서울시립대학교
UNIVERSITY OF SEOUL

프로젝트 최종 보고서

세계 테러리즘 군집 분석: 공격성을 중심으로

컴퓨터과학부

2016920004

권은진

소프트웨어응용 02분반

최종 수정일: 2019.12.10

목 차

1. 개요

1.1. 목적

본 프로젝트는 전세계에서 발생한 테러리즘을 공격적인 측면에서 군집 분석을 수행하고 그 결과를 시각화해주는 엔진을 개발하는 것을 목적으로 하며, 이러한 시각화 결과를 통해 향후 테러리즘을 예측하는데 도움이 될 수 있는 자료를 제공하는 것을 목표로 한다.

1.2. 데이터셋

분석 대상은 National Consortium for the Study of Terrorism and Responses to Terrorism (START)에 의해 관리되는 Global Terrorism Database (이하 GTD) 이며, 해당 자료는 1970 년대부터 2017 년도까지 180,000 회 이상의 테러 사건을 포함하고 있다. 데이터셋은 총 135개의 변수와 181,601개의 사건들로 구성되어 있으며, 각 변수들의 의미는 GTD에서 제공하는 Codebook을 기반으로 해석하였다.

1.3. 분석 환경

구분	이름
사용 언어	Python 3
사용 모듈	Pandas, numpy, matplotlib, Sklearn, Prince, Seaborn, folium
사용 도구	Google Colaboratory

[표 1-1] 분석 환경

2. 데이터 전처리

2.1. 기본 분석

데이터셋은 135개의 변수를 가지지만, 변수들의 의미는 중복된 것이 많았다. 예를 들어, 사용한 무기유형을 의미하는 weaptype1 변수는 범주형 변수로 13개의 값이 존재한다. 이 변수는 범주 코드번호를 저장하나 weaptype1_txt 변수는 번호가 아닌 범주명을 문자열로 저장한다. 이처럼 범주 코드번호와 범주명을 별도로 저장하기 때문에, 중복되는 변수를 제외하고 본다면 총 변수의 개수는 107개가 된다.

하나의 사건은 둘 이상의 무기를 사용할 수 있고, 공격 대상이 둘 이상일 수 있다. 예를 들어, 공격 대상을 의미하는 변수는 targtype1, targtype2, targtype3 세 가지가 존재한다. 그러나 모든 사건에서 공격 대상이 둘 이상은 아니므로 빈 값(null value)이 존재하게 된다. 이러한 missing cell의 개수가 전체 중 절반 이상을 차지할 경우 해당 변수에 대해서는 분석을 진행하지 않았다.

군집 분석에 필요한 변수는 공격 성향을 파악하는데 도움이 되는 것이어야 한다. 따라서 사건 발생 국가, 위도 및 경도, 사건 날짜 등 분석에 필요하지 않은 변수는 제외하였다. 단, 제외한 변수들은 시각화에 다시 사용될 수 있다. 다음은 실제로 사용할 변수들의 목록이다.

변수	설명
general_purpose	공격을 가한 집단이 가지는 테러리즘의 목적을 의미 (4가지 카테고리)
extended	테러 활동이 24시간 이상 지속되었으면 1, 아니면 0
suicide	테러리즘이 자살 공격을 수반하였으면 1, 아니면 0
attacktype1	테러리즘의 공격 유형 (9가지 카테고리)
targtype1	테러리즘의 공격 대상 유형 (22가지 카테고리)
targsubtype1	테러리즘의 공격 대상의 세부 유형 (111가지 카테고리)
weaptype1	테러리즘에 사용된 무기 유형 (13가지 카테고리)
weapsubtype1	테러리즘에 사용된 무기의 세부 유형 (27가지 카테고리)

[표 2-1] 변수 목록

2.2. 결측값(Missing Value) 처리

필요한 변수에도 값이 없는 경우가 존재한다. 제공한 codebook 에 따르면, 값 중에 “Unknown” 이 있어도 관련 자료를 찾지 못하면 셀을 비운다고 한다. 또는 종속적인 변수인 경우 결측값을 갖게 되는데 예를 들어 targtype1 변수에서 세부 유형이 없는 공격 대상인 경우 targsubtype1 변수는 결측값을 갖게 된다. 이런 경우 각 필드에 “Unknown” 또는 “Not exist”를 의미하는 값으로 대체하였다.

2.3. 범주형 변수(Categorical Variables) 처리

[표 2-1]의 변수 목록을 보면, 사용할 모든 변수들이 범주형 변수인 것을 알 수 있다. 군집 분석을 하기 위해서는 먼저 이 변수들을 모두 수치화(Numerical Variables 로 만드는 과정)를 해주어야 한다. 이 과정에서 원-핫 인코딩(One-Hot Encoding)을 적용하였으며, 이 방법은 범주의 크기를 벡터의 차원으로 표현하여 해당되는 범주는 1, 나머지는 0 으로 값을 부여하는 것이다. 이렇게 표현된 벡터를 원-핫 벡터(One-Hot Vector)라 부른다. 아래의 자료는 원-핫 인코딩의 예시이다.

	attacktype	targtype	weaptype
0	Bombing	NGO	Firearms
1	Hijacking	Police	Chemical
2	Infra Attack	Military	Nuclear
3	Hijacking	NGO	Nuclear

[표 2-2] 범주형 변수 처리 전

	attacktype			targtype			weaptype		
	Bombing	Hijacking	Infra Attack	NGO	Police	Military	Firearms	Chemical	Nuclear
0	1	0	0	1	0	0	1	0	0
1	0	1	0	0	1	0	0	1	0
2	0	0	1	0	0	1	0	0	1
3	0	1	0	1	0	0	0	0	1

[표 2-3] One-Hot Encoding 결과

여기서 행 0 에 대한 원-핫 벡터는 [1, 0, 0, 1, 0, 0, 1, 0, 0] 이 된다. 각 행은 하나의 사건을 의미하며 벡터로 표현이 되는데, 데이터셋은 이러한 벡터들이 모인 하나의 행렬(Matrix)로 볼 수 있다.

3. 다중 대응 분석

3.1. 개요

대응 분석(Correspondence Analysis)이란 자료의 행과 열범주를 저차원 공간상(2차원)의 점들로 동시에 나타내어, 그들의 관계를 탐구하려는 탐색적 자료 분석 기법이다. 열범주를 나타내는 변수가 2개이면 단순 대응 분석이라 하고, 변수가 3개 이상일 경우 다중 대응 분석이라 한다. 다중 대응 분석은 범주형 변수들이 많은 데이터셋을 대상으로 군집 분석을 하는데 널리 쓰이는 기법이다.

2 장에서는 데이터셋의 범주형 변수들은 원-핫 인코딩에 의해 벡터로 구성된 하나의 행렬(Matrix)로 표현되었다. 이 행렬은 8 개의 변수로부터 총 52 개의 범주를 갖기 때문에 52 차원이 된다. 그러나, 행렬의 원소들을 좌표에 나타내려면 벡터의 차원은 2 차원이 되어야 한다. 따라서 다중 대응 분석을 이용하여 차원 축소(Dimensionality Reduction)를 수행해야 한다.

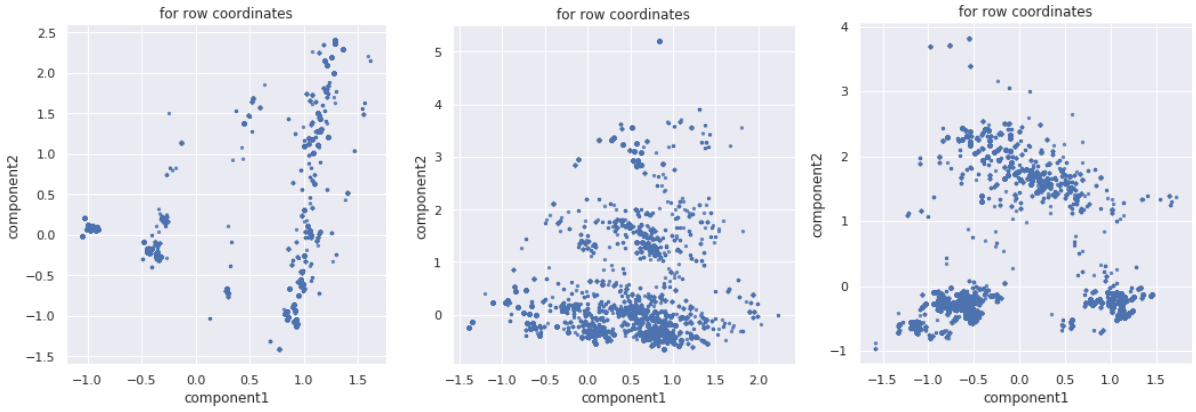
3.2. 결과

다중 대응 분석의 결과로 2 개의 계수 좌표(factor coordinates)를 얻게 되며, 각 계수 좌표를 통해 그래프 상에서 원본 데이터의 분포나 변수의 분포를 살펴볼 수 있다. 그러나 변수의 분포는 원-핫 인코딩으로 인해 기존의 변수인 attacktype1 과 weaptype1 의 관계를 살펴보는 것이 아니라, 각 변수의 범주의 관계를 살펴보게 된다. 예를 들어, attacktype1 의 Bombing 과 weaptype1 의 Firearms 의 관계를 살펴보게 되는 것이다. 이 점은 변수마다 범주의 크기가 다르며, 범주가 심하게 차이 날수록 한 쪽 변수 내 범주로 데이터가 치우치는 경향이 있어 원활한 분석이 이루어지지 않았다. 또한 데이터의 분포에서 군집 분석을 수행하는 것이 본래의 목적이었기 때문에 변수간 계수 좌표는 다루지 않았다.

[표 2-2]를 보면 총 8 개의 변수를 사용하는데, 변수를 모두 사용하거나 몇 개씩 그룹화하여 분석을 진행했을 때 데이터가 한 곳에만 집중적으로 존재하는 등 군집 분석을 수행하기 힘들다고 판단하는 그룹들은 제외하였다. 아래는 군집 분석에 적합하다고 판단한 그룹들이다.

구분	변수
Group 4	attacktype1, weaptype1, weapsubtype1
Group 10	attacktype1, weaptype1, targtype1
Group 19	weaptype1, weapsubtype1, targtype1

[표 3-1] 다중 대응 분석 적용 후 실제 군집 분석을 수행할 그룹 목록



[그림 3-1] 왼쪽부터 Group 4, Group 10, Group 19의 데이터간 계수 좌표

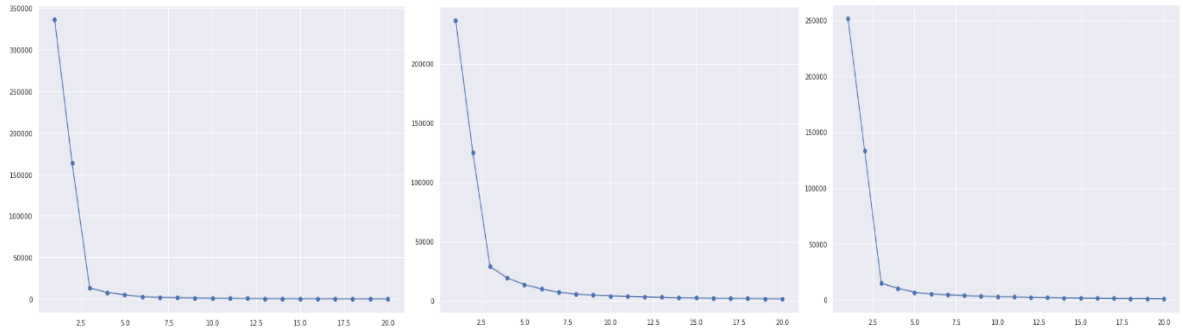
4. 군집 분석: K-Means

4.1. 클러스터 개수 설정

데이터 분포를 그래프 상에 표현할 때, 몇 개의 그룹으로 데이터를 분류해야 최적의 결과가 되는지 모호할 때가 있다. 이 문제를 해결하는 대표적인 방법으로는 엘보우 기법(Elbow Method)과 실루엣 기법(Silhouette Method)이 있다.

4.2.1. 엘보우 기법(Elbow Method)

이 기법은 클러스터 내 오차제곱합(SSE)이 최소가 되도록 클러스터의 중심을 결정하는 방법이다. 실제로 클러스터 개수에 따른 오차제곱합을 그래프 상에 나타내면, 오차제곱합이 급격히 줄어들다가 어는 순간 천천히 줄어드는 부분을 볼 수 있는데 사람의 팔에서 팔꿈치 부분에 해당되는 그 점이 적절한 클러스터의 개수이다. 아래 예시에서 가로축은 클러스터 개수이고, 세로축은 오차제곱합이다.



[그림 4-1] 왼쪽부터 Group 4, Group 10, Group 19 의 엘보우 기법 적용 결과

위 그림에서 표시된 각 점들은 클러스터의 개수를 의미한다. 팔꿈치 부분에 해당되는 점은 k = 3 일 때이지만, 실제 3 개의 클러스터에 대해 군집 분석을 수행하면 군집의 특성을 파악하기 힘들 수 있다. 이 때는 클러스터 내 응집도와 분리도를 파악하여 클러스터를 평가하는 방법인 실루엣 기법을 적용할 수 있다.

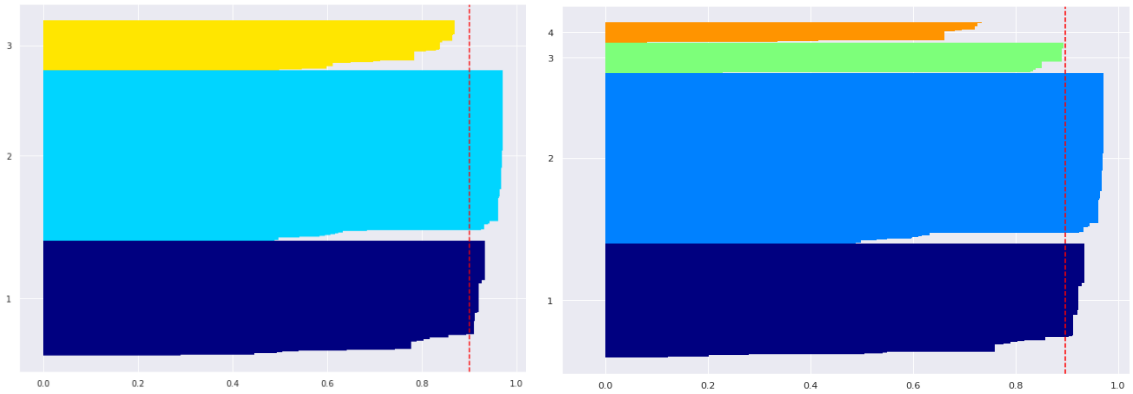
4.2.2. 실루엣 기법(Silhouette Method)

실루엣 기법은 군집 분석이 얼마나 잘 되었는지 측정하는 방법으로, 엘보우 기법의 결과가 좋지 않아 클러스터의 평가 척도를 달리할 필요가 있을 때 일반적으로 사용된다.

i 번째 데이터에 대한 실루엣 계수 s(i)는 아래와 같이 정의된다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

여기서 a(i)는 클러스터 내 데이터 응집도(cohesion)을 나타내는 값으로, i 번째 데이터와 동일한 클러스터 내의 나머지 데이터들 간의 평균 거리로 정의된다. 거리가 작을수록 응집도가 높아진다. b(i)는 클러스터간 분리도(separation)를 나타내는 값으로, i 번째 데이터와 가장 가까운 클러스터 내의 모든 데이터들 간의 평균 거리로 정의된다. 만약 클러스터 개수가 최적화되어 있다면 b(i)는 큰 값을 가지고, a(i)는 작은 값을 갖게 된다. 따라서 실루엣 계수 s(i)는 1 에 가까운 숫자가 된다.

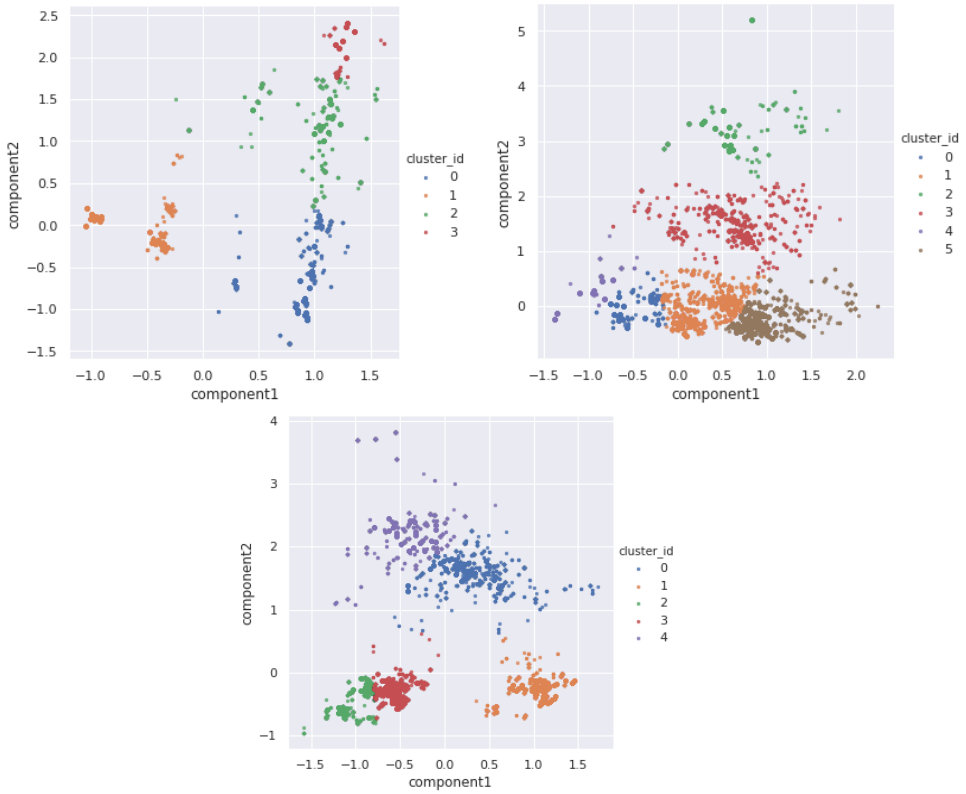


[그림 4-2] Group 4 의 실루엣 기법 적용 결과

[그림 4-2]는 3 장에서 보인 4 번째 그룹에 실루엣 기법을 적용한 결과이며, 왼쪽은 클러스터 개수가 3 일 때, 오른쪽은 4 일 때의 결과이다. 클러스터가 3 개일 때가 4 개일 때보다 실루엣 계수가 조금 더 높게 나오고, 4 개일 때는 1 개의 클러스터가 동떨어져 있고 나머지는 실루엣 계수의 평균에 가까운 값을 보인다. 처음에 보면 클러스터의 개수가 3 개일 때가 더 적합하다고 판단할 수 있지만 실제 군집 분석을 수행해보면 클러스터 3 은 데이터가 많이 없는 범주들이 섞여 있어 특징을 파악하기 힘들다. 이런 경우 해당 클러스터에 대해서만 다시 군집 분석을 수행하는 방법 등에 의해 특징을 파악할 수 있다.

4.2. 결과

다음은 3장에서 보인 각 그룹의 계수 좌표(factor coordinates)에 군집 분석을 수행한 결과이다. 클러스터 개수는 Group 4에 대해서는 3으로, Group 10에 대해서는 6으로, Group 19에 대해서는 5로 지정하였다. 이전 설명처럼 특정 클러스터에 대해서 다시 군집 분석을 수행할 수 있지만 Group 4 의 경우 클러스터 개수를 늘려주면서 클러스터 내 특징을 파악할 수 있었다.



[그림 4-3] k-means 의 적용 결과 (좌측 상단: Group 4, 우측 상단: Group 10, 하단: Group 19 의 결과)

5. 엔진 개발

5.1. 프로그램 개요

본 프로그램은 데이터베이스를 사용하지 않으며, 별도의 로그인을 요구하지 않는다. 빠른 분석을 위해 데이터셋(.csv 파일)만 받아서 결과 페이지를 전달한다. 분석을 자동화하는 과정에서 데이터베이스가 없기 때문에 서버 로컬 환경에서 필요한 데이터들을 저장한다. 데이터들은 쌓이면 백그라운드 환경에서 실행되는 프로그램(crontab)에 의해 자동으로 삭제된다. 결과 페이지는 다음과 같이 크게 3 가지가 있다.

(1) 결측값 분석 페이지: 처음에 데이터셋에 대한 결측값을 분석해주는 페이지로, 다음 페이지에서 사용자는 결과를 보고 다중 대응 분석에 사용할 변수(feature)를 선택할 수 있다. 최대 4 개의 변수그룹을 만들 수 있으므로 최대 4 번까지 군집 분석을 자동화해준다.

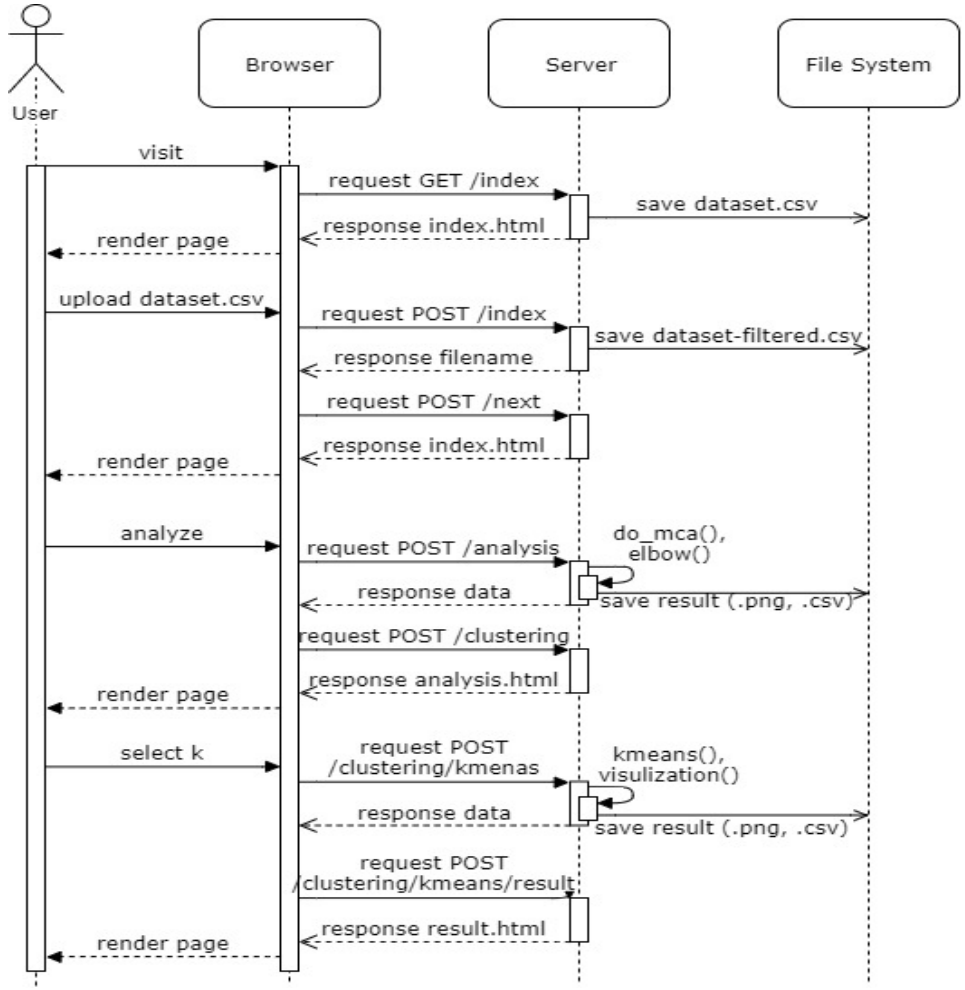
(2) 클러스터 개수 선택 페이지: 이전 페이지에서 전달받은 변수 그룹 별로 다중 대응 분석을 수행한 결과를 보여주며, 엘보우 기법 적용 결과도 함께 보여주고 각 그룹별로 클러스터 개수를 사용자가 지정할 수 있게 한다. 필요 시 실루엣 기법도 적용할 수 있도록 별도의 버튼이 구현되어 있다.

(3) 군집 분석 결과 페이지: 사용자가 지정한 클러스터 개수를 각 변수 그룹별로 적용하여 군집 분석을 수행하고, 그 결과를 전달한다. 이 페이지에서는 각 그룹별 군집 분석 전후 결과 이미지, 각 클러스터 내에서 선택한 변수의 범주 분포도, 각 클러스터 내 데이터 분포도를 지도에 마커로 표기한 뷰, 국가별, 지역별로 공격 횟수와 사망자 수를 비교하는 그래프(상위 15 개만 비교)가 포함되어 있다. 지도에 마커를 클릭할 경우 상세한 정보를 확인할 수 있다. 또한 클러스터 내 데이터를 원본(csv 파일)으로 다운로드 받아 추가 분석이 가능하다.

개발 환경은 서버의 경우 분석 환경에서 사용한 코드를 재사용하기 위해 python 3를 언어로 선택하였고, 리눅스 환경에서 Flask 프레임워크로 구축된 웹서버이며, 클라이언트의 경우 브라우저 상에서 쉽고 간편하게 분석을 진행할 수 있다.

5.2. 프로그램 모델링

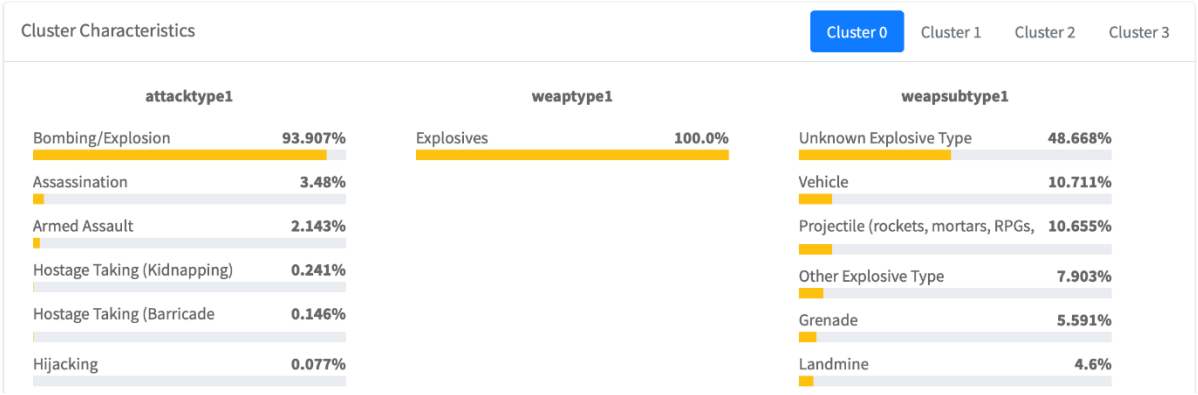
[그림 5-1]을 보면, 사용자로부터 데이터를 전달받고 바로 페이지를 렌더링을 하는게 아니라, 데이터를 처리하고 다시 요청을 받은 뒤 페이지를 렌더링한다. 이 부분은 데이터가 많을수록 사용자가 대기해야 하는 시간이 길어지는데 그 시간 동안 이전 페이지에 대한 정보를 사용자가 계속 볼 수 있도록 하기 위함이다. 또한 처음에 사용자에게 보여지는 페이지(index.html)는 데이터셋을 사용자가 업로드 했을 때와 하지 않았을 때를 구분하여 렌더링되므로, 처음 2 번의 요청에 각각 다른 결과를 보인다.



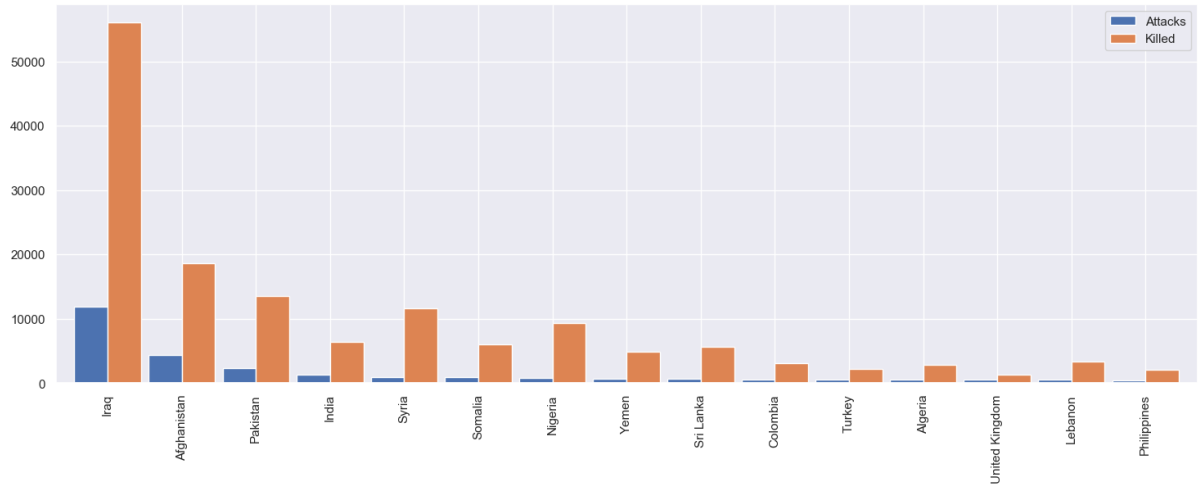
[그림 5-1] 순서 다이어그램

5.3. 실행 결과

다음은 군집 분석이 잘 되었던 3 개의 그룹 중 Group 4 의 결과 페이지 중 일부이다. Group 4 의 클러스터 4 개는 각각 전체 중 50.868%, 34.055%, 9.089%, 5.989%를 차지하며, 아래는 각 변수 내 범주 분포도와 국가별 데이터 비교 그래프이다. 범주 분포도의 경우 범주의 크기가 6 이상 이더라도 1% 미만의 결과는 이미지에서 잘려진 상태이다.

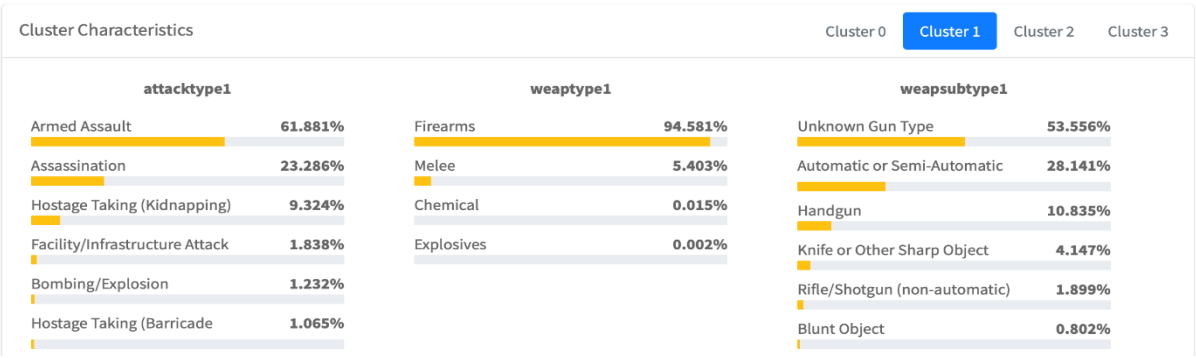


[그림 5-2] Group 4 의 Cluster 0 의 범주 분포도

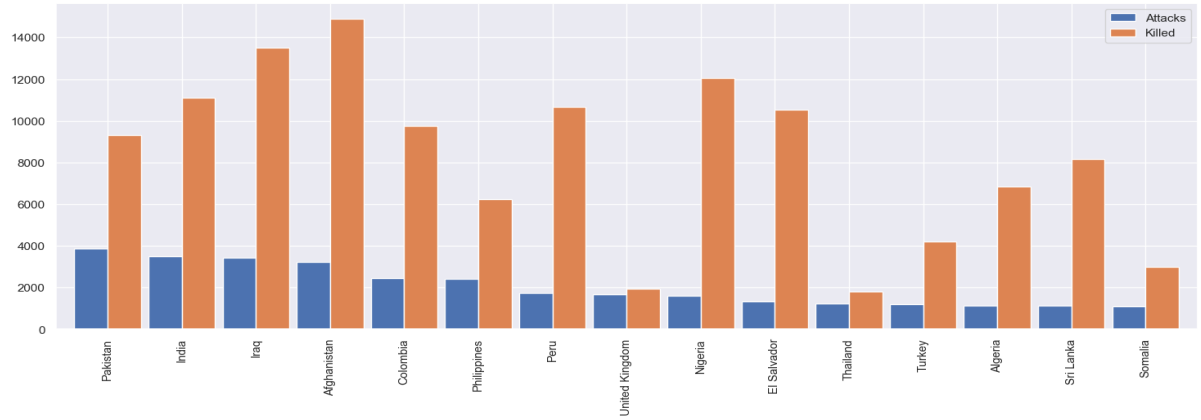


[그림 5-3] Group 4 의 Cluster 0 내 데이터의 국가별 공격 횟수 및 사망자 수 비교 그래프

첫번째 군집은 폭탄, 지뢰, 수류탄 등의 폭발형 무기를 사용한 테러 활동에 해당되는 데이터의 집합이다. 이 군집의 경우 이라크, 아프가니스탄, 파키스탄 등 중동지역에서 많이 나타난 데이터이다.

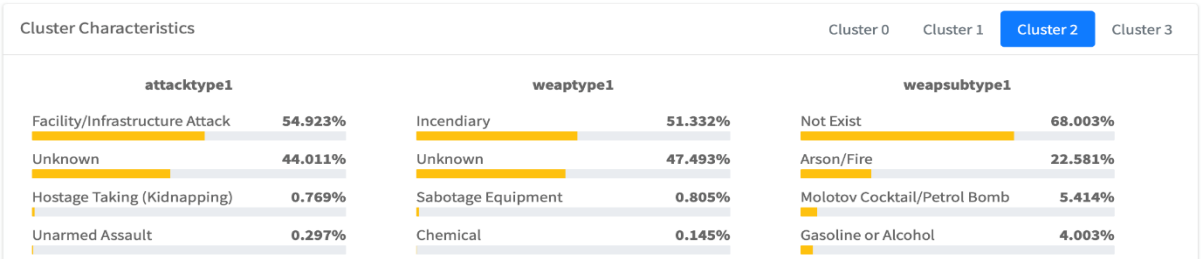


[그림 5-4] Group 4 의 Cluster 1 의 범주 분포도

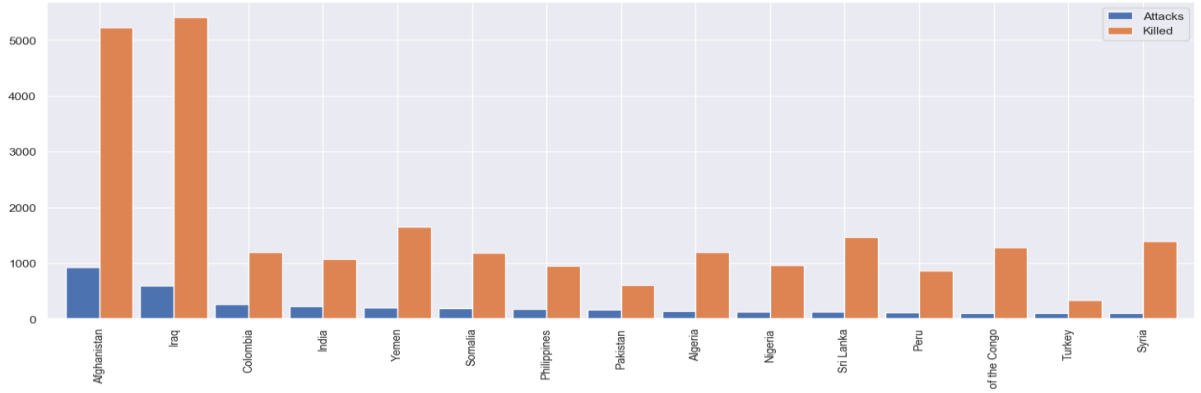


[그림 5-5] Group 4 의 Cluster 1 내 데이터의 국가별 공격 횟수 및 사망자 수 비교 그래프

두번째 군집은 총기류, 그리고 칼을 무기로 사용한 무장 공격, 암살, 납치에 해당되는 테러 활동 데이터 집합이다. 파키스탄, 인도, 이라크, 아프가니스탄 순서로 군집 내 데이터가 많이 나타난다.

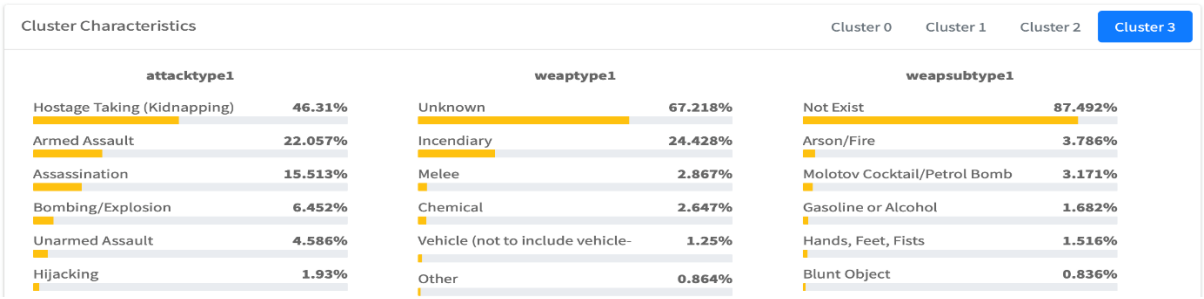


[그림 5-6] Group 4 의 Cluster 2 의 범주 분포도

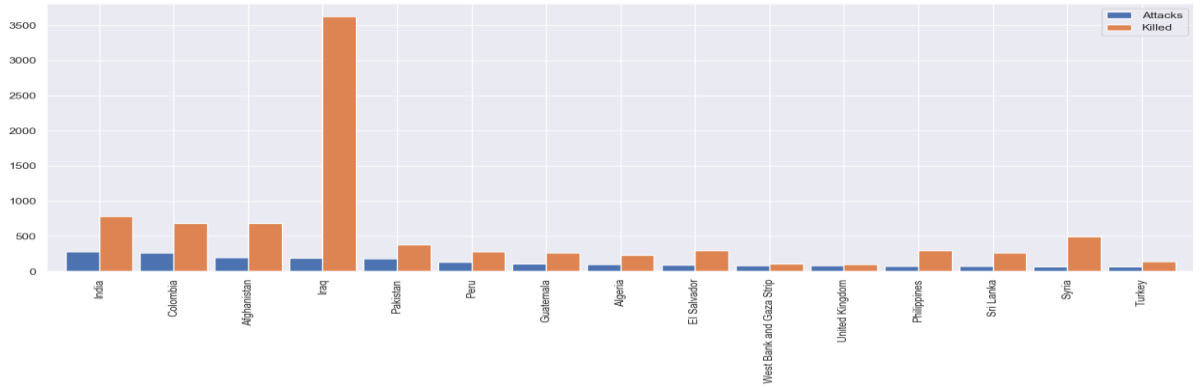


[그림 5-7] Group 4 의 Cluster 2 내 데이터의 국가별 공격 횟수 및 사망자 수 비교 그래프

세번째 군집은 소이탄, 세보타지 장비 등을 사용한 시설/인프라 공격과 알려지지 않은 공격 행위에 대한 테러 활동 데이터 집합이다. 이 군집은 전체 데이터 중 9%정도 차지하기 때문에 이전 장에서 계속 언급했던 해당되는 데이터가 적은 범주들이 모인 것이라 볼 수 있다. 클러스터 내 데이터가 적기 때문에 [그림 5-7]의 세로축을 보면 범위가 이전 그래프들보다 월등히 적다는 것을 알 수 있다. 전 클러스터와 유사하게 아프가니스탄, 이라크, 컬럼비아에서 이러한 공격 행위가 많이 발생했다.



[그림 5-8] Group 4 의 Cluster 3 의 범주 분포도

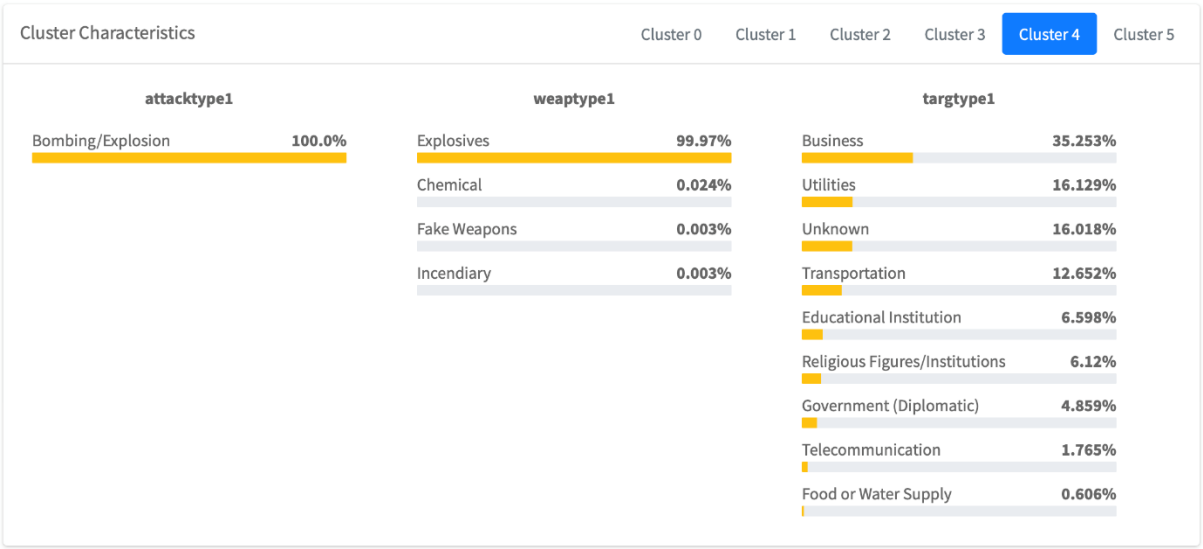


[그림 5-9] Group 4 의 Cluster 3 내 데이터의 국가별 공격 횟수 및 사망자 수 비교 그래프

마지막 군집은 전체 중 대략 6% 정도 차지하는 가장 작은 규모인데, 어디에도 속하지 못한 데이터들이 모인 듯한 느낌이다. 주먹싸움이나 소이탄을 사용한 난투(Melee)에서 발생한 납치, 무장 공격 및 암살과 관련된 테러 활동이 이 군집에 해당된다. 인도, 컬럼비아, 아프가니스탄 순서로 많이 발생했으나 파키스탄, 페루 등 공격횟수가 많이 차이가 나지 않는다. 이라크의 사망자 수가 그래프에서 유독 눈에 띄는데, 데이터를 별도로 분석해보면 사망자가 많이 발생하는 무장 공격 및 암살이 이라크에서 많이 발생했기 때문에 타 국가에 비해 사망자 수가 높게 나온 것이다. 무장 공격 행위는 두번째 군집과 유사하지만 사용한 무기가 총기류가 아니기 때문에 마지막 군집에 해당된 것으로 보인다.

다른 군집도 제작한 시각화 엔진을 통해 이와 유사하게 군집의 특징을 파악할 수 있었다. 그러나 위에서 설명했던 결과와는 다르게 그룹 내 각 군집의 특징을 모두 정확하게 정의를 내리기는 힘들었는데, 그 이유는 특정 범주가 편향되어 나타나 상대적으로 작은 비율로 분포하는 범주의 경우 모여 있어서 하나의 클러스터로 분류되기 때문이었다. 그리고 그런 작은 비율로 존재하는 범주가 많아서 클러스터의 개수를 계속 늘려서 해당 클러스터를 분할할 수는 있었으나, 전체 중 1%를 차지하는 데이터들의 모임이 어떤 특징을 가지는 클러스터로 보기는 힘들었다. 대표적인 변수가 targtype1 인데, 공격의 표적이 되는 대상들이 골고루 다양한 종류로 존재하다 보니, 범주의 크기가 작은 다른 변수(weaptype1, attacktype1)와 함께 다중 대응 분석을 할 경우 해당 변수의 범주 A 가 전체 중 50% 일 때 그 50%에 포함되는 targtype1 내 범주들이 모두 하나의 클러스터로 모이는 현상이 나타났다. 이 부분은 다시 군집 분석을 수행해서 분할해볼 수는 있겠지만, 특징을 찾기보다는 targtype1 으로 분류하는 것과 다를 바가 없을 것이라 판단하여 추가로 진행하지는 않았다.

아래는 이러한 상황 중 하나인 Group 10 의 Cluster 4 의 범주 분포도이다.



[그림 5-10] Group 10 의 Cluster 4 의 범주 분포도

이전 설명과 유사하게, targtype1 변수 내 범주 분포를 보면 비즈니스, 공공재, 교통 수단 등 다양하게 나타나고 있음을 알 수 있다.

이 외에도 2000 년대 이전 데이터들에 존재하는 결측값 때문에, 다른 변수들에 대해 군집 분석을 수행할 시 계수 좌표에서 아웃라이어(outlier)들이 많이 존재했다는 점이 해당 데이터셋이 군집 분석을 수행하는데 있어 한계를 보였다.

한편으로는 국가별로 군집이 다른 양상을 보인다는 점에서 국가별 테러활동의 특징을 살펴볼 수 있었다.