

Elasticsearch 의 한글 검색 활용

Elastic | Community Engineer | 김종민

2018. 10. 18

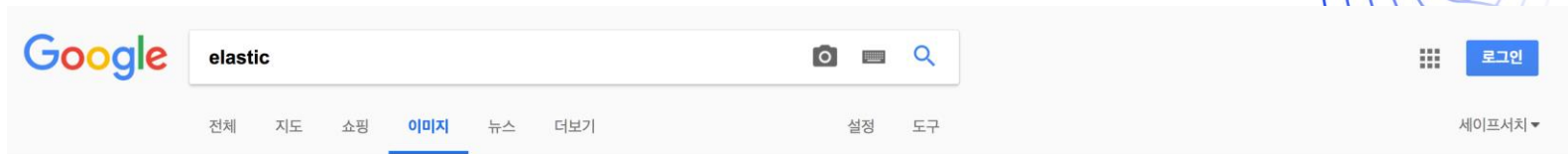
SOSCON 2018

SAMSUNG OPEN SOURCE CONFERENCE 2018



Elastic?

SOSCON 2018



elastic · GitHub
github.com



Open Source Search & Analytics · Elasticsearch | Elastic
elastic.co



Elastic to Release Source Code for X-Pack
datanami.com



Wholesale Trade Suppliers 4 cord, 3m...
supplydivision.com



Knitted Elastic Tape, Buna Hu...
indiamart.com



Tools we use: Curator for Elasticsearch | Postmark
postmarkapp.com



Elastic Band Ball | Stationery ...
bmstores.co.uk



4 inch (100mm) Wide Patterned Colored ...
strapcrafts.com



Elastic Images, Stock Photos & Vectors | ...
shutterstock.com



The Elastic Stack Download · Ge...
elastic.co



Elastic?

SOSCON 2018

- Elasticsearch 라는 검색엔진을 개발한 회사입니다.
 - (ELK Stack 으로 더 잘 알려져 있습니다.)
- 검색엔진은 우리 주변 여기저기에 있습니다.
- 요즘은 검색엔진이 데이터 분석에도 쓰입니다.

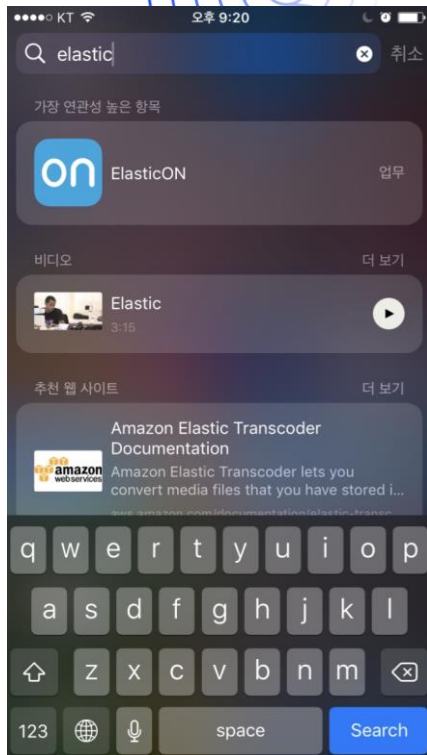


Google 검색

I'm Feeling Lucky



NAVER



SEARCH



Multilingual

The screenshot shows the Wikipedia homepage with the title "WIKIPEDIA The Free Encyclopedia" at the top. Below the title is a globe made of puzzle pieces with various characters. Surrounding the globe are links to different language editions of Wikipedia, each with the number of articles. A search bar is located at the bottom of the page, and a dropdown menu is open showing search results for the term "search".

Language	Number of Articles
English	5 430 000+ articles
日本語	1 065 000+ 記事
Deutsch	2 074 000+ Artikel
Français	1 882 000+ articles
中文	948 000+ 條目
Polski	1 229 000+ hasel
Português	971 000+ artigos
Italiano	1 365 000+ voci
Русский	1 403 000+ статей
Español	1 341 000+ artículos

Search results for "search":

- Searching**
Wikimedia disambiguation page
- Search engine optimization**
search engine optimization is the practice of increasing web page visibility in search engine results pages
- Search and rescue**
search for and provision of aid to people who are in distress or imminent danger
- Search for extraterrestrial intelligence**
effort to find civilizations not from Earth

Full Text Search

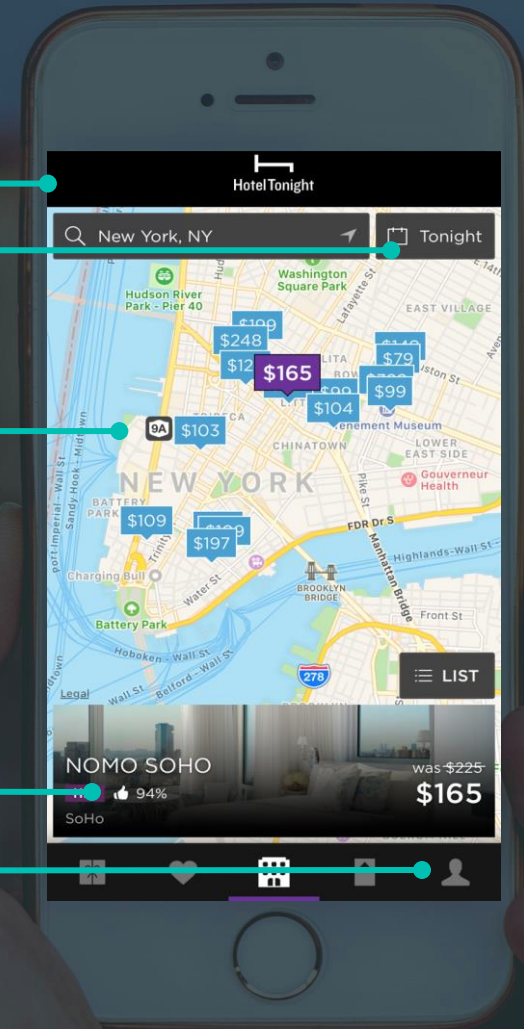
Stemming

Type ahead

Mobile
Time Range

Geo search

Influenced by Rating
Personalized Ranking



[Sign Up](#)[Restaurants](#)[Delivery](#)[Reservations](#)[Write a Review](#)[Events](#)[Talk](#)[Log In](#)

Burritos Mission District, San Francisco, CA

Showing 1-10 of 105

\$

\$\$

\$\$\$

\$\$\$\$

Open Now

Order Delivery

Order Pickup

All Filters

Sort By

Best Match[Highest Rated](#)[Most Reviewed](#)

Cities

☒ Mission☐ San Francisco[More Cities](#)

Distance

[Bird's-eye View](#)[Driving \(5 mi.\)](#)[Biking \(2 mi.\)](#)[Walking \(1 mi.\)](#)[Within 4 blocks](#)

Price

☒ \$☐ \$\$☐ \$\$\$☐ \$\$\$\$

Features

☒ Open Now 1:56 PM☐ Order Delivery☐ Order Pickup☐ Make a Reservation[More Features](#)

Category

☐ Mexican☐ Food☐ Sandwiches☐ Food Trucks[More Categories](#)

1. El Farolito

★★★★★ 4091 reviews

\$ · Mexican, Seafood, Sandwiches

On the menu: Burrito

Mission

2779 Mission St
San Francisco, CA 94110
(415) 824-7877

This is the busiest **burrito** place on Mission in my time at SF, and I think it's due to the fact that it's on Mission, birthplace of the Mission **Burrito** known worldwide, but also that... [read more](#)



2. La Taqueria

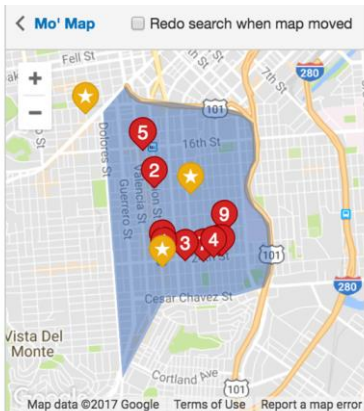
★★★★★ 2995 reviews

\$ · Mexican

Mission

2889 Mission St
San Francisco, CA 94110
(415) 285-7117

I'm not a huge fan of **burritos** (I prefer rice bowls) but this place has changed my mind. This place is known for their **burritos** and specifically without rice! When we arrived, there... [read more](#)



Search

Pagination

Time range Filter

Numeric Filter

Geo range Filter

Stemming /
Highlighting

The screenshot displays a Yelp search results page for 'Burritos' in the Mission District, San Francisco, CA. The page is annotated with red lines and boxes highlighting specific features:

- Search:** The search bar at the top contains 'Find Burritos' and 'Near Mission District, San Francisco, CA'.
- Pagination:** The top right corner shows 'Showing 1-10 of 105'.
- Time range Filter:** The 'Open Now' filter is highlighted in a red box.
- Numeric Filter:** The 'Price' filter is highlighted in a red box, showing options for \$, \$\$, \$\$\$, and \$\$\$\$.
- Geo range Filter:** The map on the right side of the page is highlighted in a red box, showing the location of the results in the Mission District.
- Stemming / Highlighting:** The word 'Burrito' is highlighted in a red box in the first restaurant listing, and 'burritos' is highlighted in a red box in the second restaurant listing.

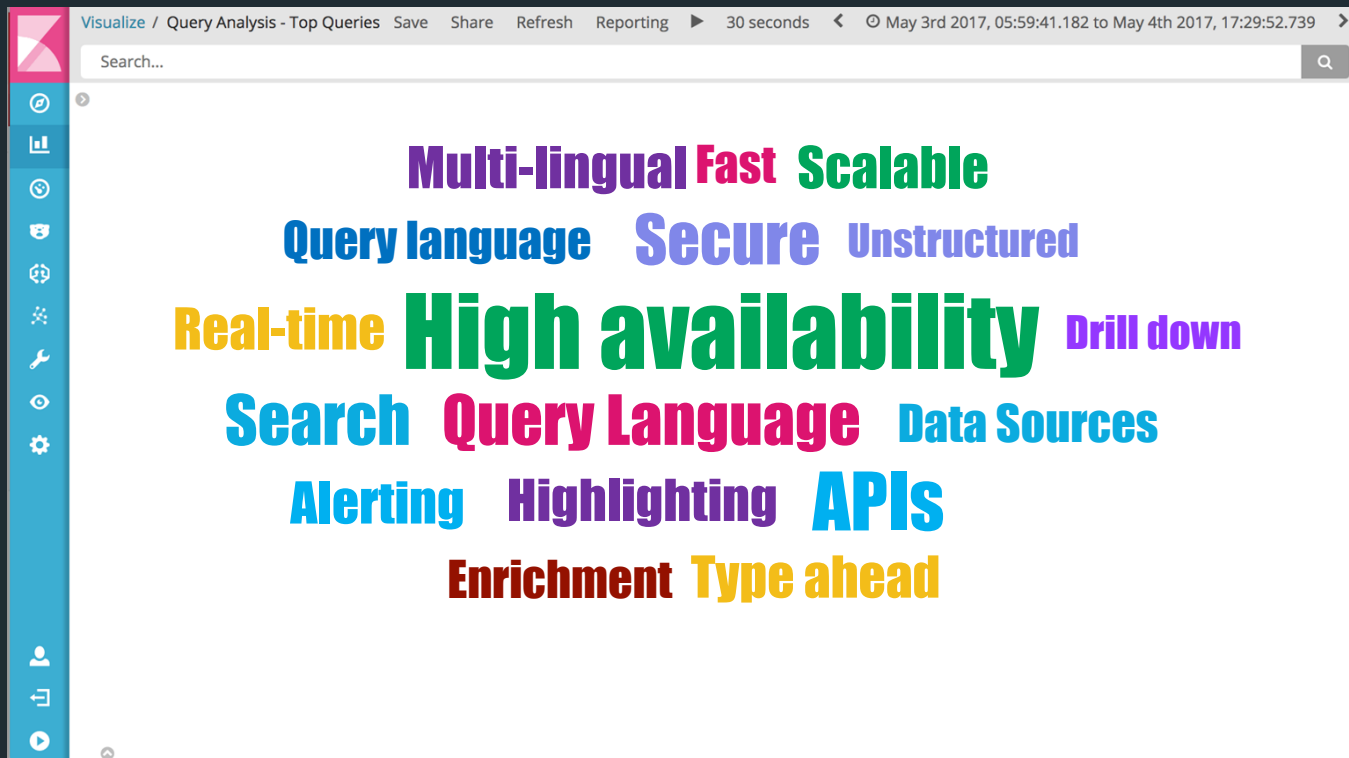
The first restaurant listed is '1. El Farolito' with 4091 reviews, located at 2779 Mission St, San Francisco, CA 94110. The second restaurant listed is '2. La Taqueria' with 2995 reviews, located at 2889 Mission St, San Francisco, CA 94110.

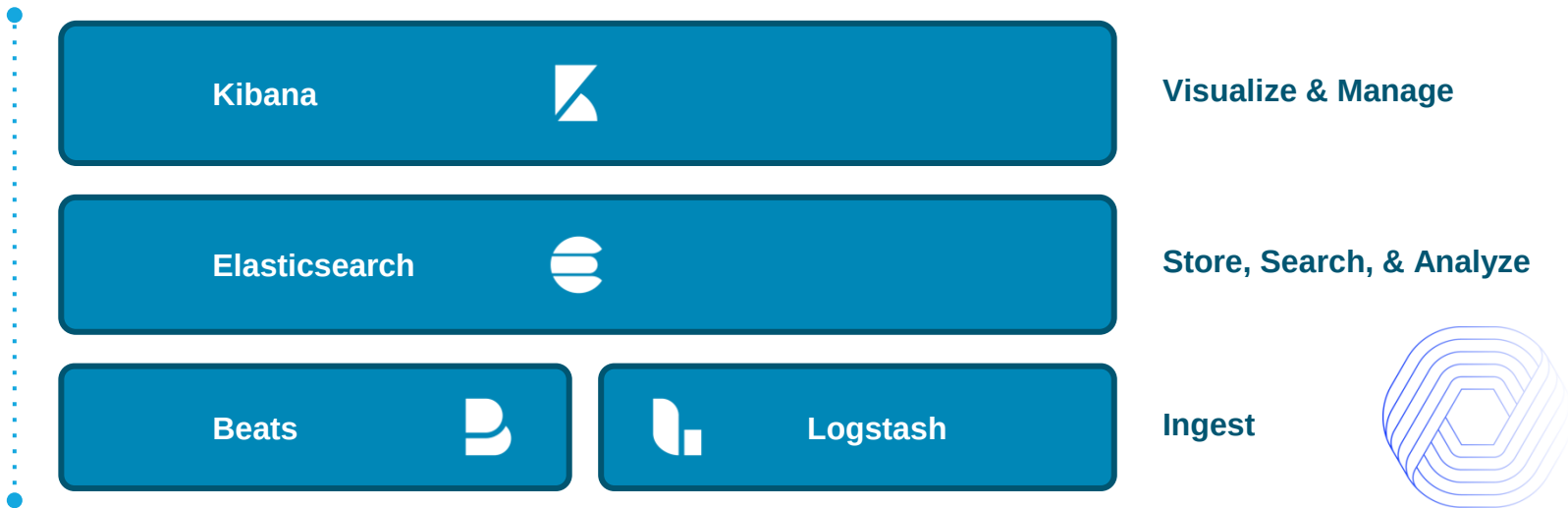
Building search needs ...

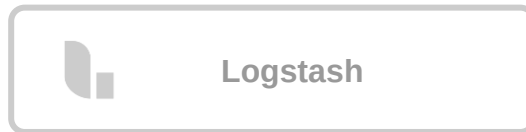
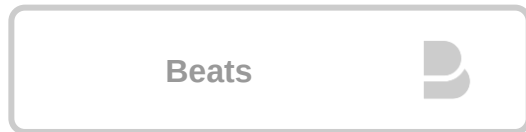
All this.

Multi-lingual Secure Type ahead
Query language *Fast* Unstructured Scalable
Real-time **High availability** Alerting
Drill down Highlighting Query Language
Search APIs Data Sources Enrichment

Elastic has all this.







SOSCON 2018 Solutions

Visualize & Manage

Store, Search, & Analyze

Ingest



Solutions



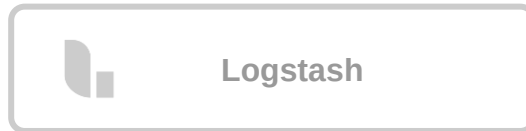
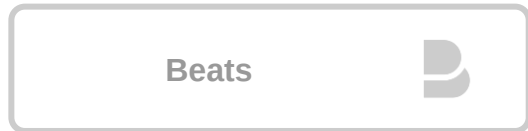
APM



Site Search



App Search



SOSCON 2018

Solutions

Visualize & Manage

Store, Search, & Analyze

Ingest



Deployment



SaaS



Elastic Cloud



Elasticsearch
Service



Site Search



App Search

Self Managed



Elastic Cloud
Enterprise



Standalone




**Elastic
Stack**



SOSCON 2018
Solutions

Visualize & Manage

Store, Search, & Analyze

Ingest



Deployment

왜 오픈 소스인가?

오픈 소스는 우리 커뮤니티와 우리 비즈니스의 강점입니다. 우리가 하는 모든 것의
중심에 오픈 소스가 있습니다. 그 이유를 알려드리겠습니다.

OPT
IN

DEFAULT

PRODUCT.ZIP

OPEN

PAID

FREE

X-PACK

OSS

Data Collection

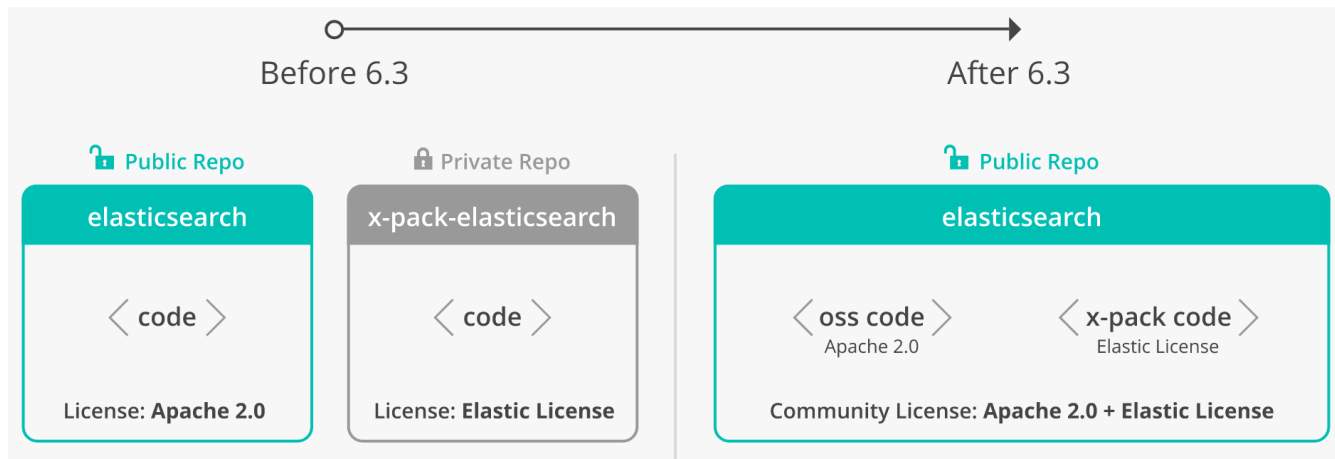
Visualizations

Enrichment

Distributed



We Opened X-Pack





elasticsearch

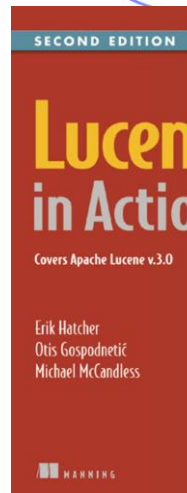
아파치 루씬 (Apache Lucene)

- Created by - Doug Cutting
- Written in - Java
- Apache Solr, Elasticsearch



SOSCON 2018

Lucene



RDBMS에서는 데이터를 테이블 형태로 저장합니다.

SOSCON 2018

열을 기준으로 인덱스를 만듭니다.
책의 맨 앞에 있는 제목 리스트와 같습니다.

DOC	TEXT
1	The quick brown fox jumps over the lazy dog
2	Fast jumping rabbits

검색엔진에서는 inverted index 라는 구조로 저장합니다.

SOSCON 2018

RDBMS 와 반대 구조입니다.

텍스트를 다 뜯어서 검색어 사전을 만듭니다.

책의 맨 뒤에 있는 페이지를 가리키는 키워드 같습니다.

TOKEN (TERM)	DOC	TOKEN (TERM)	DOC
Fast	2	jumps	1
The	1	lazy	1
brown	1	over	1
dog	1	quick	1
fox	1	rabbits	2
jumping	2	the	1

실제로는 이렇게 저장됩니다.

SOSCON 2018

텍스트를 저장할 때 몇가지 처리 과정을 거칩니다.

TOKEN (TERM)	DOC	TOKEN (TERM)	DOC
brown	1	lazi	1
dog	1	over	1
fast	1 , 2	quick	1 , 2
fox	1	rabbit	2
jump	1 , 2		

대소문자를 변환합니다.

TOKEN (TERM)	DOC	TOKEN (TERM)	DOC
Fast → fast	2	jumps	1
The → the	1	lazy	1
brown	1	over	1
dog	1	quick	1
fox	1	rabbits	2
jumping	2	the	1

토큰을 (보통 ascii 순서로) 재 정렬합니다.

TOKEN (TERM)	DOC	TOKEN (TERM)	DOC
brown	1	lazy	1
dog	1	over	1
fast	2	quick	1
fox	1	rabbits	2
jumping	2	the	1
jumps	1	the	1

불용어(stopwords, 검색어로서의 가치가 없는 단어들) 를 제거합니다.
a, an, are, at, be, but, by, do, for, i, no, the, to ... 등등

TOKEN (TERM)	DOC	TOKEN (TERM)	DOC
brown	1	lazy	1
dog	1	over	1
fast	2	quick	1
fox	1	rabbits	2
jumping	2	the	1
jumps	1	the	1

형태소 분석 과정을 거칩니다.
보통 ~s, ~ing 등을 제거하는 과정입니다.
한글은 의미 분석을 해야 해서 좀 더 복잡합니다.

TOKEN (TERM)	DOC	TOKEN (TERM)	DOC
brown	1	lazy → lazi	1
dog	1	over	1
fast	2	quick	1
fox	1	rabbits → rabbit	2
jumping → jump	2		
jumps → jump	1		

jumping, jumps가 jump 로 똑같이 바뀌었으므로 토큰을 병합 해 줍니다.

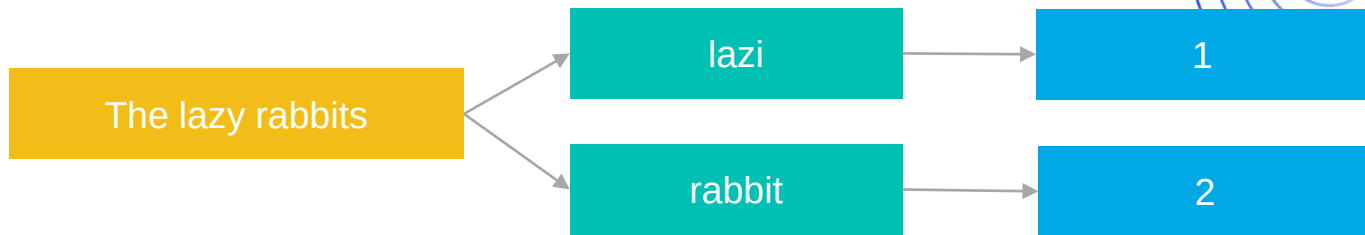
TOKEN (TERM)	DOC	TOKEN (TERM)	DOC
brown	1	lazi	1
dog	1	over	1
fast	2	quick	1
fox	1	rabbit	2
jump	1 , 2		

동의어를 처리합니다.

TOKEN (TERM)	DOC	TOKEN (TERM)	DOC
brown	1	lazi	1
dog	1	over	1
fast	1 , 2	quick	1 , 2
fox	1	rabbit	2
jump	1 , 2		

검색어도 똑같이 텍스트 처리를 합니다.

“The lazy rabbits” 라고 검색하면 “lazi”, “rabbit” 검색어를 찾습니다.



DOC	TEXT
1	The quick brown fox jumps over the lazy dog
2	Fast jumping rabbits

검색엔진과 RDBMS 비교

SOSCON 2018

RDBMS 는 쿼리 시점, 검색엔진은 색인 시점이 중요합니다.

	RDBMS	검색엔진
데이터 저장 방식	정규화	역정규화
전문(Full Text) 검색 속도	느림	빠름
의미 검색	불가능	가능
Join	가능	불가능
수정 / 삭제	빠름	느림

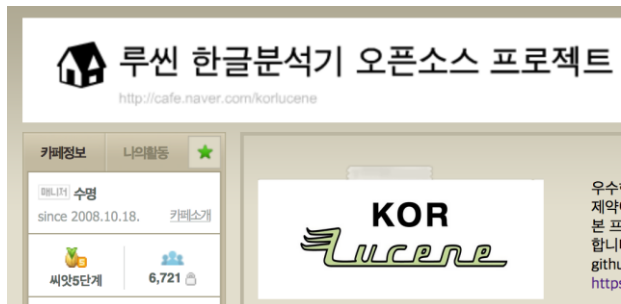
한글은 복합어로 이루어진 언어입니다.

단순한 whitespace 기반이 아닌 단어 사전 기반의 분석이 필요합니다.

- 동해물과 백두산이
 - 동해 + 물 / 백두 + 산
 - 동 + 해물 / 백 + 두산
- 대학생선교회
 - 대학생 + 선교 + 회
 - 대학 + 생선 + 교회

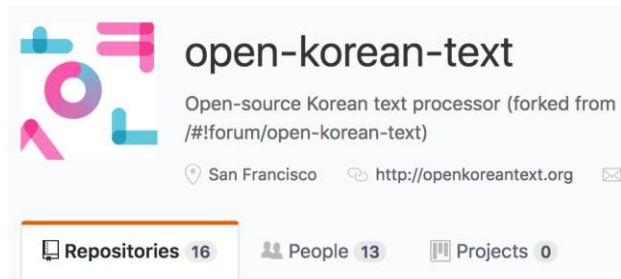


Elasticsearch 에서 사용 가능한 한글 형태소 분석기는
아래와 같은 프로젝트들이 있습니다.



은전한닢 프로젝트

검색에서 쓸만한 오픈소스 한국어 형태소 분석기를 만들자!



2018. 08 - Nori 출시

SOSCON 2018

<https://www.elastic.co/kr/blog/nori-the-official-elasticsearch-plugin-for-korean-language-analysis>



제품 Cloud 서비스 고객 [알아보기](#)

다운로드

[문의](#)



KR

블로그

[뉴스](#)

[엔지니어링](#)

[사용자 스토리](#)

[출시](#)

[아카이브](#)



2018년 8월 24일

[엔지니어링](#)

KR

EN

공식 한국어 분석 플러그인 “노리”

By Jim Ferenczi • 김기주

Share

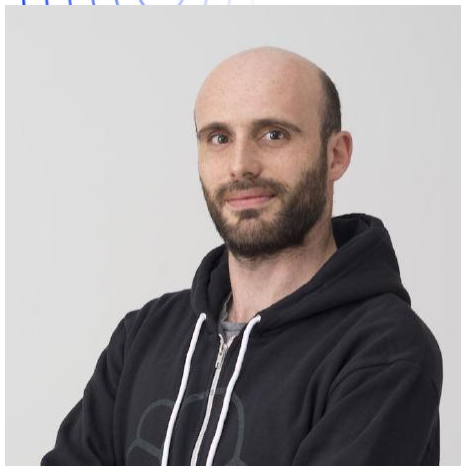


What is Nori?

SOSCON 2018

Created by **Jim Ferenczi**

- <https://github.com/jimczi>
- Creator of Kuromoji – Japanese Text Analyzer
- Contributor of MeCab
- 은전한닢의 mecab-ko-dic 사전을 사용합니다.
- Nori는 Kuromoji 엔진의 코드를 fork 해서 mecab-ko-dic 사전을 적용시킨 Elasticsearch 독자적인 형태소 분석기 입니다.



Why Nori?

SOSCON 2018

<https://issues.apache.org/jira/browse/LUCENE-8231>



[Dashboards](#)[Projects](#)[Issues](#)



Lucene - Core / LUCENE-8231



Nori, a Korean analyzer based on mecab-ko-dic



Details

Type:	 New Feature	Status:	RESOLVED
Priority:	 Major	Resolution:	Fixed
Affects Version/s:	None	Fix Version/s:	7.4, master (8.0)
Component/s:	None		
Labels:	None		
Lucene Fields:	New		

Why Nori?

SOSCON 2018

<https://issues.apache.org/jira/browse/LUCENE-8231>

 Dashboards Projects Issues



that copies the Kuromoji module and adapts it for the mecab-ko-dic. I used the same classes to build and read the dictionary but I had to make some modifications to handle the differences with the IPADIC and Japanese. The resulting binary dictionary takes 28MB on disk, it's bigger than the IPADIC but mainly because the source is bigger and there are a lot of compound and inflect terms that define a group of terms and the segmentation that can be applied. I attached the patch that contains this new Korean module called **goderi nori**. It is an adaptation of the Kuromoji module so currently the two modules don't share any code. I wanted to validate the approach first and check the relevancy of the results. I don't speak Korean so I used the relevancy tests that was added for another Korean tokenizer (<https://issues.apache.org/jira/browse/LUCENE-4956>) and tested the output against mecab-ko which is the official fork of mecab to use the mecab-ko-dic. I had to simplify the JapaneseTokenizer, my version removes the nBest output and the decomposition of too long tokens. I also





How Nori?

SOSCON 2018

<https://github.com/apache/lucene-solr/tree/master/lucene/analysis>

apache / lucene-solr

Watch 232 Star 1,729 Fork 1,411

Code Pull requests 218 Projects 0 Insights

Branch: master lucene-solr / lucene / analysis / Create new file Find file History

romseygeek

LUCENE-8237: Correct handling of position increments in sub-tokenstreams

Latest commit ab2fec1 3 days ago

..		
common	LUCENE-8237: Correct handling of position increments in sub-tokenstreams	3 days ago
icu	LUCENE-8222: Await-fix TestICUTokenizerCJK.	2 months ago
kuromoji	SOLR-12028: BadApple and AwaitsFix annotations usage	4 months ago
morfologik	SOLR-11490: Add missing @since tags	8 months ago
nori	LUCENE-8231: Add missing part of speech filter in the SPI META-INF file	2 months ago
opennlp	LUCENE-8188: Fixed bugs in OpenNLPOpsFactory that were causing InputS...	4 months ago
phonetic	SOLR-11490: Undo since tags for pre-3.1 classes	8 months ago
smarten	LUCENE-8325: Fixed the smartcn tokenizer to not split UTF-16 surrogat...	29 days ago

How Nori?

SOSCON 2018

mecab-ko-dic 사전 사이즈 : 약 200MB

```
[~/elastic/demo/nori/plugins/seunjeon/mecab-ko-dic-2.0.3-20170922]$ ls
AUTHORS          Hanja.csv        NNB.csv          Place.csv        XR.csv           install-sh
COPYING          IC.csv           NNBC.csv         Preanalysis.csv  XSA.csv          left-id.def
ChangeLog        INSTALL          NNG.csv          README           XSN.csv          matrix.def
CoinedWord.csv   Inflect.csv      NNP.csv          Symbol.csv       XSV.csv          missing
EC.csv           J.csv            NP.csv           VA.csv           autogen.sh       model.def
EF.csv           MAG.csv          NR.csv           VCN.csv          char.def         pos-id.def
EP.csv           MAJ.csv          NorthKorea.csv   VCP.csv          clean            rewrite.def
ETM.csv          MM.csv           Person-actor.csv VV.csv           configure        right-id.def
ETN.csv          Makefile.am      Person.csv       VX.csv           configure.ac     tools
Foreign.csv      Makefile.in      Place-address.csv Wikipedia.csv     dicrc            unk.def
Group.csv        NEWS             Place-station.csv XPN.csv          feature.def      user-dic
[~/elastic/demo/nori/plugins/seunjeon/mecab-ko-dic-2.0.3-20170922]$ du -h
12K    ./tools
16K    ./user-dic
217M   .
[~/elastic/demo/nori/plugins/seunjeon/mecab-ko-dic-2.0.3-20170922]$
```

How Nori?

SOSCON 2018

nori 사전 사이즈 : 24MB

```
[~/git/lucene-solr/lucene/analysis/nori/src/resources/org/apache/lucene/analysis/ko/dict]$ ls [master]
CharacterDefinition.dat      TokenInfoDictionary$fst.dat      UnknownDictionary$buffer.dat
ConnectionCosts.dat         TokenInfoDictionary$posDict.dat  UnknownDictionary$posDict.dat
TokenInfoDictionary$buffer.dat TokenInfoDictionary$targetMap.dat UnknownDictionary$targetMap.dat
[~/git/lucene-solr/lucene/analysis/nori/src/resources/org/apache/lucene/analysis/ko/dict]$ du -h [master]
24M      .
[~/git/lucene-solr/lucene/analysis/nori/src/resources/org/apache/lucene/analysis/ko/dict]$ █ [master]
```

mecab-ko-dic 사전 구조

NNG.csv

```
대학,1781,3535,1828,NNG,*,T,대학,*,*,*,*  
대학가,1781,3534,2871,NNG,*,F,대학가,Compound,*,*,대학/NNG/*+가/NNG/*  
대학교,1781,3534,2193,NNG,*,F,대학교,Compound,*,*,대/NNG/*+학교/NNG/*  
대학교수,1781,3534,3567,NNG,*,F,대학교수,Compound,*,*,대학/NNG/*+교수/NNG/*  
대학당,1781,3535,2835,NNG,*,T,대학당,Compound,*,*,대/NNG/*+학당/NNG/*  
대학모,1781,3534,2835,NNG,*,F,대학모,Compound,*,*,대학/NNG/*+모/NNG/*  
대학물,1781,3535,2835,NNG,*,T,대학물,Compound,*,*,대학/NNG/*+물/NNG/*  
대학살,1781,3535,3657,NNG,*,T,대학살,Compound,*,*,대/NNG/*+학살/NNG/*  
생,1781,3535,2674,NNG,*,T,대학생,Compound,*,*,대학/NNG/*+생/NNG/*  
대학원,1781,3535,1723,NNG,*,T,대학원,Compound,*,*,대학/NNG/*+원/NNG/*  
대학원생,1781,3535,2796,NNG,*,T,대학원생,Compound,*,*,대학/NNG/*+원/NNG/*+생/NNG/*  
대학인,1781,3535,2132,NNG,*,T,대학인,Compound,*,*,대학/NNG/*+인/NNG/*  
대학자,1781,3534,2835,NNG,*,F,대학자,Compound,*,*,대/NNG/*+학자/NNG/*
```

How Nori?

출처 : 꼬꼬마 한국어 형태소 분석기
<http://kkma.snu.ac.kr/documents/?doc=postag>

SOSCON 2018

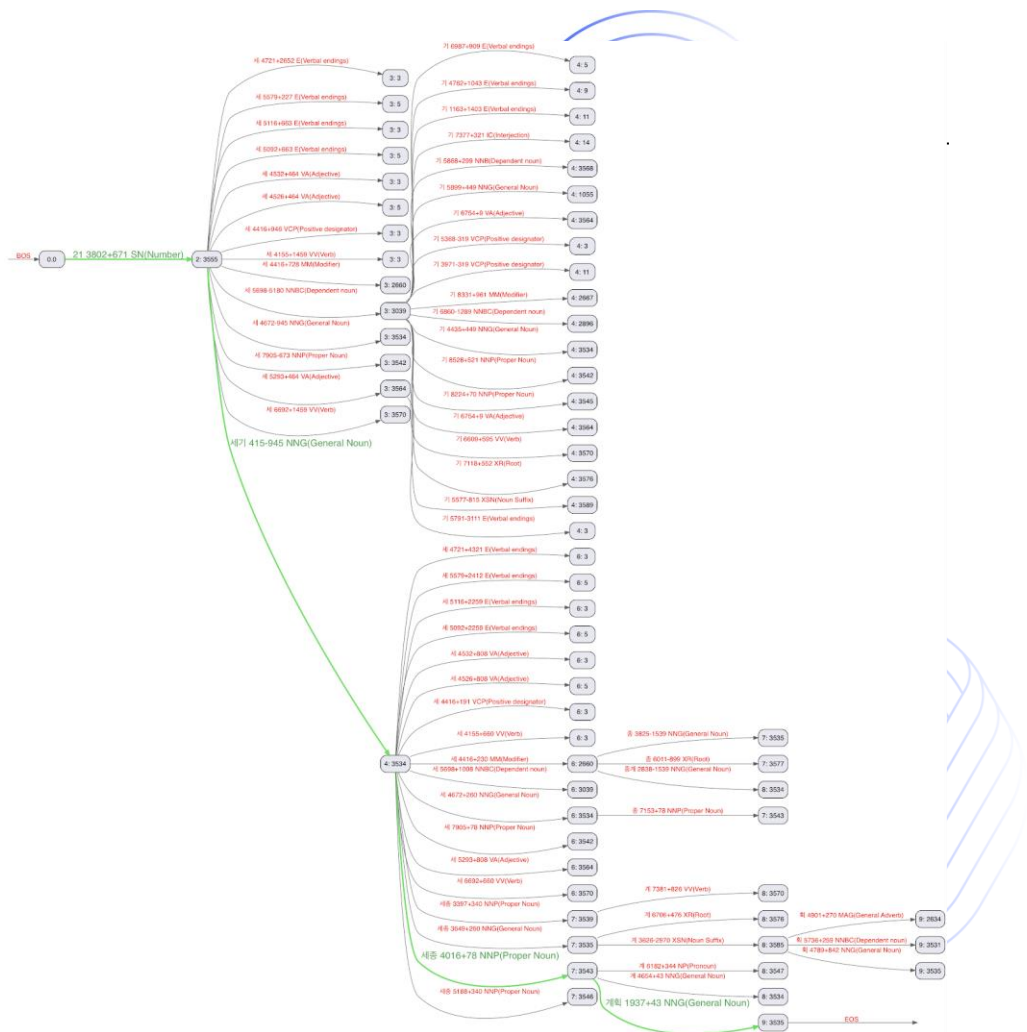
Korean POS (Part Of Speech) Table

대분류	세종 품사 태그		심광섭 품사 태그		KKMA 단일 태그 V 1.0						
	태그	설명	태그	설명	목록1	목록2	태그	설명	확률태그	저장사전	
체언	NNG	일반 명사	NN	명사	N	NN	NNG	보통 명사	NNA	noun.dic	
	NNP	고유 명사					NNP	고유 명사			
	NNB	의존 명사	NX	의존 명사			NNB	일반 의존 명사	NNB	simple.dic	
			UM	단위 명사				NNM			단위 의존 명사
	NR	수사	NU	수사		NR	NR	수사	NR		
	NP	대명사	NP	대명사		NP	NP	대명사	NP		
용언	VV	동사	VV	동사	V	VV	VV	동사	VV	verb.dic	
	VA	형용사	AJ	형용사		VA	VA	형용사	VA		
	VX	보조 용언	VX	보조 동사		VX	VXV	보조 동사	VX		
			AX	보조 형용사			VXA	보조 형용사			
	VCP	긍정 지정사	CP	서술격 조사 '이다'		VC	VCP	긍정 지정사, 서술격 조사 '이다'	VCP	raw.dic	
	VCN	부정 지정사					VCN	부정 지정사, 형용사 '아니다'	VCN		
관형사	MM	관형사	DT	일반 관형사	M	MD	MDT	일반 관형사	MD	simple.dic	
			DN	수 관형사			MDN	수 관형사			
부사	MAG	일반 부사	AD	부사		MA	MAG	일반 부사	MAG		
	MAJ	접속 부사					MAC	접속 부사	MAC		
감탄사	IC	감탄사	EX	감탄사		I	IC	IC	감탄사		IC

How Nori?

nori 사전 구조 :

- CSV(array) 형식의 텍스트로 저장된 사전을 FST (finite state transducer)로 부호화 하고 Binary 구조로 재구성.
- Matrix.def 파일 가변 길이로 변경 후 직렬화



nori 사전 구조 :

- 명사와 합성어의 부분들을 부호화 → 사전에서 중복 제거
- 오른쪽 어절이 왼쪽 어절과 다른 경우에만 저장하고 다르면 저장하지 않음. (중복 제거)

```
7.2M  6 21 11:38 analysis-nori-6.4.0-SNAPSHOT.zip
33M   6 21 11:38 elasticsearch-analysis-seunjeon-6.1.1.1.zip
```

How Nori?

SOSCON 2018

자세한 사용법은 Elastic 공식 도큐먼트에 있습니다.

<https://www.elastic.co/guide/en/elasticsearch/plugins/6.4/analysis-nori.html>

Elasticsearch Plugins and Integrations [6.4] » Analysis Plugins » Korean (nori) Analysis Plugin

« [kuromoji_number](#) token filter

[nori analyzer](#) »

Korean (nori) Analysis Plugin



The Korean (nori) Analysis plugin integrates Lucene nori analysis module into elasticsearch. It uses the [mecab-ko-dic dictionary](#) to perform morphological analysis of Korean texts.

Installation

This plugin can be installed using the plugin manager:



```
sudo bin/elasticsearch-plugin install analysis-nori
```


감사합니다

Elastic 부스에서 더 많은 데모들을 시연중이니
많이 들려주세요.

<https://www.facebook.com/groups/elasticsearch.kr/>

SOSCON 2018

SAMSUNG OPEN SOURCE CONFERENCE 2018

