

# 프로젝트 결과보고서

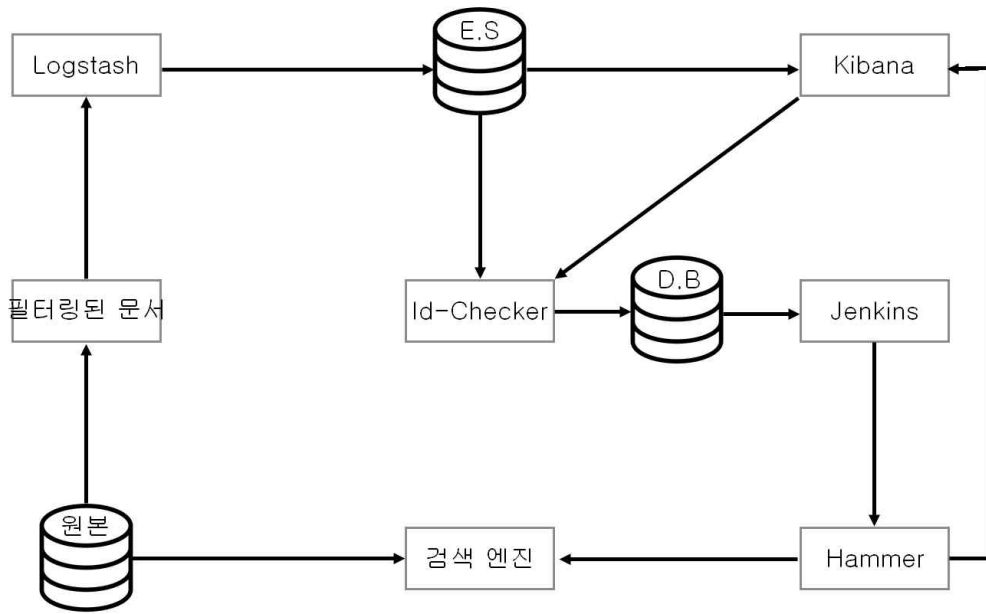
프로젝트 주제명	ID-Checker		
핵심 키워드	Scala , Play Framework , Elastic Search , Logstash , Kibana , Jenkins , Mysql , HTML5 , CSS , JavaScript		
주요 개발 내용 요약	<p>카카오 검색스팸파트는 다음에서 검색 시 스팸 문서들이 검색 결과에 노출되지 않도록 검색품질을 담당하는 곳입니다.</p> <p>이를 위해 먼저 기계학습 등을 이용하여 스팸필터링을 실시하게 됩니다.</p> <p>하지만 스팸문서인지 정상문서인지 분별이 불가능한 문서들이 존재하게 되며, 이러한 문서들을 사람이 빠르게 처리할 수 있는 운영 툴을 개발하였습니다.</p> <p>아래는 기여한 부분입니다.</p> <ol style="list-style-type: none"> <li>1. Scala, Play Framework, HTML5, CSS, JavaScript로 웹 개발.</li> <li>2. ElasticSearch 활용, 스팸 문서 검색 및 Aggregation 기능 구현</li> <li>3. 데이터는 약 5억건으로 토큰 필터링은 Lowercase를 기준으로 진행.</li> <li>4. Term 색인은 스팸 의심 id, domain, type을 기준으로 진행.</li> <li>5. Kibana로 삭제한 7일간의 문서에 대한 정보를 날짜 별로 시각화.</li> </ol> <p>스팸 의심 문서 추천 기준 - ID값의 변형, 기계학습 단어별 스팸지수 파악          예 1) 일반적인 Ascii 코드 범위가 아닌 경우 (010이 아닌 전각문자 010)          예 2) Naive Bayesian 카카오 사내 Library 활용, NB_Score 기준 추천</p>		
소속	카카오	개발자 명단	Tyler(유영호)
부서	검색스팸 파트		Min(김민규)
기간	2016.1.18~2016.2.23		

## I. 개발 내용

<p>* Github - <a href="https://github.com/pkgonan/id-checker-web">https://github.com/pkgonan/id-checker-web</a></p> <p>* 개발 환경 - Web : Scala , Play Framework , HTML5 , CSS , JavaScript - Etc : ElasticSearch, Logstash, Kibana, Mysql, Jenkins, Git, Trello, Jira, Wiki</p> <p>* 데이터 - ElasticSearch : 약 5억 건의 스팸 의심 문서</p>
---

\* Data Flow Diagram

Data Flow Diagram



II. 구현 결과

\* 스팸 의심 ID 추천 페이지(recommend.html)

ID Checker
Recommend Bucket Status

### Recommend

100 개의 스팸 id가 추천되었습니다.

id	타입	범위 / 전체	변환비율
010-8274-4949	tel	1,778 / 1,778	(100 %)
010-8815-0454	tel	1,267 / 1,267	(100 %)
010-3606-2777	tel	1,254 / 1,254	(100 %)
010-3108-8003	tel	1,191 / 1,191	(100 %)
010-8442-4191	tel	1,162 / 1,162	(100 %)
010-3516-4429	tel	1,087 / 1,087	(100 %)
011-548-8333	tel	1,070 / 1,070	(100 %)
017-274-4949	tel	934 / 934	(100 %)
010-3211-3211	tel	933 / 933	(100 %)
010-8340-4478	tel	917 / 917	(100 %)
010-8975-8333	tel	804 / 804	(100 %)
0102115886	tel	739 / 739	(100 %)
010-8908-8343	tel	722 / 722	(100 %)
010-3577-8299	tel	695 / 695	(100 %)
010-5270-0735	tel	669 / 669	(100 %)
01031187238	tel	657 / 657	(100 %)
010-8480-5770	tel	648 / 648	(100 %)
010-9922-8874	tel	598 / 598	(100 %)
010-8321-8905	tel	581 / 581	(100 %)
010-9488-3258	tel	577 / 1,431	(40 %)
010-4553-0088	tel	556 / 556	(100 %)
010-3518-4791	tel	547 / 547	(100 %)
010-5492-3341	tel	543 / 543	(100 %)
Hitnet@daum.net	email	256 / 256	(100 %)

\* 특정 ID별 Aggregation 페이지(aggregation.html)

ID Checker Recommend Bucket Status

### MetricsSpec@gmail.com

- spam type: **Unwanted**
- 악성 사이트
- 유형 사이트

\* SPAM 처리할 경우 Total count(84) 만큼의 유서가 삭제 됩니다.

번함/전체: **484 / 484 (100 %)**

id	domain	count	
1	MetricsSpec@gmail.com	funny-p.lk	99
2	MetricsSpec@gmail.com	topicconnect.com	31
3	MetricsSpec@gmail.com	ps2youtube.com	29
4	MetricsSpec@gmail.com	powvid.com	22
5	MetricsSpec@gmail.com	topicimes.com	22
6	MetricsSpec@gmail.com	valsvideo.com	16
7	MetricsSpec@gmail.com	nsjooml.com	12
8	MetricsSpec@gmail.com	pingfile.net	12
9	MetricsSpec@gmail.com	videobychoice.com	12
10	MetricsSpec@gmail.com	doovi.com	11
11	MetricsSpec@gmail.com	kingshow.com	10
12	MetricsSpec@gmail.com	pandas.com	10
13	MetricsSpec@gmail.com	bps1025.com	9
14	MetricsSpec@gmail.com	vietglatr.com	9
15	MetricsSpec@gmail.com	azerimix.com	8
16	MetricsSpec@gmail.com	marocsite.net	8
17	MetricsSpec@gmail.com	plim19.com	8
18	MetricsSpec@gmail.com	youmeda.com	8
19	MetricsSpec@gmail.com	mashpedia.com	7
20	MetricsSpec@gmail.com	playlistfan.com	7
21	MetricsSpec@gmail.com	tyoobe.com	7
22	MetricsSpec@gmail.com	aktubes.com	6

\* 특정 ID이면서 특정 Domain인 페이지(detail.html)

ID Checker Recommend Bucket Status

### MetricsSpec@gmail.com

- spam type: **Unwanted**
- 악성 사이트
- 유형 사이트

번함/전체: **99 / 99 (100 %)**

id	transform	type	domain	url	
1	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/powerucc8241186
2	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8275758
3	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8421717
4	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8251674
5	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8414568
6	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8332146
7	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/powerucc8376300
8	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8385930
9	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8350392
10	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8308999
11	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8421729
12	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8351494
13	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8426592
14	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8260328
15	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8402065
16	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8315471
17	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8250678
18	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8314444
19	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8285944
20	MetricsSpec@gmail.com	true	email	funny-p.lk	http://funny-p.lk/8354278

◀ 1 2 3 4 5 ▶

\* 사용자가 특정 ID값으로 Prefix Search한 페이지(search.html)

ID Checker Recommend Bucket Status

### Search

전체: 86925

	id	문서
1	sexblog.pw	6386
2	sex57.com	3641
3	sex.com	2614
4	sexcamsonline.mobi	2263
5	sexoffender.go.kr	2108
6	sexmuseum1@naver.com	1658
7	sexperma.org	1620
8	sex.co.kr	1325
9	sex-1.com	1296
10	sexmuseum.or.kr	1126
11	sexporno.co.kr	978
12	sexlive.tv	922
13	sexshocking.net	801
14	sexualhab.go.kr	787
15	sex100.info	778
16	sexjon.com	774
17	sexnetq.com	745
18	sexacademy.org	733
19	sexking.biz	721
20	sexhunting.net	702
21	sexqueen.biz	690
22	sexguy.co.kr	566
23	sexyour.co.kr	566
24	sex.biz	526

\* 사용자가 당일 처리한 작업들을 볼 수 있는 페이지(bucket.html)

ID Checker Recommend Bucket Status

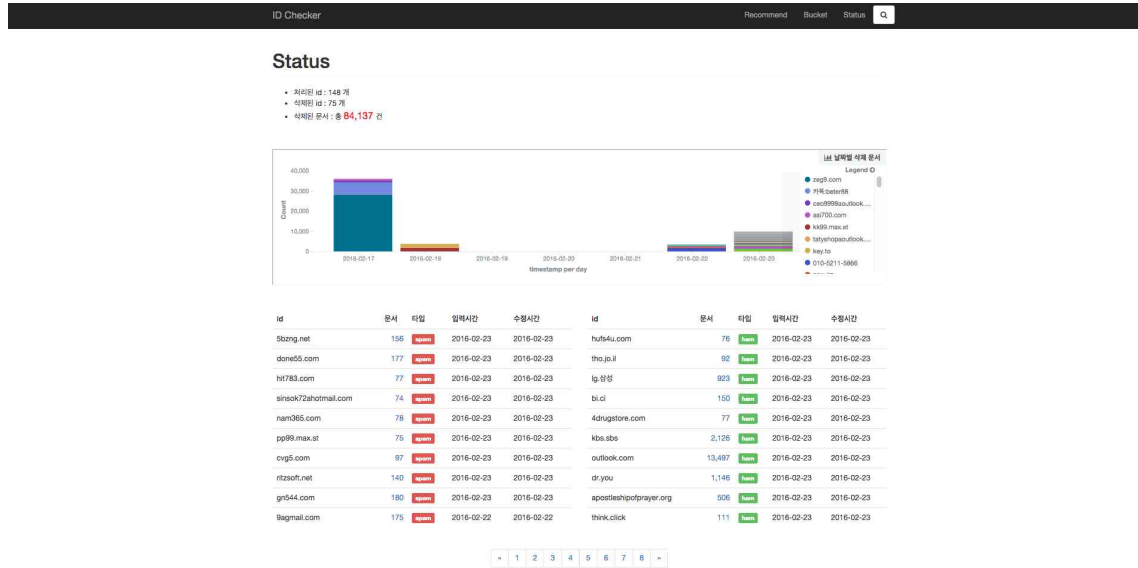
### Bucket

- 처리된 id: 130 개
- 소용러된 id: 6 개
- 삭제된 문서: 총 3,346 건

[junks](#)

id	문서	타입	id	문서	타입	id	문서	타입
VIP_C222@hotmail.com	1,024	spam	Speed554@hate.com	1,660	hold	wordpress.com	2,440,366	spam
VIP_C444@hotmail.com	1,024	spam	QW@gmail.com	1,629	hold	010-6274-4949	1,778	spam
VIP_C888@hotmail.com	773	spam	010-5658-0005	1,417	hold	Kcalano3@gmail.com	1,619	spam
VIP_C333@hotmail.com	520	spam	soman123.com	1,396	hold	YehaGuesthouse@hotmail.com	1,600	spam
running713@gmail.com	3	spam	Jinjn0225@hotmail.com	1,169	hold	010-9458-3268	1,431	spam
running7139@gmail.com	2	spam	unitel.co.kr	1,109	hold	Lee@gmail.com	1,359	spam
			rns23.com	1,056	hold	010-8815-0454	1,267	spam
			ja8.to	1,011	hold	010-3606-2777	1,254	spam
			han91.com	968	hold	010-3108-8003	1,191	spam
			p8.co.kr	926	hold	Bickertbocle@hotmail.com	1,186	spam

- \* 사용자가 지금까지 처리한 모든 작업들을 볼 수 있는 페이지(status.html)
  - Kibana로 데이터 시각화



#### IV. 기대효과

첫째, Spam 문서인지 Ham 문서인지 판단이 불가능한 문서들을 빠르게 처리할 수 있다. 이를 통해 다음 검색 서비스의 품질 향상을 기대한다.

둘째, 사람에 의해 Spam으로 판정이 된 문서들은 Mysql에 저장되어 보관하게 된다. 해당 데이터는 향후 블로그, 카페 등 다음 및 카카오의 여러 가지 서비스에서 나타나는 스팸을 처리하는데 이용될 예정이다. 이를 통해 다른 서비스의 품질 향상도 기대할 수 있을 것이다.