-비정상 이메일 특징 추출 및 송신 그룹 분류-

팀 : 동수야 일어나조

멘토 : 정좌연

팀원 : 조상혁, 문주혁, 조동수

목 차

서론

본 보고서는 1. 악성메일의 전반적 특징 추출과 2. 특징별 그룹화 후, 그룹별 특징과 공격패턴 도출 과정을 다루었으며, 이를 통해 케이스별 악성메일 예방법을 찾고자 함.

분석 방법으로는 1.탐색적 데이터 분석(EDA)를 통해 악성메일의 그룹을 나누는 기준과 그룹별 특징을 찾았으며, 2. 워드카운트와 lda 분석을 통해 그룹별 주된 주제를 도출하여 악성 메일의 목적을 파악하였음.

2. 분석환경

2.1 분석 시스템

i 분석에 사용된 시스템 정보

□ Hyper–v linux 환경

ㄴ 용도 : eml 파일의 csv파일로의 파싱 작업

□ 로컬 window 환경

ㄴ 용도 : 파싱된 csv파일의 분석 작업

ii 분석 환경

□ Jupyter Notebook

ㄴ 용도 : 데이터 분석 및 시각화

2.2 사용 언어

□ Python

3. 분석 준비

본격적인 분석에 앞서, 안전한 데이터 분석 작업을 위한 환경 세팅과 710440 개의 eml 파일들을 효율적으로 분석하기 위한 csv 파일로의 변환 작업이 이루어졌음

3.1 분석 환경 세팅

Hyper-v 을 통한 linux 환경을 세팅함. Linux 환경 안에서 스팸 메일 확인 작업이 이루어졌으며, eml 파일들을 하나의 csv 파일로 통합할 필요성을 느낌

3.2 데이터 준비 (eml 파일 파싱) 과정

a. eml 파일들의 헤더 정보 수집.

: 각각의 eml 파일들이 가지고 있는 헤더의 키값들이 달라, eml 파일들에서 등장하는 모든 헤더이름을 수집

b. 데이터 통합

: 수집한 헤더이름과 본문의 내용들을 열 값으로 가지는 csv 파일을 생성하고, 해당 csv 파일에 모든 eml 파일을 통합

c. 데이터 편집

: 헤더 정보 중 등장 빈도와 신뢰도[1]를 기준으로 MAIL_FROM(발신자), Received(수신정보), Date(발신일자)을 선별했으며, 이와 본문 정보를 가지는 csv 파일로 편집하였음.

d 데이터 확장

: 편집된 데이터에서 분석에 필요하다 생각되는 정보들을 파생시켜 분석에 사용할 csv 파일을 형성함

[1] 메일의 헤더들은 Received를 제외하고는 발신자가 쉽게 조작할 수 있음

[그림 3-1] 최종 데이터 테이블의 샘플 데이터

	file_name	MAIL_FROM	Received	Date	text_without_tag	include_url	num_of_imgs	Year	Month	Day	Time	length_of_text	main_language	Group	ip
0	059faa80-e6d3-47ff-8a45-edf0fba4e398.eml	juyoung@gmail.com	from [102.48.236.154] by 73.132.221.32 id <9543757-71176>; Mon, 16 Dec 2019 21:41:42 +0100	Mon, 16 Dec 2019 21:41:42 +0100	Content-Type: text/html;₩nContent-Transfer-Encoding: quoted-printable₩n₩n₩n₩n₩n₩n₩n안녕하세요. 회원님들~~~!₩n ₩n경기가 위축 되면서 모든게 불확실한 상황에 직면하고₩n있습니다.저 역시 하던 사업을 접고 이런 저런 상품에 기웃거리고 주식에도 ₩n손을 됐지만많은 손실을₩n보게 되었습니다.ㅠㅠ₩n ₩n그러다가 주위에 지인의 소개로 알게된 바른증권 방송₩n회원이 되면서원금회복하고 꾸준한 수익을 내고 있네요.₩n ₩n믿고 안믿고는 각자가 판단하시기 바라고요, 저도₩n힘든 시절에 도움 받아서도움을 드릴려고 정보공유합니다.₩n ₩n바른증권 방송은 확실한 전문가들로 구성되어₩n있고,업계 1위의 탄탄한 자금운영과저위험,고수익 전략, 돌발위험 대비 전략, 확실한 ₩n수익은 물론 환불보장제도를 시행하고 있습니다.₩n ₩n정보가 돈인₩n시대입니다.저 한테는 돈으로 따질수 없는 귀한 정보 공유합니다.₩n ₩n확인하는데 돈안들어₩n갑니다.>>> https://soo.gd/tBNK₩n ₩n ₩n₩n₩nyhmh mbdjxht ycvfm nzihcos	0	0	2019.0	Dec	Mon	21.0	533.0	ko	A	102.48.236.154

[표 3-1] 최종 데이터 프레임 컬럼 설명

컬럼명	내용	컬럼명	내용
File_name	eml 파일 명	* Month	발신일 기준 월
MAIL_FROM	발신자 아이디	* Day	발신일 기준 요일
Received	메일의 송신 경로, 시간	* Time	발신 기준 시간
Date	발신 기준 시간	* length_of_text	본문 길이
Text_without_tag	Html 태그를 제거한 본문 내용	*main_language	본문이 작성된 언어
* Include_url	url 포함 여부. 0 : 미포함, 1: 포함	* Group	아이디 기준 메일 보낸 횟수 A: 만번 이상 , B: 천번~만번 C: 백번~천번 , D: 백번 이하
* num_of_imgs	본문이 포함한 이미지의 개수	* ip	메일이 발신된 ip 주소
* Year	발신일 기준 연도		

(* 이 붙은 열이름은 기존 eml 파일 변수에서 파생시킨 파생변수를 의미함)

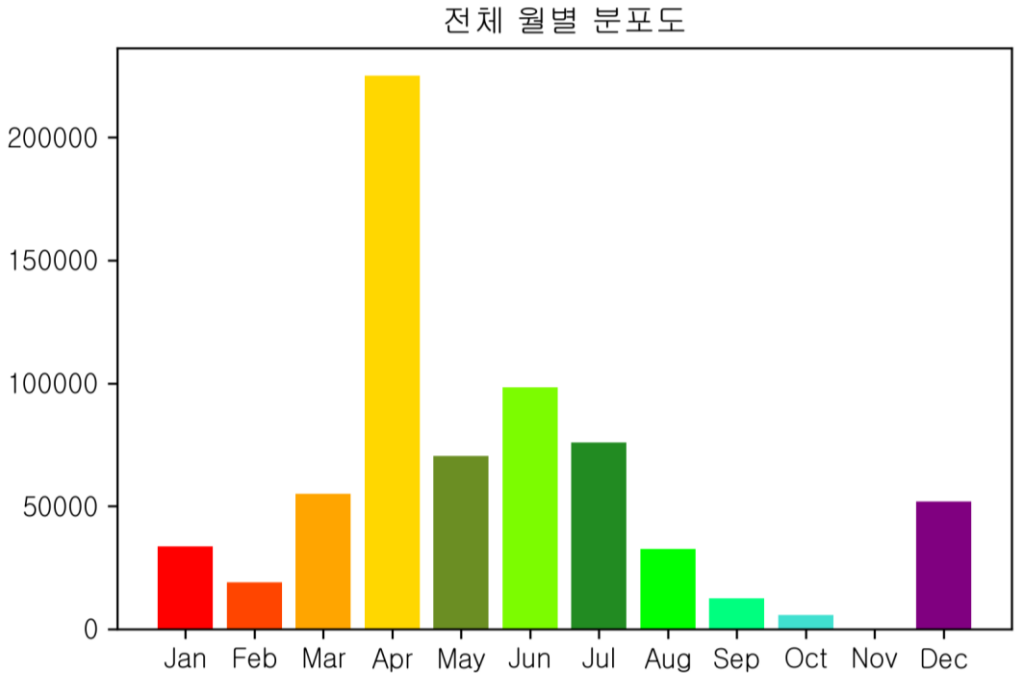
4. 비정상 메일의 속성 및 특징 분석

4.1. 전체 데이터에 대한 탐색적 데이터 분석 (EDA)

악성 메일 데이터의 전체적인 특징을 확인하기 위해 EDA 분석을 진행함. 새롭게 파생시킨 열 정보들을 기준으로 분석을 진행하였으며, 이를 통해 송신 그룹을 분류하고자 함.

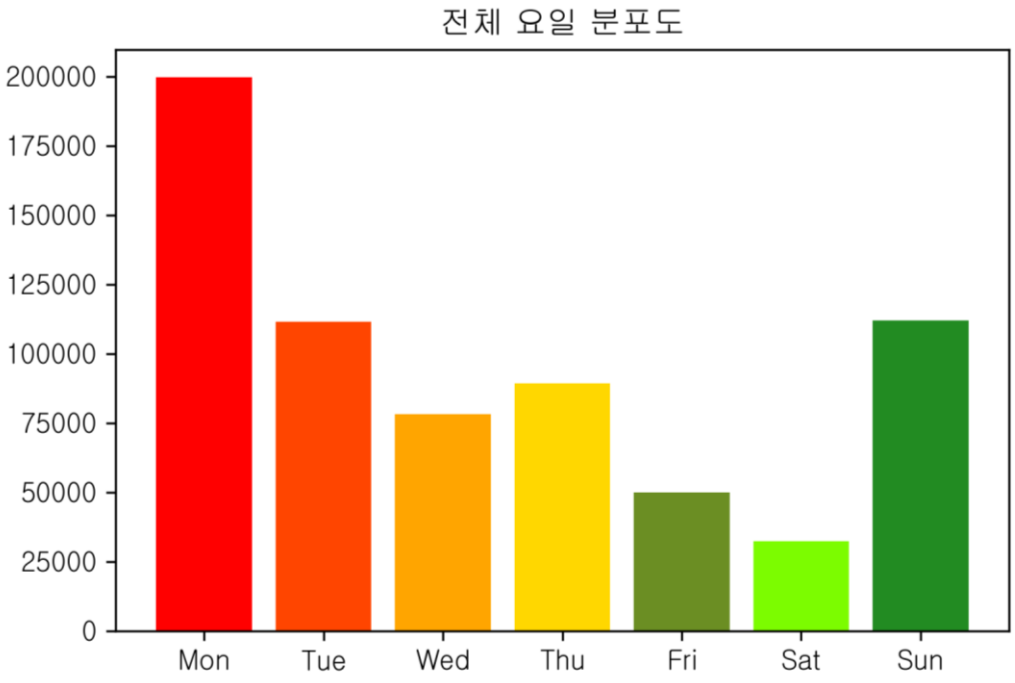
i 시간 정보 분석

[그림 4-1-1] 전체 데이터의 월별 분포도



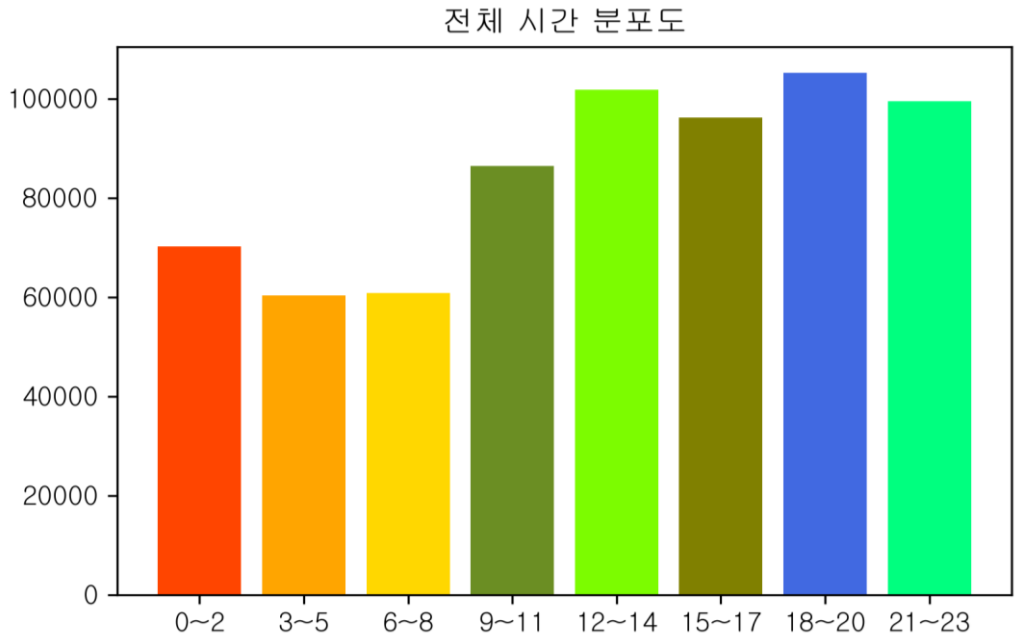
□ 본 스팸 메일들은 주로 4 월에 작성되었음

[그림 4-1-2] 전체 데이터의 요일 분포도



□ 악성 메일이 가장 활발히 작성된 요일은 월요일이며, 일주일의 후반부로 갈수록 발송이 줄어들다 일요일부터 다시 증가하는 경향을 띄고 있음

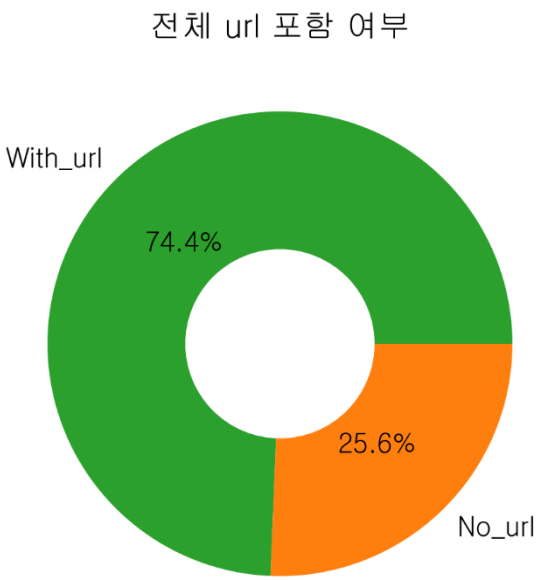
[그림 4-1-3] 전체 데이터의 시간 분포도



□ 악성 메일은 사람이 주로 활동하는 오전 9 시에서 자정까지 주로 발생되고 있으며, 새벽시간대에는 발송 건수가 줄어들었음.

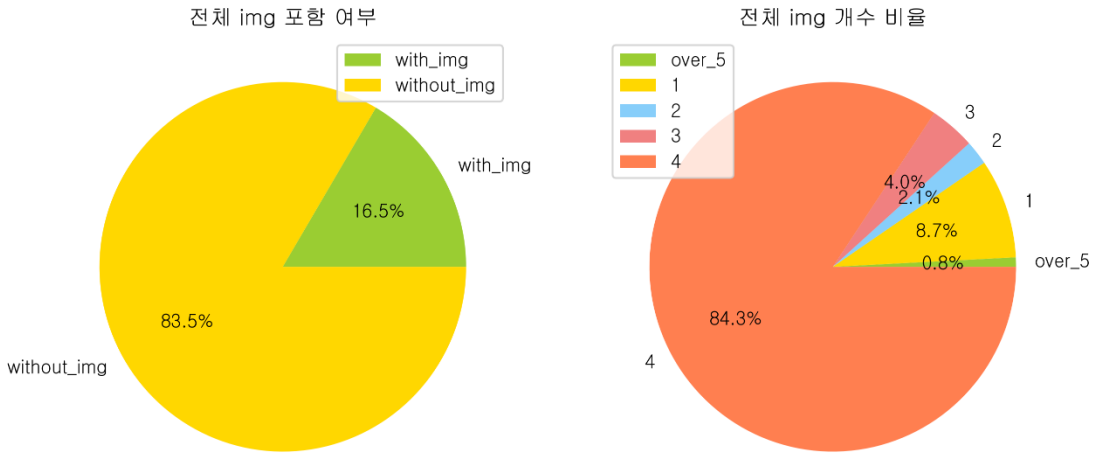
ii. 본문 정보 분석

[그림 4-1-4] 전체 데이터에서 url 링크를 포함 메일의 비율



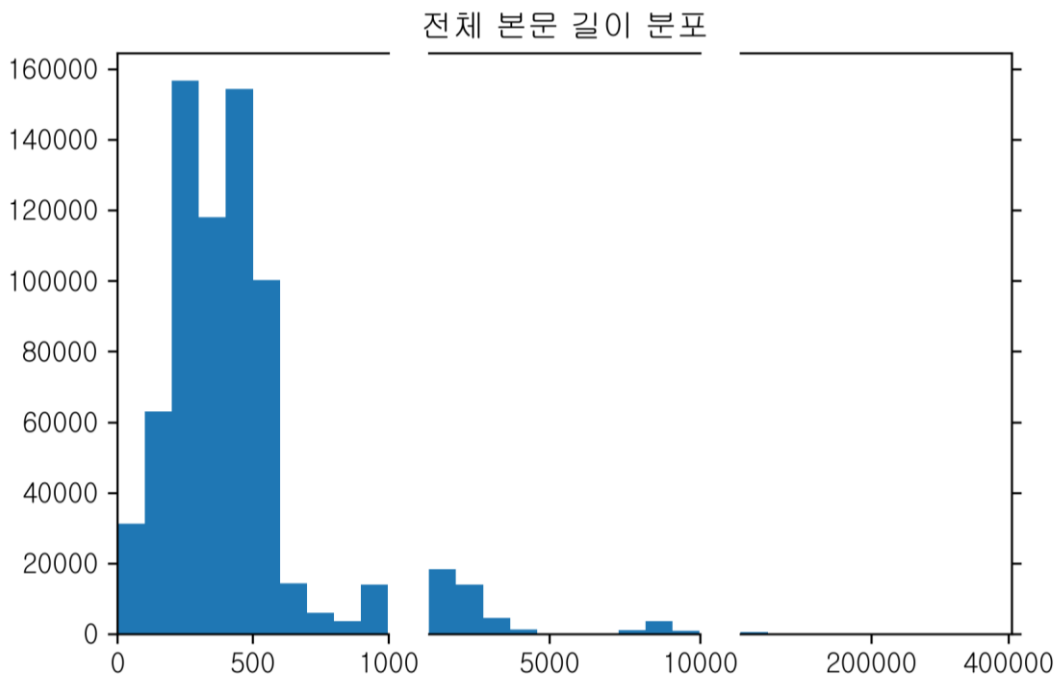
□ 전체 악성 메일 데이터에서 url 링크를 포함하고 있는 데이터는 75%를 차지하고 있음

[그림 4-1-5] 전체 데이터에서 이미지를 포함한 메일의 비율과 이미지 개수별 비율



□ 전체 스팸 메일에서 이미지를 포함하고 있는 메일의 비중은 16.5%에 불과했으며, 이미지를 포함한 메일들에선 이미지를 4 개 포함하고 있는 메일의 비중이 가장 높았음

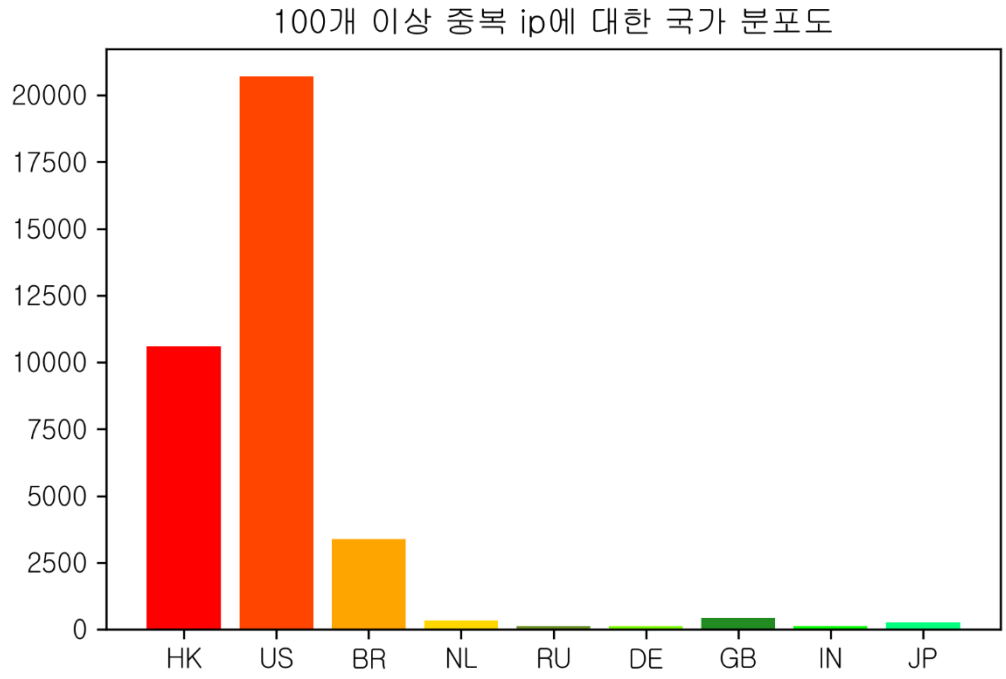
[그림 4-1-5] 전체 데이터의 본문 길이 분포도



□ 대부분의 악성 메일은 1000 자 이내로 작성되었으며, 10,000 자가 넘어가는 메일은 극소수 존재함.

iii. 발신 국가 분석

[그림 4-1-6] 중복 발신된 ip 주소에 대한 국가 추적



□ 동일 ip 주소로 100 건 이상 발송된 메일들에 대해서 발송된 국가정보를 추적하였음

□ 미국에서 발송된 악성메일이 압도적으로 많았으며, 홍콩과 브라질이 그 뒤를 따르고 있음

iv. 전체 데이터 분석 정보 요약

□ 악성 메일은 특정 기간 혹은 시기에 집중되어 발신되는 경향이 있음.

□ 대다수의 악성메일이 url 링크를 포함하고 있었으며, 이미지가 포함된 악성 메일의 경우, 이미지의 개수가 골고루 분포하지 않고 특정 개수(4 개)에 집중되어 있음.

□ 100 건 이상씩 반복적으로 악성 메일의 보내는 발신자의 경우 미국 혹은 홍콩에 집중되어 있음.

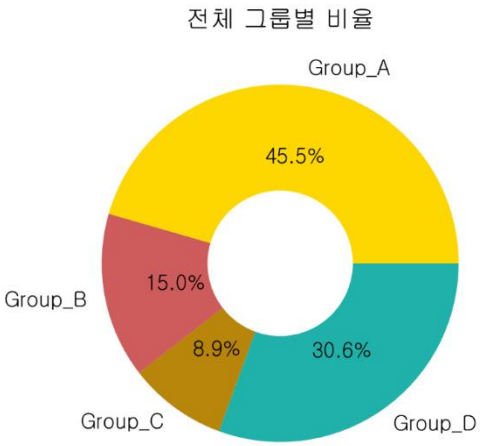
4.2. 송신 그룹 분류

ⅰ. 발신 아이디 기준 발신 빈도에 따른 송신 그룹 분류

동일 아이디로 발신된 메일들에 대해 동일 인물(혹은 단체)이 보냈다고 판단하여, 해당 아이디로 발성된 건수에 따라 그룹을 분류하였음. 그룹 A 는 만건 이상 보낸 아이디로 발송된 메일들이며, 그룹 B,C,D 는 각각 천~만 , 백~천, 백이하로 악성 메일을 발신한 아이디로 작성된 메일들임.

악성 메일의 발송 빈도는 발신자의 의도성에서 기인했다 생각할 수 있으며, 이에 발송 빈도에 따라 그룹을 나누어 분석하므로써 그룹별 특징과 의도를 파악하고자 함.

[그림 4-2-1] 메일을 보낸 횟수 기준으로 나누어진 4 개 그룹의 비율



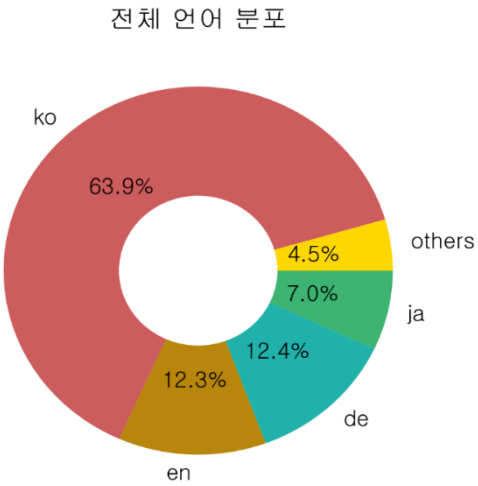
□ 악성 메일의 대략 50 퍼센트는 악성메일을 매우 빈번하게 보내는 집단에서 발송된 메일로 구성됨

□ 악성 메일을 100 건 이하로 소규모로 보내는 집단이 보낸 메일은 전체 악성 메일의 30 퍼센트를 차지하고 있음

ⅱ. 작성 언어에 따른 송신 그룹 분류

작성 언어에 따라 송신 그룹을 분류해 분석함으로써, 언어별 악성 메일의 특징과 발신 목적을 살피고자 함.

[그림 4-2-2] 전체 데이터의 언어 비율



□ 본 악성 메일 데이터는 주로 한국어로 구성되어 있음.

□ 영어와 독일어가 그 뒤로 각각 12%씩 차지하고 있으며, 그 외의 언어들이 11.5%를 구성하고 있음

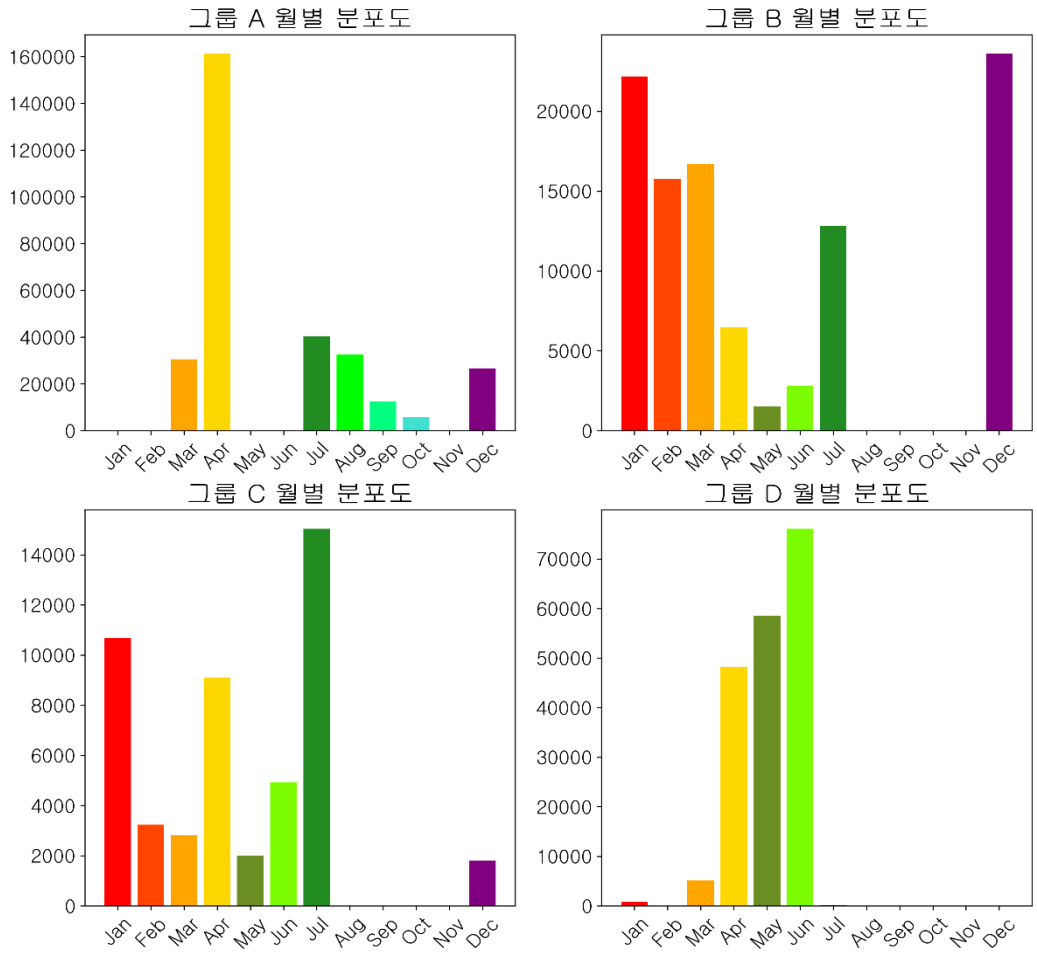
5. 송신 그룹 활동 분석

전체 데이터에 대해 진행했던 EDA 를 그룹별로 나누어서 진행함으로써, 전체 데이터의 특징을 그대로 따라가는 것이 아닌, 그룹별로 상이한 특징 패턴을 보이는지 확인하고자 함.

5.1 아이디의 발신 빈도에 따른 송신 그룹 분류

i 시간 정보 분석

[그림 5-1-1] 그룹별 데이터의 월별 분포도

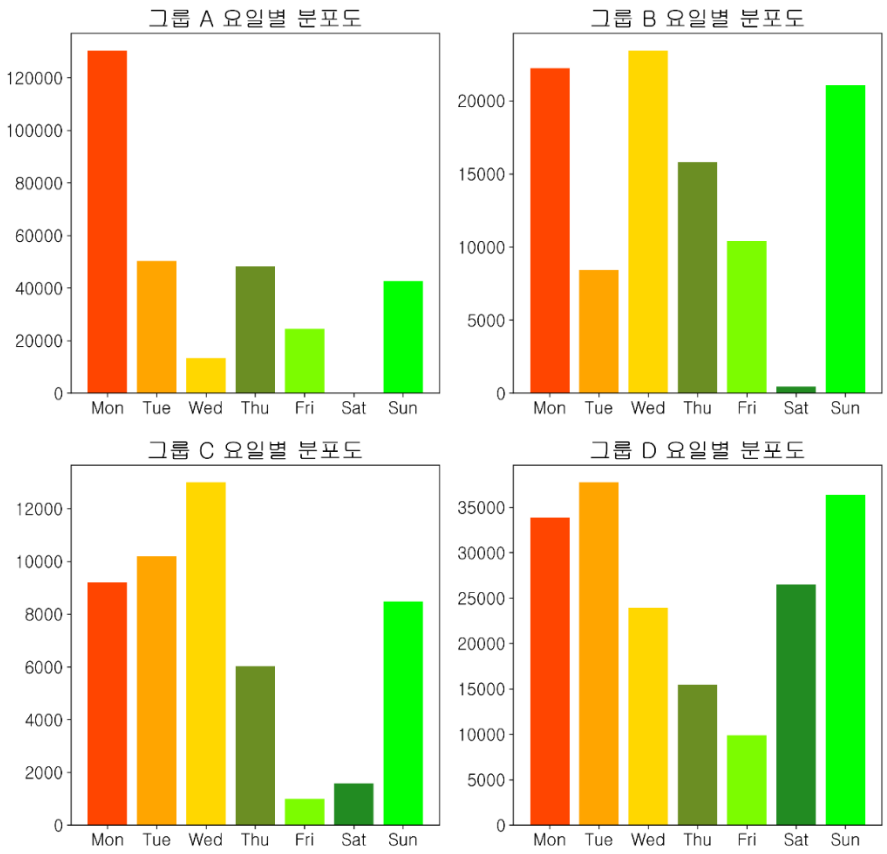


□ 그룹 A 는 악성 메일을 중복적으로 보내는 시기가 특정 월(4 월)에 집중되어 있음.

□ 그룹 D 의 경우, 상이한 아이디에 의해 발송되는 메일인 반면에, 발송 시기는 특정 기간(4,5,6 월)에 집중되어 있음.

□ 따라서, 그룹 D 의 경우도 특정 발신 주체의 존재가 의심됨

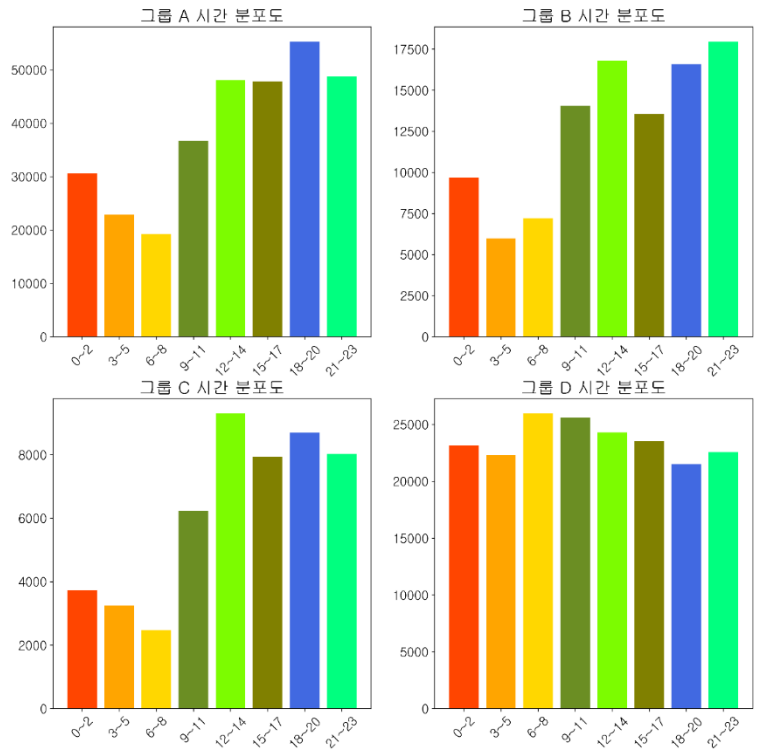
[그림 5-1-2] 그룹별 데이터의 요일 분포도



□ 그룹 A 의 경우, 발송 시기가 특정 월 뿐만 아니라 특정 요일(월요일)에 집중되어 있음

□ 그룹 B,C,D 의 경우 발송 시기가 비교적 분산되어 있으며, 주로 주초(일~수)에 많이 발송하는 경향이 있음

[그림 5-1-3] 그룹별 데이터의 시간 분포도

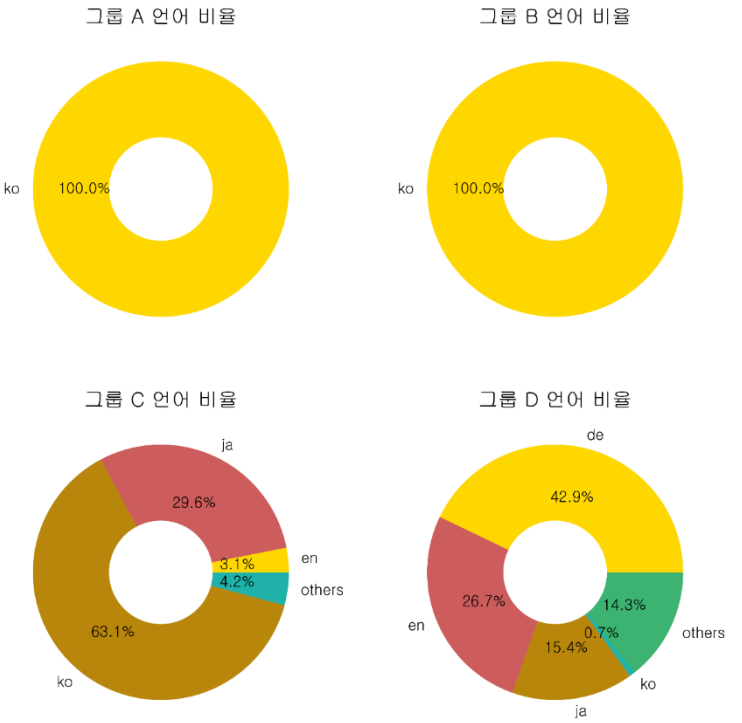


□ 그룹 A,B,C 의 경우 사람의 활동 시간인 오전 9 시~자정 사이에 발신 되었음.

□ 반면에 그룹 D 의 경우 발신 시간대가 골고루 분포되어 있음

ii. 본문 정보 분석

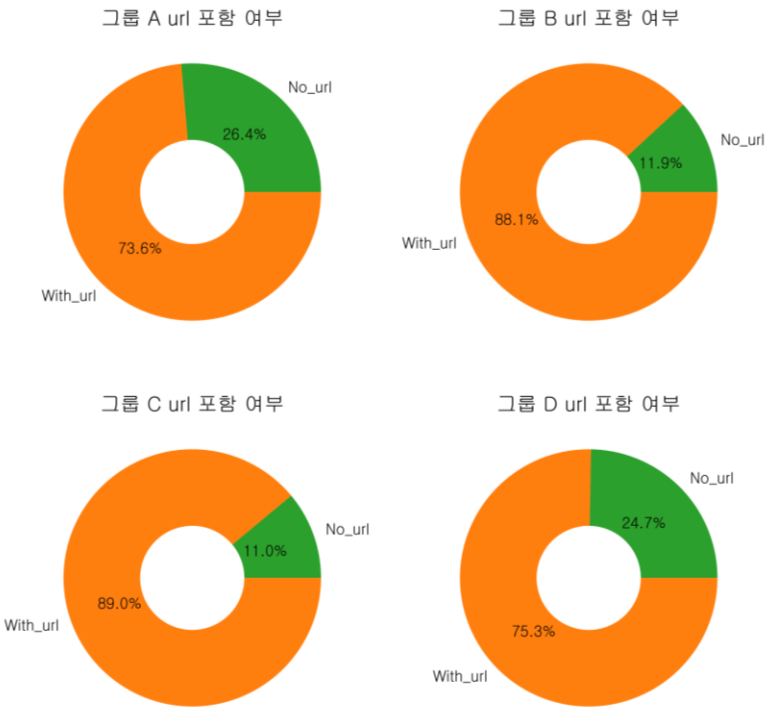
[그림 5-1-4] 그룹별 데이터의 언어비율



□ A, B 그룹의 경우 한국어 메일이 100%를 차지했으며, 그룹 C,D 의 경우 외국어의 비중이 높아짐

□ 특히 그룹 D 의 경우, 대부분의 메일이 외국어로 작성되었으며, 한국어 메일은 극소수에 불과했음.

[그림 5-1-5] 그룹별 데이터에서 url 링크를 포함한 메일의 비율



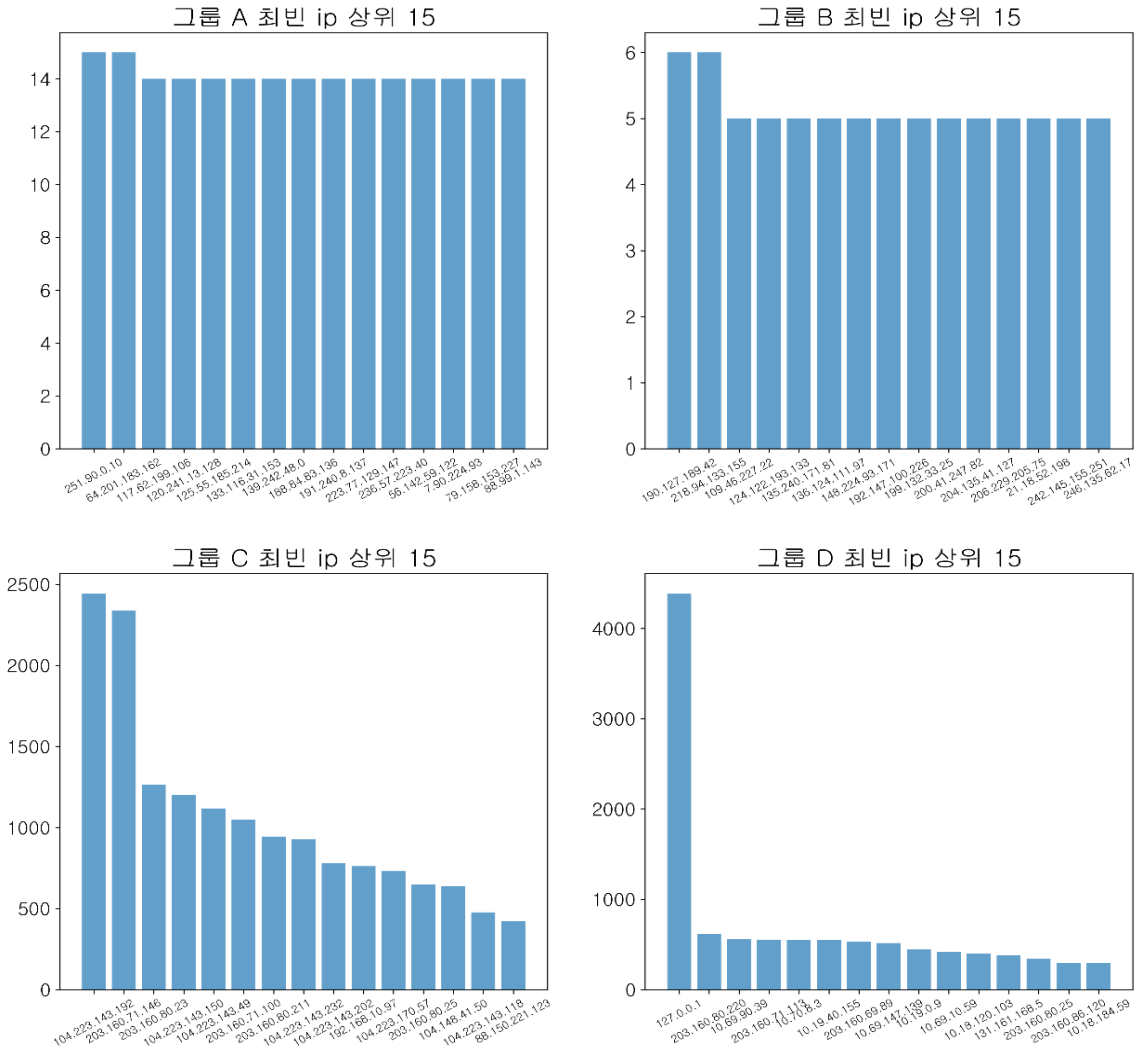
□ 모든 그룹에서 있어서 url 링크를 포함한 메일의 비율은 높게 나타나고 있음

[그림 5-1-6] 그룹별 이미지를 포함한 메일의 비율과 이미지 개수별 비율

□ 그룹 A,B,C 의 경우 거의 모든 메일이 이미지가 없는 메일이었던 반면, 그룹 D 의 경우 이미지가 있는 메일이 약 50 퍼센트였으며, 이미지의 개수도 보통 4 개로 구성됨.

iii. 그룹별 발신 정보 분석

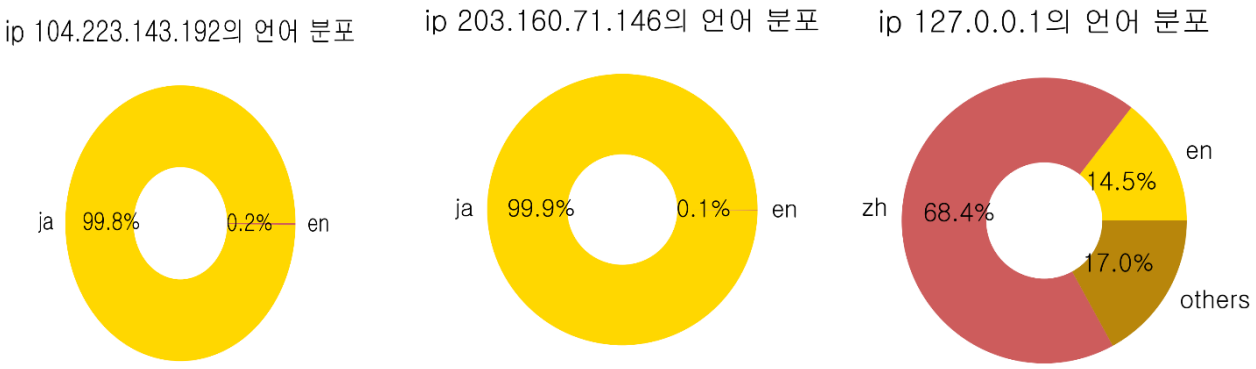
[그림 5-1-7] 그룹별 데이터의 최빈 ip 주소



□ 그룹별 최빈 ip 주소의 등장 횟수를 계산해본 결과, 동일 아이디로 다수의 악성메일을 발신한 그룹 A,B 는 여러 ip 주소들로 구성되어 있었으며, 그에 반해 아이디당 비교적 적은 수의 악성메일을 보낸 그룹 C,D 는 특정 ip 주소들에 집중되어 있음.

□ 따라서, 그룹 C,D 의 메일들은 여러 개인 혹은 단체에 의해 발송되었다는 기존의 가정은 틀렸으며, 그룹 C,D 의 메일들 또한 그룹 A,B 의 메일들과 마찬가지로 몇몇의 특정 집단에 의해 발송되었음을 확인할 수 있음.

[그림 5-1-8] 최빈 ip 주소 3 개에 대한 언어 비중



□ 또한, C,D 그룹의 최빈 ip 주소를 추적한 결과, 각 ip 주소는 주로 특정 언어(일본어, 중국어)의 글을 작성하였으며, 반복적인 내용의 악성메일을 배포하는 것을 확인할 수 있었음.

iv. 아이디의 발신 빈도에 따른 송신 그룹 분석 요약

□ 그룹 A,B,C,D 모두 주로 악성메일을 발신하는 주체가 있다 판단되며, 작성 언어에 따라 특징이 구분됨

□ 아이디의 중복이 많은 A,B 의 경우 모두 한국어로 작성되었으며, ip 주소의 중복이 많은 C,D 의 경우 특정 ip 를 추적한 결과 특정 외국어로 작성되는 경향이 있음.

□ 즉, 한국어 악성 메일의 경우 여러 ip 주소에서 소수의 아이디를 공유하며 발신되며, 외국어 악성 메일의 경우 소수의 ip 에서 여러 아이디를 통해 발신되는 경향이 있음.

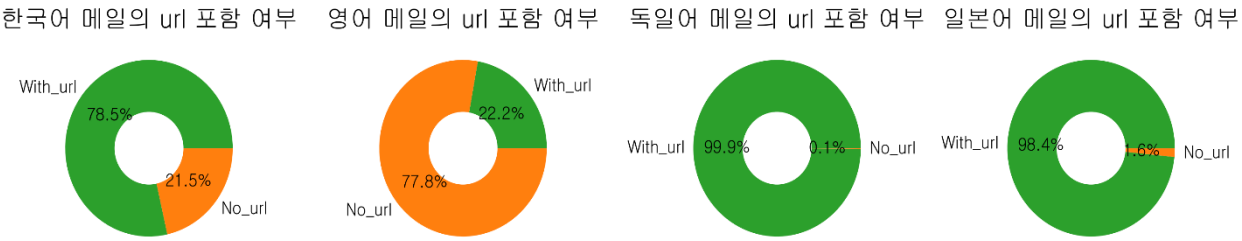
□ 그룹 D 의 경우 다른 그룹에 비해 이미지를 포함하는 메일이 많으며, 그 개수가 4 개에 집중되어 있다는 특징이 존재함.

5.2 작성 언어에 따른 송신 그룹 분류

기존의 아이디의 발신 빈도에 따른 그룹별 분석을 통해, 언어별 아이디의 중복 혹은 ip 주소의 중복이 나누어진다는 정보를 확인하였음. 이에 주요 언어별로 url 여부와 이미지 개수를 살펴봄으로써, 작성 언어별 악성 메일의 특징을 찾음. 또한 워드 카운트와 lda 모델을 통해 주요 등장 언어인 한국어, 영어, 독일어, 일본어에 대한 내용 분석을 진행함.

i. 주요 언어별 본문의 특징 분석

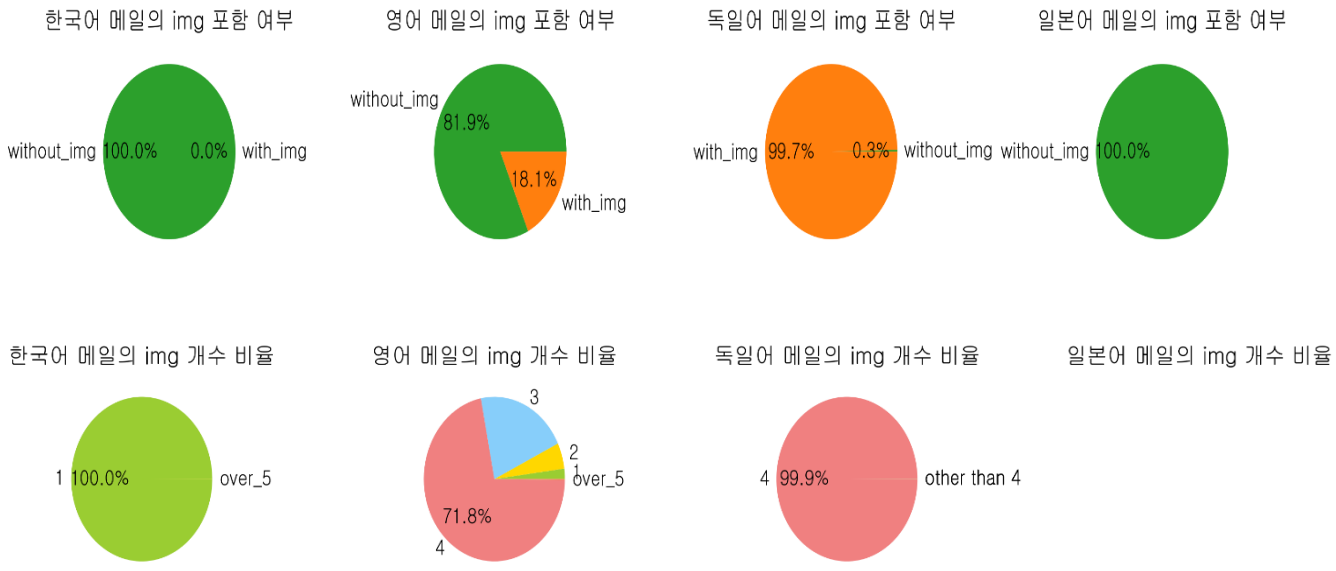
[그림 5-2-1] 언어별 url 링크 포함 여부 비율



□ 영어 메일을 제외한 언어들로 작성된 메일들은 주로 url 을 포함하고 있었으며, 특히 독일어와 일본어의 경우 거의 모든 메일이 url 링크를 포함하고 있음.

□ 한국어, 독일어, 일본어 메일의 경우, 특정 사이트로 유도하는 메일이 주를 이루는 반면, 영어 메일의 경우 단순 홍보 및 광고성 메일이 주 내용일 것이라 추정됨.

[그림 5-2-2] 언어별 이미지 포함 비율과 이미지 개수의 비율



□ 한국어와 일본어 메일의 경우 거의 모든 메일이 이미지를 포함하고 있지 않았으며, 영어 메일 또한 대부분의 메일이 이미지를 가지고 있지 않았음

□ 독일어 메일의 경우 거의 모든 메일이 이미지를 포함했으며, 포함된 이미지의 개수 또한 대부분 4 개였음.

ii. 주요 언어별 메일의 최빈 단어 분석

언어별 내용 분석을 위해 워드 카운트를 통한 언어별 최빈 단어를 분석하였음.

[그림 5-2-3] 그룹 A, B, C 의 한국어 워드 클라우드

□ 왼쪽부터 그룹 A, B, C 의 최빈 등장 단어들을 워드 클라우드로 만든 이미지

□ 공통적으로 '정보', '수익', '회원', '무료', '사업', '창업' 등의 단어들이 등장했으며, 이를 통해 수익이 필요한 사람들을 타겟으로 사람들을 유인하는 악성 메일이 작성되었음을 알 수 있음

□ 또한 '굴비', '마스크' 등 상품 광고하는 홍보성 메일도 다수 존재함.

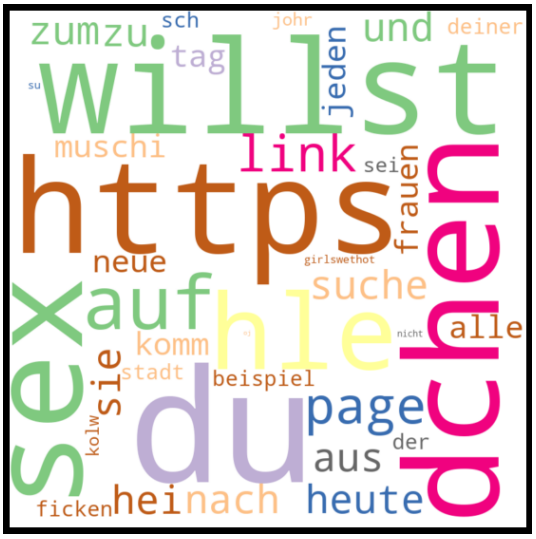
[그림 5-2-4] 그룹 C, D 의 영어 워드 클라우드

□ 왼쪽부터 그룹 C, D 의 최빈 등장 단어들을 워드 클라우드로 만든 이미지

□ 그룹 C 의 경우, 'claim' , 'account' , 'fund' , 'bank' 등의 단어들이 자주 등장했으며, 주로 대출 및 사칭을 통한 요금 청구를 요구하는 내용이 주로 등장하였음.

□ 그룹 D 의 경우, 자주 등장하는 women, woman, girls , Price 등의 단어들이 포함된 본문을 확인한 결과, 성인 사이트 혹은 성인 용품의 홍보가 다수 존재하였음

[그림 5-2-5] 독일어 워드 클라우드



□ willst(당신) , du(당신) , heute(오늘), frauen(여자들) , sex 등 성인 사이트를 연상하는 단어들과, https 와 같이 url 링크를 나타내는 단어들이 주요 키워드를 이루고 있음.

[그림 5-2-6] 일본어 워드 클라우드

□ モ ザ イ ク (모자이크) 、 リ メ イ ク (리메이크)、版(영상에서의 판)、 サ イ ト (사이트)、 マ ン コ (여성의 성기) 、動画(동영상) 이외에도 여성의 신체 부위를 상징하는 단어들이 많이 등장하는 것으로 보아, 일본어 악성 메일에는 성인 사이트 관련 내용이 있는 것으로 보임

iii. lda[2] 모델을 통한 주요 언어별 메일의 주요 토픽 분석

lda 모델을 통해, 워드 클라우드를 통해 살펴본 언어별 문맥의 특징을 확인하고자 함. 진행 과정은 다음과 같음

a 불용어 제거와 명사 추출을 통해 토큰화 진행

b Topic Coherence 라이브러리를 통해 최적의 토픽 군집수를 파악

c 최적의 토픽 수를 입력하여 모델이 모든 문서에 대해 문서의 보유 단어들과 비교하며 토픽을 할당함

d 유사한 토픽을 중심으로 단어 군집화 진행

e 군집화된 단어들을 통해 주요 토픽들을 유추함.

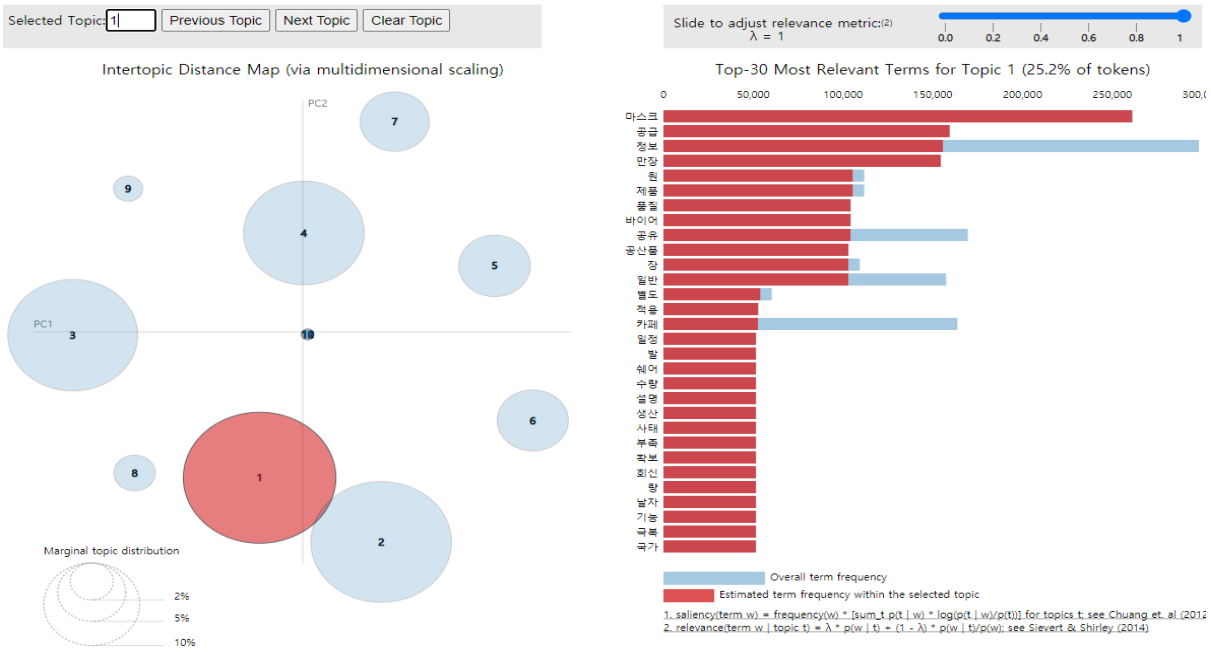
[2] Latent Dirichlet Allocation. 텍스트 마이닝 기법 중 하나로 통계적 추론을 통해 문서들의 주제(토픽)를 추출함.

lda 그래프에서 좌측의 원들은 각 토픽들을 나타며, 우측의 단어 그래프에는 Topic 생성에 사용된 단어들이 나타남.

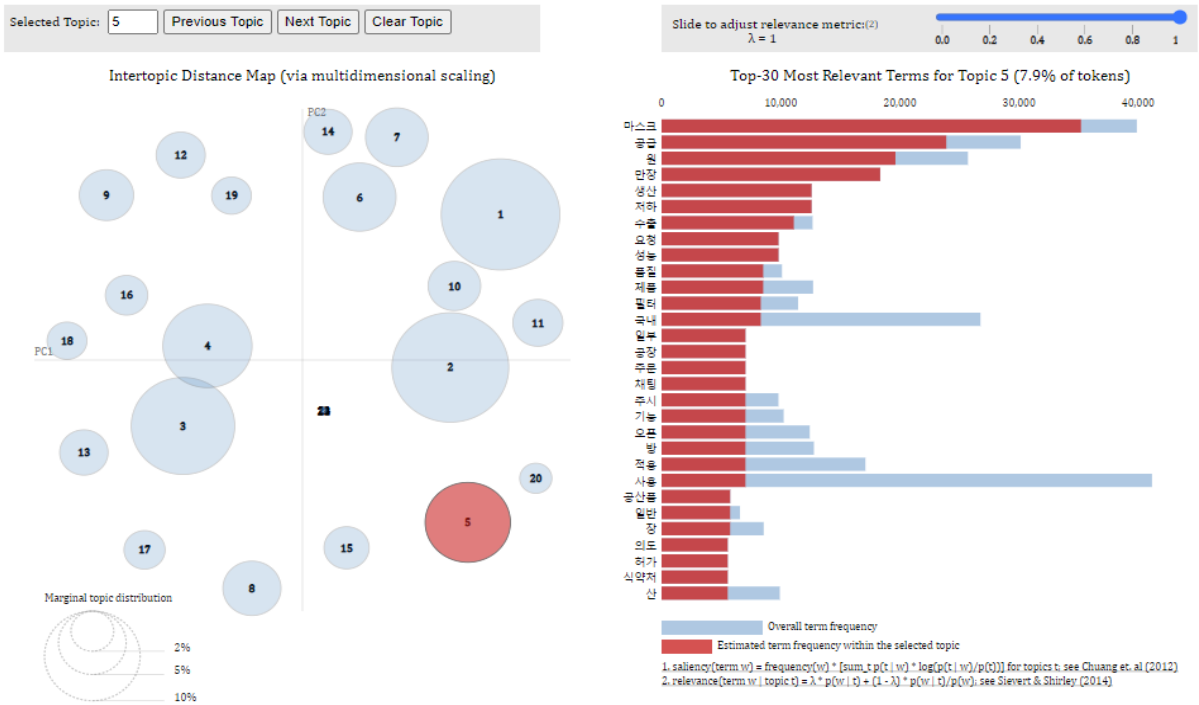
1) 한국어 이메일 LDA 분석

[그림 5-2-7] 그룹 A,B,C 의 한국어 토픽 1 : 상품 광고

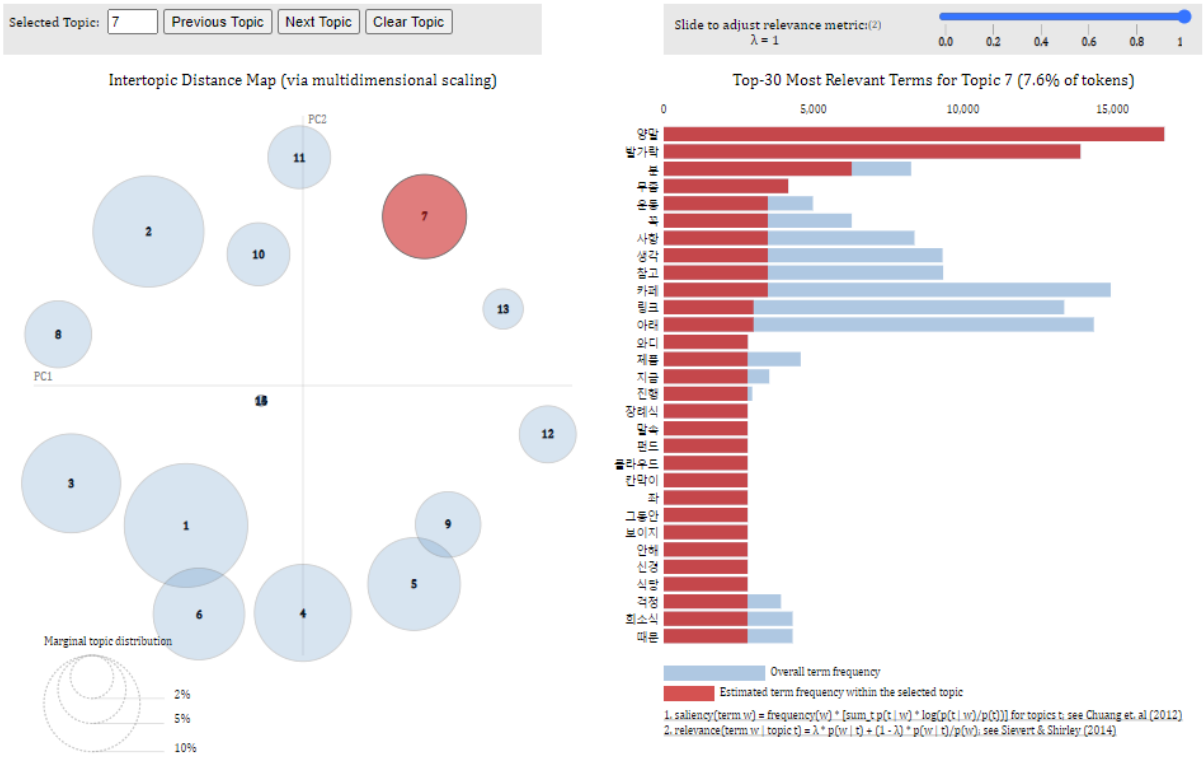
- 그룹 A -



- 그룹 B -



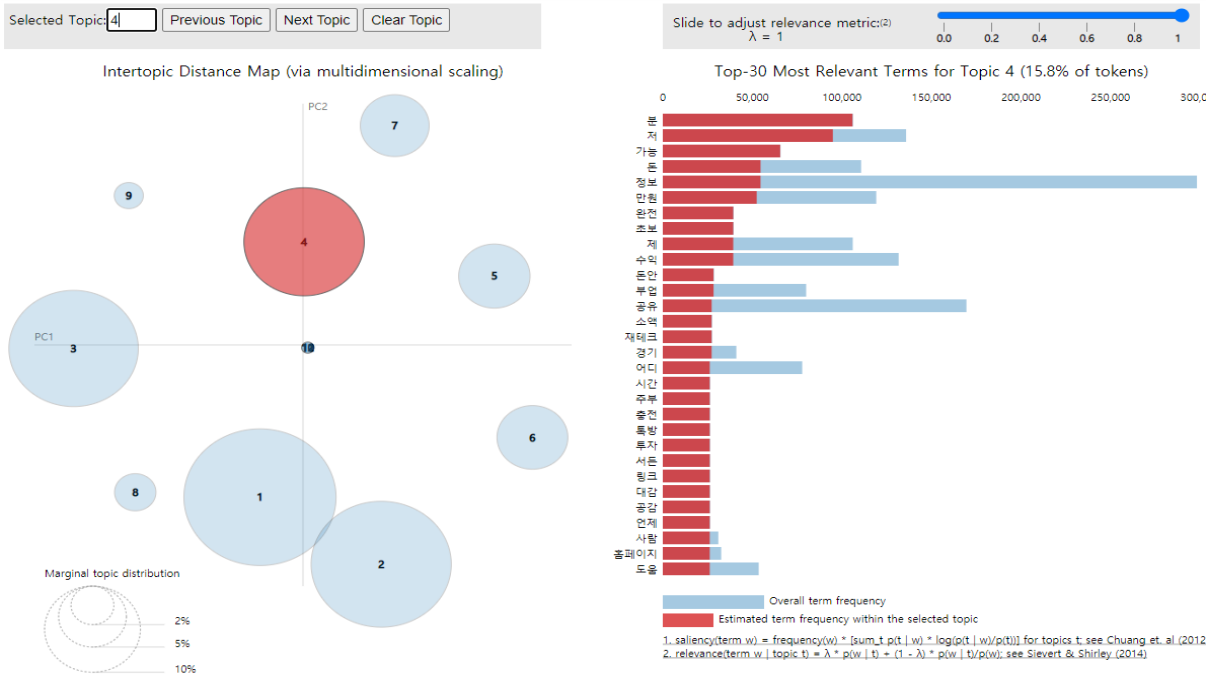
- 그룹 C –



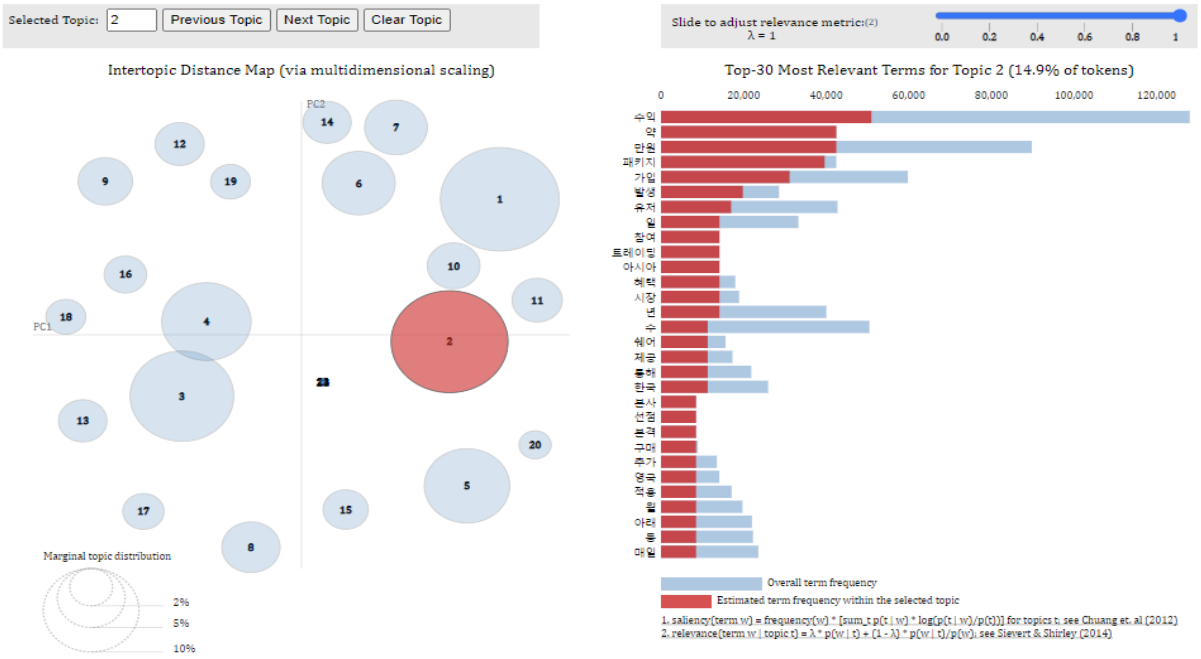
□ 그룹 A,B,C 모두 마스크, 양말, 제품 등 특정 상품 광고를 나타내는 단어들의 군집들이 존재했으며 이를 통해, 상품을 홍보하는 토픽이 존재함을 유추함.

[그림 5-2-8] 그룹 A,B,C 의 한국어 토픽 2 : 수익 상출

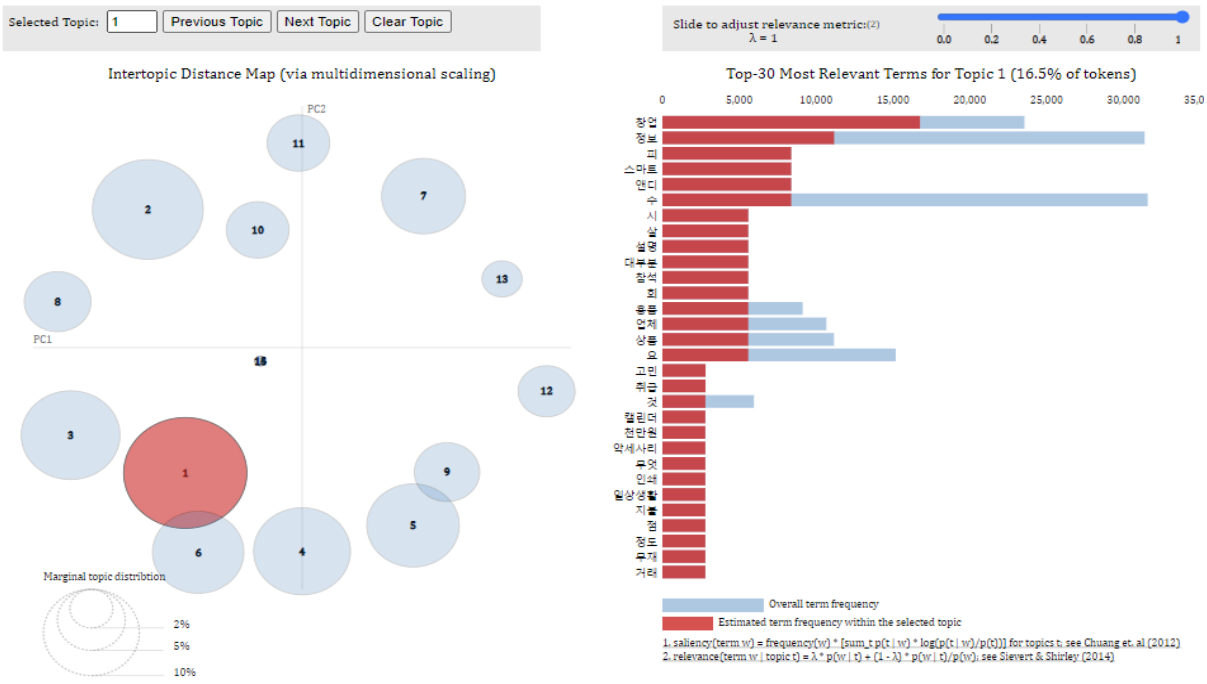
- 그룹 A -



- 그룹 B –



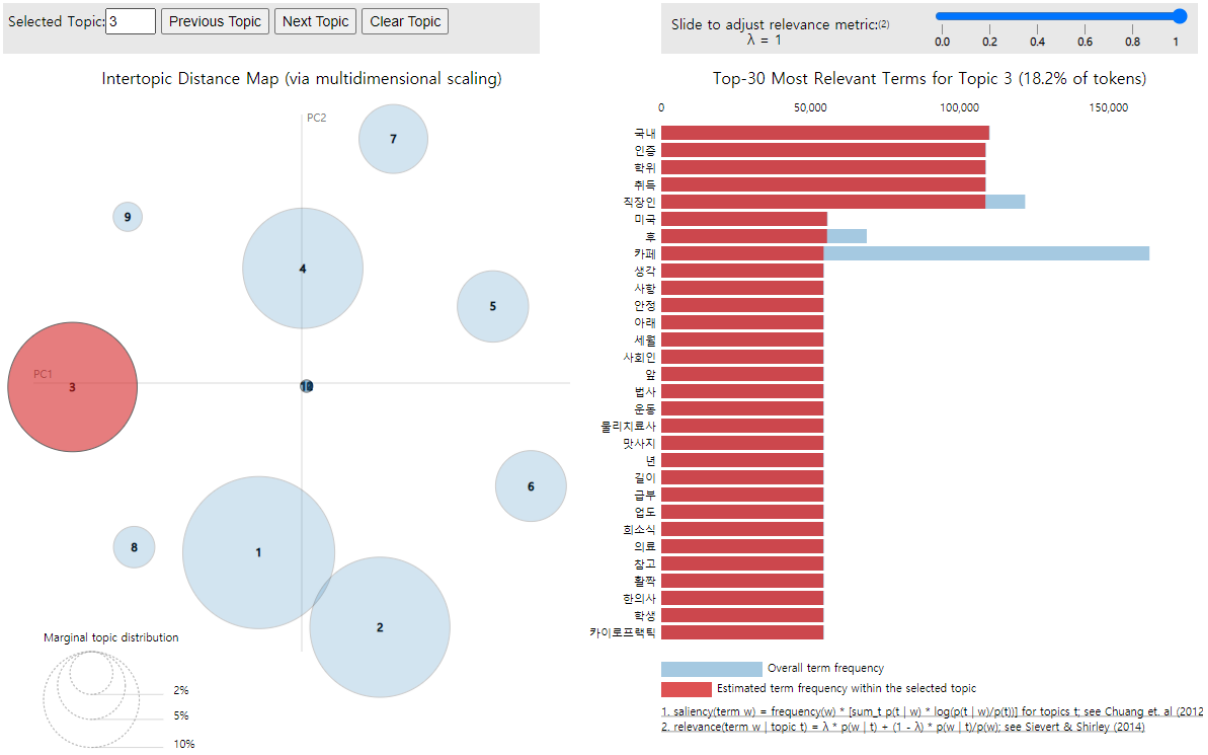
- 그룹 C-



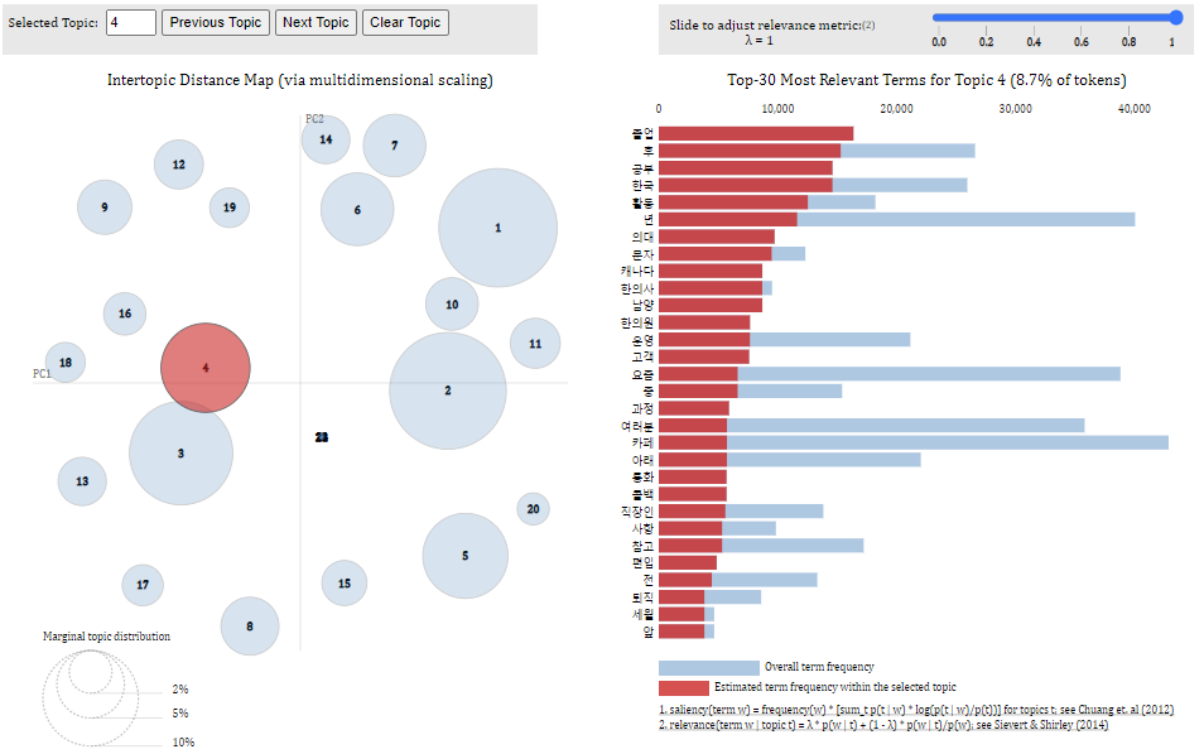
□ 돈, 수익, 패키지, 시장, 창업, 부업 등 수익 창출의 방법을 홍보하는 토픽으로 유추 가능한 단어들의 군집이 공통적으로 나타남

[그림 5-2-9] 그룹 A,B,C 의 한국어 토픽 3 : 학위 취득 유도

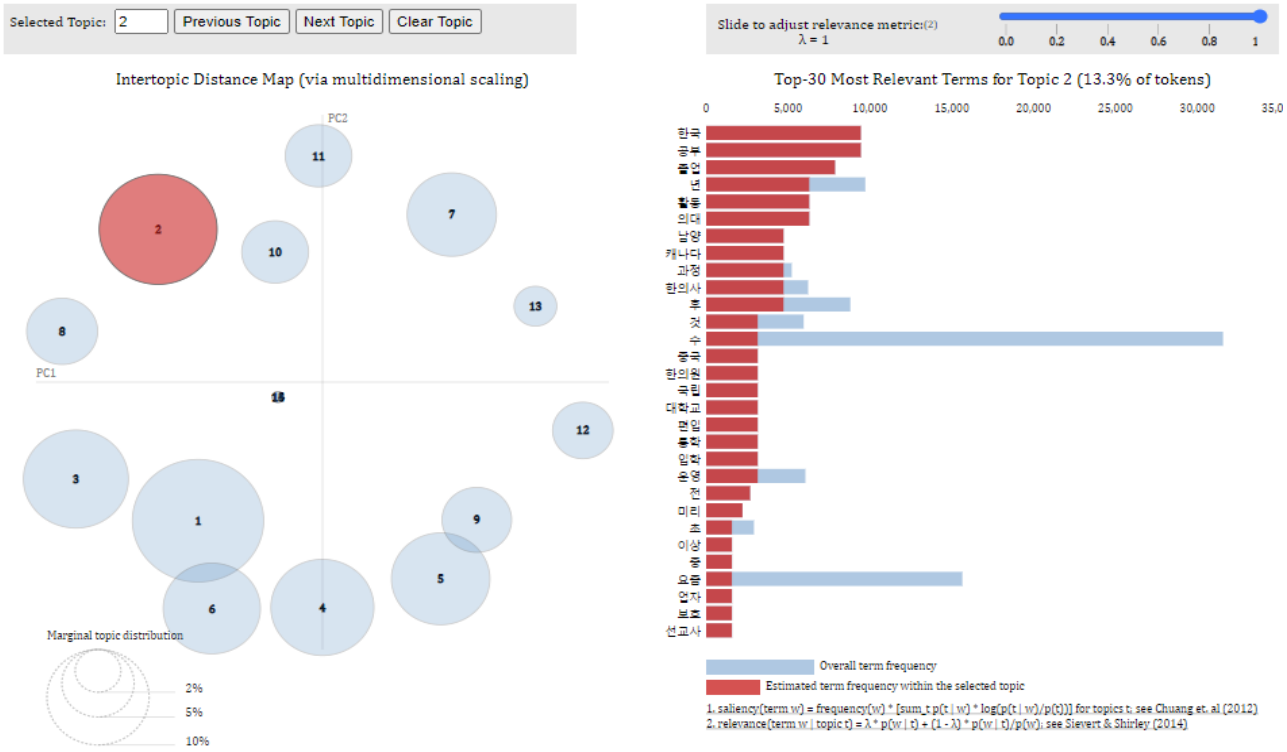
- 그룹 A -



- 그룹 B -



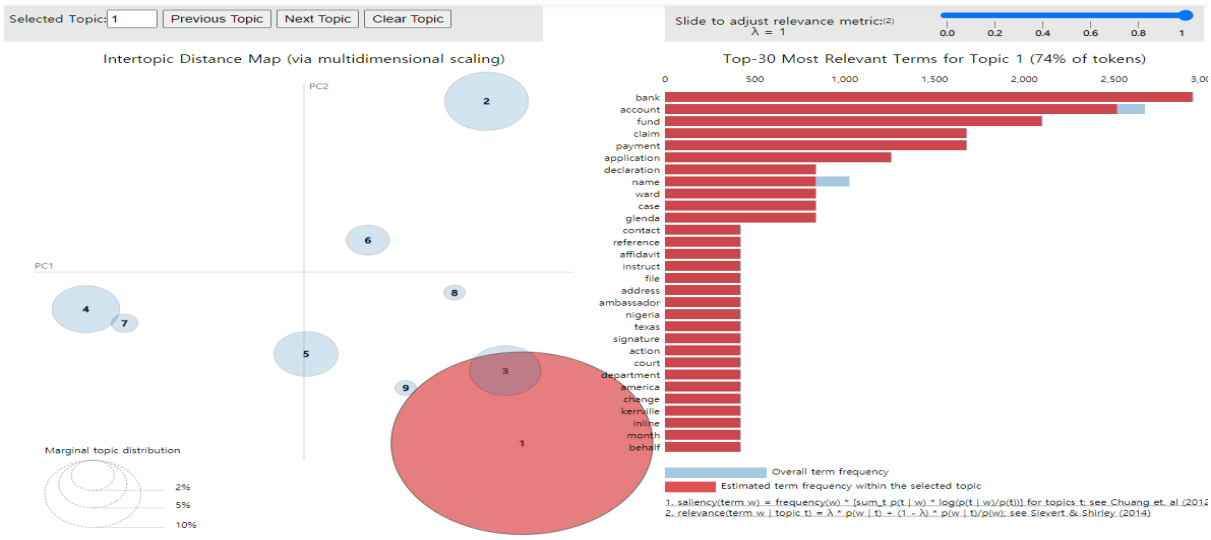
- 그룹 C –



□ 학위, 졸업, 공부, 한국 및 외국 나라 이름들 등의 단어들을 통해 자격증과 학위 취득을 유도 및 홍보하는 토픽의 글들이 존재함을 알 수 있음

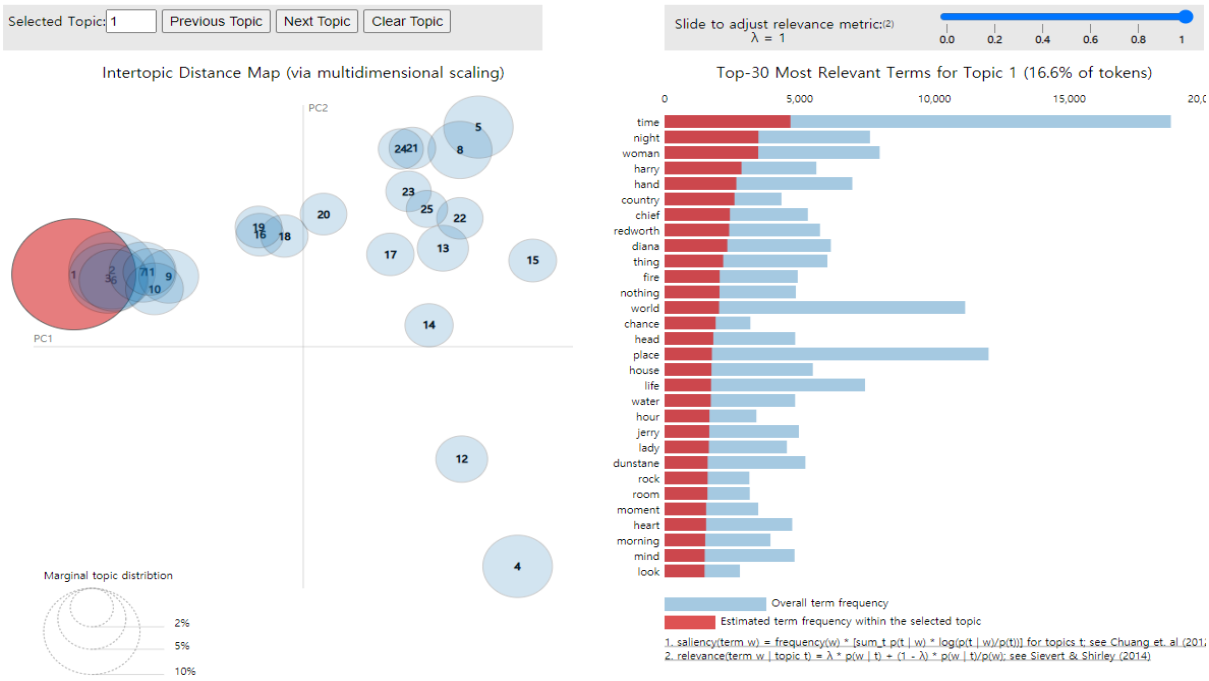
2) 영어 이메일 LDA 분석

[그림 5-2-10] 그룹 C 의 영어 토픽 : 금융



□ 그룹 C 의 영어메일의 경우 bank, account, fund, acclaim, payment 등이 커다란 군집을 이루었으며, 이를 통해 금융 관련 대출 홍보 및 거래 요구가 주요 토픽을 구성했음을 알 수 있음.

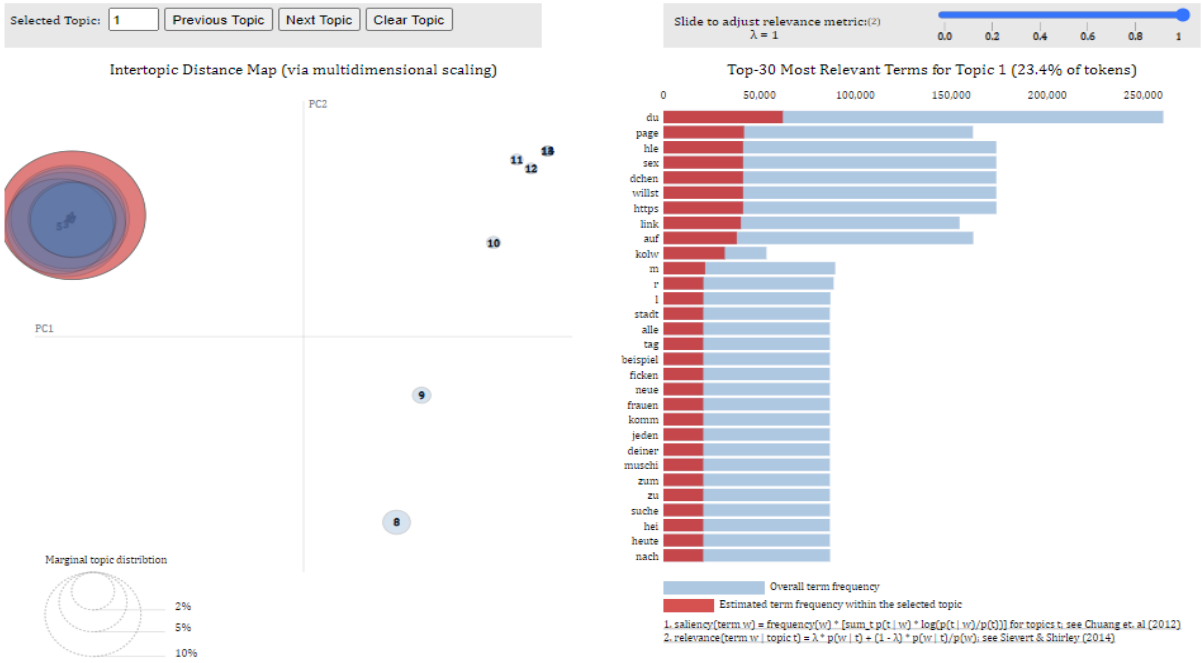
[그림 5-2-11] 그룹 D의 영어 토픽 : 성인 관련 상품 홍보



□ 그룹 D의 영어메일의 경우, woman, night, world 등의 단어들이 군집을 이루고 있으며, 이를 통해 워드 클라우드에서 살펴본 성인 사이트 및 성인 관련 상품이 주요 토픽을 이루고 있음을 확인함.

3) 독일어 이메일 LDA 분석

[그림 5-2-12] 독일어 메일 토픽 : 성인 사이트

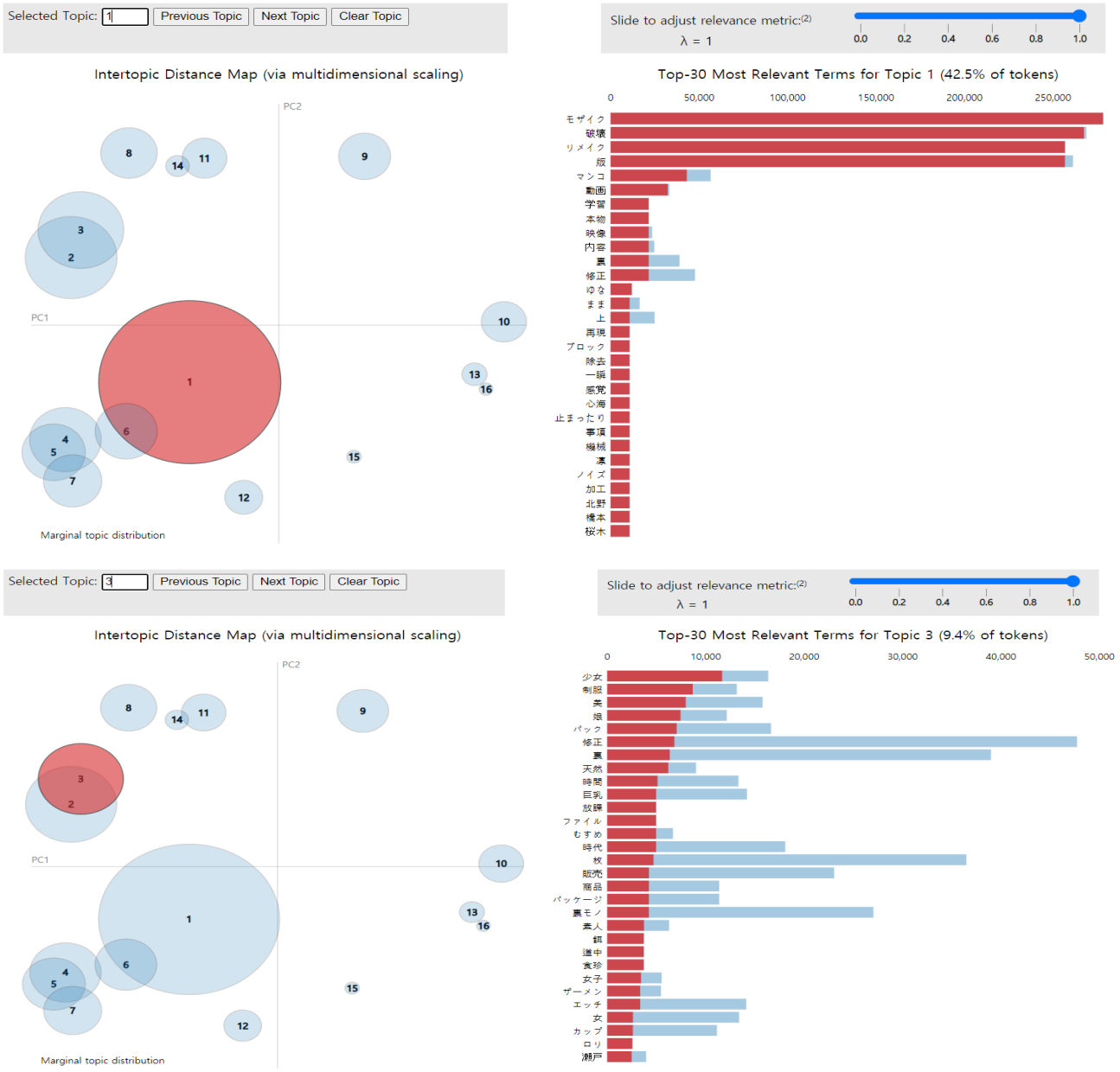


□ 단어들의 군집이 한곳에 집중되어 있으며, 이를 통해 독일어 메일의 경우 한가지 토픽의 메일들이 발송되고 있음을 알 수 있음.

□ 군집의 단어들을 살펴보면, 당신, 섹스, 원하다, 링크, 사이트 등의 단어들이 주로 있으며, 이를 통해 성인 사이트로 유도하는 메일들임을 알 수 있음.

4) 일본어 이메일 LDA 분석

[그림 5-2-13] 일본어 메일 토픽 : 성인 사이트



□ 워드 클라우드에서 살펴본 성인 사이트 관련 단어들이 커다란 군집을 이루고 있었으며, 다른 단어들의 군집 또한 성인 동영상을 지칭하는 단어로 구성됨.

□ 이를 통해, 일본어 악성 메일의 주된 주제는 성인 사이트 광고임을 알 수 있음.

ⅳ. 송신 그룹별 위험도 분류

송신 그룹의 특징에 따라 위험도를 분류하면 다음과 같이 나눌 수 있음.

위험도 上 : 독일어, 일본어 악성 메일

- 99% 정도가 url 를 포함했으며, 성인 사이트 내용이 주를 이룸

위험도 中 : 영어 악성 메일

- url 을 포함한 메일은 대략 20%에 불과했지만, 관련 내용이 성인 사이트와 금융 관련 사기성 메일이 주를 이룸

위험도 下 : 한국어 악성 메일

- 대략 80%의 메일이 url 을 포함하였지만, 관련 내용은 상품 홍보, 수익 상품 권장, 학위 유도, 등 광고성 메일이 주를 이룸.

Ⅴ. 작성 언어에 따른 송신 그룹 분석 요약

□ 한국어 메일의 경우, 제품 홍보, 수익을 미끼로한 유인, 학위 등이 주요 주제를 이루었으며, 본문에 url 링크를 포함함으로써 특정 사이트로의 방문을 유도하는 경향을 나타냄.

□ 영어 메일의 경우, 금융과 성인 관련 상품의 광고가 주로 있었으며, 이미지와 url 링크가 없는 단순 홍보성 메일이 다수 존재했다는 특징을 보임.

□ 독일어와 일본어 메일의 경우, 거의 모든 이메일이 성인 사이트 광고였으며, 해당 메일들은 url 링크를 넣어, 수신자를 성인 사이트로 유도하고 있음.

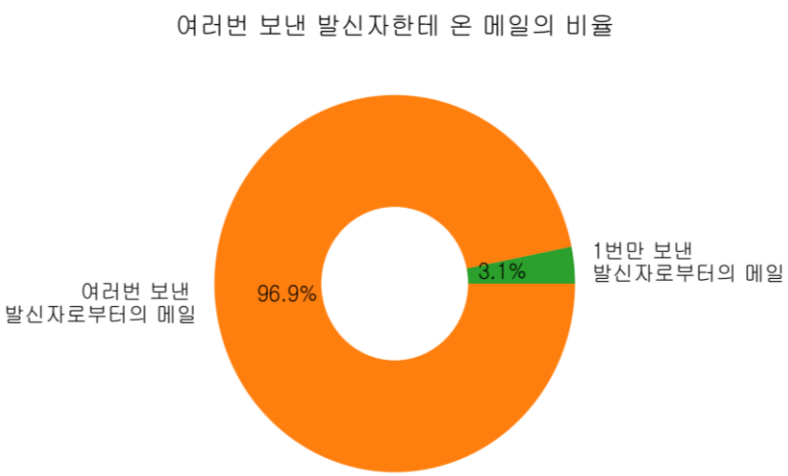
6. 송신 그룹별 이상 탐지 및 예방

아이디의 발신 빈도에 따른 송신 그룹 분석을 진행한 결과, 사실 그룹별 특정한 발송 주체 유무의 차이는 존재하지 않았으며, 아이디의 발신 빈도의 차이는 언어별 차이에 기인함을 발견하였음. 즉, 악성 메일의 송신 그룹은 작성 언어에 따라 분류하여 대응하는 것이 가장 효과적임을 알 수 있음.

우선 언어별 송신 그룹의 공통점으로는 모든 언어는 악성 메일을 주로 발송하는 특정 주체가 존재했으며, 그 발신 시기 또한 특정 시기와 기간에 집중되어 있음. 특히, 발신자 아이디와 ip 주소를 기준으로, 한 건만 보낸 발신자로부터 온 악성 메일의 비율은 전체 악성 메일에서 단,

3.1%에 불과함. 즉, 악성 메일의 송신 그룹별 특징에 따라 악성 메일의 발신자를 추적할 수 있다면, 대부분의 악성 메일을 막는 것이 가능함.

[그림 6-1-1] 여러 번 보낸 발신자한테 온 메일의 비율



언어별 송신 그룹의 차이점과 특징은 다음과 같음.

우선 한국어의 경우, ip 주소보다 발신자 아이디를 통해, 발신 주체를 추적하는 것이 효과적임. 특히, 메일의 내용이 특정 상품, 수익 창출, 학위 관련 단어들을 존재하며, url 링크를 가지고 있을 경우 악성 메일일 확률이 높다 판단됨.

영어 메일의 경우, ip 주소 추적을 통해 발신자를 추적하는 것이 효과적임. 메일의 내용이 금융과 성인 관련 단어들이 존재한다면 악성 메일이라 판단됨.

독일어와 일본어 메일의 경우, ip 주소 추적을 통해 발신자를 특징지어야 하며, 거의 99%에 달하는 악성 메일이 url 링크 혹은 이미지를 포함했으며 내용 또한 성인 사이트 관련 단어들로 구성됨. 특히, 독일어 메일의 경우 99%가 이미지를 포함하고 있음.

기타 언어로 작성된 메일의 경우, 특정 ip 주소에 발신자가 집중되어있는 특징을 가지고 있음.

따라서, 수신 메일을 언어로 분류하고, 각 언어별 악성 메일의 특징에 해당되는 메일들은 악성 의심 메일로 구분하는 과정이 필요함. 또한 악성 메일로 판단된 메일의 발신자의 경우, 발신자 정보를 기록해야 함. 추후 해당 발신자를 통한 메일을 차단할 경우, 단기간 내의 악성 메일 공격을 방어할 수 있음.

[붙임] 참고문헌

[1] Topic Modeling with Gensim[Python] , ML+ , 2020년 10월 28일 접속 ,

https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/