쉽게 접하는 빅데이터 알고리즘

오수진 주임

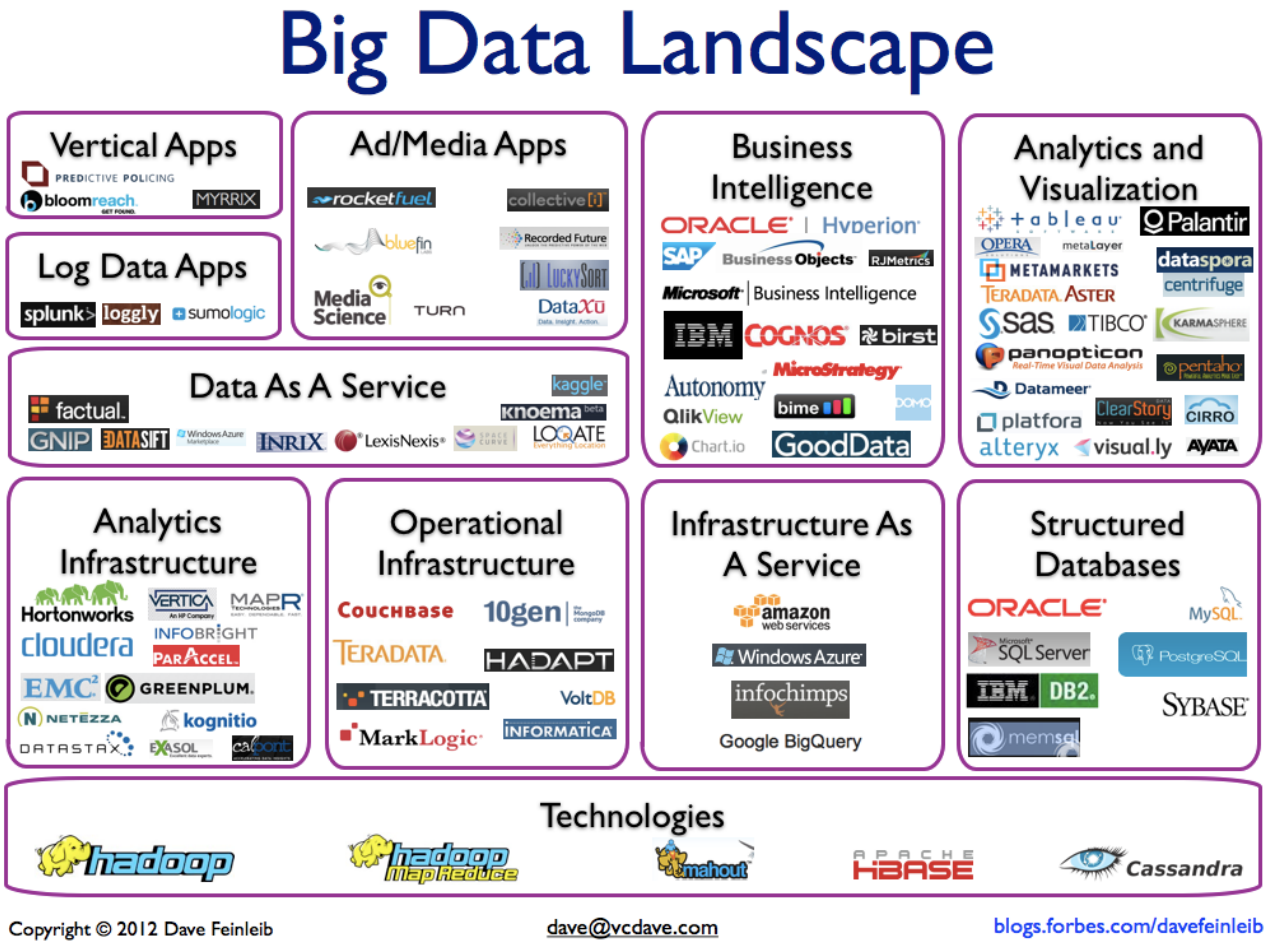
MOBIGEN
Make The Best For Smart Business

목차

1.BigData Life

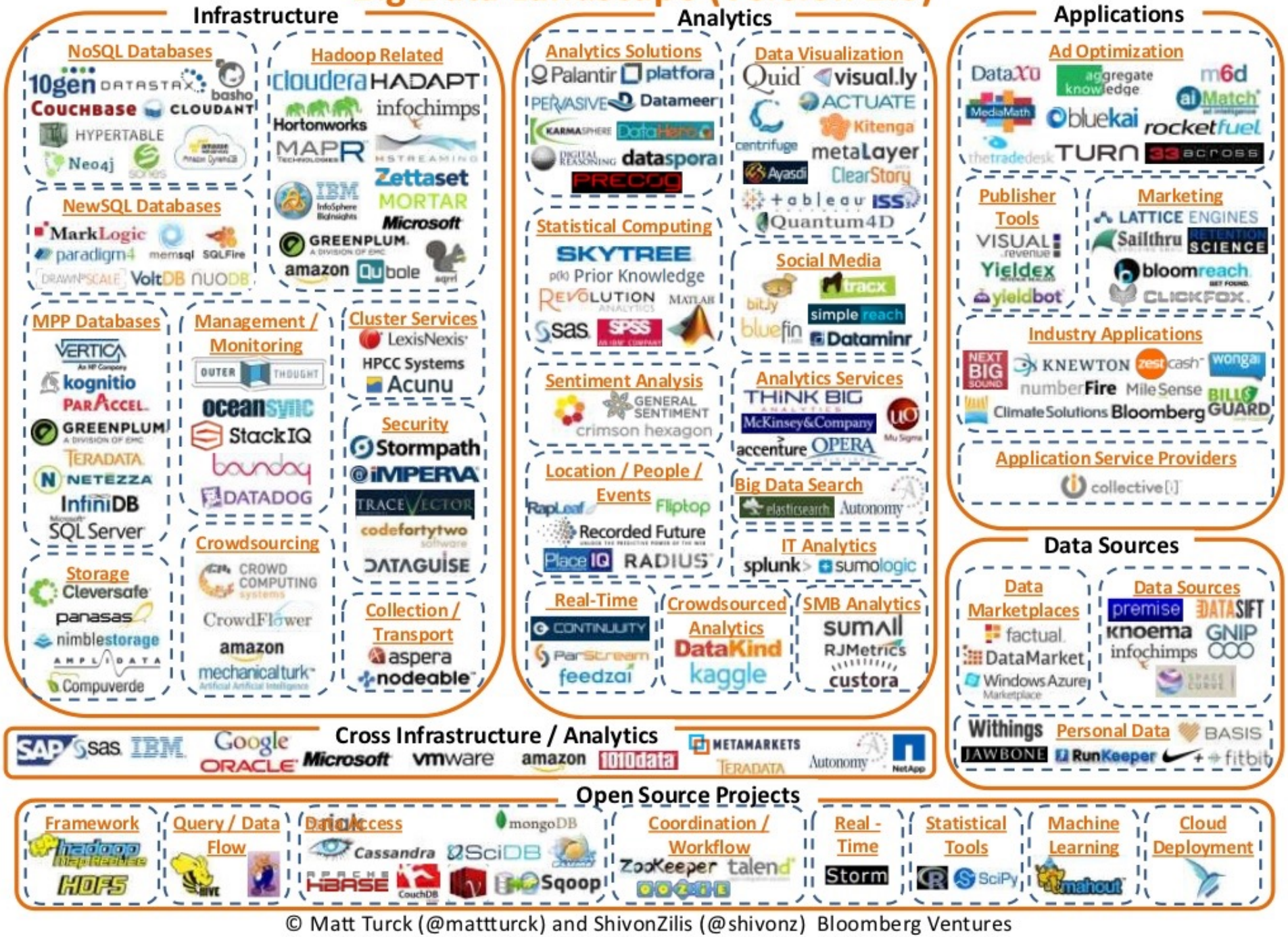
빅데이터라는 기술이 우리의 삶과 얼마나 밀접한 관련이 있을까?

1. BigData Life



빅데이터를 분석하는데에는 수많은 기술들이 있습니다.

Big Data Landscape (Version 2.0)




Big Data Landscape (Version 2.0)
Infrastructure
NoSQL Databases
10gen
DATASTAX
basho
Couchbase
CLOUDANT
HYPERTABLE
Neo4j
NewSQL Databases
MarkLogic
paradigm4
memsql
SQLFire
VoltDB
NUODB
Hadoop Related
cloudera
HADAPT
infochimps
Hortonworks
MAPR
Zettaset
IBM InfoSphere BigInsights
MORTAR
Microsoft
GREENPLUM
amazon
Qubole
MPP Databases
VERTICA
kognitio
PARACCEL
GREENPLUM
TERADATA
NETEZZA
InfiniDB
SQL Server
Storage
Cleversafe
panasas
nimblestorage
AMPLIDATA
Compuverde
StackIQ
bandy
DATADOG
Stormpath
IMPERVA
TRACEVECTOR
codefortytwo
DATAGUISE
Crowdsourcing
CROWD COMPUTING systems
CrowdFlower
amazon
mechanicalturk
Collection / Transport
aspera
nodeable
Analytics
Analytics Solutions
Palantir
platfora
PERVASIVE
Datameer
KARMASPHERE
DataHero
DIGITAL REASONING
dataspora
PRECOG
Statistical Computing
SKYTREE
p(k) Prior Knowledge
REVOLUTION
MATLAB
Data Visualization
Quid
visual.ly
ACTUATE
centrifuge
Kitenga
metaLayer
Ayasdi
ClearStory
tableau
ISS
Quantum4D
Social Media
bitly
tracx
accenture
OPERA
Location / People / Events
RapLeaf
Fliptop
Recorded Future
PlaceIQ
RADIUS
Big Data Search
elasticsearch
Autonomy
IT Analytics
splunk>
sumologic
Real-Time
CONTINUITY
ParStream
feedzai
Crowdsourced Analytics
DataKind
kaggle
SMB Analytics
sumall
RJMetrics
custora
Applications
Ad Optimization
DataXu
aggregate knowledge
m6d
MediaMath
bluekai
rocketfuel
thetradedesk
TURN
33across
Publisher Tools
VISUAL revenue
Yieldex
yieldbot
Marketing
LATTICE ENGINES
Sailthru
RETENTION SCIENCE
bloomreach
CLICKFOX
Application Service Providers
collective[i]
Data Sources
Data Marketplaces
factual
DataMarket
Windows Azure Marketplace
Data Sources
premise
DATASIFT
knoema
GNIP
infochimps
Personal Data
Withings
BASIS
JAWBONE
RunKeeper
fitbit
Cross Infrastructure / Analytics
SAP
sas
IBM
Google
ORACLE
Microsoft
vmware
amazon
1010data
METAMARKETS
TERADATA
Autonomy
NetApp
Open Source Projects
Framework
hadoop
MapReduce
HDFS
Query / Data Flow
HIVE
Data Access
mongoDB
Cassandra
SciDB
HBASE
CouchDB
Sqoop
Coordination / Workflow
ZooKeeper
talend
OOZIE
Real-Time
Storm
Statistical Tools
R
SciPy
Machine Learning
mahout
Cloud Deployment
하…. 빅데이터를 분석하는 많은 기술들이
도대체 어떻게 우리의 삶과 관련이 있다는거지….??
© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures


Big Data Landscape (Version 2.0)
Infrastructure
NoSQL Databases
10gen
DATASTAX
basho
Couchbase
CLOUDANT
HYPERTABLE
Neo4j
Hadoop Related
cloudera
HADAPT
infochimps
Hortonworks
MAPR
Zettaset
MORTAR
Microsoft
GREENPLUM
amazon
Qubole
NewSQL Databases
MarkLogic
paradigm4
memsql
SQLFire
VoltDB
NUODB
MPP Databases
VERTICA
kognitio
PARACCEL
GREENPLUM
TERADATA
NETEZZA
InfiniDB
SQL Server
Management / Monitoring
OUTER THOUGHT
oceansync
StackIQ
bunday
DATADOG
Cluster Service
LexisNex
HPCC Syste
Acur
Secur
Storm
TRACE
codefortytw
DATAGUISE
Crowdsourcing
CROWD COMPUTING
CrowdFlower
Collection /
Storage
Cleversafe
panasas
nimblestorage
AMPLIDATA
Compuverde
Analytics
Analytics Solutions
Palantir
platfora
PERVASIVE
Datameer
KARMASPHERE
DataHero
DIGITAL REASONING
dataspora
PRECOG
Data Visualization
Quid
visual.ly
ACTUATE
Kitenga
centrifuge
metaLayer
Ayasdi
ClearStory
tableau
ISS
Quantum4D
Statistical Computing
Social Media
tracx
simple reach
Dataminr
ics Services
BIG
Company
OPERA
Search
Autonomy
IT Analytics
sumologic
SMB Analytics
Framework
hadoop
MapReduce
HDFS
Applications
Ad Optimization
DataXu
aggregate knowledge
m6d
ai Match
MediaMath
bluekai
rocketfuel
thetradedesk
TURN
33across
Publisher Tools
VISUAL revenue
Yieldex
yieldbot
Marketing
LATTICE ENGINES
Sailthru
RETENTION SCIENCE
bloomreach
CLICKFOX
Industry Applications
NEXT BIG SOUND
KNEWTON
zestcash
wonga
numberFire
MileSense
BILLGUARD
Climate Solutions
Bloomberg
Application Service Providers
collective[i]
Data Sources
Data Marketplaces
Data Sources
premise
DATASIFT
knoema
GNIP
infochimps
Open Source Projects
Framework
hadoop
MapReduce
HDFS
Query / Data Flow
Hive
Data Access
mongoDB
Cassandra
SciDB
HBASE
CouchDB
Sqoop
Coordination / Workflow
ZooKeeper
talend
Oozie
Real - Time
Storm
Statistical Tools
R
SciPy
Machine Learning
mahout
Cloud Deployment
하둡이란 기술을 실생활에서
어떻게 사용 할 수 있는지 알아보도록 하겠습니다.
© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures

2. 우리가 관심있어 하는것

분명 나는 이것이 필요하다는 것을
알지만 과연 나의 친구, 가족들도
이런 기술들이 필요하다는 것을

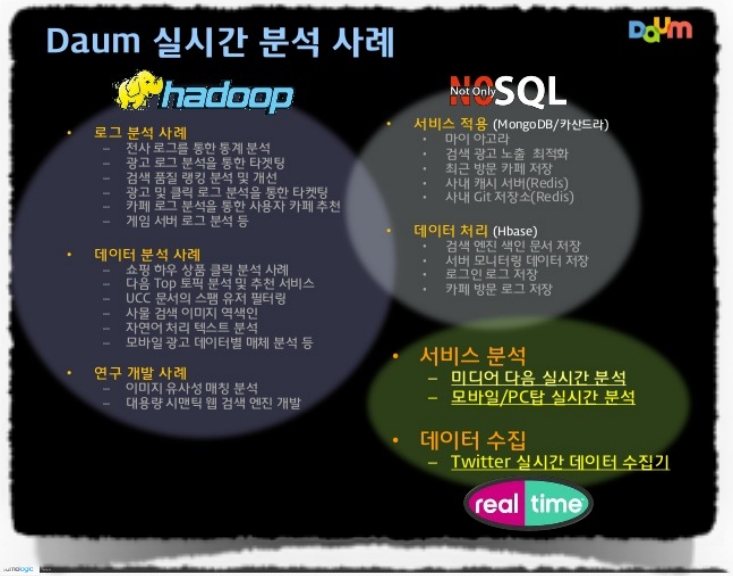
“ 공감 ”

할 수 있을까요?

[NEWS]

SK C&C, 빅데이터를 활용한
실시간(스트리밍) 통합보안로그분석 플랫폼 개발

1,800여개 사이트 대상 로그 데이터 분석으로 해킹 탐지·검색·추적까지 한번에 처리
실시간 로그 데이터 분석으로, 해킹 탐지시간 실시간(스트리밍) 데이터 처리 가능
IDS·방화벽 통합 분석 통해 해킹 이상 징후까지 해킹 탐지 정확도 향상
3TB의 빅데이터 대상 해킹 검색, 4초 內로 단축해 국내 최고 수준 평가
공개 S/W 활용으로 향후 사용 고객의 비용 절감 효과 기대



2. 우리가 관심있어 하는것

대한민국 국민이 매일매일 빼먹지 않고 하는 필수 코스

⇨ **NAVER** 뉴스보기

2. 우리가 관심있어 하는것

그런데 뉴스를 보다가 한가지 궁금증이 생겼습니다.

최신기사 정치 경제 국제 사회 문화 의학과학 사람속으로

이렇게 다양한 분야들에서


[동아알앤씨 추천매물] 서울대입구 대로변 임대수익용 빌딩 급매 52억 [2015-01-29 10:07:00]
◇ 서울 서울대입구 대로변 임대수익용 빌딩-대지730㎡ 연면적2750㎡ 7층 건물. 보증금 8억, 월 4000만에 공실 없이 임대중. 임차수요 풍부한 상권에 위치하여 투자 및 임대수익용으로 적합. 시세 이하 급매 52억원. 02-511-3773 강남 리얼티코리아 유진석 ◇ 강남…

이완구, 차남 공개 검증에 "비정한 아버지 아닌가 생각" 눈물 [2015-01-29 10:05:00]
이완구 국무총리 후보자는 29일 차남에게 제기된 병역면제 의혹에 대해 "오늘 공개검증을 실시하는 날이다, 오후로 알고 있다"고 말했다. 이 후보자는 이날 오전 인사청문회 준비를 위해 통의동 금융감독원 연수원으로 출근길에 기자들과 만나 "오늘은 좀 마음이 무겁다"며 이 같이 말했다. …

라디오스타 김구라, 방송통해 집 경매 알리며 "불행이 아니다. 헤쳐 나가는 맛이 있다" [2015-01-29 10:04:00]
라디오스타 김구라, 방송통해 집 경매 알리며 "불행이 아니다. 헤쳐 나가는 맛이 있다" '라디오스타' 김구라가 자신의 집이 경매에 들어간다고 밝혀 화제다. 28일 방송된 MBC '라디오스타' 에서는 '수요일 수요일은 라스다' 특집으로 김건모, 김성수, 이본, 김현정이 출연해 재미를…

112억년 된 항성계, '케플러 444'…지구와 비슷한 행성도 발견 [2015-01-29 10:04:00]
'112억년 된 항성계' 112억년 된 항성계가 발견됐다. 지난 27일 미국과 유럽, 호주 공동연구진은 지구와 비슷한 구조의 행성을 가진 112억 년된 항성계를 발견했다고 전했다. 연구 결과에 따르면 112억년 된 항성계는 지구로부터 117광년 떨어진 '케플러 444'라는 별…

'라디오스타' 김구라, 생활고 고백 "집 경매 넘어가지만…" [2015-01-29 10:03:46]
'라디오스타 김구라'방송인 김구라가 '라디오스타'에서 본인의 집이 경매에 들어간다고 말해 충격을 줬다.28일 방송한 MBC 예능 프로그램 '라디오스타'에서는 '수요일 수요일은 라스다' 특집으로 김건모, 김성수, 이본, 김현정이 출연해 입담을 과시했다.이날 '라디오스타' MC 윤종신이 …

이렇게 다양한 뉴스가 있는데

과연 분야별로 어떤 **키워드**가 가장 관심을 받을까요??

3. 분석하기

클라라 거짓말 모음 '끝도 없네'

본문

클라라의 거짓말이 새삼 화제다.

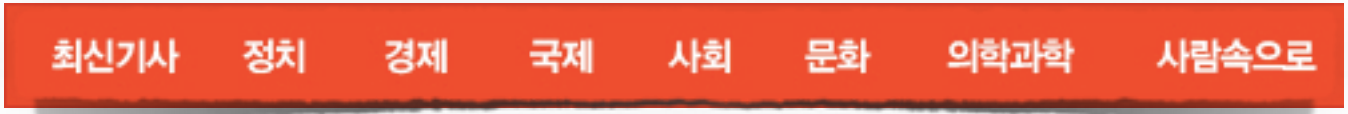
28일 방송된 SBS '한밤의 TV연예'에서는 스타들을 향한 대중의 분노와 스타들의 잘못된 대처에 대한 분석을 전했다.

최근 논란의 중심에 선 클라라는 현재 폴라리스엔터테인먼트와 갈등을 빚고 있다. 클라라는 소속사 회장으로부터 성적 수치심을 느꼈다고 주장했지만 한 매체의 문자 공개 후 여론은 급변했다.

각 기사의 본문만 발췌하여 키워드 별로 나누고 이를 분석하여 어떤 키워드가 HOT 한지 알아보고 싶었습니다.

3. 분석하기

1. 가장 먼저 rss와 xml 파서를 이용해 분야별로 기사의 링크를 수집했습니다.




```
▼<item>
   <title>클라라, 그간 했던 거짓말들 모아보니…누리꾼 '분노'</title>
  ▼<link>
     http://mbn.mk.co.kr/pages/news/newsView.php?category=mbn00012&news_seq_no=2189396
   </link>
```

2. 기사는 대부분 “~다.” 라고 끝내는 평서문의 형태로 이루어져 있습니다.

클라라의 거짓말이 새삼 화제다.

28일 방송된 SBS '한밤의 TV연예'에서는 스타들을 향한 대중의 분노와 스타들의 잘못된 대처에 대한 분석을 전했다.

최근 논란의 중심에 선 클라라는 현재 폴라리스엔터테인먼트와 갈등을 빚고 있다. 클라라는 소속사 회장으로부터 성적 수치심을 느꼈다고 주장했지만 한 매체의 문자 공개 후 여론은 급변했다.

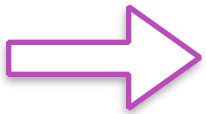
3. 분석하기

3. 본문만 가져오기

클라라의 거짓말이 새삼 화제다.

28일 방송된 SBS '한밤의 TV연예'에서는 스타들을 향한 대중의 분노와 스타들의 잘못된 대처에 대한 분석을 전했다.

최근 논란의 중심에 선 클라라는 현재 폴라리스엔터테인먼트와 갈등을 빚고 있다. 클라라는 소속사 회장으로부터 성적 수치심을 느꼈다고 주장했지만 한 매체의 문자 공개 후 여론은 급변했다.

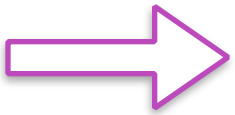


클라라의 거짓말이 새삼 화제다. 28일 방송된 SBS '한밤의 TV연예'에서는 스타들을 향한 대중의 분노와 스타들의 잘못된 대처에 대한 분석을 전했다.

4. 키워드 추출하기 – 형태소 분석

클라라Noun, 의Josa, 거짓말Noun, 이Josa, 새삼Noun, 화제Noun, 다Josa, .Punctuation, 28Number

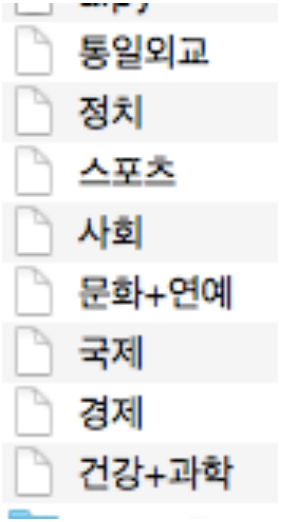
기사 본문의 **명사**만 잘라내어 분류 합니다.



클라라
나이
실제
나이
살
한국

3. 분석하기

5. 분야별 기사의 키워드를 TXT 파일로 저장합니다.



6. 수집된 데이터를 TF-IDF 알고리즘을 이용하여 분석합니다.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

3. 분석하기

1. Term Frequency : tf

단어 빈도수는 말 그대로 "단어가 그 문서에서 나타난 횟수"를 나타냅니다.

	문화 연예	사회	스포츠	정치	국제	통일 외교	경제
클라라	110	4	7	0	0	0	0
제일제당	0	0	0	0	0	0	16
한국	31	32	3	1	21	1	11

문서 내에서 각 단어의 중요도 파악 !!!

3. 분석하기

1. Term Frequency : tf

단어 빈도수는 말 그대로 "단어가 그 문서에서 나타난 횟수"를 나타냅니다.

	문화 연예	사회	스포츠	정치	국제	통일 외교	경제
클라라	110	4	7	0	0	0	0
제일제당	0	0	0	0	0	0	16
한국	31	32	3	1	21	1	11

문서 내에서 각 단어의 중요도 파악 !!!

3. 분석하기

2. Document Frequency : df

문서 빈도수는 "해당 단어가 나타난 문서의 수" 입니다.
내가 찾을 문서에선 많이 나타나고, 다른 문서에선 적게 나타날 수록 중요하다.

	문화 연예	사회	스포츠	정치	국제	통일 외교	경제
클라라	110	4	7	0	0	0	0
제일제당	0	0	0	0	0	0	16
한국	31	32	3	1	21	1	11

위 차트에서 제일제당의 df 값은 1이고 한국의 df 값은 7입니다.

3. 분석하기

2. Document Frequency : df

문서 빈도수는 "해당 단어가 나타난 문서의 수" 입니다.
내가 찾을 문서에선 많이 나타나고, 다른 문서에선 적게 나타날 수록 중요하다.

	문화 연예	사회	스포츠	정치	국제	통일 외교	경제
클라라	110	4	7	0	0	0	0
제일제당	0	0	0	0	0	0	16
한국	31	32	3	1	21	1	11

위 차트에서 제일제당의 df 값은 1이고 한국의 df 값은 7입니다.

3. 분석하기

3. Inverse Document Frequency : idf

역문서 빈도수는 df를 역수 취한 것입니다.

df값이 클 수록 중요하지 않은 단어를 나타내는 것인데,

이것을 반대로 값이 클수록 중요한 단어로 나타내기 위하여 역수를 취하는것입니다.

	tf	df	전체 문서의 개수(N)	idf
제일제당	*16*	*1*	*7*	*7/1*

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

4. 도출된 결론

	tf	df	전체 문서의 개수(N)	idf
제일제당	*16*	*1*	*7*	*7/1*

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

제일제당의 TF-IDF값 : 16*log(7/1) = 35.155

4. 도출된 결론

제일제당의 TF-IDF값 : 16*log(7/1) = 35.155

경제	오닌	35.156
경제	제일제당	35.156
경제	소상	22.181
경제	아미노산	21.972
경제	제품	21.972
경제	창업	21.972
경제	한전	20.794

각 키워드의 tf-idf값을 기준으로 정렬을 하면
분야별 **핵심 키워드**를 찾을 수 있습니다.

4. 도출된 결론

제일제당의 TF-IDF값 : 16*log(7/1) = 35.155

경제	오닌	35.156
경제	제일제당	35.156
경제	소장	22.181
경제	아미노산	21.972

CJ제일제당, 메치오닌시장 본격 공략 파이낸셜뉴스 | 22분전 | 네이버뉴스

말련 메치오닌공장 가동 "5兆 글로벌시장 1위 도약" CJ**제일제당**이 50억달러(5조4310억원) 규모의 사료용 아미노산 메치오닌 시장에 본격 진출한다. 29일 CJ**제일제당**은 프랑스 아르케마사와 손잡고 말레이시아에 총 4억달러...

- CJ**제일제당**, 5조원 규모 메치오닌 시장 본격 공략 아시아경제 | 9시간전 | 네이버뉴스
- CJ**제일제당**, 메치오닌 시장 공략 강화..'세계 1위… 이데일리 | 9시간전 | 네이버뉴스

관련뉴스 전체보기>

관련뉴스 48건 전체보기>

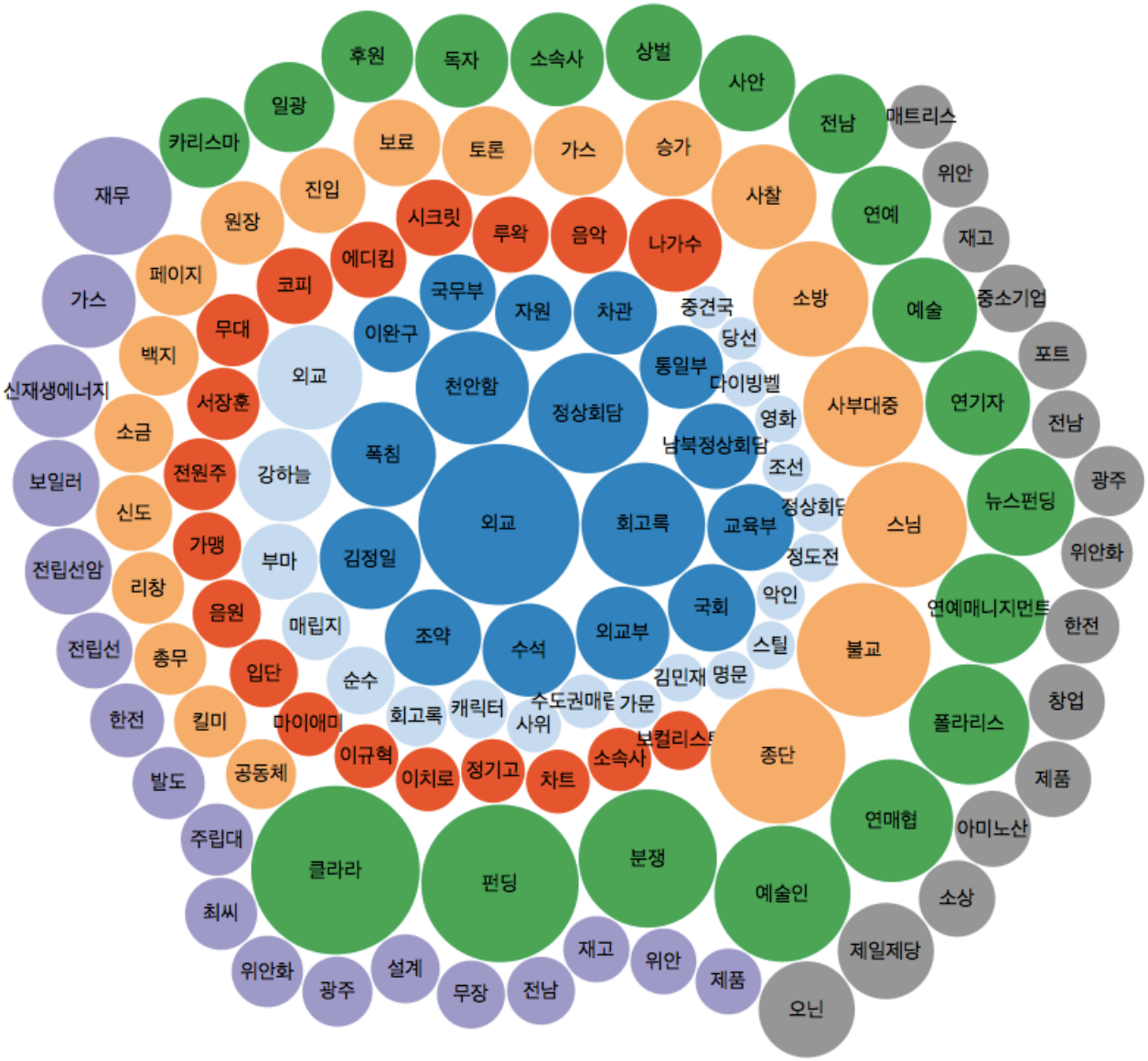
관련뉴스 18건 전체보기>

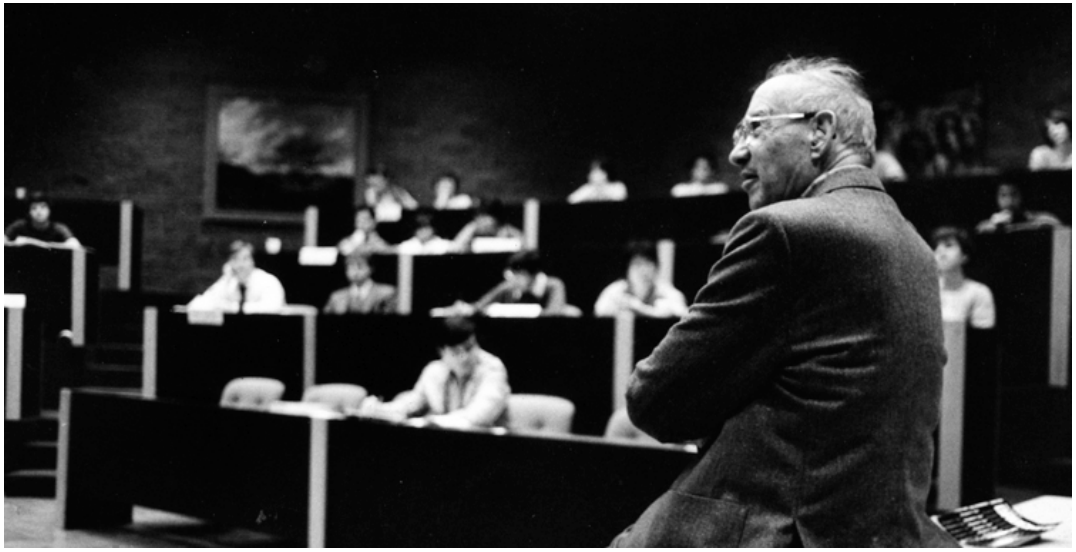
메치오닌의 형태소 분석결과 : [메Verb, 치PreEomi, 오닌Noun]

4. 도출된 결론

(D3 buble chart)

- 통일/외교
- 정치
- 스포츠
- 사회
- 문화/연예
- 국제
- 경제





미래를 예측하는 가장 좋은 방법은 미래를 창조하는 것이다.

– 피터 드러커 –

Q & A