|  |  |
| --- | --- |
| 최종 작성일 | 2019.01.23 |
| 학번 | 20144367 |
| 이름 | 이동현 |

[Report]

**Tweet Sentiment Analysis**

Tweet 데이터를 활용한 감성 분석

목차

[1. 연구 목적 및 개요 3](#_Toc536631397)

[1.1 연구 목적 3](#_Toc536631398)

[1.2 개요 3](#_Toc536631399)

[2. 연구 선행 사항 3](#_Toc536631400)

[2.1 빅데이터란 3](#_Toc536631401)

[2.2 Hadoop Eco-System이란 3](#_Toc536631402)

[2.3 CDH Cluster 구축 3](#_Toc536631403)

[3. Hadoop Eco-System 3](#_Toc536631404)

[3.1 Hadoop 3](#_Toc536631405)

[3.2 Hue 3](#_Toc536631406)

[3.3 Hive 3](#_Toc536631407)

[3.4 Impala 3](#_Toc536631408)

[3.5 Sqoop 3](#_Toc536631409)

[3.6 Oozie 3](#_Toc536631410)

[3.7 Spark 3](#_Toc536631411)

[3.8 Others 3](#_Toc536631412)

[4. Data Pipelining 4](#_Toc536631413)

[4.1 StreamSets 4](#_Toc536631414)

[4.2 NiFi 4](#_Toc536631415)

[5. Elastic Search 4](#_Toc536631416)

[6. 빅데이터 분석 4](#_Toc536631417)

[6.1 개요 4](#_Toc536631418)

[6.2 세부 연구 내용 4](#_Toc536631419)

[6.3 분석 결과 16](#_Toc536631420)

[7. 결론 16](#_Toc536631421)

[참고 자료 16](#_Toc536631422)

# 연구 목적 및 개요

본 연구를 통해 빅데이터의 기본 개념에 대해 알아보고, 빅데이터 활용에 사용되는 Hadoop Eco-system을 심도 있게 이해할 수 있도록 한다. 빅데이터의 수집, 처리, 저장, 분석, 시각화 전 과정에 걸쳐 활용되는 Hadoop Eco-system에 대하여 해당 생태계를 이루는 다양한 서비스들에 대해 알아보고 실제 활용해볼 수 있도록 한다. Hadoop의 배포판인 CDH를 활용해 실제로 클러스터를 구축하여 구축한 클러스터를 활용해 tweet 데이터를 수집하여 sentiment analysis를 진행해 봄으로써, 실생활에서 발생하는 빅데이터 활용의 전 과정을 실제로 수행해볼 수 있도록 한다.

# 연구 선행 사항

## 2.1 빅데이터란

빅데이터는 정의하는 범위에 따라, 다음과 같이 3가지로 정의된다.

*<빅데이터의 정의>*

|  |  |
| --- | --- |
| 범위에 따른 정의 | 정의 |
| 좁은 범위의 정의 | 일반적인 데이터베이스 소프트웨어로 저장/관리/분석할 수 있는 범위를 초과하는 규모의 데이터. (McKinsey, 2011) |
| 중간 범위의 정의 | 다양한 종류의 대규모 데이터로부터, 저렴한 비용으로 가치를 추출하고 데이터의 초고속 수집/발굴/분석을 지원하도록 고안된 차세대 기술 및 아키텍처. (IDC, 2011) |
| 넓은 범위의 정의 | 대용량의 데이터를 활용해 적은 용량에서는 얻을 수 없었던 새로운 통찰이나 가치를 추출해내는 일. 나아가 이를 활용해 시장, 기업 및 시민과 정부의 관계 등 많은 분야에서 변화를 가져오는 일을 의미. (Mayer-Schonberger & Cukier, 2013) |

위와 같이 정의되는 빅데이터는 크게 기업의 고객 데이터 축적, 학계의 거대 데이터 활용 과학 확산, 관련 기술의 발전을 이유로 출현하게 되었다. 우선, 기업 규모가 확장됨에 따라 기업에서 보유한 고객 데이터의 양이 폭증하게 되었고, 기업은 보유한 데이터에서 숨어있는 가치를 발굴하여 새로운 성장 동력원으로 활용하려는 움직임을 보이고 있다. 미국 상장 기업을 중심으로 대부분의 기업이 100TB 이상의 대용량의 데이터를 보유하고 있고, 이 중 상당수의 기업이 1PB가 넘는 대량의 데이터를 보유하고 있다.

학계에서는 거대 데이터를 활용한 과학이 지속적으로 확산되고 있는데, 인간 게놈 프로젝트, 대형 강입자충돌기, NASA의 기후 시뮬레이션 등을 예로 들 수 있다. 게놈 프로젝트를 통해 비약적인 시간적 / 비용적 절약을 이끌어 냈으며, 1억 5천만 개의 센서에서 4천만 번의 데이터를 수집하여 발생한 데이터를 활용하거나 32PB에 이르는 기후 관찰 정보를 활용해 시뮬레이션을 진행하는 등 대용량의 데이터를 활용한 과학이 지속적으로 발전하고 있다.

또한 대용량 데이터에 관련된 기술이 지속적으로 발전함에 따라 빅데이터가 출현하게 되었다. 아날로그의 디지털화, 저장 기술의 발전과 가격 하락, 인터넷의 발전 및 모바일 시대의 진전, 클라우드 컴퓨팅의 보편화 등 대용량의 데이터를 다루기 위한 관련 기술이 지속적으로 발전하였다.

이에 따라 빅데이터는 산업, 연구 전반에 걸쳐 확산되었으며 차세대 산업혁명에서 ‘원유’와 같은 역할을 수행함으로써, 산업 전반의 생산성을 한 단계 향상시키고, 기존에 없던 새로운 범주의 산업을 만들어낼 것으로 기대되고 있다.

# Sentiment Analysis

## 3.1 개요

앞서 연구한 내용을 바탕으로 트위터에서 실시간 데이터를 수집하여, 적절한 ETL 과정을 거쳐 데이터 분석을 진행하였다. Twitter에서 실시간으로 생성되는 tweets을 수집하여, tweet 내용을 바탕으로 감성 분석을 진행하고, 분석한 결과를 Kibana와 Hue, Excel 3D Map 등을 활용하여 시각화 하였다. 전체적인 분석 주제는 [**문서 하단의 참고 자료 URL**](#_참고_자료)을 참고하였으며, 진행한 분석 절차를 간략하게 나타내면 아래와 같다.

1. Twitter Developer API 중 Filter API를 활용하여, 분석하고자 하는 데이터에 연관된 keyword를 설정해 관련 데이터를 수집한다.
2. JSON format으로 수집되는 데이터에 StreamSets을 활용해 ETL 과정을 거친다. 결과물은 HDFS와 Hive table에 저장하는데, HDFS 내에는 JSON format으로, Hive Table 생성은 Avro format으로 저장한다.
3. Hive table로 import한 데이터에 대하여, 테이블 내 tweet 내용을 바탕으로 감성 분석을 진행한다. 후에 데이터 시각화 시에 위치와 시간 정보가 활용되므로, 데이터 가공 과정을 거친다.
4. 감성 분석을 진행하여 생성한 테이블을 Hue, Kibana, Excel 3D Map 등을 활용해 시각화 한다.

## 3.2 세부 연구 내용

1. 우선적으로, 다양하게 제공되는 Twitter Developer API 중 무료로 사용이 가능하고, 실시간 데이터 수집에 이용 가능한 [**Filter API**](#_참고_자료)를 활용한다. 해당 API를 사용하기 위해서는, Twitter Developer Platform 사이트에서 Developer 등록을 진행하여야 한다. 본 연구에서는 크리스마스와 연관된 데이터를 수집하였으며, 데이터 수집에 활용된 keyword는 아래와 같다. API 구문에서 한글을 지원하지 않으므로, url 변환을 통해 한글 keyword를 설정하였다.

*<데이터 수집에 활용된 url>*

|  |
| --- |
| https://stream.twitter.com/1.1/statuses/filter.json?track=christmas,carol,%ed%81%ac%eb%a6%ac%ec%8a%a4%eb%a7%88%ec%8a%a4,%ec%ba%90%eb%a1%a4,%ec%ba%90%eb%9f%b4,christmas.,%ed%81%ac%eb%a6%ac%ec%8a%a4%eb%a7%88%ec%8a%a4.,christmascarol,carol. |

*<Keywords>*

|  |  |
| --- | --- |
| **Topic** | **Keywords** |
| Christmas | Christmas, Christmas., carol, carol., Christmas carol, carol., 크리스마스, 크리스마스., 캐럴, 캐롤, christmascarol |

Keyword와 tweet 내 단어가 동일하면 해당 tweet을 가져오지만, 단어의 일부가 keyword와 동일하다면 가져오지 않기 때문에, 온점(.) 유무에 따라 동일한 단어에 대하여 2개의 keyword를 설정하였다.

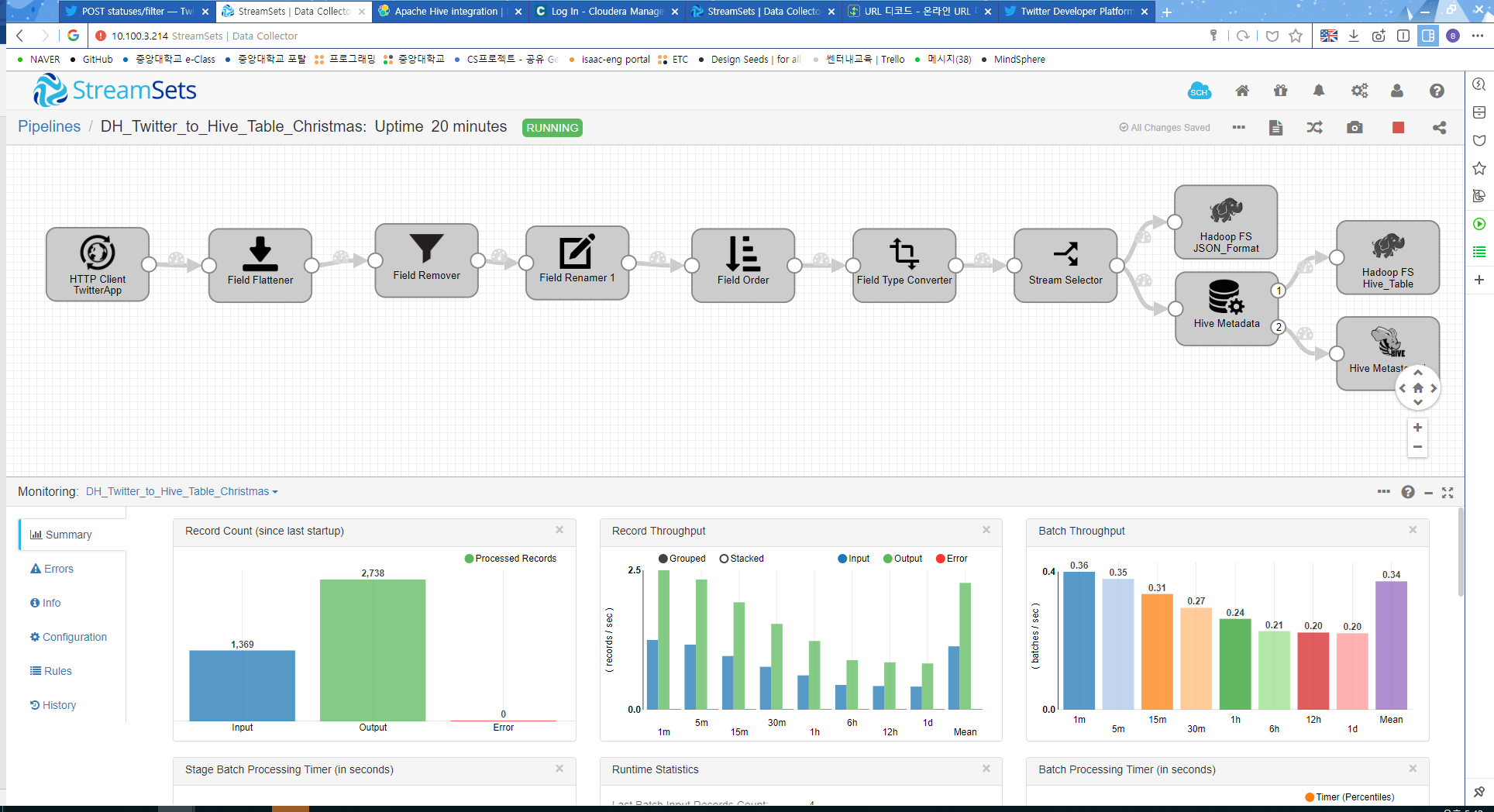
1. StreamSets Pipeline을 구축하여 설정한 keyword와 관련된 tweet을 수집하였다. Origin으로 HTTP Client stage를 선택하고, 해당 stage를 클릭하면 나오는 하단부 탭에서 Configuration > Resource URL에 위에서 언급한 URL을 삽입한다.

*<StreamSets HTTP Client Stage Configuration 화면>*



1. Filter API를 통해Tweet은 JSON format으로 수집되는데, 감성 분석에 필요한 정보를 제외한 부가적인 정보가 많으므로, nested 구조인 JSON 파일을 flatten시키고, 필요 없는 field는 제거하는 과정을 거쳐 HDFS와 Hive 내에 저장한다. 구축한 pipeline과 수집되는 데이터 현황은 아래와 같이 StreamSets Web UI를 통해 확인할 수 있다.

*<Tweet 수집을 위한 StreamSets pipeline>*



앞서 기술 했듯이, HTTP Client stage를 origin으로 두어 tweet을 json 형식으로 불러오면, 복잡한 구조의 json 파일을 가공하여 Stream Selector 이후 위쪽 stream을 통해 HDFS로, 아래쪽 stream을 통해 Hive table로 저장한다.

1. Hive Table 생성 시 Avro format으로 테이블을 생성하였기 때문에, Impala를 이용해서도 쿼리가 가능하다. 이후 데이터 분석을 위해서는 위치 정보가 필요한데, tweet을 작성한 사용자가 geo tag를 활성화 했을 때에만 위치 정보를 수집할 수 있으므로, 수집한 데이터 중 위치 정보가 없는 행을 제외한 테이블을 생성하였다. Impala에서 Avro format의 테이블에 대해서는 CTaS(Create Table as Select) 쿼리를 지원하지 않으므로, parquet 형식으로 external table을 생성하였다. 또한 tweet 내 실제 메시지는 140자를 넘는지 유무, retweet 유무에 따라 4군데에 나뉘어 저장되는데, 새로 생성한 table에 대해서는 msg field 한 곳에 저장될 수 있도록 하였다. 마지막으로, Impala에서 지원하는 형태의 timestamp format으로 변환하는 과정을 거쳤다. 테이블 생성에 사용한 쿼리 및 해당 테이블의 스키마는 아래와 같다.

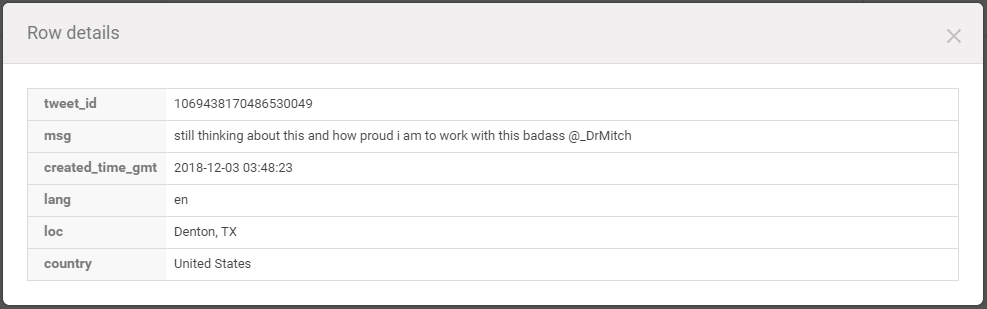
*<Impala에서 External table ‘Christmas\_tweets\_modifed’ 생성>*

|  |
| --- |
| create table christmas\_tweets\_modified stored as parquet as select  tweet\_id,  case  when rt\_msg like 'NULL' and extd\_full\_msg like 'NULL' then msg  when rt\_msg like 'NULL' and extd\_full\_msg not like 'NULL' then extd\_full\_msg  when rt\_msg not like 'NULL' and extd\_full\_msg like 'NULL' then rt\_msg  when rt\_msg not like 'NULL' and extd\_full\_msg not like 'NULL' then rt\_full\_msg end as msg,  cast (from\_unixtime( unix\_timestamp(concat( '2018 ', substring(created\_time,5,15)), 'yyyy MMM dd HH:mm:ss')) as timestamp) as created\_time\_GMT,  lang,  `location` as loc,  country  from christmas\_raw\_data  where `location` not like 'NULL' and country not like 'NULL'; |

*<Christmas\_tweets\_modified 스키마>*

|  |  |  |
| --- | --- | --- |
| Name | Type | Description |
| Tweet\_id | String | 각 tweet 별 고유 식별자 |
| Msg | String | 실제 tweet 내 내용을 담은 메시지 |
| Created\_time\_gmt | Timestamp | GMT 기준 tweet 작성 시간 |
| Lang | String | tweet이 작성된 언어(ex. us, fr, ko) |
| Loc | String | Tweet의 세부 위치정보 |
| Country | String | Tweet이 작성된 국가명 |

*<Example : Christmas\_tweets\_modified>*



1. 수집 결과 Christmas\_tweets\_modified 테이블에는 총 433,565 개의 tweet이 저장되었다. 해당 tweet에 대하여, dictionary table을 이용해 tweet 내 단어들을 분석하였다. Dictionary table에는 8천여개의 영어 단어에 대하여 해당 단어가 긍정, 부정, 중립 중 어떤 성격을 띄는지에 대한 정보를 담고 있다. 사용된 dictionary table의 schema 및 예제는 아래와 같다.

*<Table Schema : Dictionary>*

|  |  |  |
| --- | --- | --- |
| Name | Type | Description |
| Type | String | strongsubj / weaksubj |
| Length | Int | 단어 수 (1) |
| Word | String | 단어 |
| Pos | String | 동사, 명사 등 단어의 품사 |
| Stemmed | String | N / Y |
| polarity | String | Negative / Positive / Neutral |

*<Example : Dictionary table>*



1. 생성한 dictionary table을 이용해 수집한 tweet 내 단어의 긍정/부정 여부를 판단하여, 해당 tweet의 감성 분석을 시행한다. 감성 분석에 사용된 쿼리문 일부는 Impala에서 지원하지 않으므로, hive에서 진행하였다.

*<감성 분석에 사용될 view l1, l2, l3, table tweets\_sentiment를 생성하는 쿼리문>*

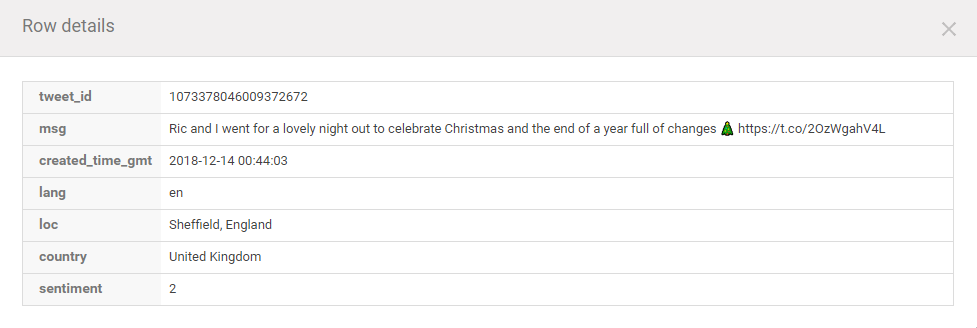
|  |
| --- |
| --create view l1  create view IF NOT EXISTS l1 as select  tweet\_id,  words  from Christmas\_tweets\_modified lateral view explode(sentences(lower(msg))) dummy as words;  --create view l2  create view IF NOT EXISTS l2 as select tweet\_id, word from l1 lateral view explode( words ) dummy as word;  --create view l3  create view IF NOT EXISTS l3 as select  tweet\_id,  l2.word,  case d.polarity  when 'negative' then -1  when 'positive' then 1  else 0 end as polarity  from l2 left outer join dictionary d on l2.word = d.word;  --create table tweets\_sentiment  create table IF NOT EXISTS tweets\_sentiment stored as parquet as select  tweet\_id,  case  when sum( polarity ) > 0 then 'positive'  when sum( polarity ) < 0 then 'negative'  else 'neutral' end as sentiment  from l3 group by tweet\_id; |

*<각 view/table 별 간단한 설명>*

|  |  |
| --- | --- |
| View / Table | Description |
| L1 | Tweet을 explode 하여 문장 내 포함된 단어 각각을 배열(words)에 담음. |
| L2 | 배열 내 단어들을 각각의 row로 explode함. |
| L3 | 행 별로 dictionary를 이용해 각 단어의 긍정/부정 여부를 판단함 |
| Tweets\_sentiment | 단어 별로 진행한 감성 분석 내용을 바탕으로 tweet의 긍정/부정 여부 판단 |

1. 각 tweet별 감성 분석이 완료되면, tweet\_id를 이용해 기존의 christmas\_tweets\_modified와 join하여 시각화에 사용될 tweetsbi 테이블을 생성한다. Christmas\_tweets\_modified에 sentiment column만 추가 되었으므로, 관련 쿼리문 및 설명은 생략한다. 아래는 생성된 tweetsbi 테이블의 row 예시이다. Sentiment 가 2 이면 긍정, 1이면 중립, 0이면 부정을 의미한다.

*<Example : tweetsbi Table>*

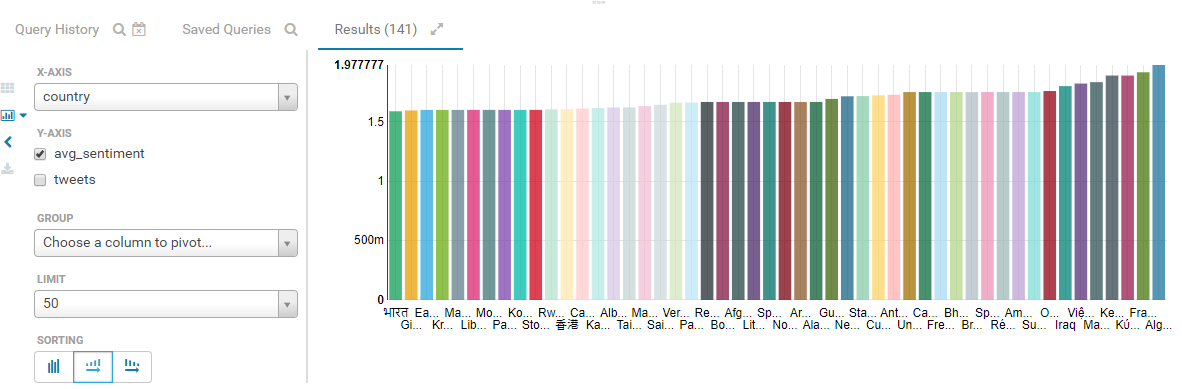


1. 생성된 tweetsbi 테이블에 대하여 hue에서 간단한 쿼리 및 시각화를 진행하였다. 각 국가별 감성 지수의 평균을 구하여 Christmas 관련 tweet 내 국가별 행복도를 시각화 하였다. 아래는 시각화에 사용된 쿼리와 결과물이다.

*<Hue visualization에 사용된 쿼리문>*

|  |
| --- |
| select country, avg(sentiment) as avg\_sentiment, count(\*) as tweets  from tweetsbi  where sentiment is not null  group by country  order by avg\_sentiment desc; |

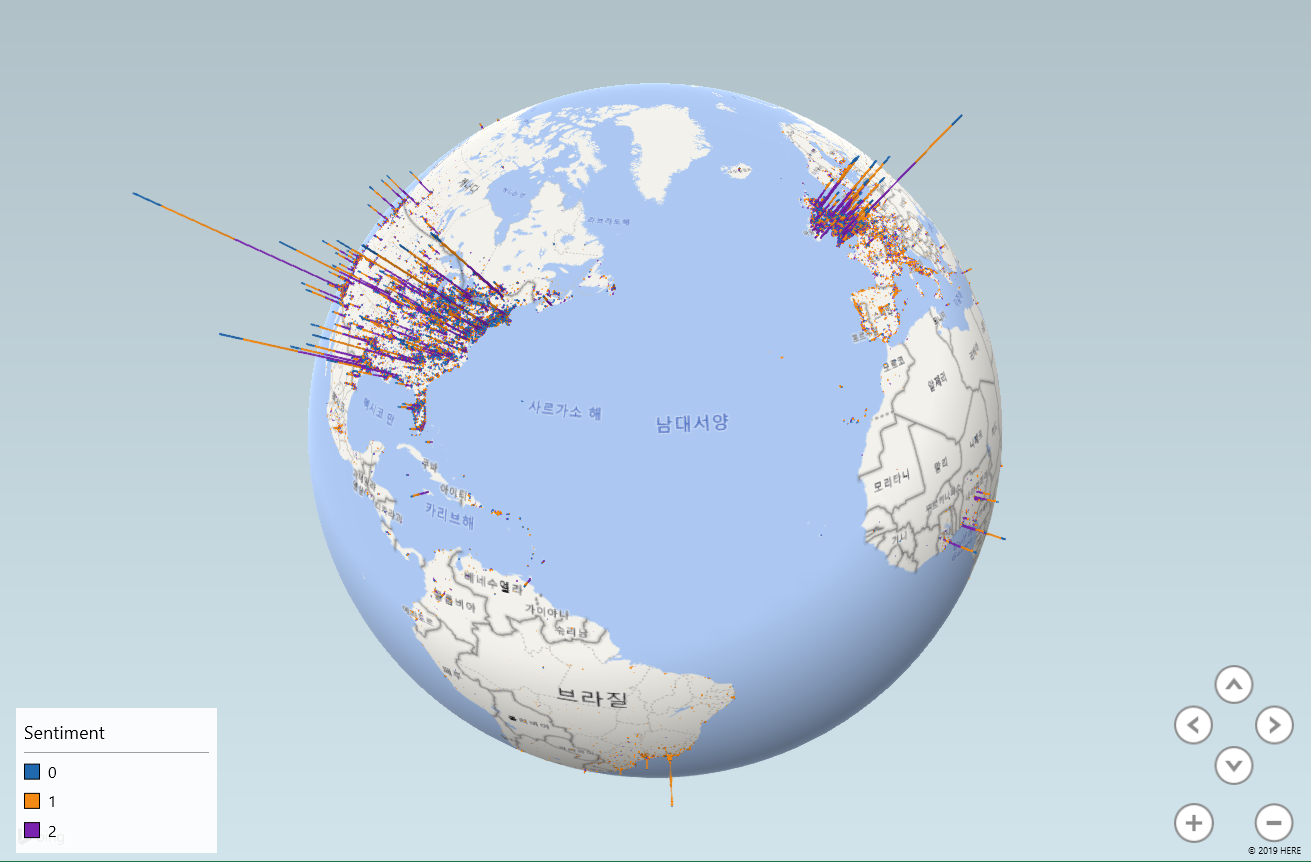
*<Hue visualization 결과>*



국가별로 수집한 tweet의 양이 달라 정확한 결과라고 말하기는 어렵지만, 국가별 행복도 순위를 알 수 있었고, 전반적으로 높은 행복도를 보임을 그래프를 통해 확인할 수 있었다.

1. Excel 내 3D Map 기능을 활용하여, 각 국가 및 지역별 감성 분석을 시각화 하였다. ODBC driver을 활용하여 생성한 tweetsbi hive table을 import하여 시각화에 활용하였다. Table 내 지역 정보를 활용하여 지역별로 발생한 tweet 수를 표시하고, 시간 정보를 활용해 시간대별 생성되는 tweet을 애니메이션을 통해 확인할 수 있었다. 막대그래프 외에도 거품형, 열지도 등의 3D 지도를 지원한다. 영문 단어를 기준으로 감성 분석을 진행하다 보니, 영어권 국가에서 발생하는 tweet 수가 압도적으로 많음을 확인할 수 있었다.

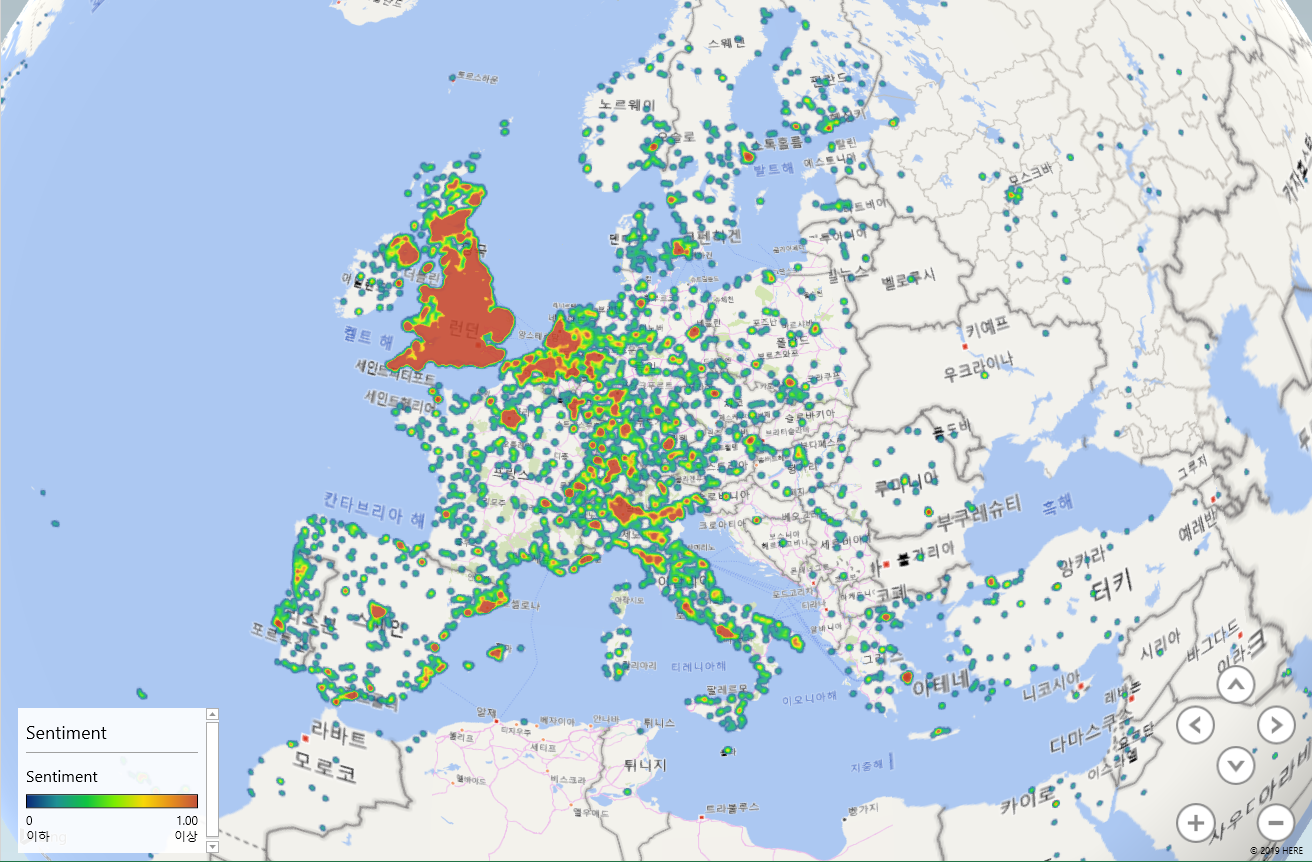
*<Excel 3D Map : Bar graph>*



*<Excel 3D Map : Heat map\_1>*



*<Excel 3D Map : Heat map\_2>*



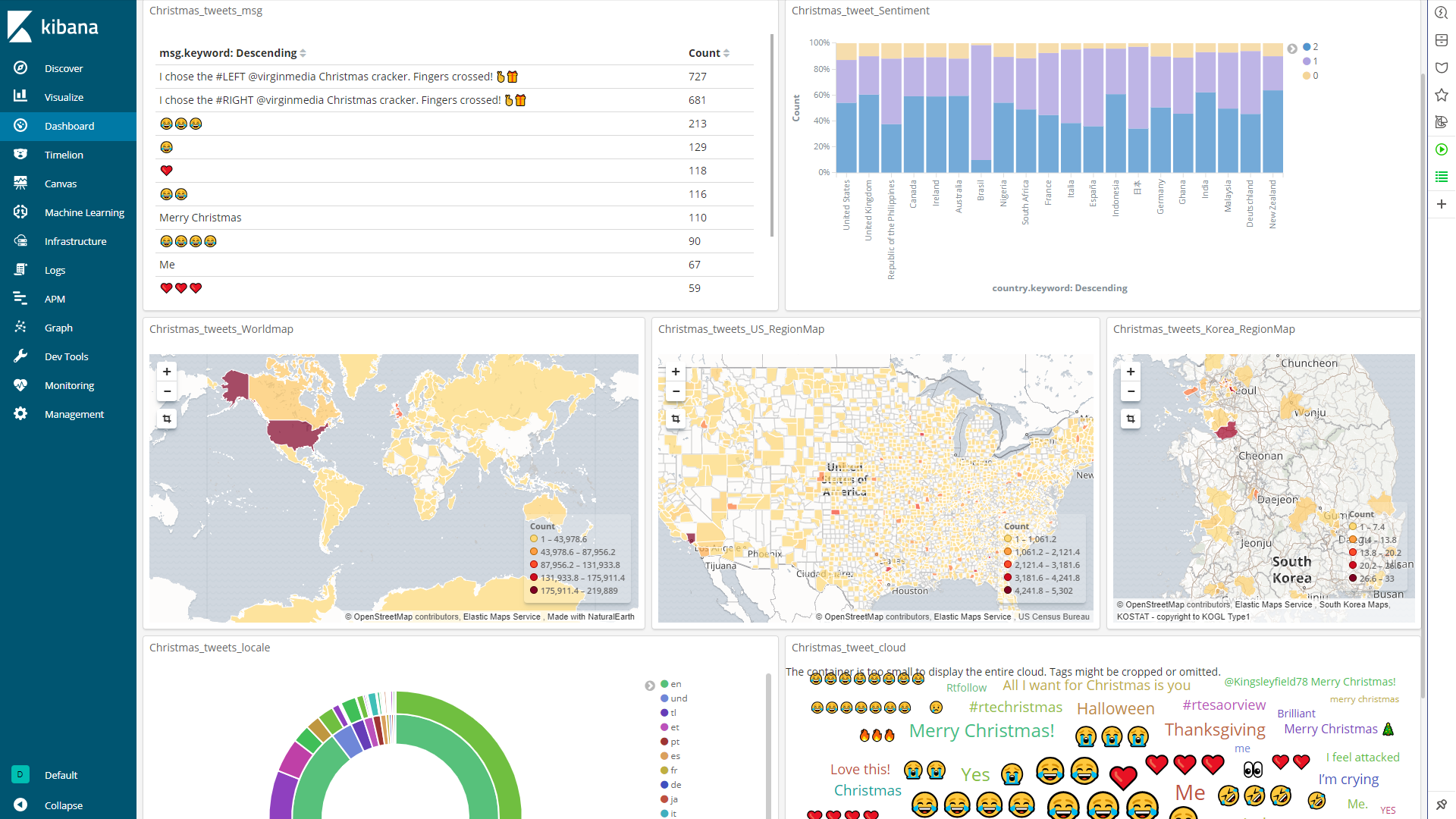
1. ElasticSearch에서 시각화하기 위해서는, ElasticSearch의 index에 데이터를 우선적으로 저장해야 한다. Hadoop과 ElasticSearch간 connector인 ElasticSearch-Hadoop.jar파일을 클러스터 내 노드에 다운받고, hive에서 쿼리문 작성을 통해 ElasticSearch 내 index 생성이 가능하다. Index 생성 및 데이터 삽입을 위한 hive 쿼리는 아래와 같다.

*<ElasticSearch Index 생성을 위한 Hive 쿼리>*

|  |
| --- |
| --hadoop-hive connector 추가  add jar /opt/elasticsearch-hadoop-6.5.0/dist/elasticsearch-hadoop-hive-6.5.0.jar;  --ElasticSearch에 Index/type tweets\_christmas/tweets 생성  CREATE EXTERNAL TABLE tweetsbi\_kibana (  tweet\_id string,  msg string,  created\_time\_gmt TIMESTAMP,  lang string,  loc string,  country string,  sentiment bigint  ) STORED BY ' org.elasticsearch.hadoop.hive.EsStorageHandler'  TBLPROPERTIES('es.resource' = 'tweets\_christmas/tweets','es.nodes'='10.100.3.210');  --데이터 삽입  INSERT OVERWRITE TABLE tweetsbi\_kibana SELECT \* FROM tweetsbi; |

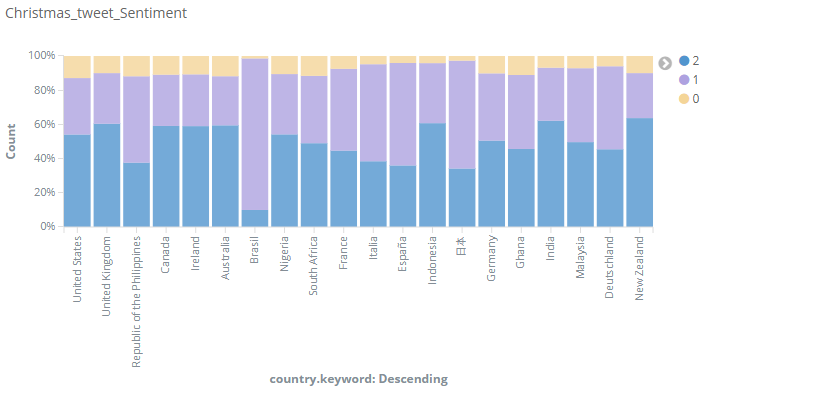
1. Visualize 및 Dashboard 구성을 위해 Kibana에 index pattern을 추가하고, 해당 index pattern을 이용해 visualize chart 및 dashboard를 구성한다. 구성한 dashboard 및 visualize chart는 아래와 같다.

*<Kibana Dashboard>*



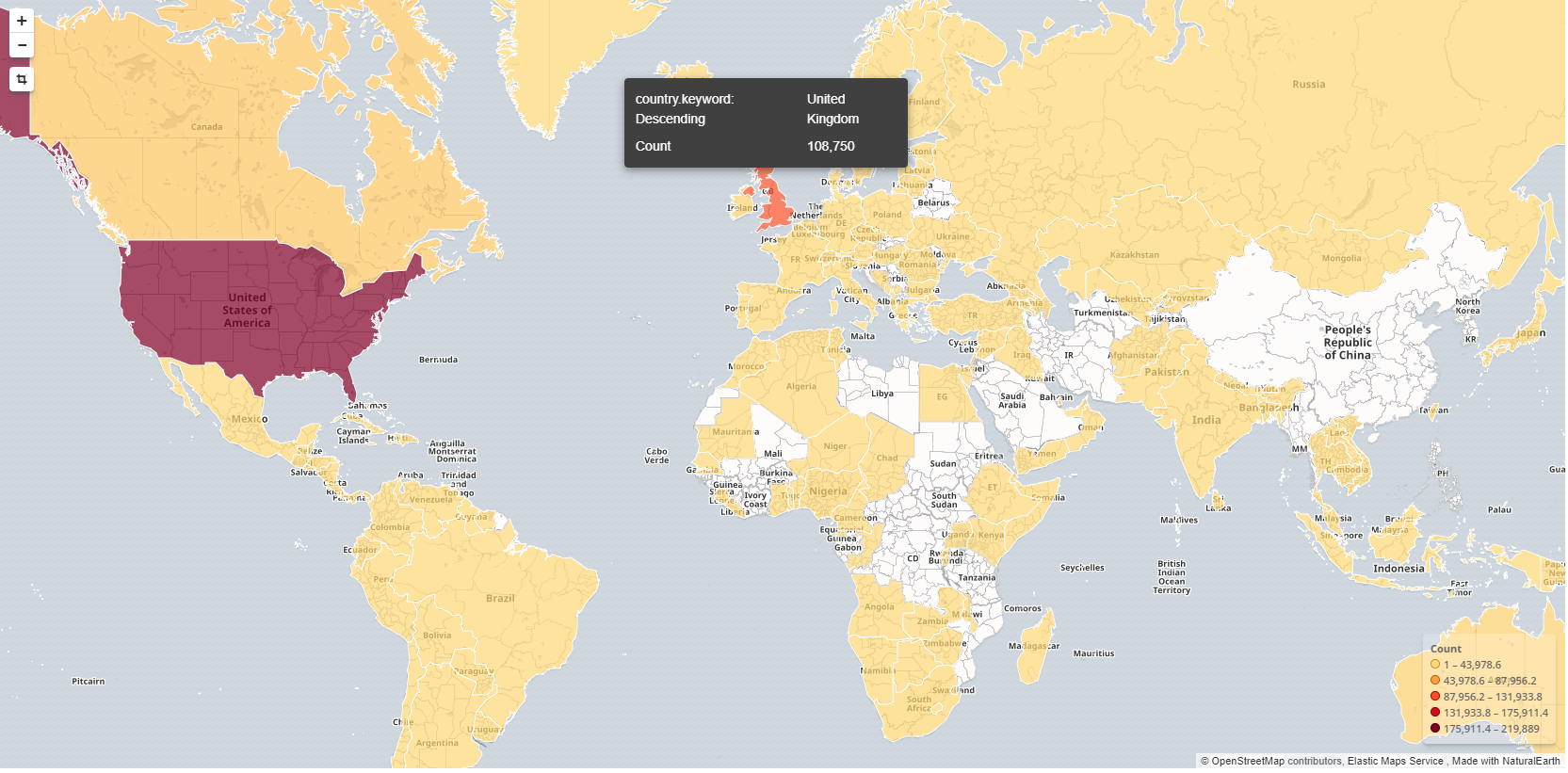
Dashboard는 Christmas ①키워드에 관련된 트윗 중 가장 많이 작성된 트윗, ②국가별 sentiment analysis를 시각화 한 차트, ③국가별, 국가 내 지ㅇ역별 작성된 트윗 수(④⑤), ⑥트윗을 작성한 언어 및 국가, ⑦트윗에 포함된 단어 중 가장 많이 사용된 단어 등으로 구성되어 있다. Filter를 이용해 전체 dashboard에 대하여 특정 기간, 지역 등 특정 조건에 대한 결과값을 확인할 수 있다. 또한 자동으로 dashboard에 연결된 index pattern을 동기화 하여, dashboard를 갱신할 수 있다.

*<Visualize Chart\_*② *: 국가별 Sentiment Analysis>*



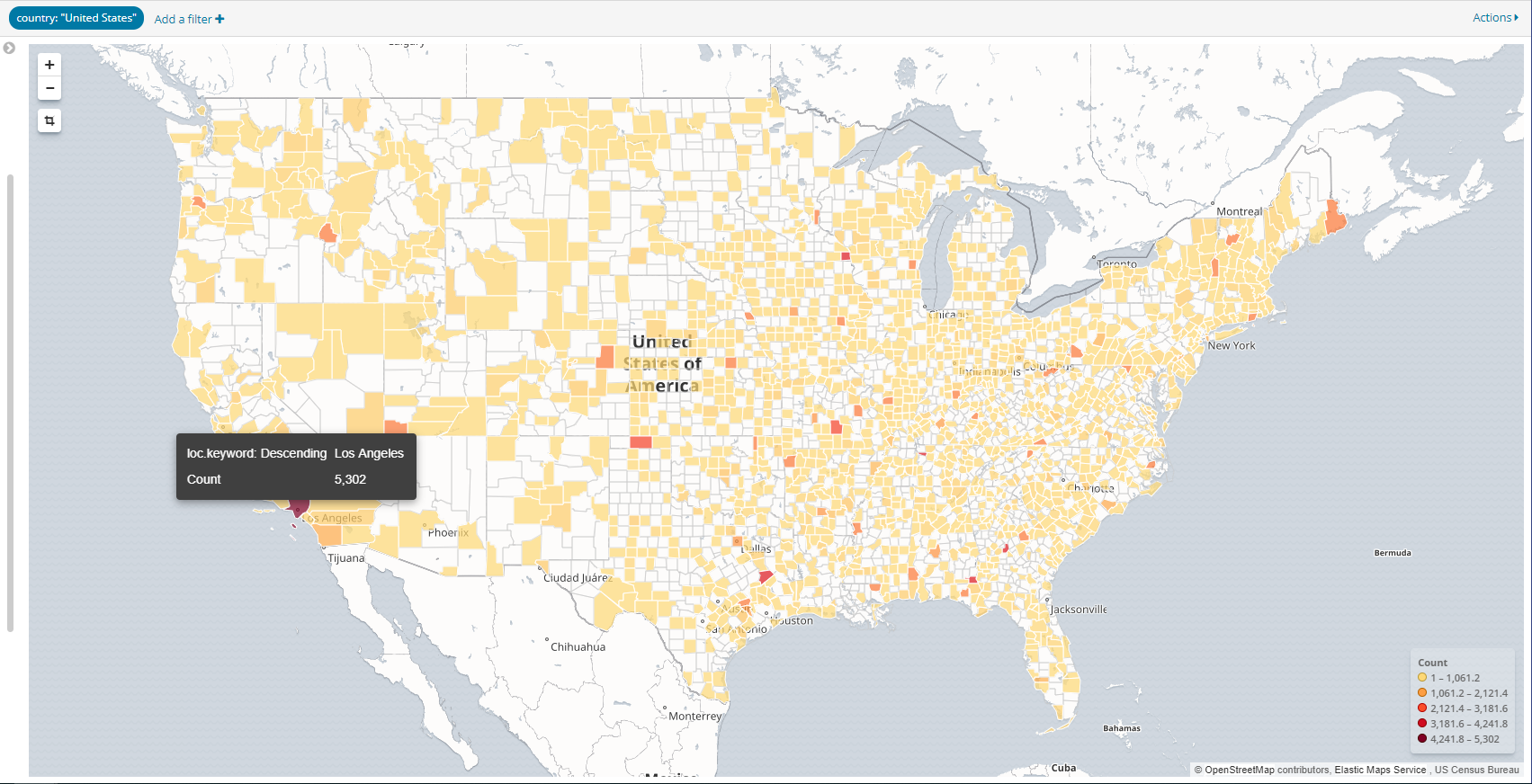
감성 분석을 진행한 결과물에 대하여 각 국가별 분석 결과를 시각화한 차트이다. 감성 분석에 사용된 사전이 영문 단어에 대해서만 정의가 되어 있어, 정의되지 않은 단어에 중립 점수를 매기는 방식에 따라 비 영어권 국가에 대해서는 비교적 중립 상태(1)의 tweet이 많은 것을 확인할 수 있다. 앞서 언급했듯이, 2는 긍정, 1은 중립, 0은 부정의 감정을 의미한다.

*<Visualize Chart\_*③ *: World Map>*



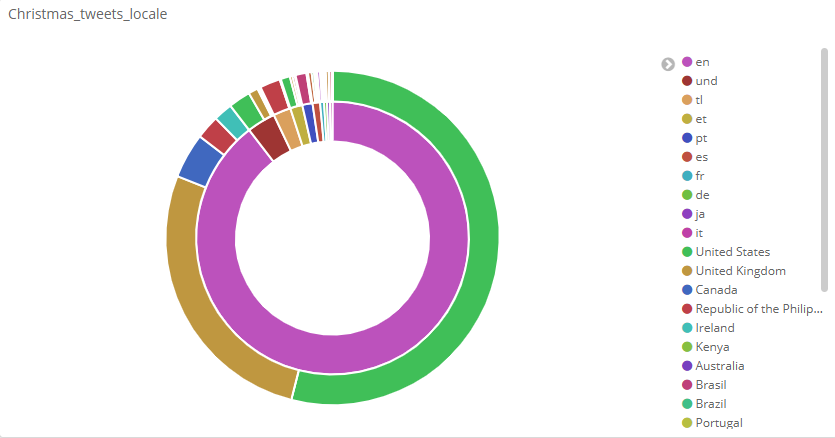
Tweet 내 포함된 위치 정보를 활용해 Region map visualize chart를 구성한 것이다. 미국, 영국 등 영어권 국가에서 발생한 tweet이 많음을 확인할 수 있는 동시에, 전 세계적으로 트윗이 발생함을 확인할 수 있다.

*<Visualize Chart\_*④ *: United States Region map>*



마찬가지로, tweet내 포함된 지명으로 표시된 위치 정보를 활용하여, 미국 내에 상세 지역별 tweet발생 수를 표시하였다. LA, 뉴욕 등 도심 지역을 중심으로 많은 양의 tweet이 발생하였음을 확인할 수 있었다.

*<Visualize Chart\_*⑥ *: 국가별 / 언어별Christmas tweets>*



Tweet에 사용된 언어별, 국가별 트윗 수를 나타낸 차트이다. 분홍색으로 표기된 부분을 통해, 영어로 90%가 넘는 양의 tweet이 작성되었고, 그 중 대부분의 tweet이 미국과 영국에서 작성되었음을 확인할 수 있었다. 다음으로는 캐나다, 필리핀 등에서 많은 양의 tweet을 작성했음을 확인하였다.

*<Visualize Chart\_*⑦ *: Tweet에 가장 많이 포함된 단어들>*



Christmas 라는 keyword와 관련하여 수집한 tweet 내에, 가장 많이 포함된 단어를 보여주는 cloud visualize chart이다. 단어의 크기가 클수록 많이 사용된 단어임을 의미한다.

## 6.3 분석 결과

위의 Tweet 감성 분석을 위해 StreamSets을 활용해 ETL 작업을 진행하고, Hive를 이용해 간단한 데이터 분석 및 가공을 진행하였다. Hive 및 HDFS 내 저장된 파일을 보기 위해 Hue를 사용하였으며, Hue를 통해 작성한 Hive 쿼리를 이용해, Hue에서 제공되는 시각화 기능도 이용해 보았다. 뿐만 아니라 ElasticSearch Hadoop connector 및 ElasticSearch, Kibana를 이용하여 수집한 tweet 데이터를 분석하여 차트 및 dashboard로 시각화를 진행해 보았다. 이 외에도 Oozie를 활용해 주기적으로 StreamSets 을 통해 추가된 hive table의 행들을 ES로 넘겨주는 Coordinator 작성을 통해 갱신된 데이터에 대해서도 시각화를 진행하거나, Sqoop을 이용해 수집한 테이블을 RDB로 export하기, 수집한 데이터를 바로 HDFS나 Hive table에 넣지 않고 Kafka의 topic을 활용하여 부하 줄이기 등 하둡 내 다양한 서비스를 활용해 볼 수 있다. 감성 분석 자체가 의미를 가지기 보다는, 실생활에서 발생하는 대용량의 데이터를 Hadoop 내로 import하여 처리 및 분석, 시각화 하였다는 점에 의의를 가진다. 영어 단어만을 가지고 감성 분석을 진행하는데 그치지 않고, 한국어 tweet에 대해서도 n-gram 분석을 통해 감성 분석을 진행하고, 위도 및 경도 데이터의 활용을 통해 더욱 정확한 결과치를 내는 것을 다음 목표로 할 것이다. 실제 데이터를 이용해 의미 있는 분석 및 시각화를 진행했다는 점에서 현 데이터 분석은 의미를 가진다.

# 7. 결론

다양한 이유로 세계적으로 높은 관심을 불러 일으킨 빅데이터에 대하여 빅데이터의 기본적인 개념 및 배경부터, 빅데이터 처리 및 분석에 활용되는 Hadoop Eco-system에 대한 이해, Hadoop Eco-system을 이루는 각각의 서비스들의 개념 및 아키텍처 파악, 마지막으로 실제 tweet 데이터를 이용한 감성 분석을 진행해 봄으로써, 전반적인 빅데이터에 대한 이해도를 높일 수 있었다. 앞서 파악한 Hadoop 내 서비스들에 대한 이해와 tweet 데이터를 이용해 진행해 본 데이터 분석 및 시각화 경험을 바탕으로, tweet 데이터 외에 Facebook, Kakao, Samsung Health 등에서 지원하는 다양한 API 들을 통해 수집한 데이터를 처리하고, 분석할 수 있다. 더 나아가, json 형태의 반정형 데이터인 tweet 분석에 그치지 않고, 사진, 동영상 등의 비정형 데이터에 대해서도 데이터 분석을 진행할 수 있다.

# 참고 자료

* Twitter Filter API : <https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html>
* 데이터 분석 참고 URL : <https://github.com/star7357/tutorials/blob/hdp/tutorials/hortonworks/analyzing-social-media-sentiment-data/tutorial.md>