김종윤
2023-28318
협동과정 인공지능 전공

소셜컴퓨팅
과제 2 - 3주차

1. 선택한 키워드
“애플페이”, “OpenAI”

2. 한글 stemming과 stopwords 처리
우선, 선택한 키워드 중에 “애플페이”는 한글로 되어 있으며, 모든 트윗이 한글로 작성되어있음을 지난주 과제에서 밝혔기 때문에 한글 stemming과 stopword처리가 따로 필요하다.

두 키워드의 언어 차이로 인해 한글, 영어 모두 분석할 수 있도록 `Tokenizer`라는 클래스를 작성했다. tokenize 메소드는 입력으로 들어온 문자열을 (영어: `word_tokenize`, 한글: `okt.morphs`) 형태소에 맞게 토크나이징했다. 이후 stopword와 punctuation을 제거하여 regularize했다. 이렇게 나온 terms(`List[str]`) 들은 pos_tag를 활용해 명사, 동사, 형용사, 부사로 분류했다.

한글 stopword는 “https://www.ranks.nl/stopwords/korean” 에서 제공한 리스트를 활용했으며 깃헙에 함께 txt 파일로 코드와 함께 업로드했다.

3. 트윗 내용 분석
3.1.단어 빈도 (Word Cloud)
단어의 형태소는 2.에서 언급한 Tokenizer가 분석해서 return해준다. 이때, 영어는 nltk가 제공해준 pos-tag, 한글은 KoNLPy에서 제공해준 Okt로 분석된 pos-tag 를 이용해 분류된다.

3.1.1.전체 단어
전체 단어 중 가장 많이 사용된 단어는 애플페이 키워드에서는 “애플” 과 “페이”이고, OpenAI 키워드에서는 “openai”이다. 다만, 이모티콘이나 말줄임표(…), unicode 상으로 파이썬 `string.punctuation`에 없는 작은 따옴표가 stopwords와 punctuation 제거 로직에서 필터링 되지 않아 결과가 조금 오염되었다. 나아가 애플페이 키워드 트윗 분석은 한글 위주로 진행된 터라

영어에서만 working하는 python에서 제공하는 `string.lower()`메소드를 사용할 수 없었다. 이런 이유로 'RT' 또한 애플페이 키워드에서 나타난 것이다.

애플페이 키워드에서 눈여겨 볼만한 다른 단어로는 "쿠팡", "파트너"가 있다. 이 두 단어는 애플페이라는 화제성 키워드에 편승하여 트윗에 광고를 한 경우로 해석된다. 이는 지난 주 과제에서도 언급한 것처럼 2022년~2023년에 계정이 많이 생성된 점은 어뷰징성 계정이 많이 생성되었다고 약한 의심을 해볼 수 있다.

WordCloud를 보면 애플페이 키워드에서는 "서비스", "아이폰", "21일" (애플페이 서비스 시작일), "할인" 등 애플페이와 관련한 여러 단어들이 나타나고, OpenAI 키워드에서는 인공지능 모델과 관련된 "model", "multimodal", "capabilities"가 보이며, "GPT4"를 발표했기 때문에 "announcing"도 확인할 수 있다.

애플페이	등장 횟수	OpenAI	등장 횟수
애플	1218	openai	558
페이	1143	gpt4	367
RT	708	ai	139
…	624	chatgpt	137
👇	514	model	133
✅	513	’	116
은	366	new	86
스	299	large	85
쿠팡	272	multimodal	84
파트너	272	capabilities	72

Table1: 전체 단어 중 10개의 최빈 단어

Figure 1: 전체 단어에 대한 WordCloud (좌: 애플페이, 우: OpenAI)

3.1.2.명사

애플페이 키워드에서는 “애플”과 “페이”가 모두 명사로 사용되었다. (3.1.1-Table1 결과와 비교) 하지만 OpenAI 키워드에서 “openai”는 일부만 명사로 인식되었다. 이런 이유는 openai가 고유명사이지만 최근에 만들어진 단어라는 점 그리고 ‘open’이라는 동사가 포함되어 nltk tokenizer 혼동이 온 것으로 생각된다.

애플페이	등장 횟수	OpenAI	등장 횟수
애플	1218	openai	181
페이	1143	model	133
은	366	chatgpt	89
스	299	’	89
쿠팡	272	capabilities	72
파트너	272	ai	67
활동	271	results	66
일환	271	h…	65
정액	268	gpt-4	56
수수료	261	today	36

Table2: 최빈 10개의 명사 단어

Figure 2: 명사에 대한 WordCloud (좌: 애플페이, 우: OpenAI)

3.1.3.동사

애플페이 키워드에서 “보러가기”, “받을” 이 가장 많이 사용되었다. 이는 앞서 지난주 과제에서 언급했고 3.1.1에서 언급한 것과 같이 광고 트윗이 적지 않았다는 것에 연관성이 있다. 쿠팡에서는 제품 링크를 공유할 때, 쿠팡 파트너로 링크를 생성할 수 있다. 이 경우, 해당 링

크를 클릭할 때마다 링크를 공유한 사람이 일정 금액을 버는 구조로 되어있다. 이때, 링크 공유 시 자동으로 생성되는 멘트는 다음과 같다. "👇보러가기👇 {쿠팡 연결 링크} 이 글은 쿠팡파트너스 활동의 일환으로, 일정액의 수수료를 제공 받을 수 있습니다 {제품 명, 키워드 등}" 이 멘트로 인해 "보러가기"와 "받을"이 동사에서 높은 등장 횟수를 보이는 것이다.

OpenAI 키워드에서는 "openai"가 동사로 오인식된 경우와 함께 'GPT'의 발표를 "announcing", "launched", "released", "announced"라는 동사를 사용해 설명한 트윗이 많다는 것을 생각할 수 있다. 의아한 점은 "alignment", 정렬이라는 뜻을 가진 명사가 동사로 분류된 것이다. ai-alginment (인공지능이 목적한 대로 작동하는 지를 정렬)하는 것과 관련이 있어 보이고, 문장에서 동사 위치 혹은 동사처럼 사용되어 동사로 분류된 것으로 생각된다.

애플페이	등장 횟수	OpenAI	등장 횟수
보러가기	257	openai	136
받을	240	announcing	65
는	141	alignment	65
대	70	know	46
늘었네요	67	launched	42
한	66	chatgpt	35
되고	51	ai	35
사는	48	think	24
됩니다	43	released	20
하고	35	announced	18

Table3: 최빈 10개의 동사 단어

Figure 3: 동사에 대한 WordCloud (좌: 애플페이, 우: OpenAI)

3.1.4. 형용사

형용사의 경우, 애플페이에서는 “빠르다”, “좋지”, “필요없이” 등 애플페이 서비스에 대한 묘사가 주를 이뤘다. OpenAI에서는 이전 챕터에서 분석한 것처럼 형태소를 잘못 분석한 신조어 “gpt-4”, “openai”가 많이 등장했다. 그 뒤로는 GPT4가 새로 발표된 Large-scale Language Model (이미지도 입력받을 수 있어져서 multimodal이다.)이기 때문에 new, large, multimodal 또한 여러번 사용되었다.

애플페이	등장 횟수	OpenAI	등장 횟수
있습니다	220	gpt-4	311
안되어서	67	openai	195
가능해지나	48	new	86
필요없이	22	large	85
이랬지	19	multimodal	65
있	19	best-ever	65
빠르다	16	ai	32
좋지	14	’	19
안남았나보다	14	capable	19
없는게	12	biggest	17

Table4: 최빈 10개의 형용사 단어

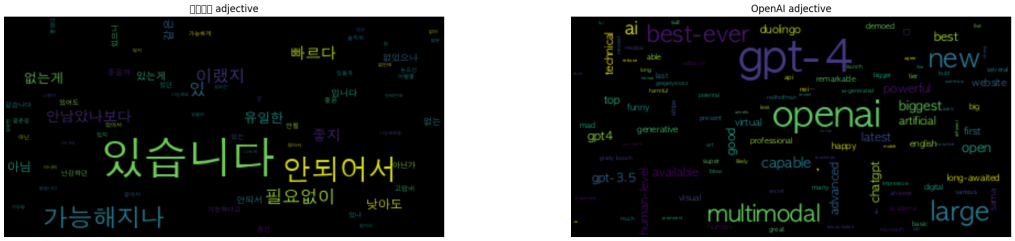

Figure 4: 형용사에 대한 WordCloud (좌: 애플페이, 우: OpenAI)

3.1.5.부사

부사의 경우, 명사, 동사, 형용사에 비해 사용 빈도가 매우 낮다. 애플페이 키워드에서는 애플페이는 데이터 “없이” 사용할 수 있다는 점에서 “없이”가 많이 사용되었다. 나아가 사람들이 오랜기간 바라오던 서비스였더터라 “드디어”, “빨리” 같은 부사도 많이 사용되었다.

OpenAI 키워드에서는 최근의 인공지능 발전을 말하는 트윗에서 주로 “recently”, “successfully”, “really” 등이 많이 사용됨을 알 수 있다. 다만 haven’t와 같은 줄인 표현에 붙는 “n’t”가 부사로 분류된 점이 의아하다. nltk 토크나이저가 pos-tag 오류를 꽤많이 일으키는 것 같다.

애플페이	등장 횟수	OpenAI	등장 횟수
없이	23	n't	21
근데	13	also	18
아무리	13	openai	14
다	12	even	13
드디어	11	recently	12
빨리	8	yet	12
너무	6	still	12
이미	6	really	10
많이	5	successfully	9
걍	4	literally	6

Table5: 최빈 10개의 부사 단어

Figure 5: 부사에 대한 WordCloud (좌: 애플페이, 우: OpenAI)

4. Additional Analysis (선행 연구 참고)

트윗 본문에 대한 분석에서 나아가 강의에서 언급된 선행 연구들을 기반으로 GLM, PCA, FactorMap을 만들어보았다.

아래 분석에 사용된 피쳐는 아래와 같다. (모든 시간은 UNIX timestamp으로 변환 후, 데이터상 최초를 0으로 처리해서 정규화했다.), (`~~여부`로 적혀 있는 피쳐는 1: 있음, 0: 없음)

피쳐: “트윗 생성 시간”, “유저 생성 시간”, “해쉬태그 여부”, “맨션 여부”, “팔로워 수”, “해당 유저가 작성한 트윗 개수”, “친구수”, “favourites 수”

예측 라벨: “리트윗 수”

4.1.Generalised Linear Model

`statsmodels`에 GLM이 구현되어 있어 쉽게 분석할 수 있었다. 아래 해석에서는 유의 수준 0.05를 기준으로 해석된다.

Dep. Variable:	rt_cnt	No. Observations:	1100
Model:	GLM	Df Residuals:	1091
Model Family:	Gaussian	Df Model:	8
Link Function:	identity	Scale:	31279.
Method:	IRLS	Log-Likelihood:	-7249.2
Date:	Sat, 25 Mar 2023	Deviance:	3.4125e+07
Time:	00:43:13	Pearson chi2:	3.41e+07
No. Iterations:	3	Pseudo R-squ. (CS):	0.3591
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	195.3967	23.968	8.152	0.000	148.420	242.373
twt_created	-0.0018	0.000	-15.778	0.000	-0.002	-0.002
user_created	-1.671e-07	4.78e-08	-3.495	0.000	-2.61e-07	-7.34e-08
hashtag	-97.6162	30.589	-3.191	0.001	-157.569	-37.663
mention	156.7661	12.541	12.500	0.000	132.186	181.346
followers	0.0010	0.001	0.957	0.338	-0.001	0.003
twts_user	-5.915e-05	5.52e-05	-1.072	0.284	-0.000	4.9e-05
friends	-0.0036	0.004	-0.997	0.319	-0.011	0.004
favorites	-0.2362	2.604	-0.091	0.928	-5.340	4.867

Dep. Variable:	rt_cnt	No. Observations:	1083
Model:	GLM	Df Residuals:	1074
Model Family:	Gaussian	Df Model:	8
Link Function:	identity	Scale:	1.4129e+07
Method:	IRLS	Log-Likelihood:	-10447.
Date:	Sat, 25 Mar 2023	Deviance:	1.5175e+10
Time:	00:43:14	Pearson chi2:	1.52e+10
No. Iterations:	3	Pseudo R-squ. (CS):	0.06745
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-596.6991	383.961	-1.554	0.120	-1349.249	155.851
twt_created	0.0054	0.238	0.022	0.982	-0.462	0.472
user_created	3.192e-06	7.28e-07	4.382	0.000	1.76e-06	4.62e-06
hashtag	-961.6042	329.602	-2.917	0.004	-1607.611	-315.597
mention	1506.3995	260.385	5.785	0.000	996.055	2016.744
followers	0.0002	0.000	0.684	0.494	-0.000	0.001
twts_user	-0.0006	0.001	-0.937	0.349	-0.002	0.001
friends	-0.0317	0.036	-0.886	0.376	-0.102	0.038
favorites	-72.0427	83.152	-0.866	0.386	-235.018	90.932

Figure 6: GLM Regression 결과 (좌: 애플페이, 우: OpenAI)

애플페이, OpenAI 키워드 모두에서 팔로워수, 유저가 작성한 트윗 개수, 친구수, 그리고 favourites 수가 리트윗 수와 큰 관련성이 있다는 것을 알 수 있다.

나아가 OpenAI에서는 트윗 생성 시간역시 큰 관련성이 있음을 보인다.

4.2.PCA 분석

2 factor 에 대한 PCA 분석 결과이다. scikit-learn에 PCA가 구현되어 있어 쉽게 결과를 생성할 수 있었다. 데이터 스케일이 다른 문제를 해결하기 위해 `StandardScaler` 를 이용해 정규화했다.

PCA	0	1
twt_created	0.189959	-0.280625
user_created	-0.256574	-0.270472
hashtag	0.011824	-0.137666
mention	-0.248849	0.492758
followers	0.590523	0.198060
twts_user	0.169821	0.325610
friends	0.585488	0.233606
favorites	0.266019	-0.341010
rt_cnt	-0.211898	0.519830

PCA	0	1
twt_created	0.010813	-0.122545
user_created	-0.351513	-0.365030
hashtag	0.143849	-0.298984
mention	-0.324353	0.450714
followers	0.465093	-0.159301
twts_user	0.420625	0.347364
friends	0.302043	0.540601
favorites	0.404097	-0.319463
rt_cnt	-0.320411	0.137321

Figure 7: 2 factor PCA analysis 결과 (좌: 애플페이, 우: OpenAI)

4.3.Factor Map

4.2에서의 PCA 분석을 기반으로 factor map을 그리면 아래와 같다. 두 키워드 공통으로 친구수와 팔로워수는 orthogonal 한 관계를 보이고 있다. 리트윗 수는 팔로워 수에 반하는 방향으로 영향을 주고 있다. 멘션수 역시 친구수와 favourites 수에 orthogonal하다. 해쉬태그의 영향력 역시 두 키워드 공통으로 비슷하게, mention, followers와는 orthogonal 하고, favourites와 friends와는 반대방향의 영향을 주고 있다.

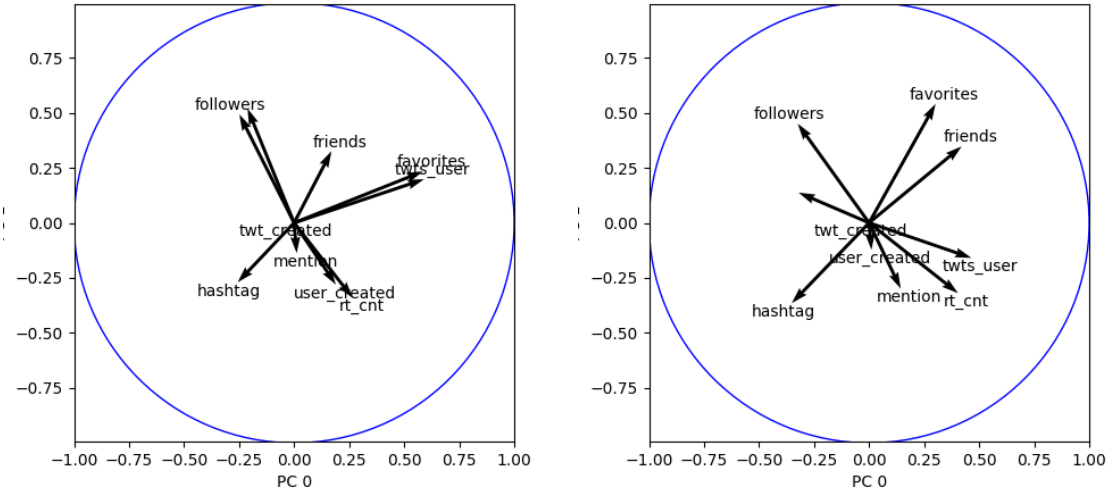

Figure 8: 2 factor PCA analysis로 그린 factor map (좌: 애플페이, 우: OpenAI)