프로젝트 결과보고서

프로젝트 주제명	ID-Checker		
핵심 키워드	Scala , Play Framework , Elastic Search , Logstash , Kibana , Jenkins , Mysql , HTML5 , CSS , JavaScript		
주요 개발 내용 요약	카카오 검색스팸파트는 다음에서 검색 시 스팸 문서들이 검색 결과에 노출되지 않도록 검색품질을 담당하는 곳입니다. 이를 위해 먼저 기계학습 등을 이용하여 스팸필터링을 실시하게 됩니다. 하지만 스팸문서인지 정상문서인지 분별이 불가한 문서들이 존재하게 되며, 이러한 문서들을 사람이 빠르게 처리할 수 있는 운영 툴을 개발하였습니다. 아래는 기여한 부분입니다. 1. Scala, Play Framework, HTML5, CSS, JavaScript로 웹 개발. 2. ElasticSearch 활용, 스팸 문서 검색 및 Aggregation 기능 구현 3. 데이터는 약 5억건으로 토큰 필터링은 Lowercase를 기준으로 진행. 4. Term 색인은 스팸 의심 id, domain, type을 기준으로 진행. 5. Kibana로 삭제한 7일간의 문서에 대한 정보를 날짜 별로 시각화. 스팸 의심 문서 추천 기준 - ID값의 변형, 기계학습 단어별 스팸지수 파악 예 1) 일반적인 Ascii 코드 범위가 아닌 경우 (010이 아닌 전각문자 0 10) 예 2) Naive Bayesian 카카오 사내 Library 활용, NB_Score 기준 추천		
소속	카카오	개발자 명단	Tyler(유영호)
부서	검색스팸파트		Min(김민규)
기간	2016.1.18~2016.2.23		

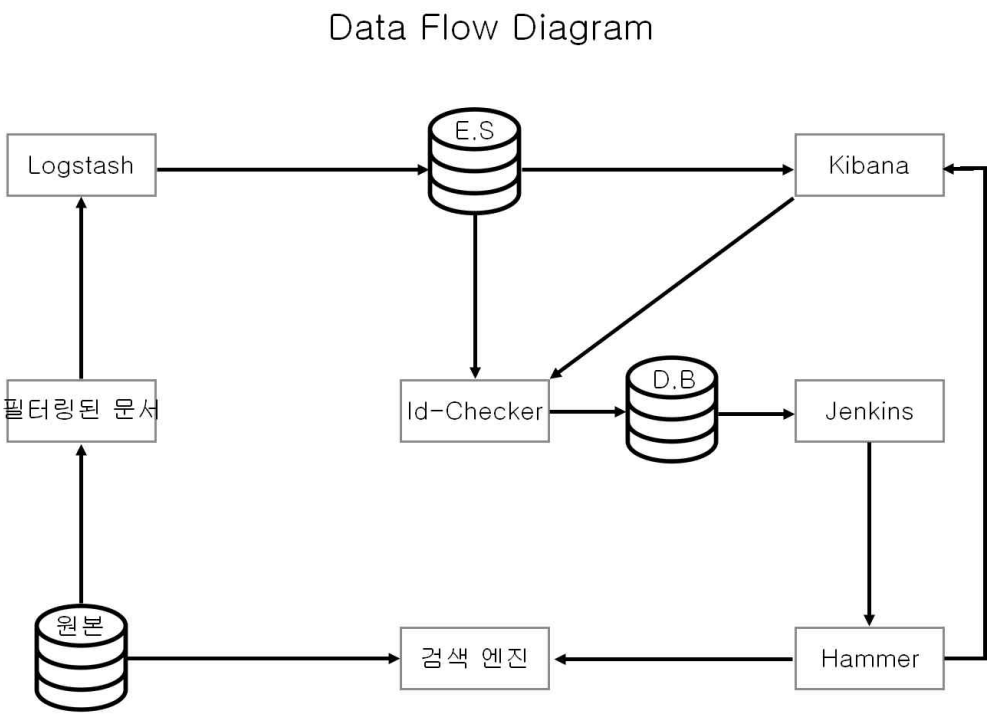
Ⅰ. 개발 내용

* Github
- https://github.com/pkgonan/id-checker-web

* 개발 환경
- Web : Scala , Play Framework , HTML5 , CSS , JavaScript
- Etc : ElasticSearch, Logstash, Kibana, Mysql, Jenkins, Git, Trello, Jira, Wiki

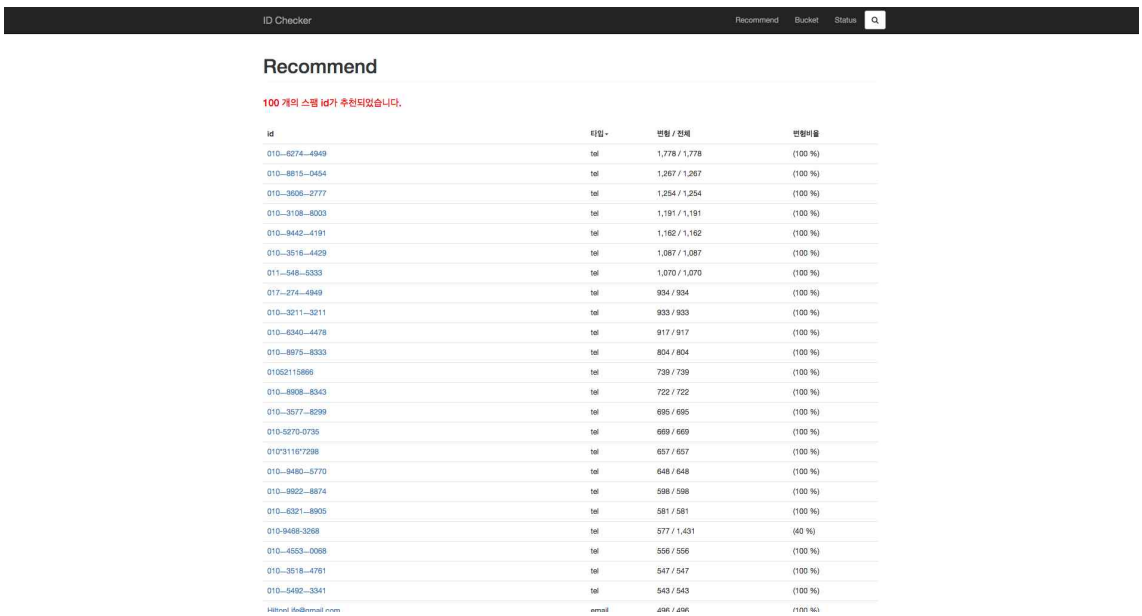
* 데이터
- ElasticSearch : 약 5억 건의 스팸 의심 문서

* Data Flow Diagram

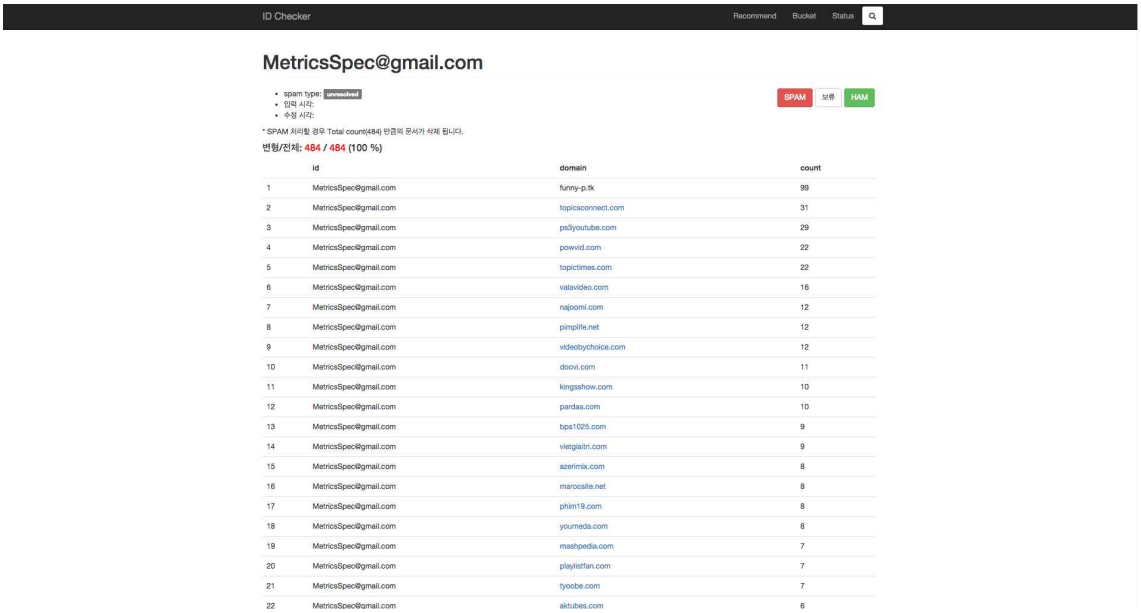


Ⅱ. 구현 결과

* 스팸 의심 ID 추천 페이지(recommend.html)

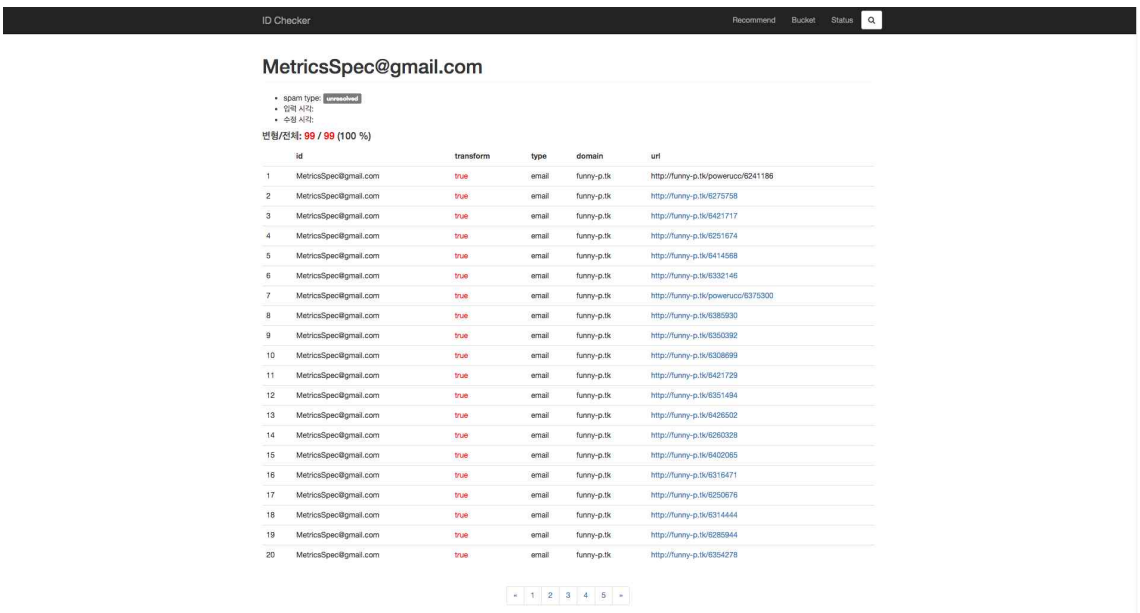


* 특정 ID별 Aggregation 페이지(aggregation.html)



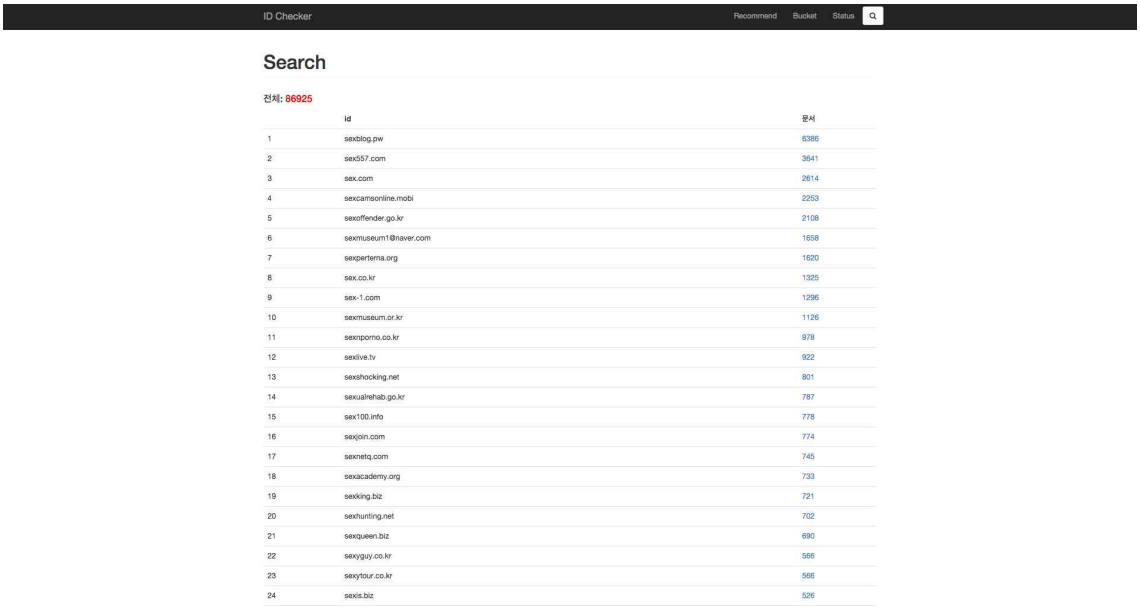
	id	domain	count
1	MetricsSpec@gmail.com	funny-p.tk	99
2	MetricsSpec@gmail.com	topicsconnect.com	31
3	MetricsSpec@gmail.com	ps3youtube.com	29
4	MetricsSpec@gmail.com	powvid.com	22
5	MetricsSpec@gmail.com	topictimes.com	22
6	MetricsSpec@gmail.com	valavideo.com	16
7	MetricsSpec@gmail.com	najoomi.com	12
8	MetricsSpec@gmail.com	pimplife.net	12
9	MetricsSpec@gmail.com	videobychoice.com	12
10	MetricsSpec@gmail.com	doovi.com	11
11	MetricsSpec@gmail.com	kingsshow.com	10
12	MetricsSpec@gmail.com	pardaa.com	10
13	MetricsSpec@gmail.com	bpa1025.com	9
14	MetricsSpec@gmail.com	vietgiaitri.com	9
15	MetricsSpec@gmail.com	azerimix.com	8
16	MetricsSpec@gmail.com	marocsite.net	8
17	MetricsSpec@gmail.com	phim19.com	8
18	MetricsSpec@gmail.com	yourmeda.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)

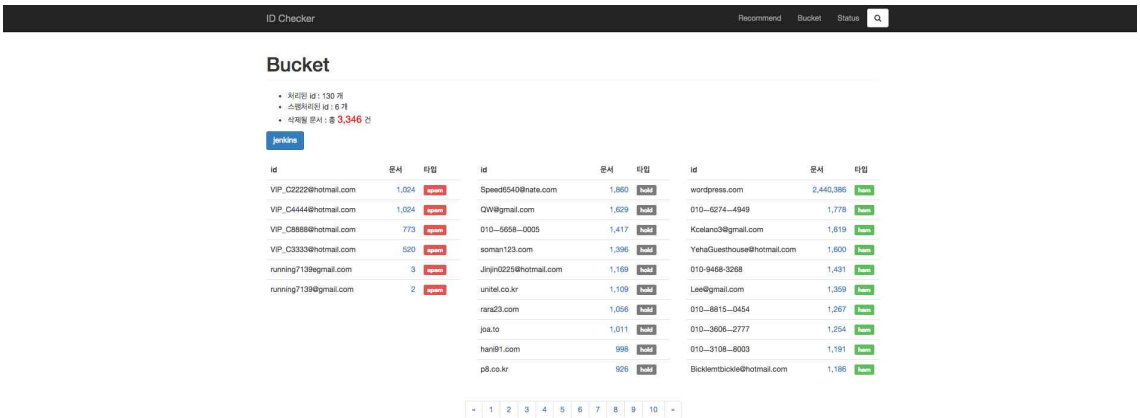


	id	transform	type	domain	url
1	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/powerucc/6241186
2	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6275758
3	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6421717
4	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6251674
5	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6414568
6	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6332146
7	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/powerucc/6375300
8	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6385930
9	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6350392
10	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6308699
11	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6421729
12	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6351494
13	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6426502
14	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6260328
15	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6402065
16	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6315471
17	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6250676
18	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6314444
19	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6285944
20	MetricsSpec@gmail.com	true	email	funny-p.tk	http://funny-p.tk/6354278

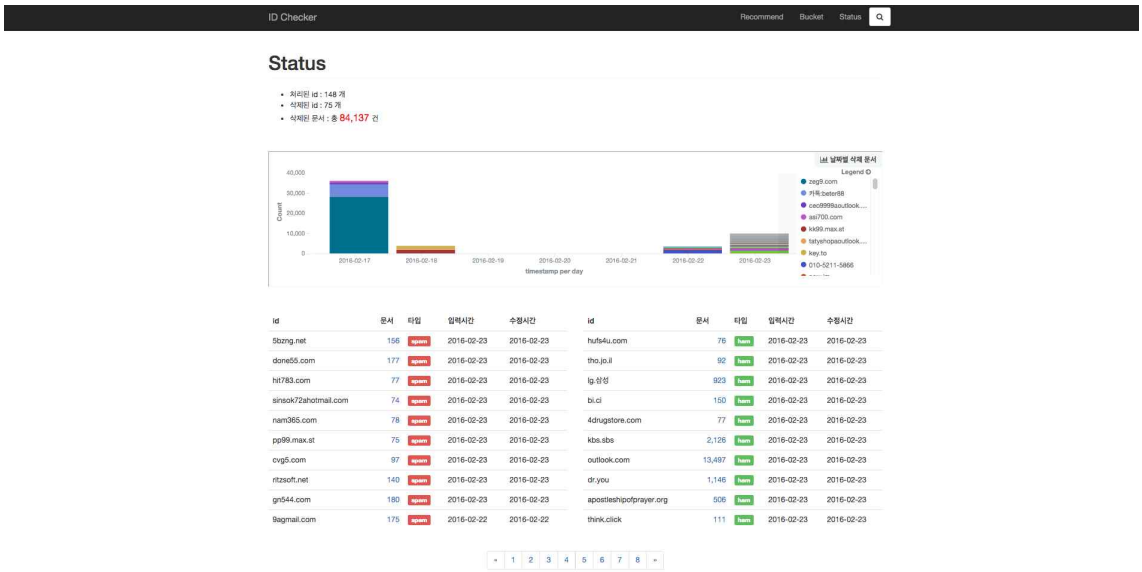
* 사용자가 특정 ID값으로 Prefix Search한 페이지(search.html)



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



* 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, 까페 등 다음 및 카카오의 여러 가지 서비스에서 나타나는 스팸을 처리하는데 이용될 예정이다. 이를 통해 다른 서비스의 품질 향상도 기대할 수 있을 것이다.