈앤

2017.10.11

뉴스검색 R

http://search.naver.com/search.naver?where=news&**query=skt**&ie=utf8&sm=tab\_opt&sort=0&photo=0&field=0&reporter\_article=&pd=3&**ds=2017.10.01**&**de=2017.10.02**

# Crawling

연관검색어 ? edg skt skt 고객센터 sk skt t1 tworld skt월드 2017 롤드컵 skt 티월드  
skt 역전 입몰 롤드컵 skt edg skt vs edg skt 번호변경 skt 대리점 skt 멤버십 더보기

뉴스 1-10 / 17건 [뉴스검색가이드](#)

✔ 관련도순 ✔ 최신순 ✔ 오래된순

검색결과 자동고침 시작 ▶



[SKT, 장애청소년 ICT 메이커톤 대회 개최](#) 머니투데이 | 2017.10.01. | 네이버뉴스 | [🔗](#)  
SK텔레콤은 한국장애인단체총연맹과 공동으로 지난달 28~29일 이틀간 '2017 SKT 장애청소년 ICT(정보통신기술) 메이커톤 대회'를 개최했다고 1일 밝혔다. 이번 대회는 장애청소년 112명으로 이뤄진 28개팀이 참가해...

- SKT, 장애 청소년과 메이커톤 대회 개최 ZDNet Korea | 2017.10.01. | 네이버뉴스
- SKT 장애청소년 ICT메이커톤 대회 연합뉴스 | 2017.10.01. | 네이버뉴스
- SKT, '2017 ICT메이커톤 대회' 개최 뉴스워치 | 2017.10.01.
- SKT, '2017 ICT메이커톤 대회' 개최 아크로팬 | 2017.10.01.

[관련뉴스 24건 전체보기 >](#)



[SKT, '티움'에 국내 거주 외국인 유학생 초청](#) ZDNet Korea | 2017.10.01. | 네이버뉴스 | [🔗](#)  
외국인 유학생 20명에 티움 투어 기회 제공 (지디넷코리아=박수형 기자)SK텔레콤(대표 박정호)은 추석연

## 뉴스토픽

- 1 코스
- 2 에이
- 3 개성
- 4 학교
- 5 세탁
- 6 4차
- 7 박찬
- 8 경전
- 9 수원
- 10 홈앤

2017.10.11

뉴스검색R

```
<html lang="ko" data-useragent="mozilla/5.0 (macintosh; intel mac os x 10_13_0) applewebkit/537.36 (KHTML, like
gecko) chrome/61.0.3163.100 safari/537.36" data-platform="macintel">
  ▶ #shadow-root (open)
  ▶ <head>_</head>
  ▼ <body class>
    <div id="nxtt_div" style="display:none;position:absolute;border-width:0;z-index:11000"></div>
    ▶ <div id="u_skip">_</div>
    ▼ <div id="wrap">
      ▶ <div id="header_wrap" role="heading">_</div>
      ▶ <script type="text/javascript">_</script>
      ▼ <div id="container" role="main">
        ▼ <div id="content" class="pack_group">
          <h1 class="blind">000 검색결과 시작</h1>
          ▼ <div id="main_pack" class="main_pack">
            <script type="text/javascript">var nx_cr_area_info=[{ n:"tab",r:1 }];</script>
            ▶ <div id="nx_related_keywords" class="sp_keyword section">_</div>
            ▶ <script type="text/javascript">_</script>
            <script>g_crt+="";</script>
            ▶ <script type="text/javascript">_</script>
            ... ▶ <div class="news mynews section">_</div> == $0
            ▶ <div class="paging" style="display: block;">_</div>
            ▶ <script type="text/javascript">_</script>
            ▶ <script type="text/javascript">_</script>
            <script type="text/javascript" src="https://ssl.pstatic.net/sstatic/search/js/news/
            rev31 news office action-170612.js"></script>
            ▶ <script type="text/javascript">_</script>
            <div class="collection"></div>
            </div>
            ▶ <div id="sub_pack" class="sub_pack">_</div>
            <div class="lv_dimmed"></div>
html body div#wrap div#container div#content.pack_group div#main_pack.main_pack div.news.mynews.section
```

# Crawling



**newspaper3k 0.1.5**

*Simplified python article discovery & extraction.*

**Latest Version:** [0.2.5](#)



**Crawling 실습**

# 데이터 전처리

SK텔레콤 (266,500원 ▼1000 -0.4%)은 사물인터넷(IoT) 진흥주년을 맞아 11일부터 사흘간 코엑스 SK텔레콤 (266,500원 ▼1000 -0.4%)은 사물인터넷(IoT) 진흥주년을 맞아 11일부터 사흘간 코엑스 SK텔레콤 (266,500원 ▼1000 -0.4%)은 사물인터넷(IoT) 진흥주년을 맞아 11일부터 사흘간 코엑스 SK텔레콤 (266,500원 ▼1000 -0.4%)은 사물인터넷(IoT) 진흥주년을 맞아 11일부터 사흘간 코엑스에서 열리는 '사물인터넷 국제전시회'에 부스를 마련하고 가정과 일터, 도시와 농장 등 우리 일상생활 전반에 적용된 다양한 IoT 제품과 서비스를 선보인다고 밝혔다.

137평(약 459 m<sup>2</sup>) 규모로 선보이는 이번 전시의 주제는 '진짜 IoT(True IoT with SK Telecom)'다. 전시부스는 우리의 일상생활 전반을 상징하는 '가정'과 '일터', '농장+도시', '자동차'의 4개 구역과 SK텔레콤의 IoT 플랫폼과 네트워크를 소개하는 '트루 IoT 존'으로 구성된다.

가정 구역에선 에어컨과 로봇청소기, 온도조절기, CCTV, 가스경보기, 레인지후드, 정수기, 밥솥, 공기질센서, 공기청정기, 제습기, 세탁기, 조명 등 SK텔레콤의 IoT와 결합된 다양한 가전 제품들이 소개된다. SK텔레콤은 현재 70여 제조사와 손잡고 300여 모델을 시장에 출시했다.

자동차 구역에선 SK네트웍스와 함께 IoT를 활용한 법인 자동차 운행관리 서비스를 선보이며 일터 구역에선 로라(LoRa)망을 활용하는 가스와 수도 검침, 시설물 위험감지 시스템, 고정형 가스감지기 등 각종 제품과 서비스가 소개된다.

또한, 농장+도시 구역에선 가축이나 농장의 각종 데이터를 IoT망을 통해 확인하고 관리 효율을 높이는 '라이브케어'와 '수목생장관리', 도시 생활에서 해마다 관심이 높아지는 '미세먼지 모니터링 서비스'가 전시된다.



parsing, pos tagging,  
spam 제거, word indexing 등...

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix



## KoNLPy



511

Note:

You are not using the most up to date version of the library. [0.4.4](#) is the newest version.

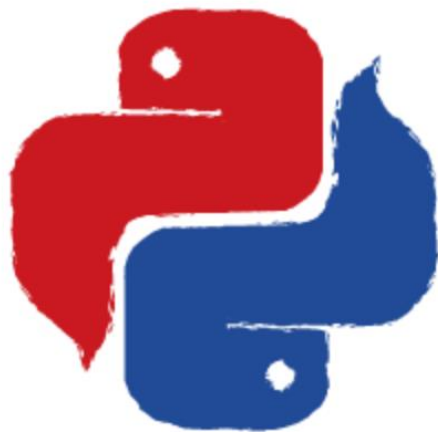
## KoNLPy: 파이썬 한국어 NLP

build failing docs passing

KoNLPy(“코엔엘파이”라고 읽습니다)는 한국어 정보처리를 위한 파이썬 패키지입니다. 설치법은 [이 곳을](#) 참고해주세요.

NLP를 처음 시작하시는 분들은 [시작하기](#)에서 가볍게 기본 지식을 습득할 수 있으며, KoNLPy의 사용법 가이드는 [사용하기](#), 각 모듈의 상세사항은 [API](#) 문서에서 보실 수 있습니다.

# 데이터 전처리



KoNLPy

Star 511

```
In [48]: from konlpy.tag import Twitter
```

```
t = Twitter()
```

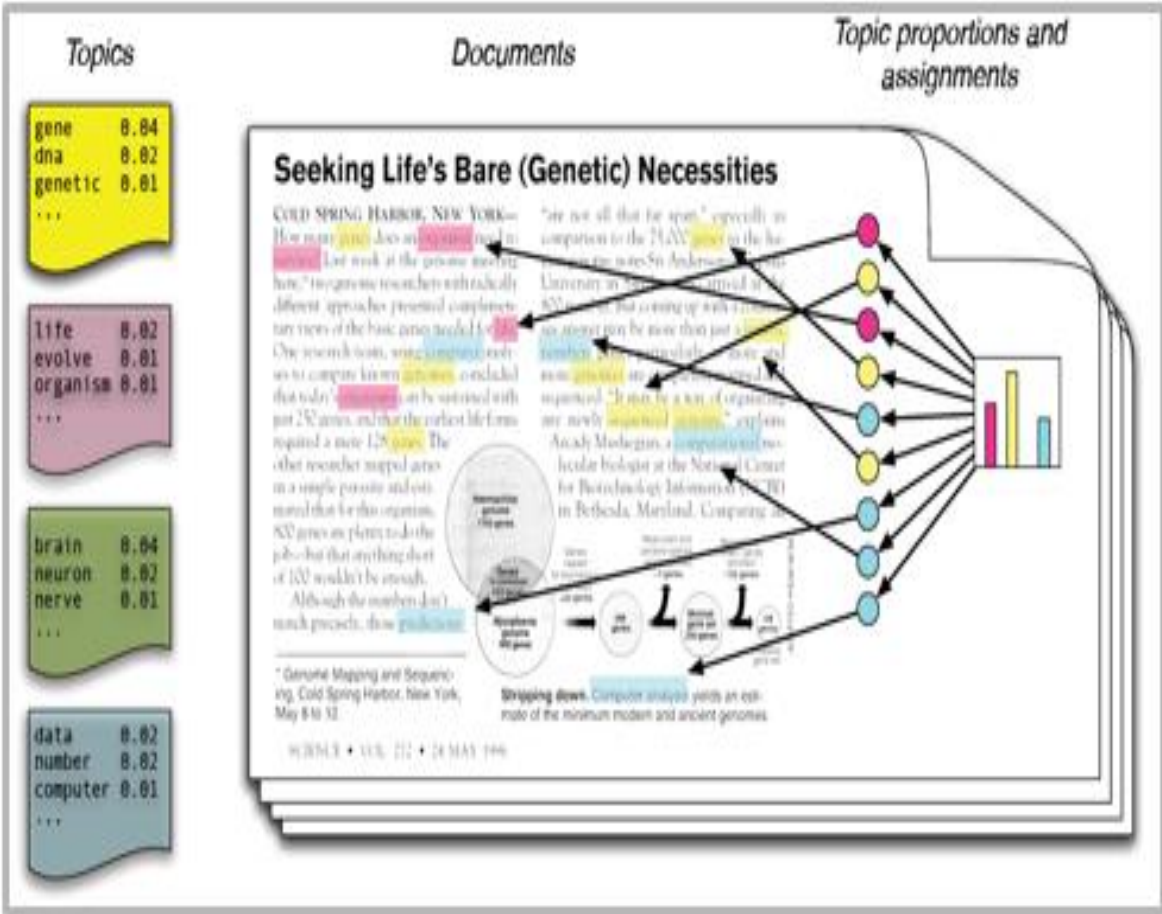
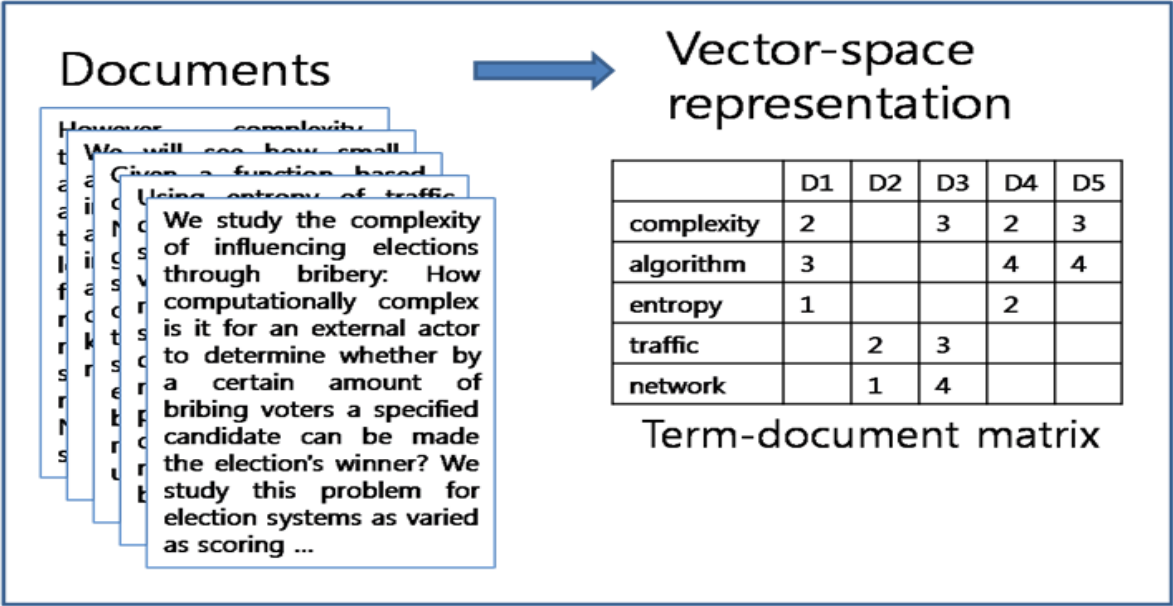
executed in 488ms, finished 16:34:59 2017-10-11

```
In [49]: t.pos("안녕하세요. 반갑습니다. 오늘 아침은 무엇을 먹었나요??")
```

executed in 2.03s, finished 16:35:26 2017-10-11

```
Out[49]: [('안녕하세', 'Adjective'),  
          ('요', 'Eomi'),  
          ('.', 'Punctuation'),  
          ('반갑', 'Adjective'),  
          ('습니다', 'Eomi'),  
          ('.', 'Punctuation'),  
          ('오늘', 'Noun'),  
          ('아침', 'Noun'),  
          ('은', 'Josa'),  
          ('무엇', 'Noun'),  
          ('을', 'Josa'),  
          ('먹었', 'Verb'),  
          ('나요', 'Eomi'),  
          ('??', 'Punctuation')]
```

# 토픽모델링



## Term-document matrix (TF-IDF)

Vector-space  
representation

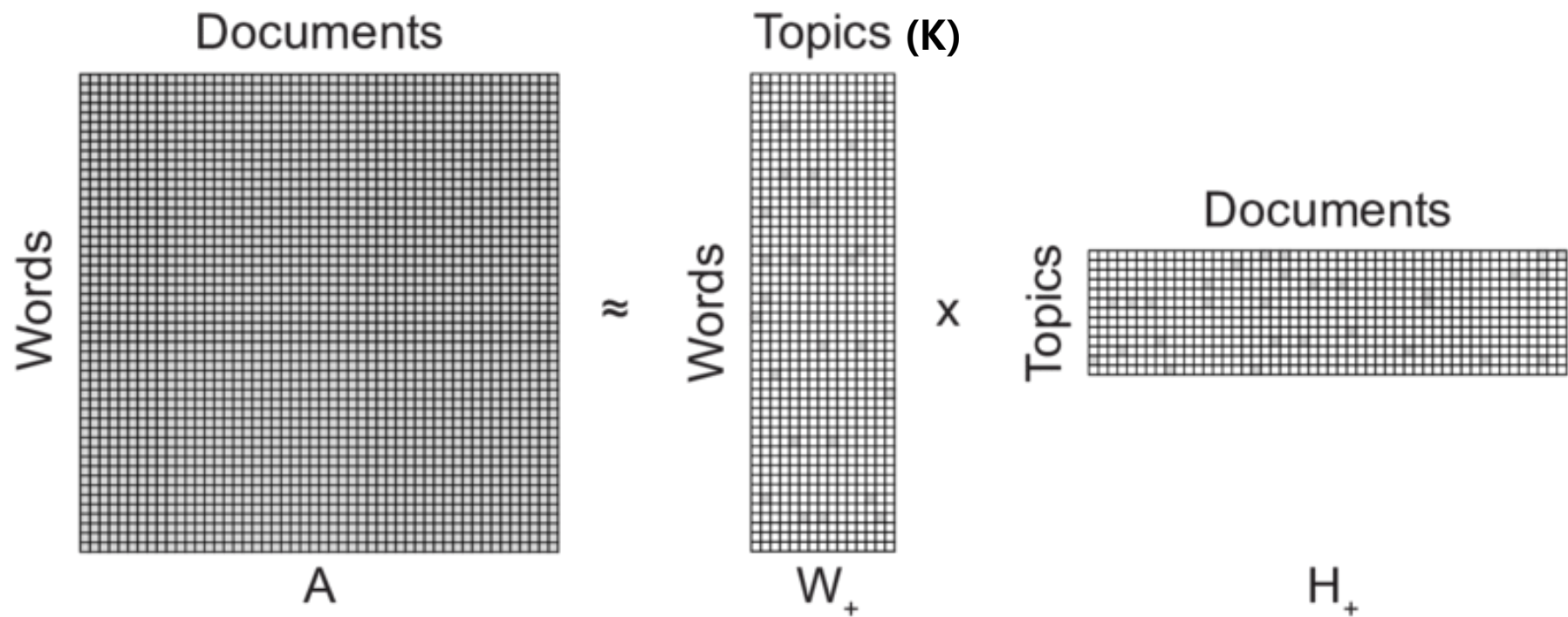
	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

- $tf$ : term frequency
  - 문서 내에서 각 단어의 수
- $df$ : document frequency
  - 해당 단어가 나오는 문서의 수

토픽모델링 NMF (non-negative matrix factorization)



Term-document matrix

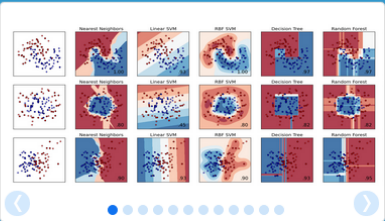
# 토픽모델링

scikit  
learn

Home Installation Documentation ▾ Examples

Google™ Custom Search Search x

Fork me on GitHub



# scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

## Classification

Identifying to which set of categories a new observation belong to.

**Applications:** Spam detection, Image recognition.

**Algorithms:** *SVM, nearest neighbors, random forest, ...* — Examples

## Regression

Predicting a continuous value for a new example.

**Applications:** Drug response, Stock prices.

**Algorithms:** *SVR, ridge regression, Lasso, ...* — Examples

## Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** *k-Means, spectral clustering, mean-shift, ...* — Examples

## Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency

**Algorithms:** *PCA, Isomap, non-negative matrix factorization.* — Examples

## Model selection

Comparing, validating and choosing parameters and models.

**Goal:** Improved accuracy via parameter tuning

**Modules:** *grid search, cross validation, metrics.* — Examples

## Preprocessing

Feature extraction and normalization.

**Application:** Transforming input data such as text for use with machine learning algorithms.

**Modules:** *preprocessing, feature extraction.* — Examples

```
from sklearn.decomposition import NMF
from sklearn.decomposition import LatentDirichletAllocation
```



# 토픽모델링

```
▼ # LDA
K = 10
lda = LatentDirichletAllocation(n_components=K, learning_method='batch')
W = lda.fit_transform(tdm)
H = lda.components_

# 각 토픽별 키워드 출력
▼ for k in range(K):
    print(f"{k}th topic")
▼     for index in H[k].argsort()[::-1][:20]:
        print(index2voca[index], end=" ")
    print("\n")
```

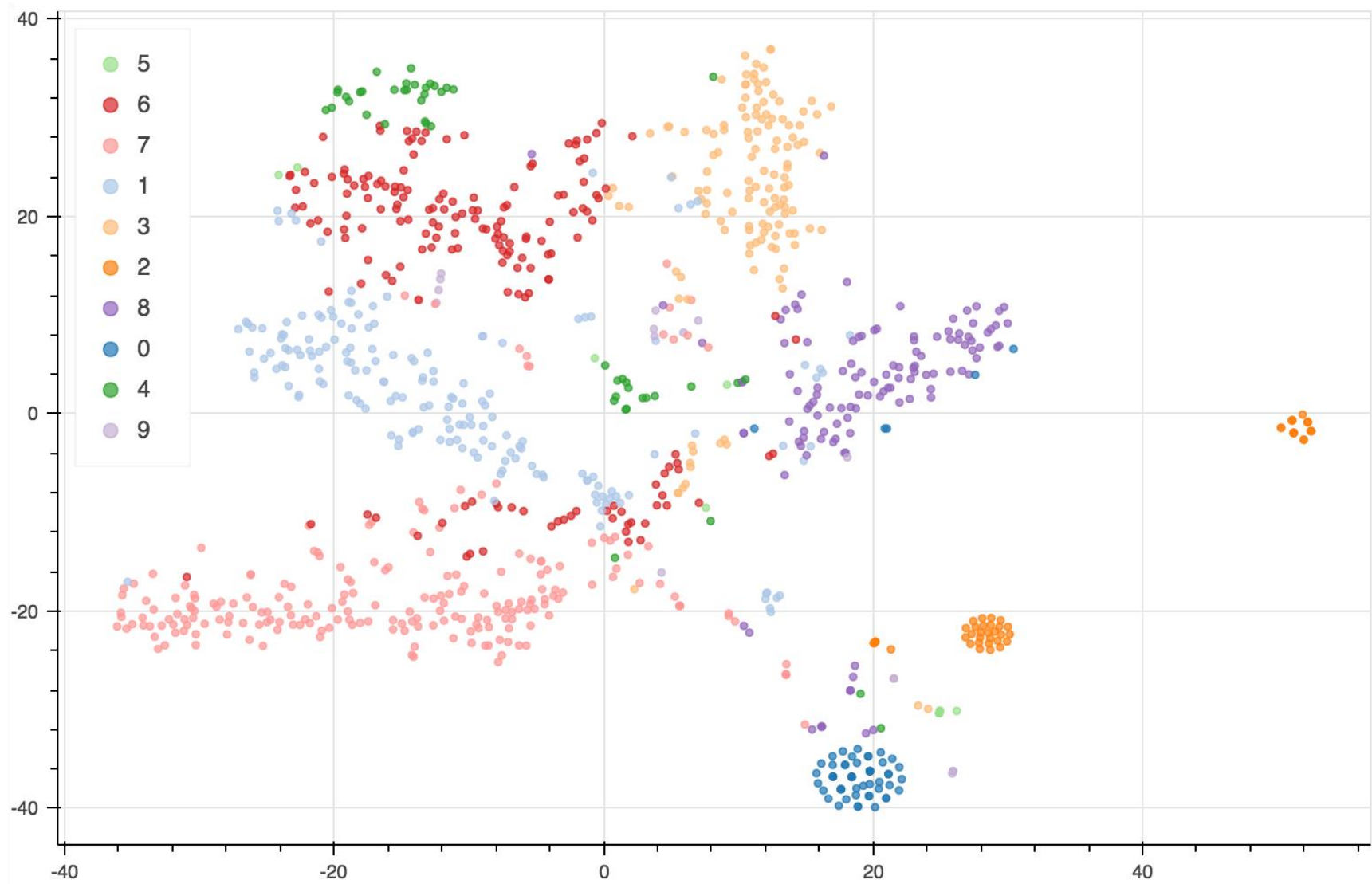
execution queued 17:28:49 2017-10-11

```
▼ # NMF
K = 10
nmf = NMF(n_components=K)
W = nmf.fit_transform(tdm)
H = nmf.components_

# 각 토픽별 키워드 출력
▼ for k in range(K):
    print(f"{k}th topic")
▼     for index in H[k].argsort()[::-1][:20]:
        print(index2voca[index], end=" ")
    print("\n")
```

executed in 5.92s, finished 15:37:57 2017-10-11

토픽모델링 -LDA



# 토픽모델링 - NMF

executed in 16.0s, finished 18:09:30 2017-10-11

## 0th topic

중국 우리 지금 사드 진행 운기 있는 배치 한국 미국 교수 때문 문제 생각 가지 어떻 경제 그렇 하고 아니

## 1th topic

탄핵 헌재 심판 선고 결정 재판관 대통령 인용 헌법재판소 기각 대행 국회 권한 의견 각하 결과 변론 오전 정미 절차

## 2th topic

시간 대선 지원 출마 억원 보통사람 정부 청춘 황교안 인상 소환 검찰 금리 공개 변경 주택 문재인 한국 서울 사업

## 3th topic

대표 후보 대선 민주당 의원 정당 있는 지사 국민 문재인 경선 바른 개헌 보수 주자 탄핵 자유 한국 안희정 국민의당

## 4th topic

수사 특검 검찰 있는 부분 지금 사실 결과 때문 대해서 수석 경우 최순실 대한 조사 같은 보면 아니 그렇 어떻

## 5th topic

탄핵 대통령 헌법 입니 사유 사건 위반 법률 행위 아니 청구인 위배 국회 최순실 하여 범죄 주장 소추 추장 입증

## 6th topic

대통령 박근혜 청와대 파면 결정 삼성동 사저 권한 때문 대한 있는 대행 의원 탄핵 대선 정치 메시지 보수 국정 헌법재판소

## 7th topic

집회 탄핵 태극기 박근혜 국민 광장 촛불 서울 시민 경찰 퇴진 행동 촛불집회 우리 광화문 오후 참가자 반대 단체 대한민국

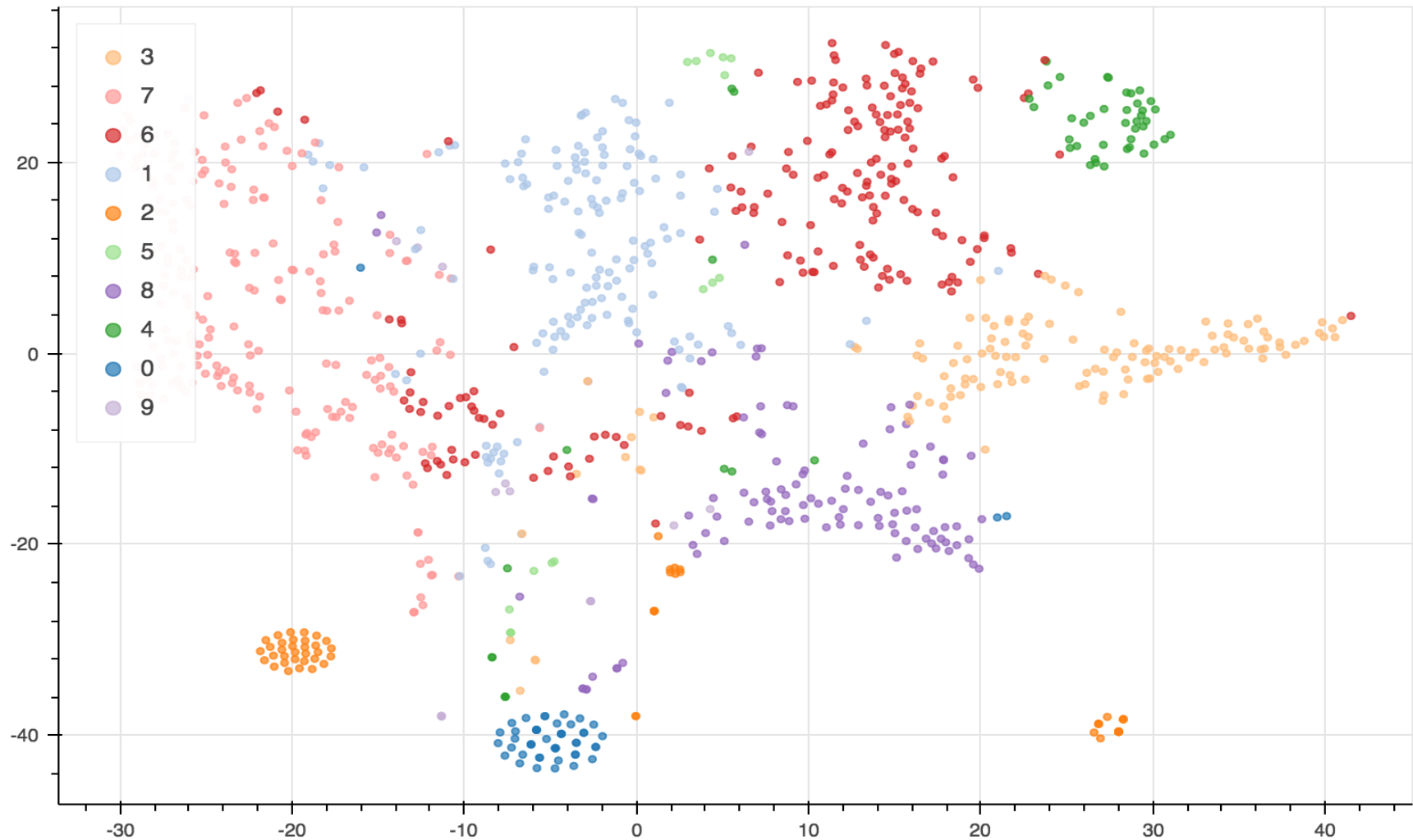
## 8th topic

국정원 한국 북한 정부 사드 대한 배치 정보 정치 문제 대해 경제 정책 개혁 중국 말했 통해 사찰 국회 국가

## 9th topic

상황 청구인 국민 대통령 국가 구조 세월호 의무 위기 보고 헌법 안전 국가안보실 있는 생명 재판관 심각 직책 수행 재난

토픽모델링 - NMF

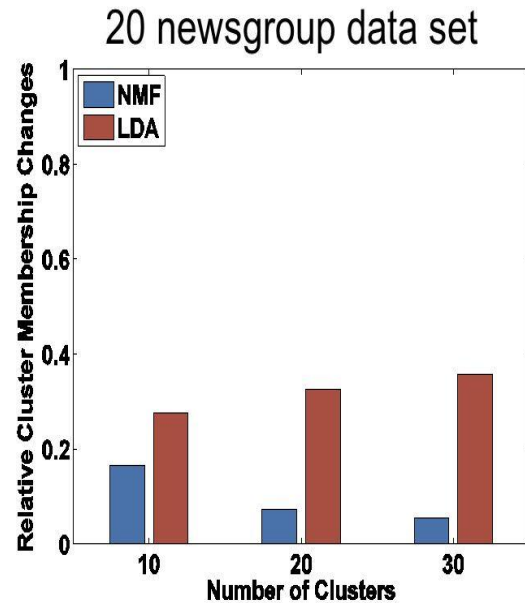
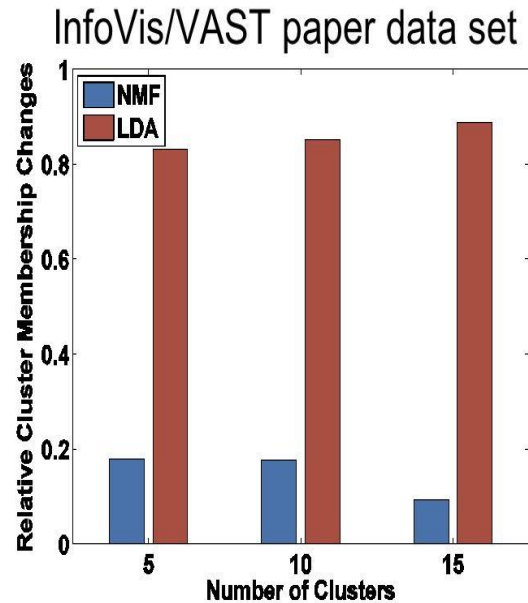


# 토픽모델링 – LDA, NMF 비교

## NMF vs. LDA

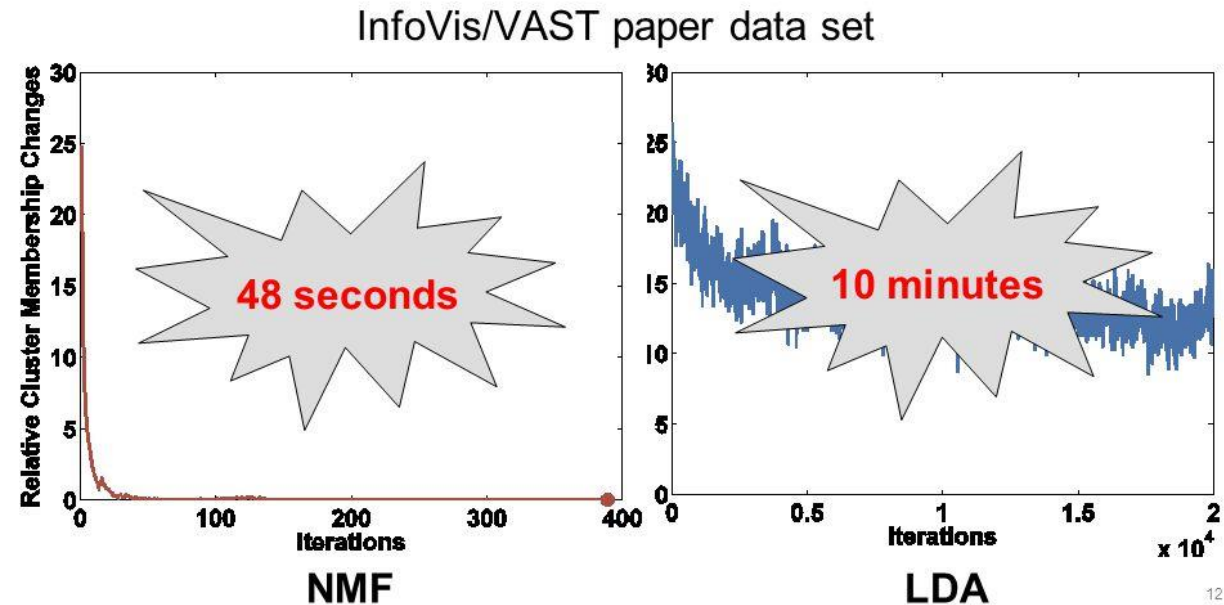
### Consistency from Multiple Runs

Documents' topical membership changes among 10 runs

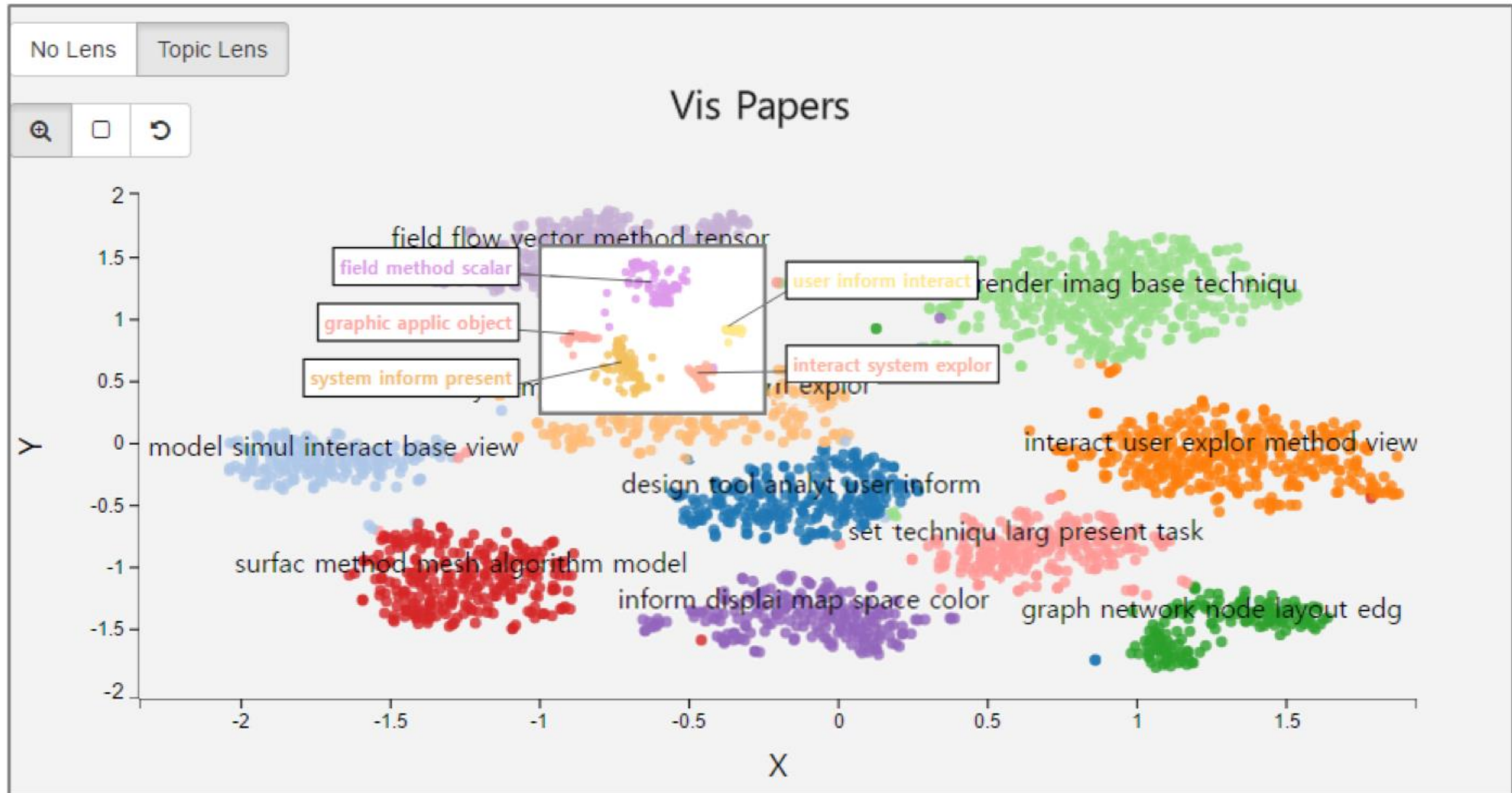


## Why NMF (instead of LDA)? Empirical Convergence

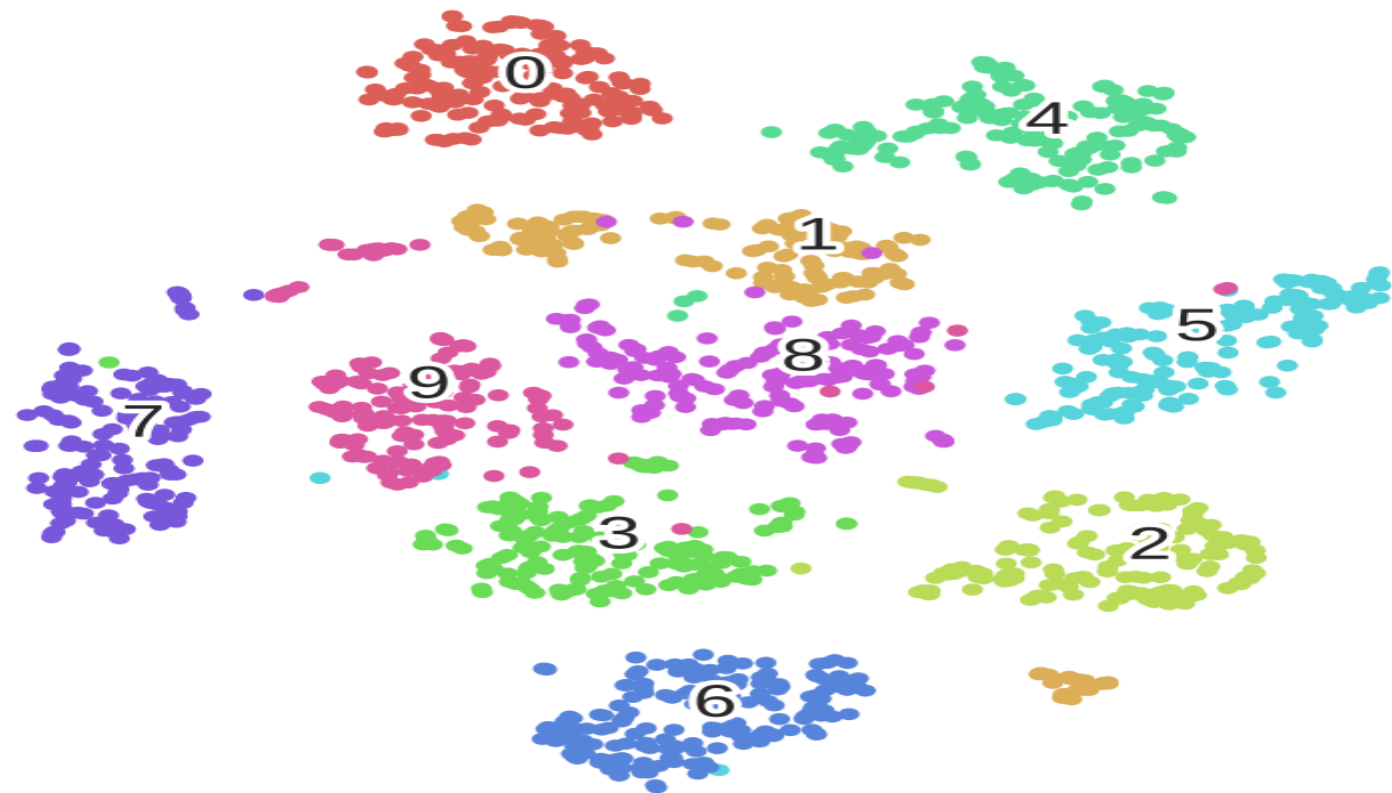
Documents' topical membership changes between iterations



## 토픽모델링 예



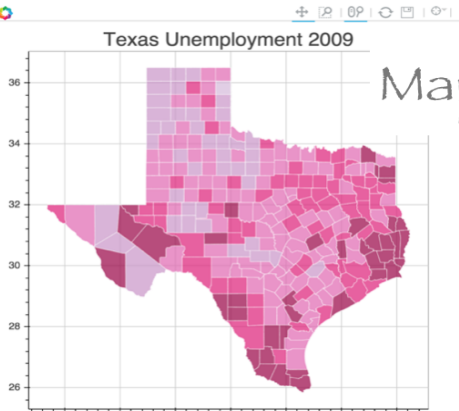
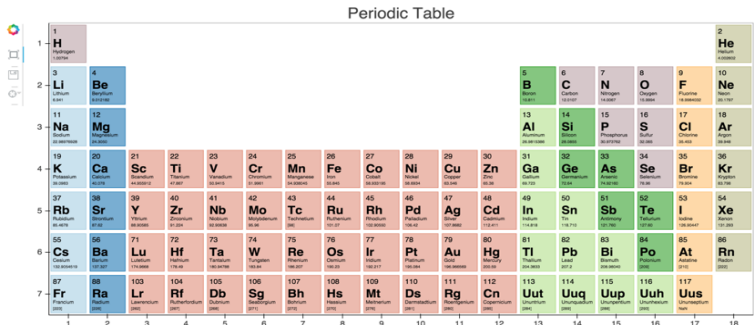
Visualization - T-SNE



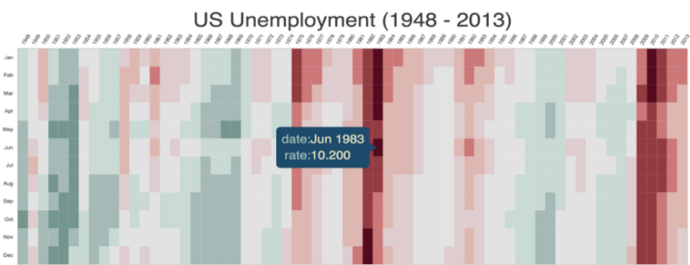


# Visualization - Bokeh

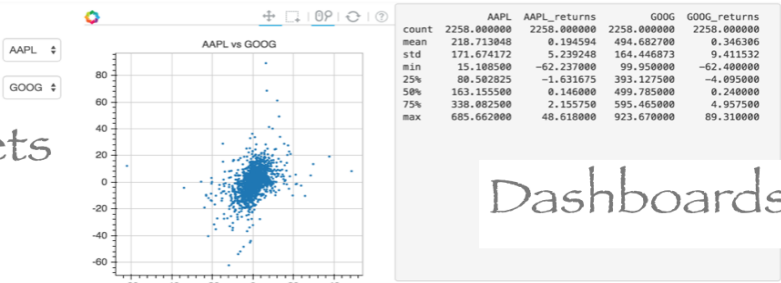
## Custom visualizations



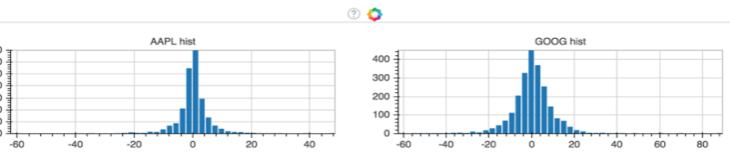
Hover



Widgets

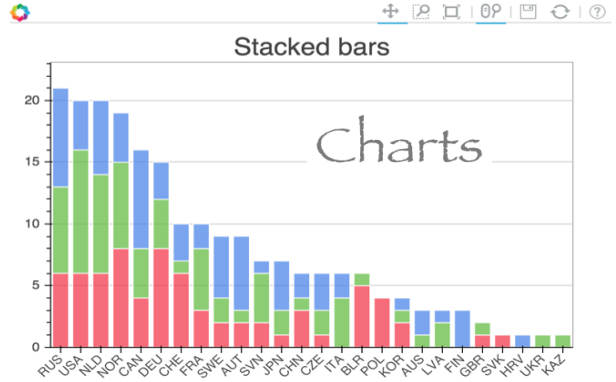


Dashboards

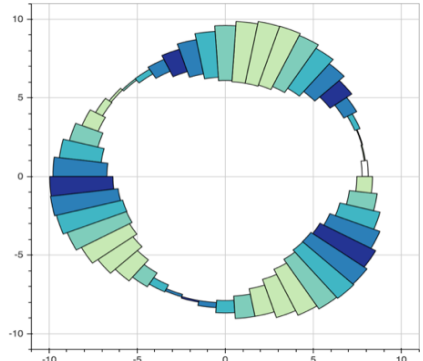


Bokeh

Tools



Charts

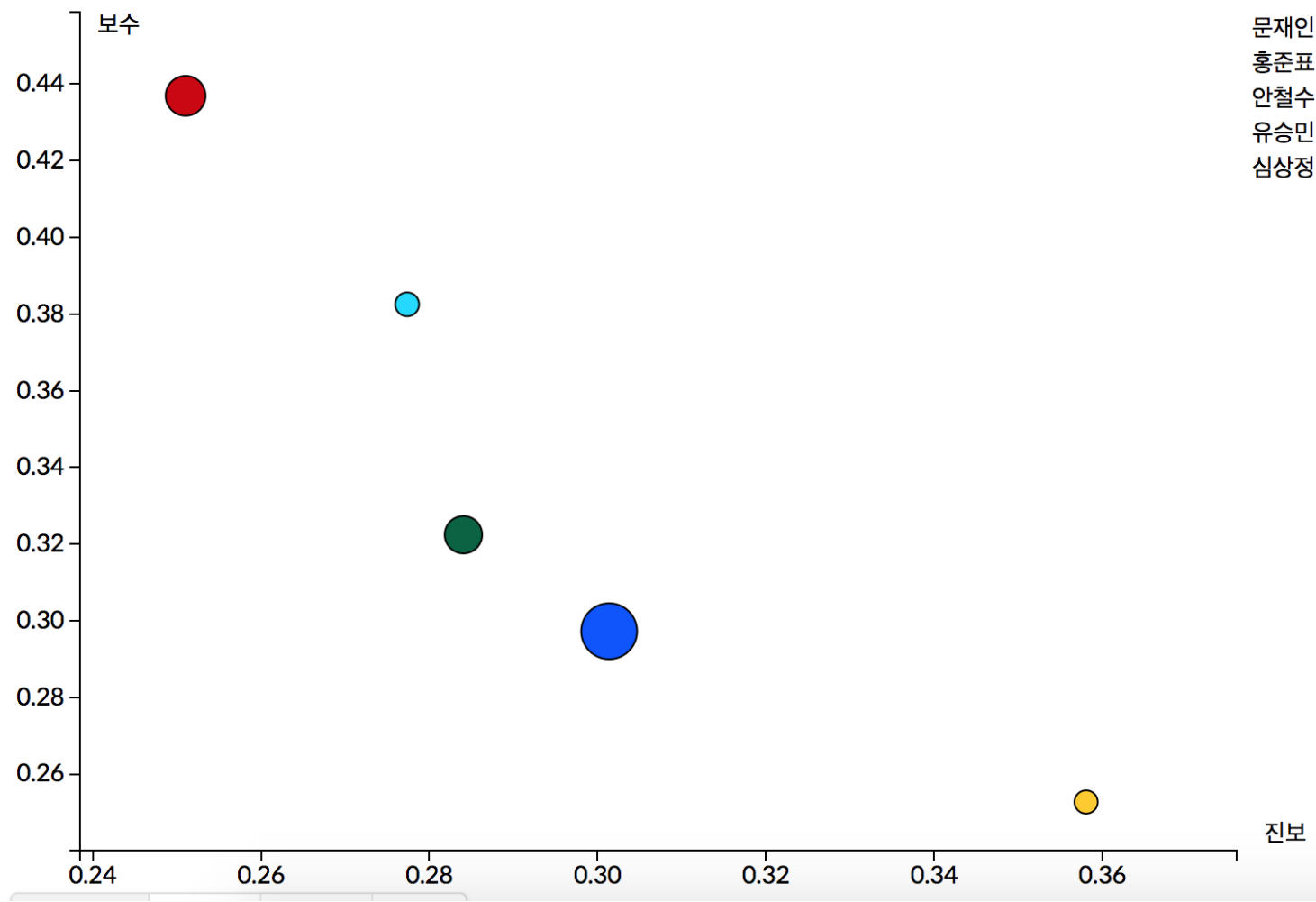


Streaming/  
Animations



# Visualization – Example

CandiVis



- 문재인
- 홍준표
- 안철수
- 유승민
- 심상정

좌표 단어를 입력하세요

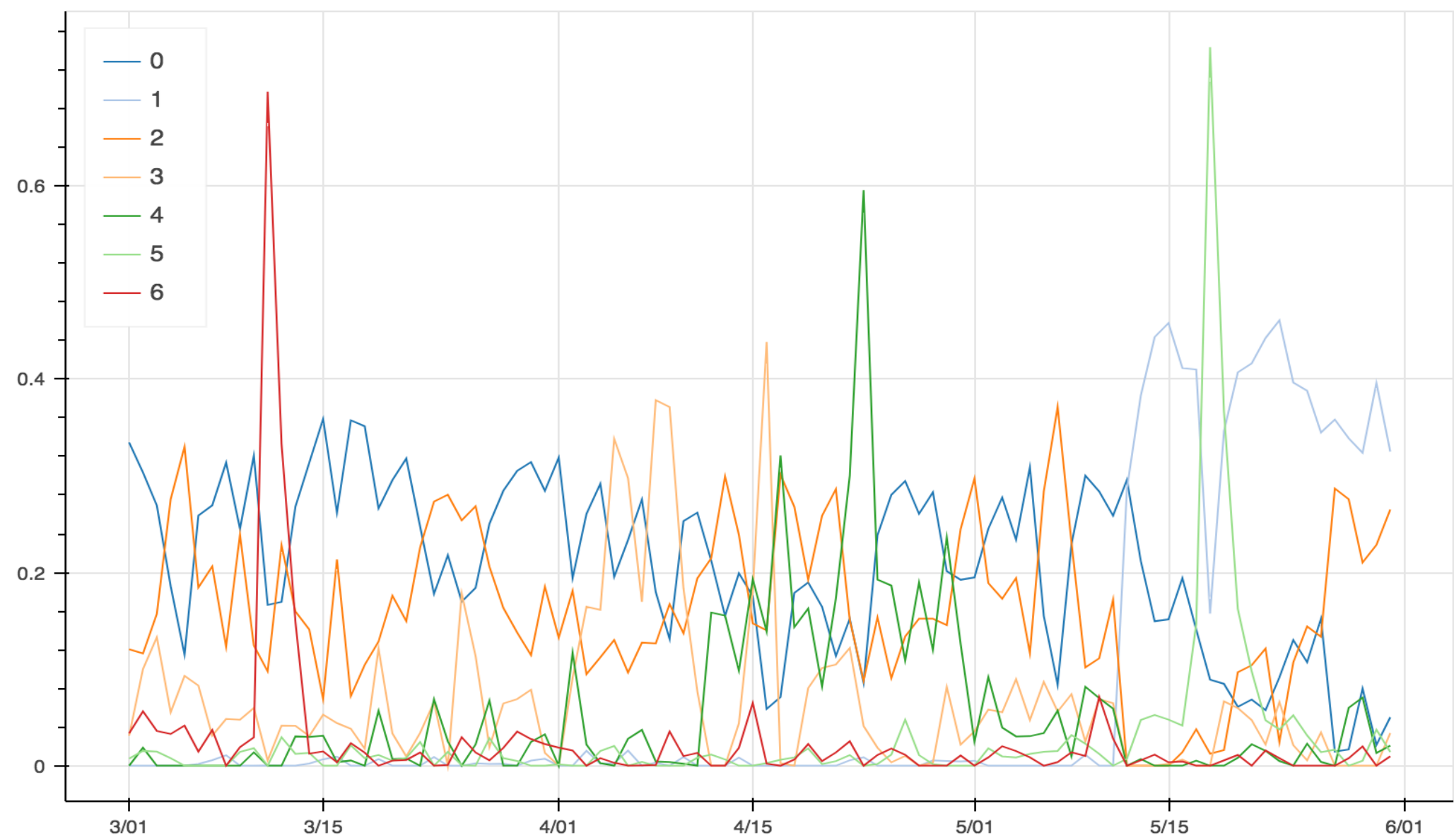
Apply

대선 후보 정보

각 점(후보)를 클릭해주세요!

관련 기사

# Twitter topic modeling - wannacry



# Twitter topic modeling - wannacry



## 4번 토픽 워드클라우드

**감사합니다**

**Any Questions?**

**yumere7833@gmail.com**