Agora, Silk Road 다크웹 마켓 데이터 분석

Kaggle Dark Web Market Data 분석 toy project

강 규 욱

2024. 1. 25.

목차

과제 분석 내용 요약

과제 목적

다크웹 포럼(Agora, Silk Road)의 마약, 무기, 서적, 각종 불법 서비스 등 불법거래 데이터에 대한 분석 및 자연어처리(NLP)를 통한 판매물건 카테고리 예측 모델 구축.

분석 내용

전처리 및 EDA

Agora, Skill Road 두개의 Raw Data에 대하여 전처리 및 EDA 진행. 데이터 탐색 및 정제는 기본적인 오픈소스 라이브러리(Pandas, Numpy, Re, Spacy 등)을 이용하여 진행하였다. 데이터 탐색은 프로그래매틱 탐색과 시각적 탐색을 병행하였고, 데이터 품질 문제가 다수 발견되어 수정하였다. 시각화는 Seaborn, Matplotlib, Plotly등을 이용하였다.

포럼에서 판매 중인 물품에 대한 분석

전처리한 내용을 바탕으로 파이차트를 활용해 현재 판매중인 물품의 트렌드에 대하여 분석하고 Skill Road 데이터를 시계열 분석을 활용해 판매 물품의 카테고리별 시세 변화 예측을 하였다.

다크웹 포럼 유저 분석

Elbow Method와 Silhouette method를 활용하여 판매자(Vendor)에 대한 군집분석 (Clustering)을 진행하였고, 이를 통해 유저간 연관성을 분석하였다. Clustering 에 사용된 Feature는 판매자의 총판매액, 거래물품수, 평균평점이었고, 이에 따라 Cluster별로 활동지수를 구분하였다.

두 포럼의 유사성 및 차별성 분석

EDA를 통해 분석한 내용을 바탕으로 Agora, Skill Road 두 포럼의 유사성과 차별성을 분석하였다. 두 포럼의 유사성은 총판매액과 거래량이 마약에 큰 비중이 있다는 점이었으며, Agora는 다양한 문자형 데이터가 풍부하여 자연어 처리에 용이하단 특징이, Skill Road는 US Dollar, 시간데이터, 리뷰 수 데이터와 같은 수치형 데이터가 있어서 통계분석과 시계열분석이 용이하단 특징이 있다.

전처리 및 EDA

데이터 탐색

데이터탐색을 통해 중복값, 결측치, 부정확한 데이터를 확인하여 품질의 문제를 파악하고, 변수, 관측치, 관측단위 등을 확인하여 데이터 구조문제를 파악.

탐색

데이터프레임을 직접 불러와서 데이터 시각적 탐색. .info(), .descrive(), .value_counts() 등의 코드를 사용하여 프로그래매틱 탐색. Agora의 데이터사이즈는 109689건, Skill Road의 데이터사이즈는 566564건이다.

표 1. Agora Data

	Vendor	Category	Item	Item Description	Price	Origin	Destination	Rating	Remarks
0	CheapPayTV	Services/Hacking	12 Month HuluPlus gift Code	12-Month HuluPlus Codes for $25. They are wort...	0.05027025666666667 BTC	Torland	NaN	4.96/5	NaN
1	CheapPayTV	Services/Hacking	Pay TV Sky UK Sky Germany HD TV and much mor...	Hi we offer a World Wide CCcam Service for En...	0.152419585 BTC	Torland	NaN	4.96/5	NaN
2	KryptykOG	Services/Hacking	OFFICIAL Account Creator Extreme 4.2	Tagged Submission Fix Bebo Submission Fix Adju...	0.007000000000000005 BTC	Torland	NaN	4.93/5	NaN

표 2. Agora Data Info

RangeIndex: 109689 entries, 0 to 109688
Data columns (total 9 columns):

#	Column	Non-Null Count	Dtype
0	Vendor	109689 non-null	object
1	Category	109689 non-null	object
2	Item	109685 non-null	object
3	Item Description	109662 non-null	object
4	PriceBTC	109677 non-null	float64
5	Origin	99807 non-null	object
6	Destination	60528 non-null	object
7	Rating	109674 non-null	object
8	Remarks	12616 non-null	object

dtypes: float64(1), object(8)

표 3. Skill Road Data

	Title	Sellerid	PriceUSD	PriceBTC	Rating	Reviews	Origin	Destination	Category	Subcategory	Market	Date
0	Ray Ban Tech RB3460 001 Aviator/Flip Out/ Sung...	FoxyGirl	61.1615418288	0.0982320000	NaN	NaN	China	Worldwide	Apparel	NaN	SilkRoad2	2014-07-17
1	Ray Ban RB3025 Aviator Classic Sunglasses Replica	FoxyGirl	37.9625939448	0.0609720000	NaN	NaN	China	Worldwide	Apparel	NaN	SilkRoad2	2014-07-17
2	Rolex - Watch Box (AAA Grade Replica)	RepAAA	100.4297770434	0.1613010000	NaN	NaN	Hong Kong, (China)	Worldwide	Apparel	NaN	SilkRoad2	2014-07-17

표 4. Skill Road Data Info

Index: 566564 entries, 0 to 567218
Data columns (total 14 columns):

#	Column	Non-Null Count	Dtype
0	Item	566564 non-null	object
1	Vendor	566340 non-null	object
2	PriceUSD	566564 non-null	float64
3	PriceBTC	566564 non-null	float64
4	Rating	566564 non-null	float64
5	Reviews	566564 non-null	float64
6	Origin	566341 non-null	object
7	Destination	566341 non-null	object
8	Category	364491 non-null	object
9	Date	566564 non-null	datetime64[ns]

dtypes: datetime64[ns](1), float64(4), object(9)

데이터 품질 문제 정의

중복값 제거 및 결측치와 부정확한 데이터에 대한 처리 필요. 가격(Price, PriceUSD, PriceBTC)이 천문학적으로 높거나, 물품의 평균적인 시세에 비해 비정상적으로 높게 입력된 데이터를 다수 식별. 입력값이 적절하지 않은 Column에 입력되어 있는 문제 해결 필요. 지역에 대한 Data (Origin, Destination) 입력값 통일 필요(USA, us, United States, worldwide prior ship 등). 두 포럼 간 Data에 대한 통일 필요 (Column명, Category분류).

표 5. Trash Data (Price ...)

	Vendor	Category	Item	Item Description	PriceBTC	Origin	Destination	Rating	Remarks
22603	HAPPYholland	Drugs/Stimulants/Speed	2g clean pure speed paste in snow seal 14â□¬ ...	â□□â□□â□□â□□â□□â□□â□□â□□â□□â□□â□□â□□â□□â□□â□□â...	130396.0896890602	belgium	worldwide prior ship	4.857/5	
104536	EdWestwick	Drugs/Cannabis/Concentrates	56 grams of dabs	Tested 3 different batches they all range bet...	122500.0000000000	usa	NaN	4.97/5	
104534	William_Shatner	Drugs/Cannabis/Concentrates	QP of FIRE Purp and Sour Diesel Mix Shatter! G...	William Shatner is back from the cosmos bringi...	120004.6873169000	usa	NaN	5.00/5	

데이터 정제

데이터 품질 문제 해결

비정상적인 가격 문제는 시각적 탐색과 온라인 정보 검색을 통해 키워드들을 직접 수동으로 매핑하여 해결하였다. 'lsd', 'cocaine', 'kg', 'kilogram', 'kilograms'등 과 같이 고가에 해당하는 키워드에 해당하지 않으면서 높은 가격(200BTC 이상)인 데이터는 삭제하고, 키워드에 해당하는 경우는 .sort_values()메소드를 통해 고액부터 육안으로 확인하면서 삭제하였다.

Item, Item Description, Origin, Destination, Rating에 대한 텍스트 전처리는 정규표현식과 수동매핑, Numpy, Spacy라이브러리등을 활용해 전처리하였으며, 특수기호 처리, 소문자화와 Item, Item Description 의 경우 분석의 편의를 위하여 불용어들을 처리하였다.

KNN예측 모델을 이용한 Category 결측치 처리

상대적으로 결측치가 없이 잘 정리 되어있는 Agora Data의 Category에 비해 Skill Road Data의 Category는 같은 판매자의 같은 종류의 물건이 다르게 입력되어 있거나 결측치가 많아 부정확한 Data가 많았다. 이를 해결하기 위하여 Agora Data의 Item에 입력된 문자들을 Bag-of-Words(TF-IDF)방식으로 벡터화하고 KNN 모델을 학습하였다. 학습된 모델은 Skill Road의 Item값을 입력하면 입력값과 가장 유사한 Agora Data의 Index값을 출력하며, 이를 이용하여 Skill Road의 Category를 예측하였다.

표 6. KNN을 이용한 Skill Road Data Category Nan값 예측

	Item	Category
55550	social engineering art human hacking	Info/eBooks
55551	expert making small talks	Services/Other
55552	construction secret hiding places	Info/eBooks
55553	psychedelic chemistry	Chemicals
55554	total synthesis ii ecstasy amphetamines synthesis	Info/eBooks
55555	psilocybin mushroom handbook easy indoor outdo...	Info/eBooks
55556	figure second income start grow successful onl...	Info/eBooks
55557	1000dollar day ebay method legal unethical	Data/Accounts
55558	adult affiliate network real money adult traffic	Services/Other
55559	arrest proof	Info/eBooks
55560	anarchist cookbook ver. 2000	Info/eBooks
55561	basics hacking penetration testing ethical hac...	Info/eBooks
55562	wifi hacking	Info/Guides
55563	portable trades occupations grandpa success gu...	Info/eBooks
55564	lockpicking book collection	Info/eBooks
55569	bitcoin beginner book	Drugs/Psychedelics/Mushrooms
55570	gentleman timely guide timeless manners	Info/eBooks/Making money
55571	beat system steal book	Services/Other
55572	blow mind illustrated guide orgasmic oral sex ...	Info/eBooks/Relationships/Sex
55573	million years think like tycoon	Info/eBooks
55574	influence psychology persuasion	Drugs/Benzos
55579	twitter pays dollars	Counterfeits/Accessories
55580	female orgasm black book	Data/Accounts

표 6은 Skill Road에서 기존 Category가 결측값이었던 데이터들로, Item에 입력된 문자를 유사도 측정 방법(KNN)을 이용하여 Agora Data의 Item과 가장 유사한 Index의 Category로 반환하여 채워졌다.

Data시각화 및 기술통계량

Data시각화

전처리한 Data를 기반으로 시각화를 진행.

그림 3. Agora Item x Origin

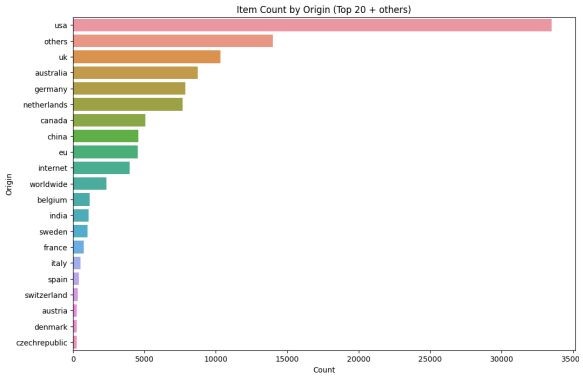


그림 1. Agora TotalSales x Origin

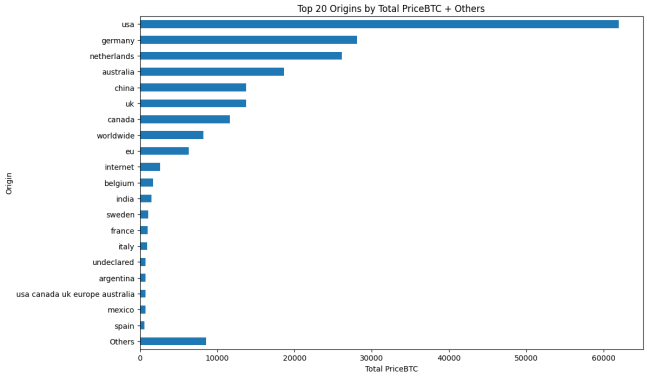


그림 4. Skill Road Item x Origin

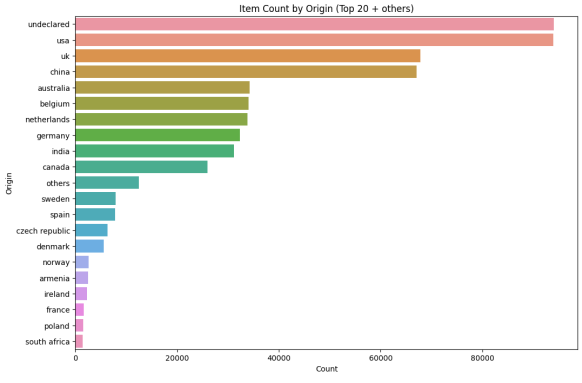


그림 2. Skill Road TotalSales x Origin

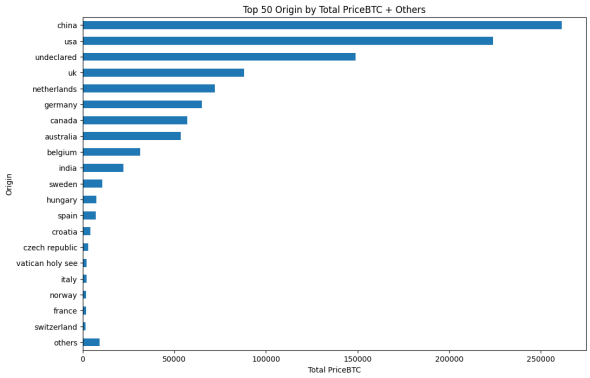


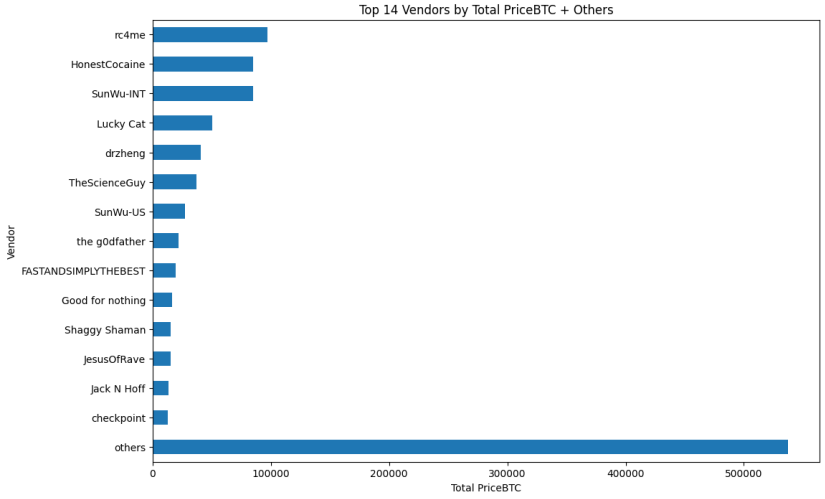
그림 3과 그림 4는 각각 Agora, Skill Road 두 포럼의 물품이 가장 많이 생산되는 지역 상위 20개를 Bar Plot으로 나타낸 그래프이다. 그림 1과 그림 2는 두 포럼의 물품 거래액이 가장 높은 생산지 상위 20개를 나타낸 그래프이다.

표 7. Agora, Skill Road Origin Data

	Agora - Item	Agora - TotalSales	Skill - Item	Skill - TotalSales
0	Usa 33493	Usa 61995	Undeclared 94048	China 261604
1	Uk 10321	Germany 28065	Usa 93975	Usa 223842
2	Australia 8722	Netherlands 26130	Uk 67794	Undeclared 148797
3	Germany 7874	Australia 18612	China 67089	Uk 87940
4	Netherlands 7668	China 13776	Australia 34277	Netherlands 71997

그림 5는 SkillRoad Data에서 총판매액 상위14인과 나머지에 대한 데이터이다. 전체 1334인 중에서 상위 14인의 누적판매액과 나머지의 누적판매액이 일치한다. Agora의 경우 전체 3183인의 판매자중 상위 142인의 누적판매액이 나머지의 누적판매액과 일치한다.

그림 5. SkillRoad Top50 Selling Vendor



기술통계량

표 8. Agora 기술통계량 (float type)

	PriceBTC	Rating
count	108692.0000000000	100455.0000000000
mean	2.1039918797	4.8880693047
std	12.5947108731	0.3544324803
min	0.0015037500	0.0000000000
25%	0.1245574060	4.8970000000
50%	0.3826712086	4.9700000000
75%	1.3121418382	5.0000000000
max	2760.9295858833	5.0000000000

표 9. Agora 기술통계량 (Object type)

	Vendor	Category	Item	Item Description	Origin	Destination
count	108692	108692	108692	108665	98978	60093
unique	3183	104	103719	68070	314	689
top	optiman	Drugs/Cannabis/Weed	custom	terms...	usa	worldwide
freq	874	21201	45	291	33493	23561

표 10. Skill Road 기술통계량 (Object type)

	PriceUSD	PriceBTC	Rating	Reviews
count	566564.0000000000	566564.0000000000	566564.0000000000	566564.0000000000
mean	869.6978381028	1.8966169363	2.0719751873	15.9211474785
std	6578.2610085707	14.7540898423	2.2281907650	78.1515598543
min	0.9007101072	0.0011440000	0.0000000000	0.0000000000
25%	41.6149443590	0.0875210000	0.0000000000	0.0000000000
50%	111.5188456655	0.2400800000	0.0000000000	0.0000000000
75%	384.4499393569	0.8372190000	4.5500000000	1.0000000000
max	582979.9248000000	1376.2730570000	4.9000000000	2483.0000000000

표 11. Skill Road 기술통계량 (Float type)

	Item	Vendor	Origin	Destination
count	566564	566340	566341	566341
unique	31754	1334	59	25
top	free viagra cialis levitra 1000 non controlled...	rc4me	undeclared	worldwide
freq	688	27975	94048	332715

포럼에서 판매 중인 물품에 대한 분석

판매되는 물품들에 대한 동향 분석

Agora에서 판매되는 물품들을 Category 별 총판매액을 Pie그래프를 통해 시각화하였다. 총판매액은 약 228,687 BTC이고 최상위 Category에 대한 판매액은 표 12과 같다.

표 12. Agora TotalSales via Category

Level1	PriceBTC
Drugs	213248.6704677381
Services	4221.1139614314
Other	2792.5060115058
Counterfeits	2085.0799177559
Weapons	1662.6785524326
Electronics	1210.2795404967
Forgeries	1140.7444018210
Info	1073.3369410942
Data	656.2276855750
Jewelry	277.9229135763
Drug paraphernalia	127.6262220219
Tobacco	98.3872669038
Chemicals	92.5115017290

그림 6. Agora TotalSales Via Category

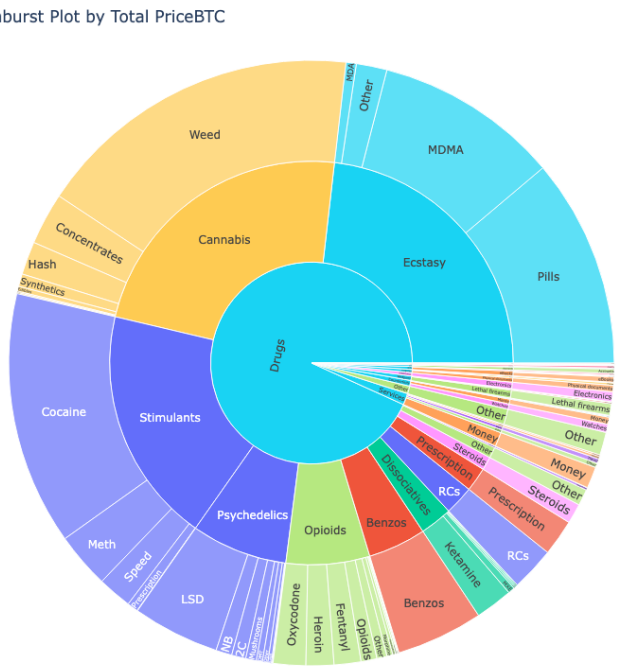
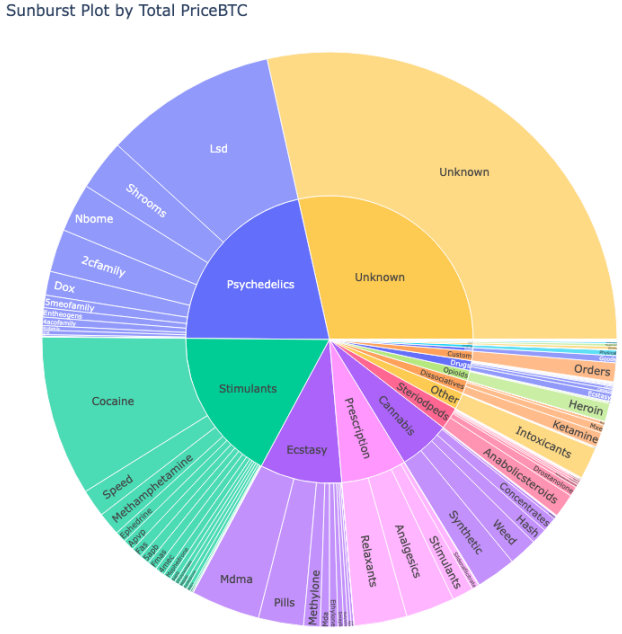


그림 7. Skill Road TotalSales via Category



Skill Road에서 판매되는 물품들의 Category 별 총판매액은 모두 약 1,074,554BTC이고 최상위 Category에 대한 판매액은 표 13과 같다.

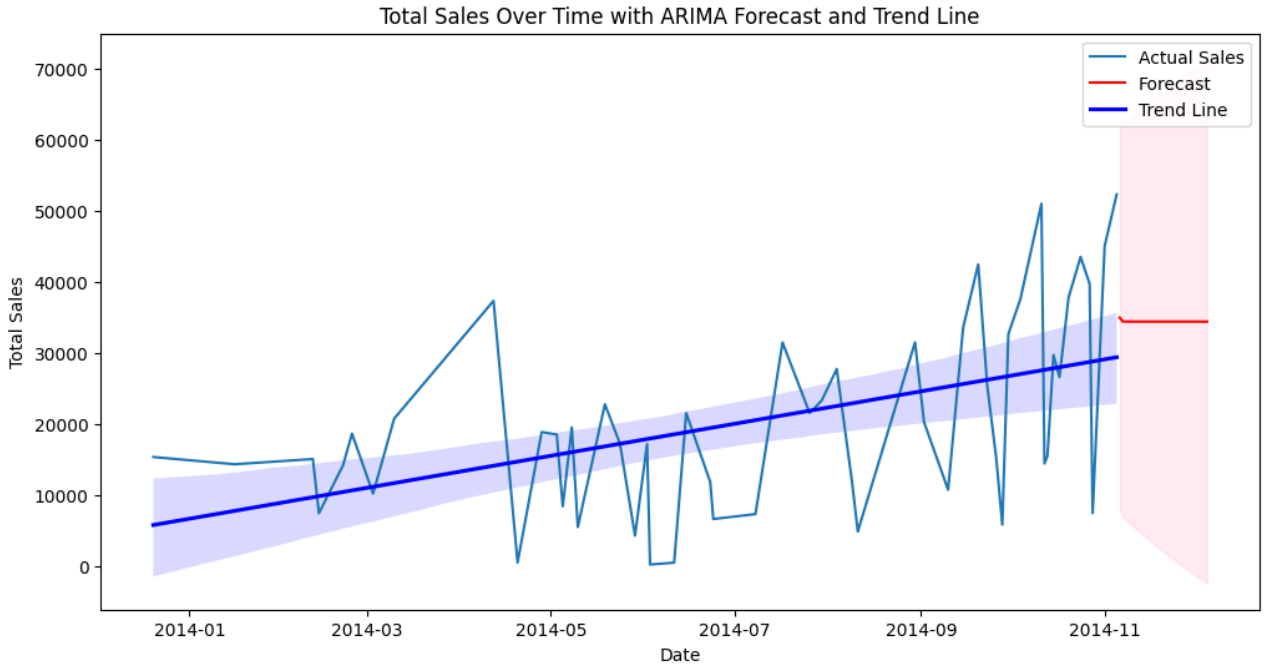
표 13. Skill Road TotalSales via Category

Level1	PriceBTC
Unknown	306117.9393560000
Psychedelics	229707.8047810000
Stimulants	185630.0495870000
Ecstasy	99243.1220570000
Prescription	78993.5226540000
Cannabis	62193.6988740000
Steriodpeds	26855.5754090000
Other	20281.8769500000
Dissociatives	15907.0723130000
Opioids	12455.0195310000
Drugs	11831.7382440000
Custom	11462.5116490000

시세 변화 및 예측

SkillRoad Data를 이용하여 날짜 별 총판매액에 대한 추세선을 그리고 ARIMA를 활용한 시계열분석으로 2014년 11월 이후 30일간 SkillRoad 다크웹 총 물품 거래액에 대한 예측. 불법거래 물품(마약, 총기류, 불법소프트웨어, 위조문서, 가품 등) 특성상 명확한 추세나 계절성이 존재하지 않을것으로 예상되어 ARIMA 모델 사용.

그림 8. Skill Road ARIMA



다크웹 포럼 유저 분석

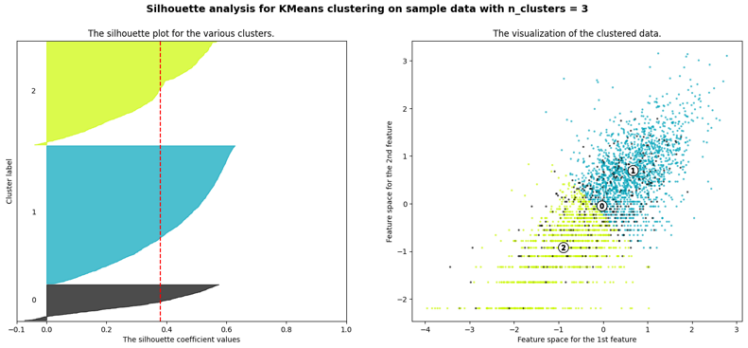
유저 간 연관성 분석

군집 분석을 이용하여 유저 간 연관성을 분석하였다. Cluster의 개수를 결정하는 평가지표는 Elbow Method와 Silhouette Method를 활용하였다. 유저(Vendor)별 Clustering을 하기위해 Feature Engineering을 통해 Vendor를 Unique Value로 갖고, 총판매액(TS), 총판매물품수(NoI), 평균평점(Rating)을 Column으로 갖는 새로운 데이터프레임을 만들었다.

표 14. Agora Vendor_df_for_Clustering

	Vendor	TS	NoI	Rating
0	#NAME?	4.6299321950	5	6.0000000000
1	-BIGG-BALLs-	2.4911464063	9	5.8062500000
2	-Euphoria-	12.0388506203	13	6.0000000000

그림 10. Agora Siluhoutte Score



For n_clusters = 2 The average silhouette_score is : 0.3561382365881271
For n_clusters = 3 The average silhouette_score is : 0.38023830756225513
For n_clusters = 4 The average silhouette_score is : 0.33053353303449906
For n_clusters = 5 The average silhouette_score is : 0.31059203438039484
For n_clusters = 6 The average silhouette_score is : 0.3142132518661735
For n_clusters = 7 The average silhouette_score is : 0.31842544618111324
For n_clusters = 8 The average silhouette_score is : 0.3118406729295943
For n_clusters = 9 The average silhouette_score is : 0.30512308957889633

그림 9. Agora Elbow Method

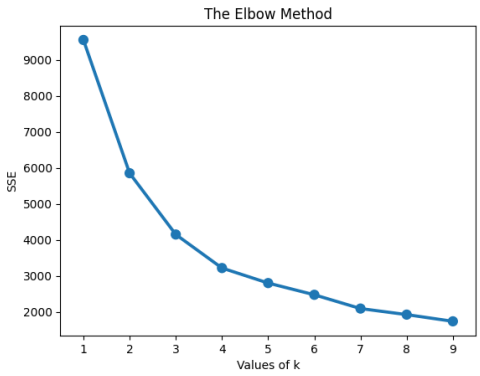


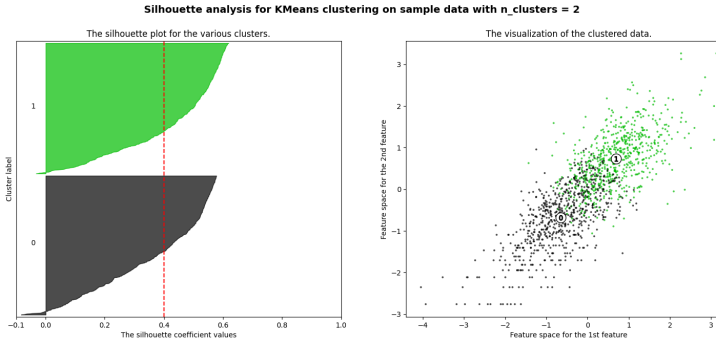
그림 10은 Cluster 개수 별 Siluhoutte Score와 산점도를 나타내주는 평가지표로, Cluster가 3일 때 가장 점수가 높게 나타났다. 또한 그림 9에서 그래프가 k값이 3일때 꺾이므로, Agora Data는 3개의 클러스터로 나누어 주었다. 클러스터링의 성능을 높이기 위해 Right Skewed Feature에 log를 취하였고, Right Skewed Feature에 Box-Cox기법을 적용하였다. 또한 stansardscalrer를 이용해 표준화를 하여 스케일링을 진행하였다. 클러스터링을 진행한 후 각 클러스터 별 Feature 값의 평균은 표 15와 같았다

표 15. Agora Clustering

Cluster	TS mean	NoI mean	Rating mean	NoC mean	count
0	4.0269723778	6.6952789700	5.9361210057	1.9158798283	1165
1	39.1356492303	23.8888888889	5.0133722208	3.7494089835	423
2	130.0572306831	56.9197492163	5.9505526648	4.8351097179	1595

클러스터별로 총판매량, 총판매아이템 갯수의 평균에서 차이가 남에 따라 클러스터를 유저의 활동지수(ActivityIndex)로 설정하고, Cluster 0에 속하는 판매자의 활동지수는 'low', Cluster 1 에 속하는 판매자의 활동지수는 'mid', Cluster 2에 속하는 판매자의 활동지수는 'high'로 변환하였다.

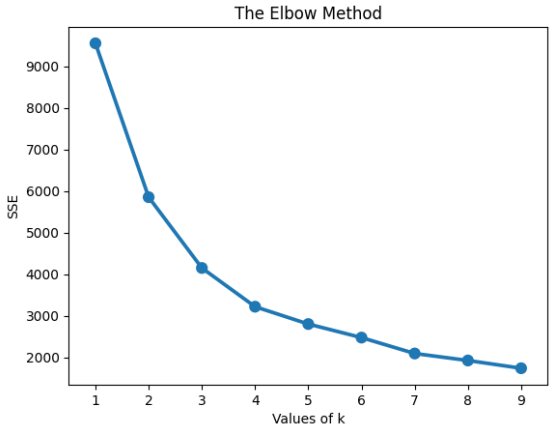
그림 11. Skill Road Siluhoutte Score



For n_clusters = 2 The average silhouette_score is : 0.4001109692434405

For n_clusters = 3 The average silhouette_score is : 0.36635628988820684

그림 12. Skill Road Elbow Method



반면 SkillRoad Data에서는 Cluster가 2일 때 silhouette score가 0.4로 가장 높았고 이에따라 k=2 Clustering을 진행하였다. 결과는 표 16과 같다.

표 16. SkillRoad Clustering

Cluster	TS mean	NoI mean	Rating mean	NoC mean	count
0	43.6245958735	66.1482558140	1.6204611183	2.7994186047	688
1	1616.4213081424	806.2383900929	3.4702718292	6.8436532508	646

마찬가지로 , Cluster 0에 속하는 판매자의 활동지수는 'low', Cluster 1 에 속하는 판매자의 활동지수는 'high'로 설정하였다. 그림 14는 Agora의 Clustering을 3D 산점도로 표현한것이고, 그림 13은 SkillRoad의 Clustering을 3D 산점도로 표현한것이다.

그림 14. Agora Clustering Scatter 3D

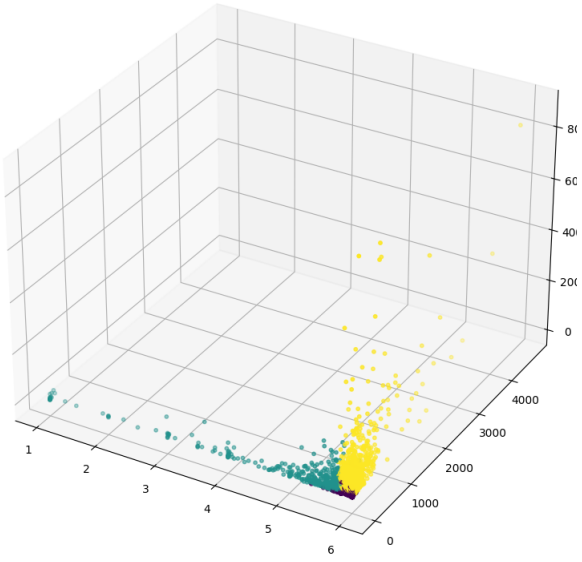
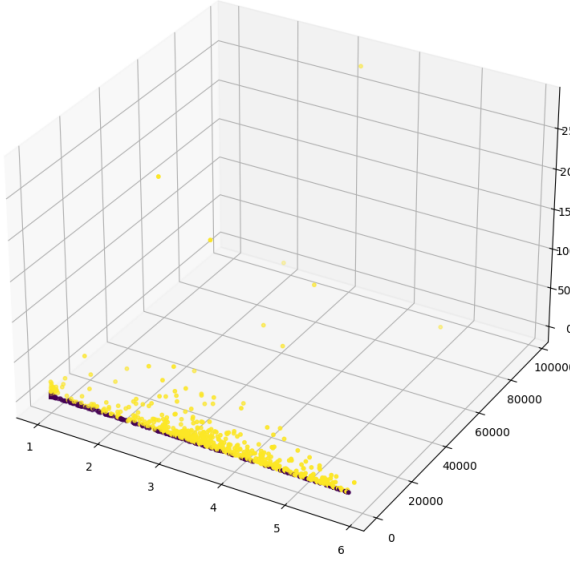


그림 13. SkillRoad Clustering Scatter 3D



결론

두 포럼의 유사성 및 차별성

표 17. 포럼간 유사점 및 차이점

	Agora	SkillRoad
차이점	- 카테고리 분류가 Skill Road에 비해 비교적 정밀하게 되어있음. - Item과 Item Description 칼럼이 있어서 자연어 처리를 통한 분석에 용이 - review 개수에 따른 데이터 부족	- 카테고리 분류가 정밀하게 되어있지 않음 - 광고성 데이터가 많음 - 시간 데이터가 존재하여 시계열 분석 가능 - review 개수에 관한 데이터 분석 활용 가능 - USD 와 BTC 자료가 함께있어서 시세파악과 데이터 훼손 여부 파악에 유리
유사점	- 마약류에 대한 거래가 가장 큰 비중을 차지함	

보완점

NLP모델의 활용

자연어처리(NLP)를 통한 판매물건 카테고리 예측 모델 구축을 성공적으로 확인하였으나, SkillRoad의 데이터수가 50만개가 넘어 카테고리 분류를 모두 적용하려면 맥북 M1pro 프로세서 기준 학습에 약 100시간 넘게 필요하였다. 이로 인해 본 보고서에는 SkillrRoad의 카테고리 재분류 이후 분석을 통해 얻을 수 있는 인사이트를 다루지 못하게 되었다.

수치형 데이터 활용

자연어처리와 군집분석에 집중하다가 상대적으로 분석이 용이한 SkillRoad의 수치형 데이터에 대한 분석 내용을 담지 못하였음. (상관분석, 회귀분석 등)

DL 활용

신경망을 활용한 분석이나 BERT같은 언어 모델을 활용해보면 좋을 것 같음.

References

참고 문헌 및 출처

- 유원준, 안상준 (2023) 딥 러닝을 이용한 자연어 처리 입문

- Statista. Average price for a dose of illicit psychoactive substances in France in 2020, by type of drug (2020) https://www.statista.com/statistics/1173002/price-way-drugs-france/

- 공돌이의 수학정리노트 (Angelo's Math Notes). 고윳값과 고유벡터 (2019) https://angeloyeo.github.io/2019/07/17/eigen_vector.html#fn:2

- Kaggle. Kmeans clustering with Elbow Method and Silhouette (2019) https://www.kaggle.com/code/abhishekyadav5/kmeans-clustering-with-elbow-method-and-silhouette

- Youtube. [딥러닝 자연어처리] TF-IDF (Minsuk Heo 허민석) (2019) https://www.youtube.com/watch?v=meEchvkdB1U