Portfolio

Fake reviewer detection

신재만

과제 개요

- Fake reviewer를 탐지하는 방법에 대한 과제를 진행
- yelp의 이용자 프로필을 데이터로 활용
- 이용자들의 프로필로부터 행동 심리적인 특징을 추출

» 정상적인 리뷰를 쓰는 집단과 fake 리뷰를 쓰는 집단을 나누고 행동 심리적인 특징을 기반으로 두 집단간의 기하학적 차이에 대해 분석

» 이를 기반으로 새로운 특징을 추출하여 탐지 모델의 성능 향상이 과제의 목적

※ Fake reviewer란 상업적인 목적으로 제품을 사용하지 않고 자사의 제품을 홍보하기 위해 리뷰를 작성하는 리뷰어, 경쟁사의 제품을 폄하하기 위해 리뷰를 작성하는 리뷰어, 그 외 도움이 되지 않는 리뷰를 작성하는 리뷰어를 뜻함

※ yelp란 크라우드 소싱 리뷰 포럼으로 미국의 다국적 기업에 의해 지원되는 지역 검색 서비스이다.

데이터에 대한 추론

» Fake reviewer들은 정상적인 리뷰어들을 모방하려는 경향이 있음

» 또한 상업적인 목적으로 리뷰를 작성하기에 이들의 행동패턴이 정형화되기도함

» 이를 기반으로 다음과 같은 가설을 설정

"Fake reviewer들은 정형화된 행동패턴을 가지고 있으므로 이들로 구성된 집단을 기하학적으로 표현하였을 때 이들의 기하학적 모형은 상대적으로 단순한 모형을 보일 것이다."

활용한 방법과 도구

» RandomForest
» Topological Data Analysis
» Python : Numpy, Pandas, Sci-kit learn, Giotto-tda

과제 수행을 위한 기법

» 활용 기법 : Topological Data Anaylsis(TDA)

» Topology는 데이터간의 거리나 좌표의 영향을 받지 않고 대상의 불변하는 기하학적 성질을 추출

» 활용 지표 : persistent entropy

※ Persistent entropy는 대상의 기하학적 무질서도를 나타낼 수 있어 가설을 검증하기에 적합

활용한 데이터

○리뷰를 게재할 수 있는 사이트 yelp의 데이터를 활용
○호텔 데이터와 식당 데이터로 구성되어 있으며 리뷰어, 업체, 제품 및 서비스에 대한 정보를 포함

	reviewerID	name	location	yelpJoinDate	friendCount	reviewCount	firstCount	usefulCount	coolCount	funnyCount
	필터	필터	필터	필터	필터	필터	필터	필터	필터	필터
1	yevHGEUQGmMBXhJ8B5A	Kevin T.	Oconomowoc, WI	May 2011	4	85	8	129	47	81
2	yoD_PYCHjnPjh70ATA0Jgw	Veronica B.	Saint Paul, MN	January 2010	5	48	5	60	34	21
3	XrFCsg4ewW5cts9QXokWPe	Paul The Commander M.	Saint Louis, MO	August 2008	15	135	29	235	85	102
4	y5ptsWnvGEAf0QaiFhBcg	Stella BraveTart J.	Lexington-Fayette, KY	September 2009	49	104	36	282	83	92
5	uUVZJm9yxNI5FBsXbt4WEg	Ginger 'where's my meds' w.	San Francisco, CA	August 2008	22	34	2	81	32	52
6	ZCRYAGLLTi8ZRPi1P7Cw	Johan Johanne S.	San Leandro, CA	August 2010	0	34	0	20	3	3
7	uCFIY6vBBp7O5un0CLggTw	Daniel Don Quijote K.	Honolulu, HI	June 2011	72	51	0	57	19	16
8	tdE3_J2ot_nL3M3sy0MC	Jen Mme Federer C.	New York, NY	December 2008	136	238	37	200	103	40
9	Uu-qEGsSb72ngIQUF85rDO	Anna P.	New York, NY	March 2010	1	11	0	9	4	3
10	zyk-YPhfFZK5kkqozEKrWw	Kara P.	Covington, KY	December 2009	66	48	2	51	19	14
11	V2OQaeZvePoCkbZ-Uye6KA	Rachel here's lookin at U.	Philadelphia, PA	October 2007	52	234	21	554	402	192
12	TYSP5swMHOnqgBUpQH7GO	Jeffrey A.	Los Angeles, CA	June 2011	2	8	0	3	1	1
13	wB_o7yYlA1bnjgo3aPokg	Craig H.	Gridley, IL	February 2006	0	20	6	11	6	1
14	wDWGjailHD5rTF49cJlg	penny h.	Chicago, IL	February 2007	0	52	1	22	7	4
15	WRMbs8iXsVZT9GHYCLTA	Esther P.	Chicago, IL	December 2011	2	10	0	4	0	2
16	UqYndXp9pbFoacaonileTw	Amy w.	Richardson, TX	February 2009	8	43	1	42	10	10
17	TIlh76nA82VBpXK_XucOBg	c m.	Chicago, IL	May 2010	0	22	0	14	2	4
18	x1uo32SAD_XMq0JoC8_OQ	Shadow shadow K.	San Diego, CA	April 2011	0	2	0	0	0	0
19	Zld8ZxcPSCXvDfMZTBXVZC	Emily W.	Berkeley, CA	December 2006	13	7	0	7	3	0
20	TtvhYwNGqDv6AM2EnYZE0g	Michael The best is good enough for me S.	Denver, CO	January 2007	511	838	175	4146	3440	3411
21	wIFeLz74XNf7ZBNnfxDzWQ	Tanya thrifty splurger S.	River Forest, IL	August 2007	7	129	16	126	34	29
22	UWT8_eEQVLsFaQfqUmX-Mw	Tina M.	Chicago, IL	October 2010	31	85	10	64	15	10
23	yhTclCkoPGD_BRE9DmXQ	Kate S.	Chicago, IL	April 2008	0	11	0	5	1	0
24	yCykhVHx9HTYE9Cp23oug	K. N.	Boston, MA	September 2009	0	12	1	3	0	0
25	tpV0gtaIY54F4y2bv-gDow	Garrett deadbody L.	Minneapolis, MN	November 2010	3	32	6	4	4	5
26	xqBMehJzMhUEI33Eun4OCw	Pete B.	Chicago, IL	October 2009	19	88	2	77	6	12
27	-Zj7AvbUsQqnG5zDZ4E9Hw	Maybelyn L.	Madison, WI	June 2010	1	45	0	8	5	5
28	UvN5nsbB9EG-0rhNkY0gpw	John M.	Hawthorne, CA	December 2011	1	18	1	7	2	3
29	UM0hJuVRR27cTRLrQU9EA	Dee M.	Gurnee, IL	April 2012	0	3	1	3	1	0
30	We19fTNDqsoKmbdIo1E5A	Karim Friendbear V.	Seattle, WA	July 2011	22	84	0	61	21	18
31	WJIRyXYyEwUrNh8swNCQ	Steve shanks s.	MA	January 2007	114	361	18	656	352	301
32	IHYBy5NwFgexq54iA446Q	chewy c.	San Francisco, CA	November 2007	6	39	2	19	10	7
33	VMNtDvOrzWL6taFm7IKWw	Kimberley Survival of the Fittest K.	New York, NY	April 2009	61	131	17	408	143	172
34	uT5Ie4EvfILWP5olbIM4Q	Jon J. S.	Chicago, IL	July 2008	23	26	1	34	10	24

전처리

» 리뷰어의 프로필 정보로부터 다음과 같은 특징값을 추출

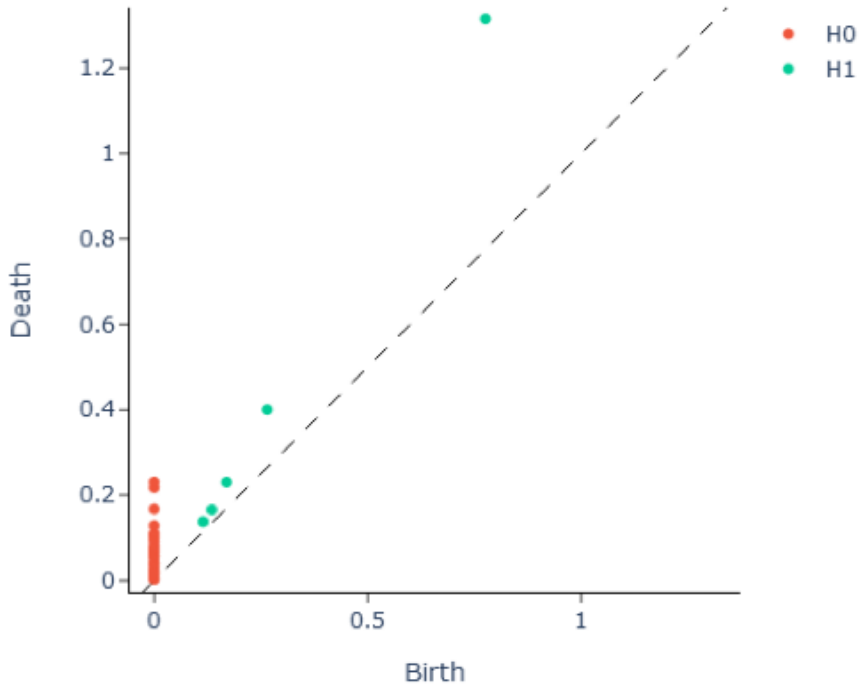
1. 하루동안 작성한 리뷰 개수의 최댓값
2. 전체 평점 중 긍정적인 평점의 비율
3. 작성한 리뷰글들의 길이의 평균
4. 해당 리뷰어가 남긴 평점과 다른 리뷰어들이 남긴 평점간의 편차
5. 이전에 쓴 글과의 유사도(TF-IDF와 Cosine similarity 활용)

Persistent homology 계산

Topological data analysis는 persistent homology를 계산하는 것에서 시작

다음과 같은 과정을 거침

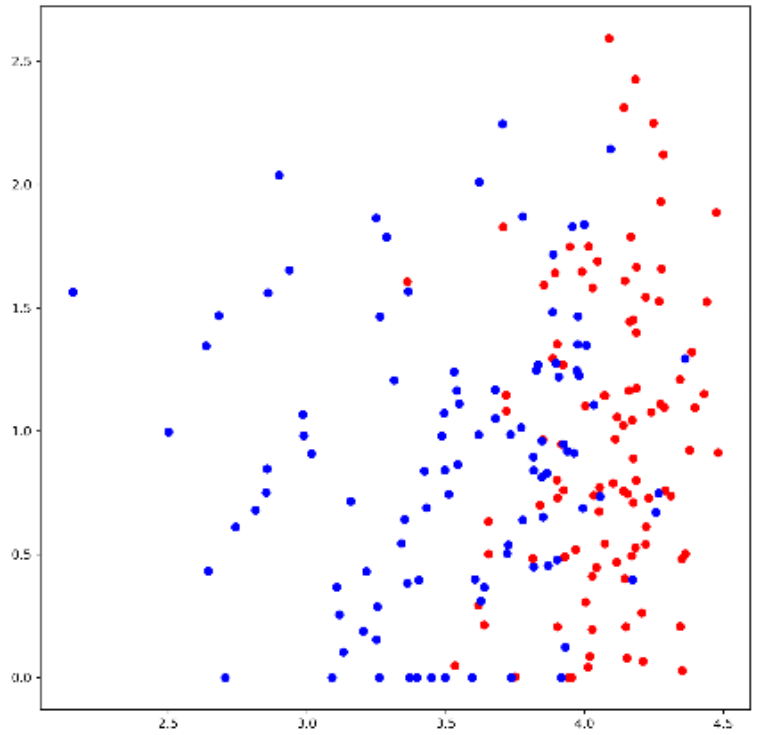
1. 용이한 분석을 위해 PCA를 활용한 차원축소
2. 호텔 데이터와 식당 데이터를 각각 정상 리뷰어 집단과 가짜 리뷰어 집단으로 분류
3. 각 집단으로부터 임의로 30명의 표본을 n회 추출
4. 30명으로 이루어진 각각의 집합들에 대해 persistent homology를 계산
5. persistent diagram으로 표현(오른쪽 그림)



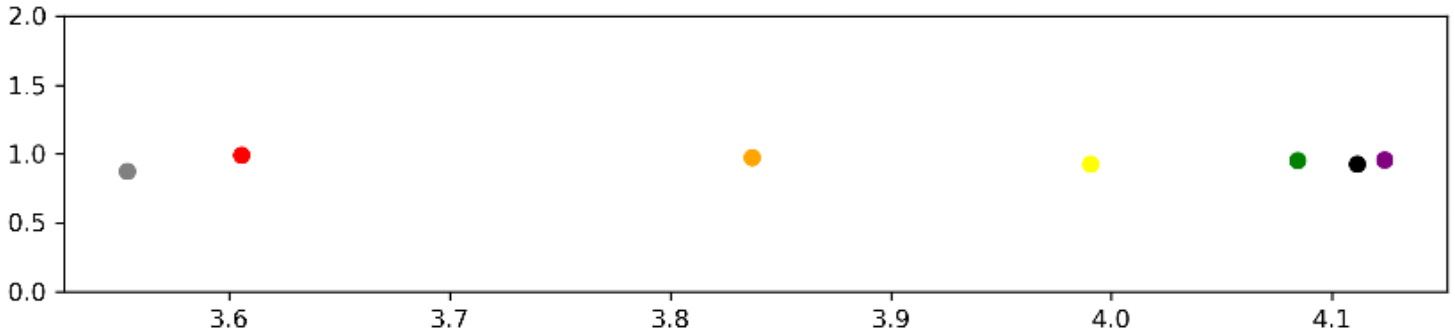
Persistent entropy를 활용한 분석1

Persistent homology를 계산한 이후에는 다음과 같은 과정을 수행

1. persistent homology를 계산한 각 집단에 대해 persistent entropy를 계산
2. entropy는 diagram에서 H0와 H1에 대해 각각 계산
3. 각 집단의 persistent entropy를 (H0, H1)으로 두고 시각화하여 확인(오른쪽 그림)
4. 파란색 점들은 가짜 리뷰어 집단의 entropy를 뜻하며 빨간색 점들은 정상 리뷰어 집단의 entropy를 뜻함.
5. 지표로 확인한 결과 가짜 리뷰어 집단의 entropy가 정상 리뷰어 집단의 entropy와 구별됨을 확인할 수 있음.



Persistent entropy를 활용한 분석2



앞선 결과를 바탕으로 다음 실험을 진행

1. 가짜 리뷰어 집단과 정상 리뷰어 집단이 섞여있을 때 각 집합들의 entropy 평균을 확인
2. 30명으로 이루어진 각각의 가짜 리뷰어 집단에서 임의로 n명의 가짜 리뷰어를 정상 리뷰어로 대체
3. n이 커짐이 따라 entropy의 분포의 변화를 확인
4. 위 그림은 n의 값이 5씩 커짐에 따른 각 entropy 분포의 평균을 시각화
5. 회색 -> 빨강 -> 주황 -> 노랑 -> 초록 -> 보라 -> 검정 순으로 분포가 이동함을 확인

분석 결과

» 분석 결과 다음과 같은 사실을 확인

1. 가짜 리뷰어 집단의 persistent entropy 값은 정상 리뷰어 집단보다 낮은 경향이 있음
2. 가짜 리뷰어 집단속에서 정상 리뷰어들의 비율을 늘릴수록 기하학적으로 정상 리뷰어 집단에 가까운 모양을 띄움

» 분석 결과를 응용하여 다음과 같은 사실을 추론

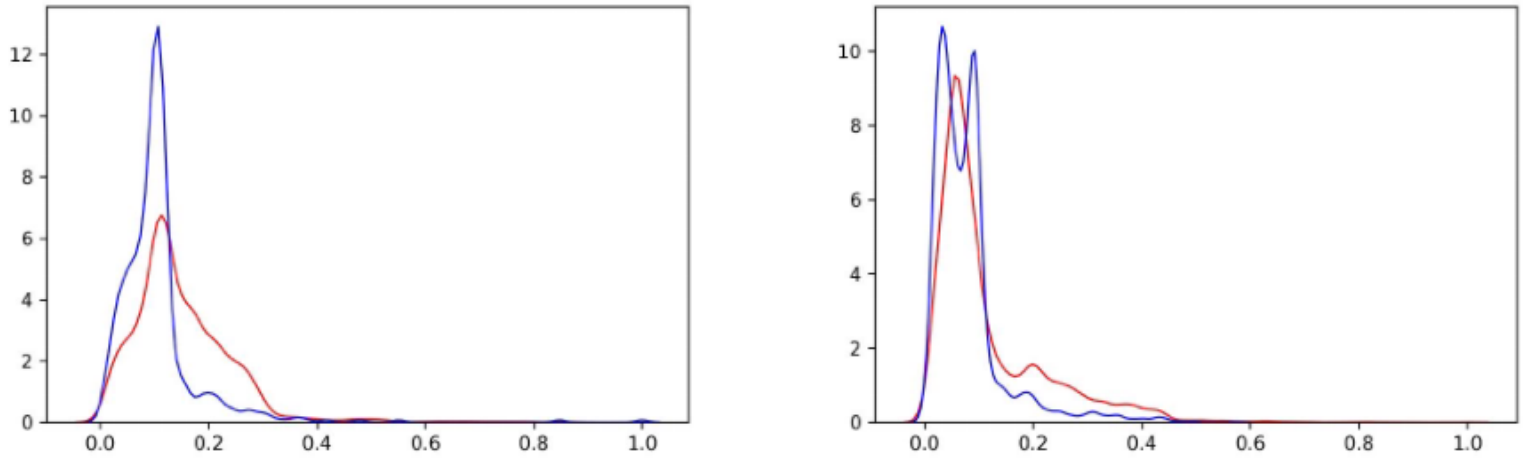
"가짜 리뷰어 집단과 정상 리뷰어 집단이 적절하게 섞인 집합에 임의의 샘플 r을 더했을 때 entropy의 변화가 정상 리뷰어 집단에 가까워 진다면 샘플 r은 정상 리뷰어 집단으로 추론할 수 있으며 가짜 리뷰어 집단에 가까워 진다면 샘플 r은 가짜 리뷰어로 추론할 수 있다."

예측 모델의 성능 향상을 위한 제안

» 분석 결과를 기반으로 다음과 같은 방법을 제안

1. 주어진 리뷰어 데이터를 train set, test set, 그리고 또다른 집합 T로 분류
2. T로부터 가짜 리뷰어 집단과 정상 리뷰어 집단이 적절하게 섞였으며 29명으로 구성된 여러개의 집합을 생성하고 persistent entropy를 계산
3. train set과 test set의 각 샘플을 위 집합에 추가하고 다시 persistent entropy를 계산
4. entropy의 변화량을 각 샘플의 새로운 특징 값으로 정의

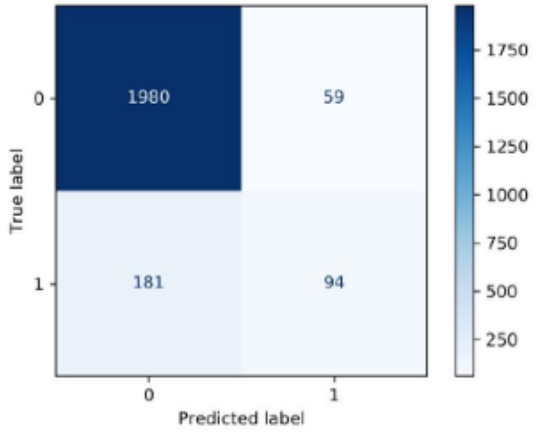
새로운 특징값의 확률분포

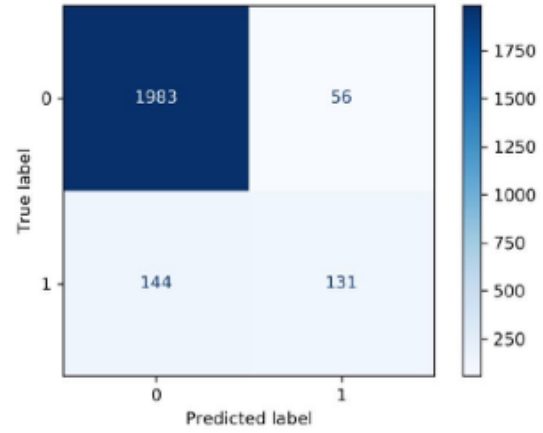


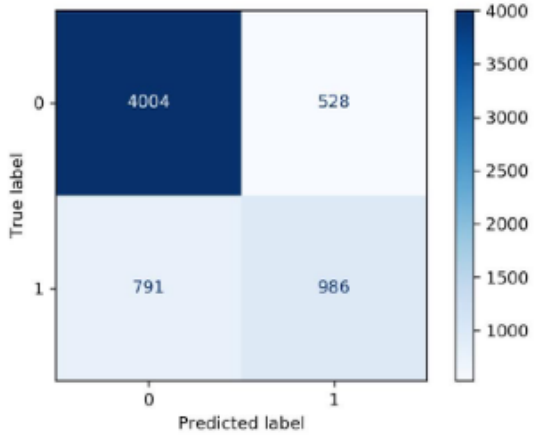
위 그림은 가짜 리뷰어와 정상 리뷰어에대한 새로운 특징 값에 대한 분포를 나타냄.
빨강색 선은 정상 리뷰어의 확률분포를 나타내며 파랑색 선은 가짜 리뷰어의 확률분포를 나타냄

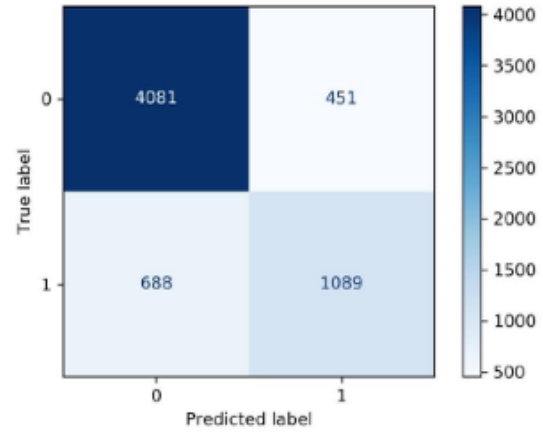
예측모델의 탐지 성능 요약

1.새로운 특징값이 모델의 성능향상에 도움이 되는지를 확인하기 위해 예측모델을 학습하고 결과를 비교
2. 예측모델로 Randomforest를 사용함
3. 오른쪽 그림은 test set을 예측할 때의 confusion matrix를 시각화한 것
4. 왼쪽의 2개의 matrix는 새로운 특징 값을 추가하기 전의 예측 결과,
오른쪽 2개의 matirx는 새로운 특징 값을 추가한 후의 예측 결과
5. 상단은 호텔 리뷰어의 예측 결과이며 하단은 식당 리뷰어의 예측 결과









총평

- 해당 연구의 방법론은 상대적으로 패턴이 명확할 것으로 보이는 집단과 그렇지 않은 집단을 분류할 때 범용적으로 유효할 수 있다고 생각됨
- 따라서 fraud detection이나 anomaly detection 분야등에 적용할 수 있을 것이라 생각됨.