

# Elasticsearch Fulltext 검색 서비스 적용기

기술 VS 서비스

아프리카TV | 정경석

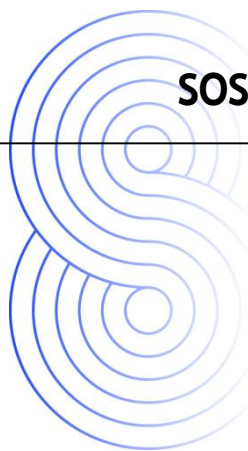
2018.10.18

**SOSCON 2018**

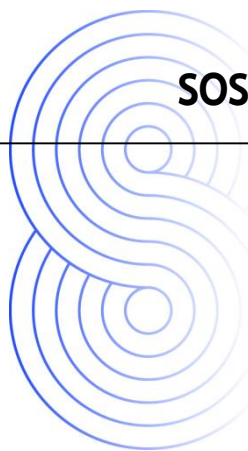
SAMSUNG OPEN SOURCE CONFERENCE 2018



- 아프리카TV 검색/캐시 클러스터
- 기술서 저자
- 2015(netty), 2017(cache) soscon 발표
- 명실상부한 공라인
  - Redis
  - Netty
  - Elasticsearch



## 기술 VS 서비스



# 첫 번째 이야기

게시물 본문검색!

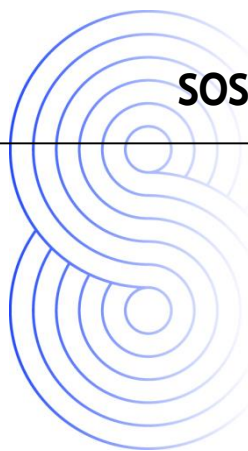
**SOSCON 2018**

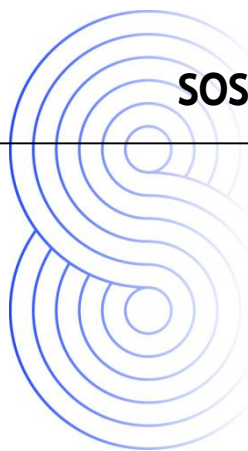
SAMSUNG OPEN SOURCE CONFERENCE 2018



게시물의 본문 내용이 검색 되면 돼!!!

Like 쿼리 쓰면 되겠네. 간단하지?





시간이 좀 걸려도 제대로 된 방법을 찾을 것인가?

이상과 현실의 괴리... 일정!!



개발자 친화적인 방법



- Select no, title, ... from bbs\_table  
where contents like '%배틀그라운드%'
  - 데이터가 많아질수록...
  - 사용자가 많아질수록...

Disk IO, CPU 리소스 Full!!!

### 개발자 친화적인 방법 2



- `Select no, title, ...`  
`from bbs_table`  
`where match(content) against( '배틀그라운드' )`
  - 성능과 결과는 괜찮다. 아직까지는...
  - 데이터가 많아지고 검색할 필드가 더 많아지면?

수평 확장 불가. 대안이 필요하다!!

검색 엔진 도입 Elasticsearch



elastic

- Apache Solr 대비 검색 트렌드 상승세
- 다양한 플러그인과 에코시스템
- 실시간 인덱싱
- 잘 정리된 공식문서
- 수평확장과 분산처리

가즈아!!!

## 설치와 실행



- `wget https://artifacts.elastic.co/downloads ...`
- `tar xvf elasticsearch-XXX ...`
- `vi config/elasticsearch.yml`
  - `bootstrap.system_call_filter: false`
  - ...
- `bin/elasticsearch -d`

쉽군!!

## 데이터 인덱싱

- <https://www.elastic.co/guide/en/elasticsearch/client/java-api/current/transport-client.html>
- `org.elasticsearch.client.transport.TransportClient`
- `org.elasticsearch.action.bulk.BulkRequest`
- `org.elasticsearch.action.bulk.BulkProcessor`
- Java, JavaScript, Groovy, .NET, PHP, Python ...

예제가 많아서 허들이 적음(Ctrl+C, Ctrl+V)

## 질의 - Rest API

- POST bbs\_index/\_search  
{"query":{"term":{"contents":"배틀그라운드"}}, "from":0, "size":10}
- {"took":159, "timed\_out":false, "\_shards":{"total":5, "successful":5, "failed":0}, "hits":{"total":12755, ...}}

데이터가 잘 나오는군!!

그러나...

- POST bbs\_index/\_search  
{"query":{"term":{"contents":"배틀"}}, "from":0, "size":10}
- {"took":62, "timed\_out":false, "\_shards":{"total":5, "successful":5, "failed":0}, "hits":{"total":425, ...

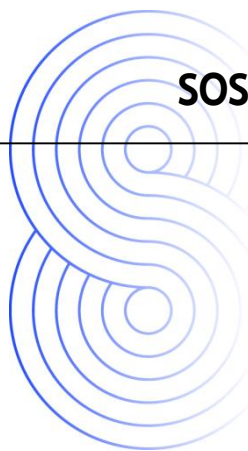
어?!!

## 형태소 분석

‘의미를 가지는 최소 단위’로 잘라내는 행위

- 감나무에 감이 절로 열렸다.
  - 감나무에 / 감이 / 절로 / 열렸다
  - 감(어근)/ 나무(어근) / 예(조사) / 감(어근) / 이(조사) / 절로 (어근) / 열(어근) / 리(피동접사) / 었(선어말 어미) / 다(어말어미)

나는 개발자인가 국어학자인가?!



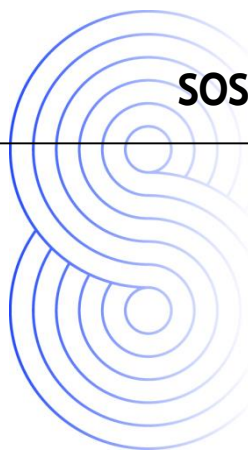
검색과 형태소 분석 왜?

- “감나무에 감이”를 검색하면
  - 옆집 감나무에 감이, 대추나무에 대추가 열린 것을 보니...
  - 감나무에 달린 감이 가을을 재촉하네요.
  - 행복한감나무농장의 감이 익어갑니다.
  - 매장 뒤뜰 감나무에 감이 주렁주렁
  - ...

네이버 검색 결과

## Built-in

- Standard Analyzer
- Whitespace Analyzer
- Stop Analyzer
- Pattern Analyzer
  
- Custom analyzer
  - Combined, third-party-plugin  
(아리랑, 은전한닢, open korean text, Nori)



## 아리랑 형태소 분석기

- Github : <https://github.com/HowookJeong/elasticsearch-analysis-arirang>
- <http://jjeong.tistory.com/category/Elastic/Elasticsearch>

The screenshot shows the GitHub repository page for 'HowookJeong / elasticsearch-analysis-arirang'. The repository is described as 'korean analyzer (lucene analyzer kr arirang)'. It has 14 commits and 11 branches. The current branch is 'master'. A recent commit by 'henry\_정호욱' is shown with the message 'clean.'.

jjeong

실무 예제로 배우는 Elasticsearch 검색엔진

Home

Location

Tags

Media

Guestbook

Admin

New Post

📖 도서 정보 보기 (2018)

## 'Elastic/Elasticsearch'에 해당되는

1. 2018.10.04 [Elasticsearch] \_analyze 예제 - 특수문자 제
2. 2018.08.01 [Elasticsearch] id max length
3. 2018.07.25 [Elasticsearch] Dynamic template - \_default
4. 2018.07.05 [Elasticsearch] Default field type 값.
5. 2018.07.04 [Elasticsearch] 그냥 settings, mappings, tem
6. 2018.07.03 [Elasticsearch] Nested Query 는 Script Fields
7. 2018.06.28 [Elasticsearch] HighLevelRestClient 를 이용
8. 2018.06.27 [Elasticsearch] \_all field 사용하지 마세요.
9. 2018.06.27 [Elasticsearch] Move TransportClient to High
10. 2018.06.27 [Elasticsearch] 6.3 설치 변경 내용

## 아리랑 형태소 분석기

- POST bbs\_index/\_search  
{"query":{"term":{"contents":"배틀그라운드"}}, "from":0, "size":10}
- {"took":29, "timed\_out":false, "\_shards":{"total":5, "successful":5, "failed":0}, "hits":{"total":12375, ...}}

Good...

## 아리랑 형태소 분석기

- POST bbs\_index/\_search  
{"query":{"term":{"contents":"배틀"}}, "from":0, "size":10}
- {"took":29, "timed\_out":false, "\_shards":{"total":5, "successful":5, "failed":0}, "hits":{"total":18351, ...

관찰군...

## 아리랑 형태소 분석기

- POST bbs\_index/\_search  
{"query":{"term":{"contents":"불닭먹방"}}, "from":0, "size":10}
- {"took":29, "timed\_out":false, "\_shards":{"total":5, "successful":5, "failed":0}, "hits":{"total":8, ...

응?

```
POST _analyze
{"analyzer":"arirang_analyzer", "text": "불닭먹방" }

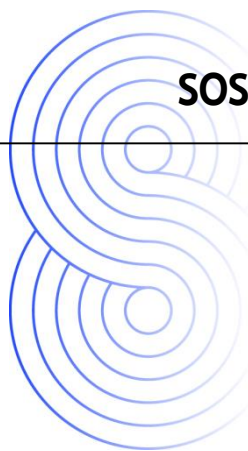
{
  "tokens": [
    {
      "token": "불닭먹방",
      "start_offset": 0,
      "end_offset": 4,
      "type": "korean",
      "position": 0
    }
  ]
}
```

불닭, 먹방이 사전에 등록되어 있지 않음.

쿼리 결과 이상의 원인은 사전

- 아리랑 형태소 사전
  - total.dic
  - synonym.dic
  - extension.dic
  - compounds.dic
  - josa.dic
  - ...

검색되지 않는 단어가 또 나오면?



# 기획자가 원했던건...

SOSCON 2018



강은비.(kscry12) 2018-10-04 18:26 | 조회수 358

## 10월 4일 목요일 오후 9시~30분 사이 방송 시작!!

은녕하세요 ~ 강은비입니다 !! 와우 ~~ 큰일이다 큰일 .. 낚시 갈 줄 알고 맘준비 단단히하고 있었는데 태풍으로 낚시 담  
주로 미뤄져서 뭔가 맘이 답답하여 건담베이스 갔는데 .....



65



38



공유하기



강은비.(kscry12) 2018-07-02 23:03 | 조회수 1,297

## 기부스 & 축구 중계 방송 공지 입니다

안녕하세요~ 강은비 입니다! 기부스는 격주로 인사드려요 ^^ 기부스 방송국으로  
블러오세요! <http://www.afreecatv.com/gibo...>



85



30



공유하기

내용

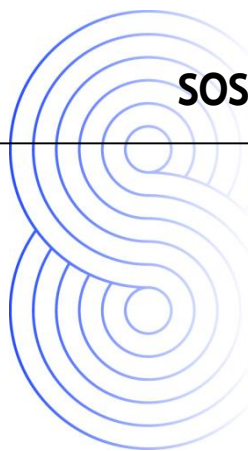
주로

검색

Like 검색

N-gram

- 1-gram, 2-gram, 3-gram
- 아버지가 방에 들어가신다  
➔ 아버, 버지, 지가, 방에, 들어, 어가, 가신, 신다



## N-gram

- POST bbs\_index/\_search  
{"query":{"**match**":{"contents":"**불닭먹방**"}},"from":0,"size":10}
- {"took":29,"timed\_out":false,"\_shards":{"total":5,"successful":5  
,"failed":0},"**hitstotal**":**421**, ...}

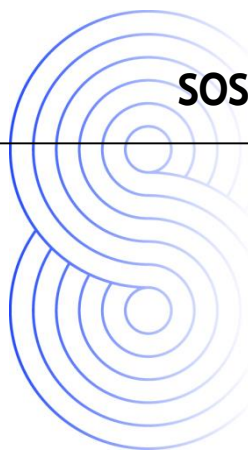
기획자가 원했던 결과가 출력됨.. 오버하지 말자...

핵불닭볶음면 먹방

- slop : 형태소간의 거리

```
{ "query": {  
  "match_phrase": {  
    "content": {  
      "query": "불닭먹방",  
      "slop": 3  
    }  
  }  
}}
```

적당한 값은 항상 테스트를 통해서.

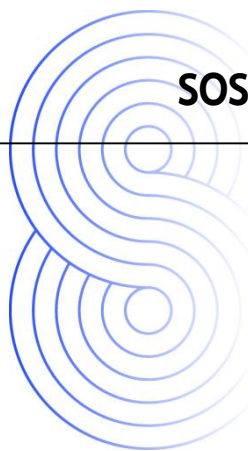


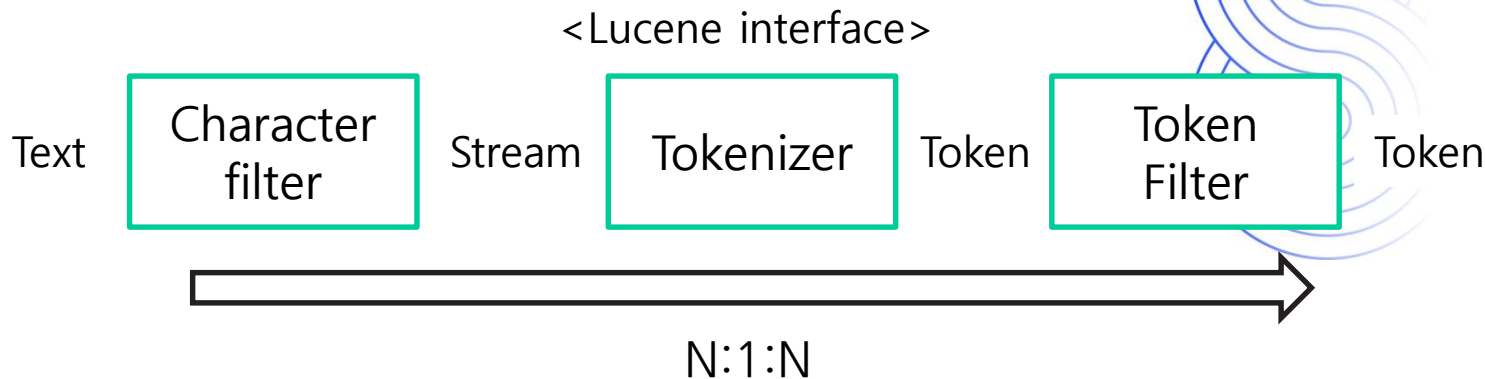
- 자연어 검색이란
  - 의미가 같은 글을 검색할 수 있도록 하는 일련의 행위  
문을 달는다 ➔ 문(명사) 달다 (품사 분해)  
시피유 가격 ➔ CPU 가격, 시피유 가격, 씨피유 가격(동의어 처리)  
건물 사이길 ➔ 건물 사잇길, 건물 샛길(맞춤법 교정)
- 서비스에 맞는 기술 선택인가?
  - 게시판 검색, 아이디 검색

적정기술!

- 형태소 분석기
  - 문장을 적절하게(?!) 분해 및 Tagging
  - 한글 형태소 분석과 사전(불닭, 먹방)
  - 분석기 내부구조
- 매핑(Schemaless?)
  - 필드 타입, 형태소 분석기 바인딩
- 쿼리
  - term, match, match\_phrase, regexp...
- 형태소 분석과 검색
  - 인덱싱, 쿼리

더 많은 기능들은 모두 레퍼런스 문서에!!





- Standard Analyzer
  - Character filter : None
  - Tokenizer : StandardTokenizer
  - Token Filter : StandardFilter
  - Token Filter : LowerCaseFilter
  - Token Filter : StopFilter

Filter chain, Template method 패턴

<Elasticsearch plugin interface>

<Lucene interface>

Character  
filter

Tokenizer

Token  
Filter

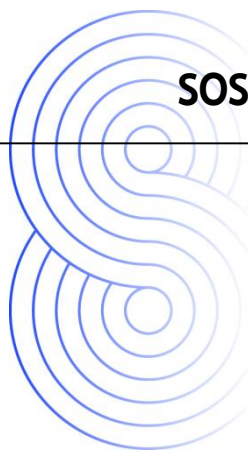
<Arirang, MeCab(메카부)>

Custom 형태소 분석기 작성시 2~3개의 프로젝트 필요!

## 필드 Type, 형태소 분석기

```
{  "mappings": {  
    "bbs": {  
        "properties": {  
            "contents": {  
                "type": "text",  
                "analyzer": "my_ngrams"  
            },  
            "date": {  
                "type": "date"  
            },  
            "name": {  
                "type": "keyword"  
            },  
            "user_id": {  
                "type": "long"  
            }  
        }  
    }  
}
```

string vs keyword vs text



Elasticsearch의 SQL

- match\_all
- term
- exists
- prefix
- match
- match\_phrase
- ...

<https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl.html>

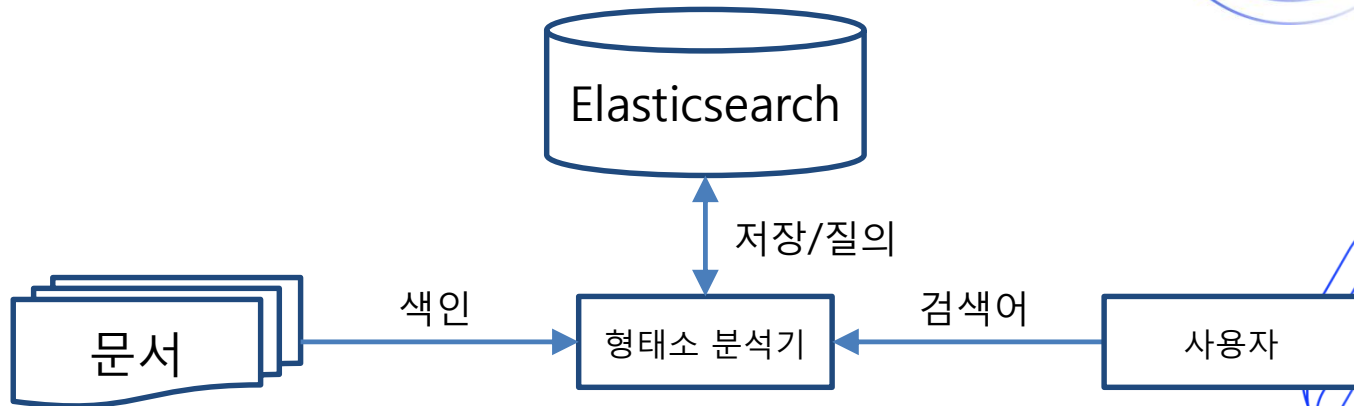
## 형태소 Tagging

- 형태소 분석기 별 tag의 이름 값이 다름.

| LA다저스 선발출장 |        |      | 검색 |
|------------|--------|------|----|
| Sever      | Toeken | Type |    |
| test01     | la     |      |    |
|            | COMMON |      |    |
|            | 다저스    |      |    |
|            | COMMON |      |    |
|            | 선발     |      |    |
|            | COMMON |      |    |
|            | 출장     |      |    |
|            | COMMON |      |    |

Tokenfilter, 검색 의도 파악

- 분석기와 검색의 관계
  - 키워드와 문서 모두 형태소 분석



# 두 번째 이야기

## 동의어 검색

**SOSCON 2018**

SAMSUNG OPEN SOURCE CONFERENCE 2018



검색

배그



통합검색

생방송

VOD

BJ

연관검색어 [스타](#) [오버워치](#) [배그](#) [김성태](#) [서든](#) [깨박이](#) [피파](#) [lol](#) [시참](#) [리니지m](#) [뜨뜨뜨뜨](#) [남캠](#) [소통](#)

정확도순

시청인원순

최신순

고화질순

고해상도순



30개씩 보기

생방송 ( 332건 )

LIVE



[\[생\]김성태 VS 훈수출 오늘 배틀그라운드](#)

BJ정보 BJ명 [항상#김성태\(rvvv17\)](#) | 방송국명 BJ김성태

사형정보 👤 **13,357** | 방송시작 2018-09-27 19:52:20 | 화질/해상도 8000K/ 1920x1080



[\[생\]상어녀 15킬 치킨 배틀그라운드](#)

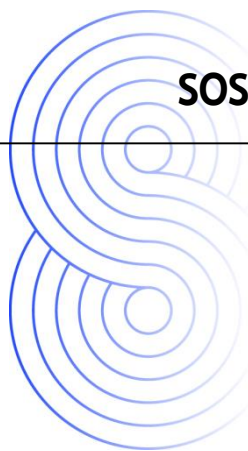
BJ정보 BJ명 [상어녀\(595935\)](#) | 방송국명 상어녀

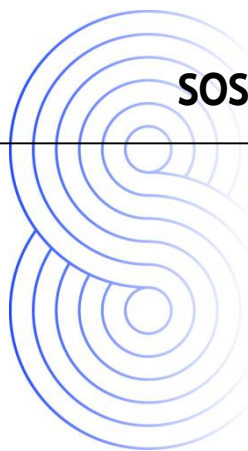
사형정보 👤 **1,041** | 방송시작 2018-09-27 10:21:26 | 화질/해상도 8000K/ 1920x1080

## 동의어 사전 등록

- 글자는 다르지만 같은 의미를 가진 단어
  - 배그 ➔ 배틀그라운드
  - 가난 ➔ 빈곤
  - 칭구 ➔ 친구
  - 불소 ➔ 블레이드소울
  - 땡땡이 ➔ 멍멍이(?)
  - 땡작 ➔ 명작(?)

사전 관리 필요!





싱크홀을 검색 했는데 검색 결과가 없음 CS.

- 기대결과 : “한나가 들려주는 싱크홀 이야기”



네이버는?



**네이버**

싱크홀

통합검색 이미지

연관검색어 ? 24번국도

싱크홀

시간대와 관심사에 맞춘 **컨텍스트 자동완성** ? ON OFF

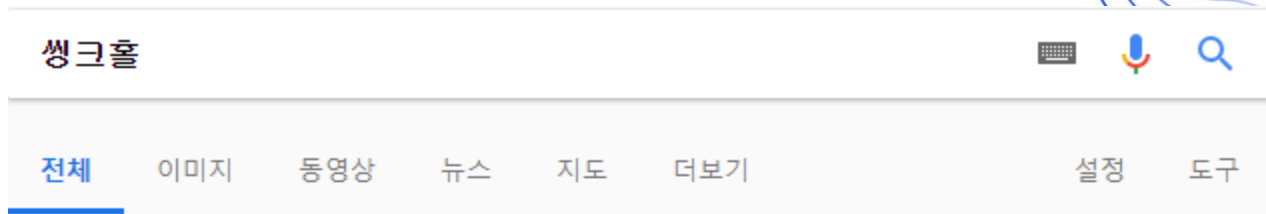
도움말 신고 자동완성 끄기

**지식백과**

**싱크홀** 지하 암석이 용해되거나 기존의 동굴이 붕괴되어 생긴 움덩이

 **싱크홀** 땅이 원통 혹은 원뿔 모양으로 꺼지는 현상으로, 균열대(지층이 머긔나 균열이 생

구글은?



검색결과 약 4,620,000개 (0.32초)

수정된 검색어에 대한 결과: **싱크홀**

다음 검색어로 대신 검색: **싱크홀**

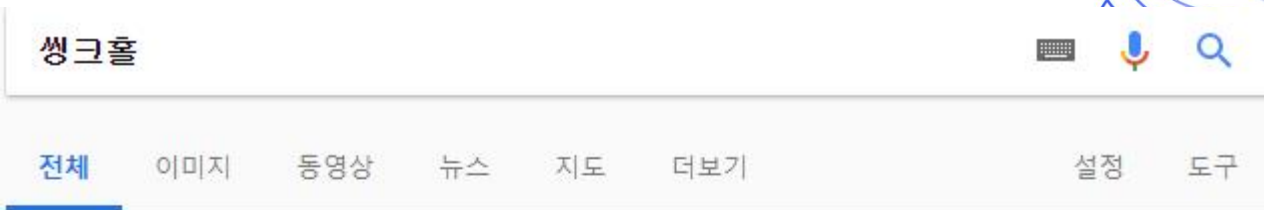
**싱크홀 - 나무위키**

<https://namu.wiki/w/싱크홀> ▼

2018. 9. 10. - 해외의 용어인 싱크홀을 그대로 사용했다가 국립국어원에서 '땅꺼짐'이라는 순화어를 제시한 뒤 2015년 하반기 기준으로 언론에서는 '땅꺼짐 현상' ...



구글은?



이것을 찾으셨나요? **싱크홀**



국카스텐 - 싱크홀(EBS 스페이스 공감 494회) - YouTube

아프리카TV는?

afreecaTV

검색

씽크홀



통합검색

생방송

VOD

BJ

정확도순

시청인원순

최신순



| 30개씩 보기



| 1달

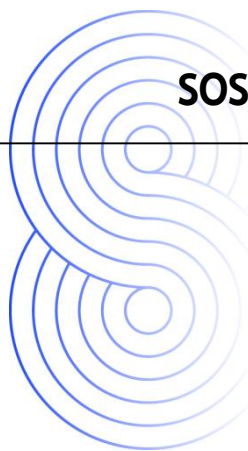
입력하신 검색어 '씽크홀'에 대한 검색결과가 없습니다.

검색어의 단어 수를 줄이거나, 특수문자를 제외하고 검색해 보세요.

- 네이버 : 싱크홀, 씩크홀 두 결과를 합쳐서 출력
- 구글 : 검색어를 싱크홀로 변경한 결과와 변경 전 결과 출력(케바케)
- 아프리카TV : 오타자 검증 시스템이 없으므로 동의어 처리 결정.

사실은...

- 국립국어원 : 씹크홀(X), 싱크홀(O)
- 네이버 국어사전 : 씹크홀(X), 싱크홀(O)
- 구글 : 씹크홀을 오타자로 인식 ➔ correction!
- 네이버 : 씹크홀을 싱크홀의 동의어로 처리



그래서 아프리카TV는?

afreecaTV

검색

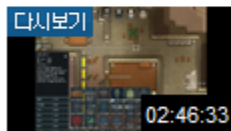
싱크홀



통합검색 생방송 VOD BJ

"싱크홀"에 대한 통합검색 결과입니다. (1,981건)

VOD ( 1,979건 )



[\[생\]싱크 림월드 /  
우주생존시물](#)

싱크

2018-10-07 22:52



[한나가 들려주는 싱  
크홀 이야기](#)

[2]

핫동TV



[\[생\]싱크 3분이 다크  
엘프!](#)

싱크

2018-10-05 18:44



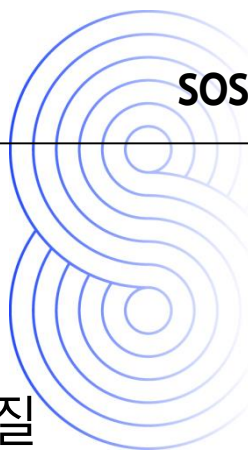
[\[생\]싱크 리니지 투  
사회장30장 원정러쉬](#)

싱크

2018-10-05 18:15



- 형태소 분석기만 바뀌서는 드라마틱한 변화 없음.
- 형태소 분석기의 품질 == 사전과 Tokenizer의 품질
- ES 6.4의 플러그인 작성
  - ES 소스코드의 gradle esplugin
- Split brain
  - S/W down시 발생
- 사전관리의 실패는 서비스 장애
  - 장애 방지를 위한 방안 수립 필요



**감사합니다.**

**SOSCON 2018**

SAMSUNG OPEN SOURCE CONFERENCE 2018

