



# 2조 최종발표

온라인 게임 채팅 데이터를 이용한 자연어 처리


고려대학교 정보보호대학원

2017572003 이재웅

2017572005 김혜민

2017572006 이수연

# Introduction - Title



채팅 내용과  
유저 타입과의  
관계

주제 1 : 유저 분류에 따른 채팅 내용

ex) 초보 vs 고수

피드백 반영 후 재실험

주제 2 : 채팅 내용에 따른 유저 분류

채팅 데이터를 이용하여 Clustering 진행 후  
각 Cluster 별 유저 특징 추출

새로 진행

# 주제 1

-classification-

# Experiment - Methodology

## 데이터 전처리

사용한 라이브러리: KoNLPy, twitter class

## 정규화

ex. 입니달ㅋㅋ → 입니다ㅋㅋ

## 불필요한 대화 내용 제거

같은 문장 중복 시 제거  
ASCII 코드 제외 (영어, 숫자, 특수문자)  
-님 제외 (사람을 부르는 호칭)

## 불필요한 형태소 삭제

명사, 동사, 형용사, Korean Particle 제외 삭제  
욕설은 삭제하지 않음



**Data**  
(ID, Corpus)



**Preprocessing**  
(ID, {word, \*pos, Frequency})



**Classification**

twitter pos	
Noun(명사)	Conjunction(접속사)
Verb(동사)	Exclamation(감탄사)
Adjective(형용사)	Josa(조사)
Determiner(관형사)	PreEomi(선어말어미)
Adverb(부사)	Eomi(어미)
Alpha(알파벳)	Suffix(접미사)
Foreign(외국어, 한자)	Unknown(미등록어)

# Experiment - Methodology

## 유저 분류 및 라벨링

채팅을 500자 이상한 유저 중 게임에 적응 완료한 유저와 게임에 적응하지 못한 유저 분류



적어도 1시간이상 게임을 플레이 했으나  
접속 일수가 3일 이하인 유저  
(이탈)

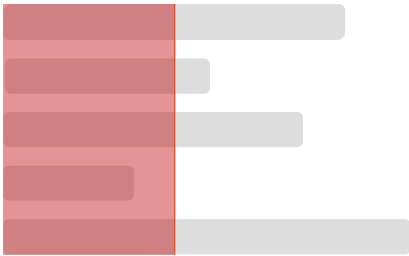


10일 이상 게임을 플레이 했으며  
레벨이 40 이상인 유저  
(고레벨)

# Experiment - Methodology

## Bag of Words (BOW)

1



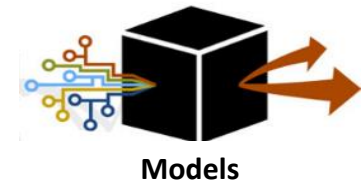
첫 접속 후 1일(2일, 3일) 동안의 대화 추출  
(데이터 공평성)

2

	w1	w2	w3	...	w <sub>m</sub>
U1	0	0	0	...	5
U2	0	1	0	..	2
U3	0	2	0	...	0
...	...	...	...	...	...
U <sub>k</sub>	2	1	0	...	0

Bag of Words 생성 및 차원 축소

3



분류 알고리즘을 이용한 성능 평가

# Experiment - Results

## Bag of Words (BOW) 기반 Classification 결과 (Accuracy)

Train set과 Test set을 7:3 비율로 나눈 뒤 평가, 모든 알고리즘은 scikit 라이브러리의 디폴트 설정 사용  
TF 10 미만 단어 삭제

	1일 대화	2일 대화	3일 대화
고수	279 (37.2%)	392 (43.36%)	438 (46.11%)
초보	471 (62.8%)	512 (56.64%)	512 (53.89%)
Number of Terms	3,836	5,679	7,280
SGD	69.33%	71.69%	86.11%
Naïve Bayes	<b>71.11%</b>	<b>84.19%</b>	<b>87.50%</b>
Logistic Regression	70.22%	79.04%	84.03%
<b>Average</b>	70.22%	78.31%	<b>85.88%</b>

많은 일 수의 대화를 기반으로 피처를 만들수록 정확도가 증가

# Experiment - Results

Bag of Words (BOW) 기반 Classification 성능 비교

채팅이 아닌 인게임 변수를 이용하여 classification 모델 생성 이후 성능 비교

## 사용 변수 (총 11가지)

퀘스트 완료 수, 사망 수, 채집 수, 개인 인던 입장 수,  
소셜 인던 입장 수, 레벨, 길드 활동 수, 친구 활동 수, 범죄 행위 수

	1일 대화	1일 활동
Number of Features	3,836	11
SGD	69.33%	65.59%
Naïve Bayes	71.11%	75.33%
Logistic Regression	70.22%	82.95%
<b>Average</b>	70.22%	<b>74.62%</b>

유저의 활동량을 직접적으로 알 수 있는 인게임 변수가 채팅 데이터보다 정확도가 높은 편이며 피쳐 수도 적음



# Experiment - Results

Bag of Words (BOW) 기반 Classification 성능 비교

1일 대화 예측과 1일 활동 예측 비교

## 각 분석 Confusion Matrix

1일 대화	예측 고수	예측 초보
실제 고수	53	28
실제 초보	34	107

1일 활동	예측 고수	예측 초보
실제 고수	69	12
실제 초보	40	101

종합	대화 고수 라벨링	대화 초보 라벨링
활동 고수 라벨링	65 (44)	44 (25)
활동 초보 라벨링	22 (9)	91 (3)

( )안의 숫자는 실제 고수 유저 수

# Experiment - Results

Bag of Words (BOW) 피쳐 중요도

Logistic Regression Coefficient 값 Top 100의 일부



날아감	탱이될런
헉있음	부탁드려
까자	원лак
싸우냐	병임
쫓깐	오섯음
피똥	꺼정
개쩌는갈	또찾어
개색허들	이속
개쉬움	몰랐찌



ㅇㅁㅇ	팟죤
어오렉	진행하구
다되섯어	도베르만
버스	신고
말하세	비싸젓
채굴	슴똥
약초	부탁할
님신	안심하시
암채짓하	사람죽엿으

초보 유저 대화에 욕설이 더 많음

고수 대화에는 아키에이지 게임 내 독특한 시스템인 '범죄' 관련 용어가 많이 등장함  
채굴이나 약초와 같은 게임 용어가 많이 등장함

# Conclusion

1

## 인게임 피쳐 사용과 유사한 정확도

실제 이탈 연구에 많이 사용되는 인게임 피쳐와 채팅 피쳐가 분류 면에서 유사한 성능을 보임  
단순한 게임 활동량으로 알 수 없는 이탈을 예측할 수 있음  
-> 게임 활동을 많이 한 유저가 무조건 이탈을 하지 않을 것이라는 보장은 없음

2

## 직관적

인게임 변수에 비해 값이 직관적으로 이해됨  
게임의 흐름 및 분위기를 이해할 수 있음 (게시판 모니터링과 유사)



게임 채팅 데이터로 유저 분류가 어느 정도 가능하다!!

# 주제 2

## -clustering-

# Experiment - Methodology

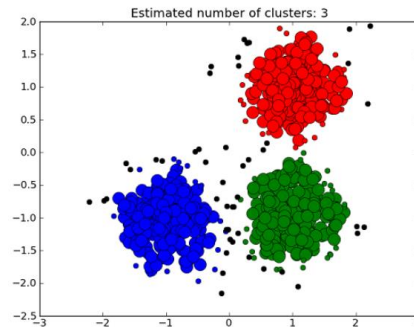
채팅 내용에 따른 유저 분류

1

	w1	w2	w3	...	w <sub>m</sub>
U1	0	0	0	...	5
U2	0	1	0	..	2
U3	0	2	0	...	0
...	...	...	...	...	...
U <sub>k</sub>	2	1	0	...	0

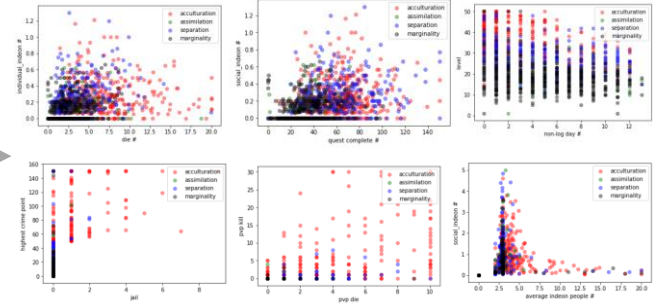
유저 별 벡터 추출

2



유저 클러스터링 진행

3



클러스터 별 유저 특징 파악

# Experiment - Results

앞서 생성한 단어 단위의 vector를 사용하여 클러스터링 진행

TF로 차원 축소 후 진행  
전체 유저 수 2,917명

1

## K-means (k=4)

Cluster 밀도의 차이가 심함

Clustered Instances

0	1	( 0%)
1	2252	( 77%)
2	82	( 3%)
3	582	( 20%)

2

## DBscan

vector들이 뭉치지 않아 전부 outlier로 계산됨

group counts

0	-1	2917
---	----	------

3

## Hierarchical

cluster가 생성되지 않음

Clustered Instances

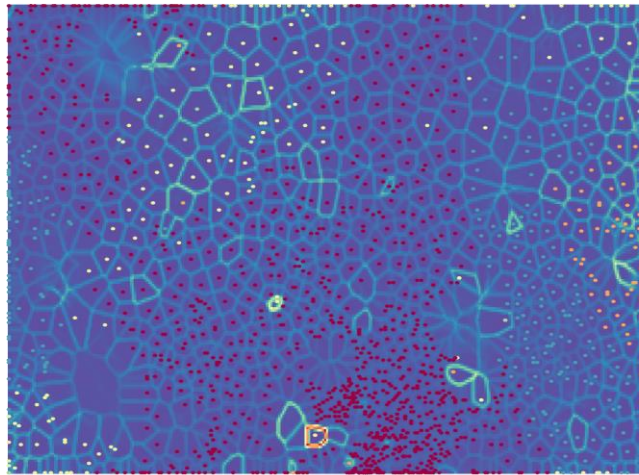
0	2916	(100%)
1	1	( 0%)

# Experiment - Results

## 4 SOM(Self-Organizing Maps)

유저가 말한 단어의 기반으로 SOM 클러스터링 진행

단어 50개 이상 유저

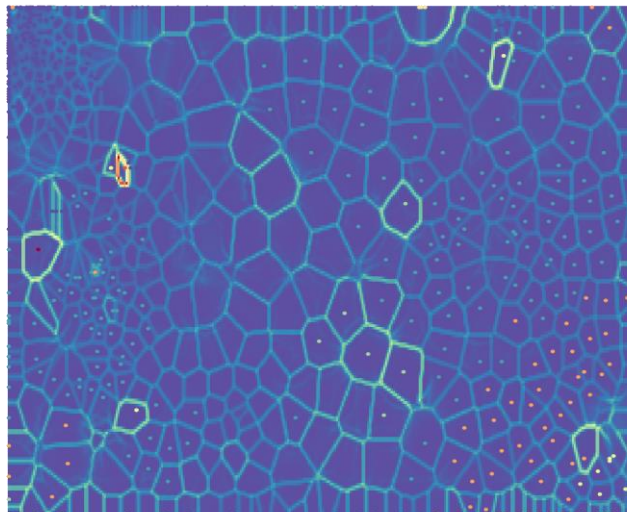


SOM클러스터링 결과

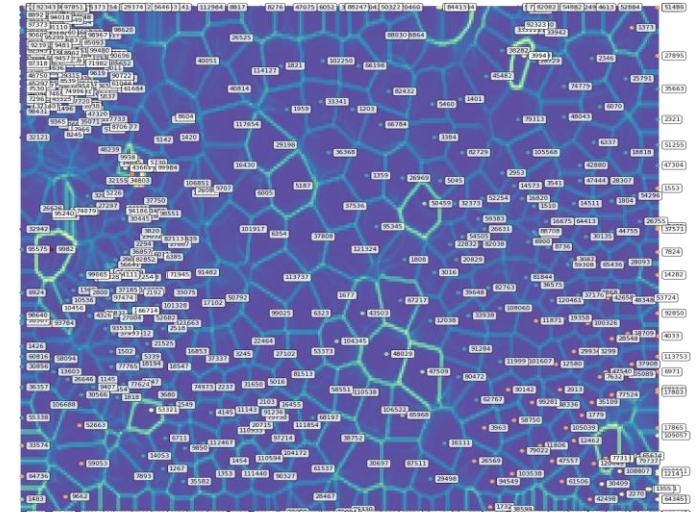


유저 아이디

단어 100개 이상 유저



SOM클러스터링 결과



유저 아이디

# Experiment - Results

## 4 SOM(Self-Organizing Maps)

클러스터링 결과 기반으로 유저의 채팅을 살펴보았지만 별다른 특징이 안들어남

yellow

['아직', '애드', '디젯', '오심', '에프', '제발', '컴퓨터', '자', '백날', '백프로', '쓰세', '진행했', '님꺼임', '고사', '청', '요깅', '파이팅', '있느', '타오르는', 'ㅋㅋ', '멤버', '어딴거', '의소', '풀힐', '그모', '드감', '밀기', '올스텝', '저가서', '키감', '동물학대', '조', '조질자식', '삼각형', '길드초대됨', '힐임', '지전임', '테러리스트', '답답함', '함선관', '쌀뭉', '글픽더좋게센', '쫓망', '만족', '글픽용', '낮제', '돌려줄', '새끼', '로맨스', '일본도', '일과', '남게하는', '미젠빨', '무등', '하루엣', '스페아습', '멤시씨', '미춘윤갈', '락팻', '컷말주세요뿌임', '파셔', '내가컨', '골좁봐', '회복됨', '증가하구', '질렸', '안되겠음', '백배', '달리냐', '지케보',

red

['오심', '제발', '갓어영', '시땀시', '상항', '마', '소중한', '면하는', '잡은', '활주', '절로', '복오너', '복실', '무시한', '님하', '만주', '무끄러워', '지전임', '니꼴', '일등', '할리', '세이프', '딜차', '때려주', '일본도', '구분하자', '년겜', '기릿', '소변', '우월밀', '라파엘로', '내가컨', '우쭈쭈서', '등했었음', '오르점', '올라탐', '짹았', '매춘부', '컷어', '피드', '그르데노', '안갯음', '하샤', '권신리', '하인어딴', '아구렸', '너렙', '몬료', '시작하신

mint

['오심', '대충', '제발', '컴퓨터', '아영', '상항', '니뒤로', '모션', '쓰세', '고사', '정문', '황폐', '또죽', '바구군', '드라이브', '오카마', '소중한', '격앙', '잡은', '신촌', '앞이닌까', '하나비', '완함', '게발', '라디온', '없을꼴', '프로야', '털거', '안텀', '역작', '점박이', '원루브크라프스', '애꽃', '가져가서', '때쟁', '아이폰', '내조놈', '다지엇', '토끼풀', '알려졌', '유리할', '애하', '페멘타킬', '어딴더', '한철', '불완', '타레기여어', '쌀임사', '완하심', '자라노', '허락할', '박지호', '딩몬', '상일락', '렙정', '야지전', '들켰나봄', '낯두', '속하잖', '구등', '무나서', '썩워', '겐쳐', '괘추하', '블써', '개센니', '부차

orange

['아직', '오심', '대충', '제발', '만나려', '메디슨', '갓어영', '다냐', '블루', '기록', '여러분', '글쿤', '어중간', '참가자', '황폐', '통발', '우어', '복학', '모래톱', '전맵', '빨간색', '타오르는', '맹아', '원기', '꼴', '똥개', '입지', '아녕', '어딴거', '의소', '제작하실', '경기', '사슬갑옷', '음탕', '구', '옥해도', '망하는', '강호동', '북쪽', '컷할게', '윙글', '질', '포함된', '렙벨', '내려지', '만주', '님두졌음', '조질자', '형', '트래픽', '레어', '시녀', '두마', '옛날', '라차', '조놈', '워', '개새키', '가장딜체', '홀림', '거켜봐', '안녕애드', '페드', '똥똥하냐', '미친놈', '구관', '쌀뭉', '글픽더좋게센', '내꺼팻', '레새끼', '방어율', '포르놈', '커먼', '루키', '우미', '케삭했

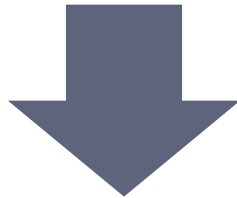


# Experiment - Methodology

오타가 많다

BOW 벡터는 차원이 크고 sparse 하다

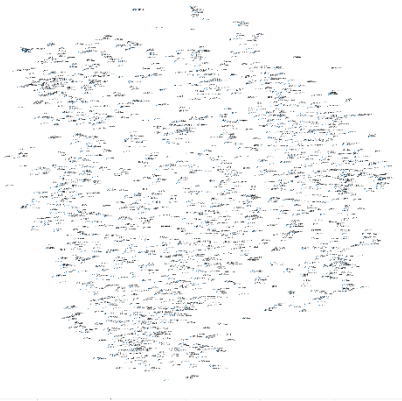
TF로 차원을 축소하는 것도 한계가 있다



word2vec의 결과를 클러스터링 하자

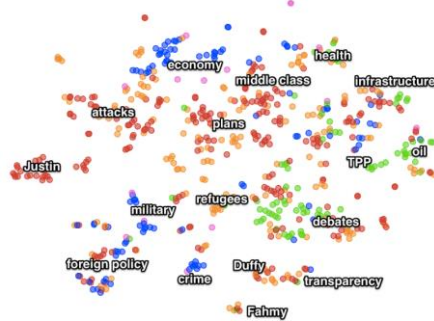
# Experiment - Methodology

1



word2vec 생성

2



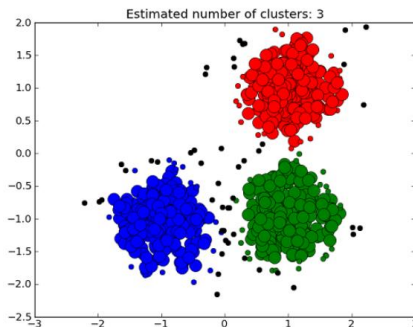
단어 클러스터링

3

	c1	c2	c3	...	cn
U1	1	5	12	...	5
U2	2	1	16	..	2
U3	4	2	2	...	0
...	...	...	...	...	...
Uk	2	1	0	...	0

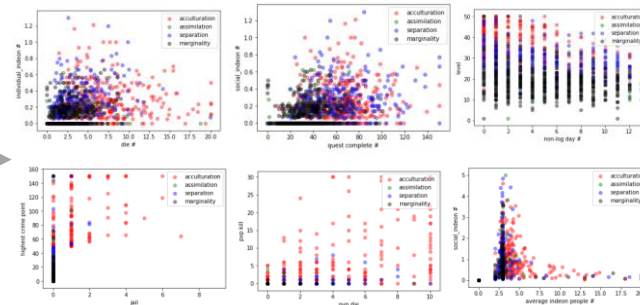
유저 별 단어 클러스터 벡터 생성

4



유저 클러스터링 진행

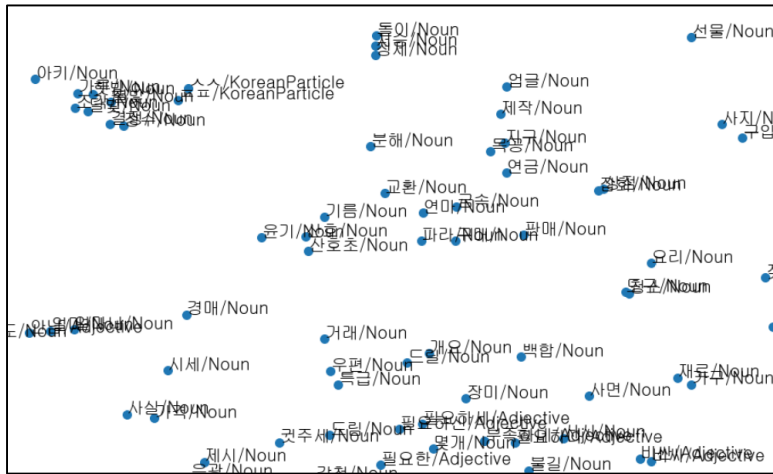
5



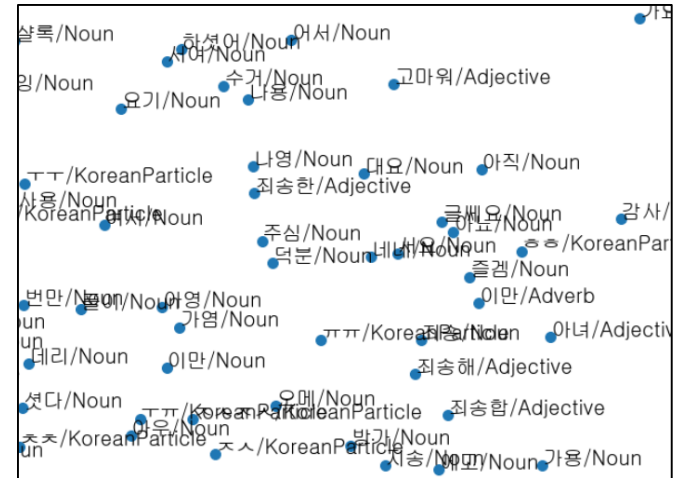
클러스터 별 유저 특징 파악

# Experiment - Results

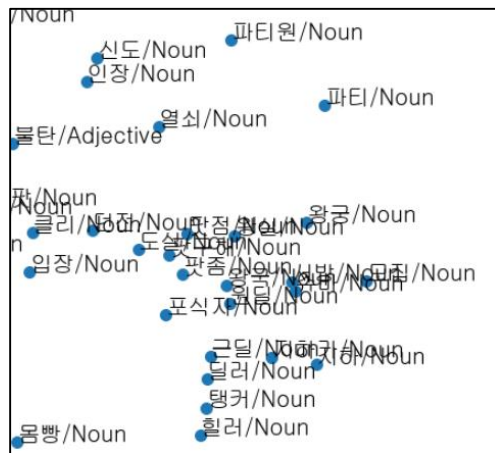
word2vec 결과



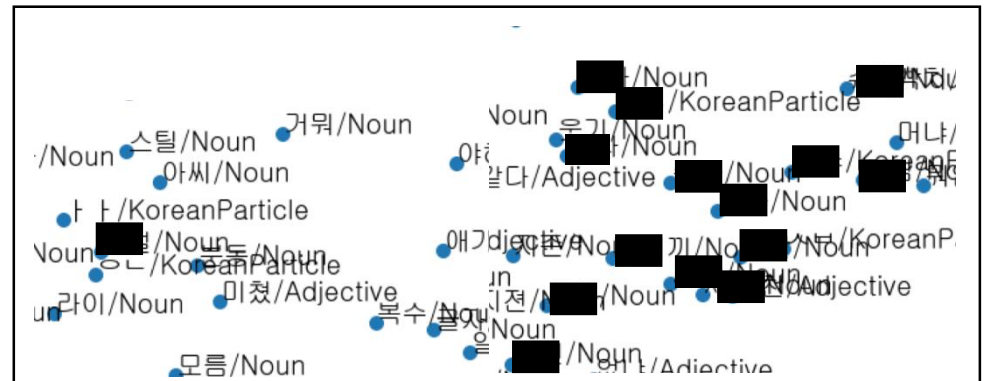
거래 & 채집 & 생산 관련



정중한 대화



파티 관련



욕설 및 부정적 대화

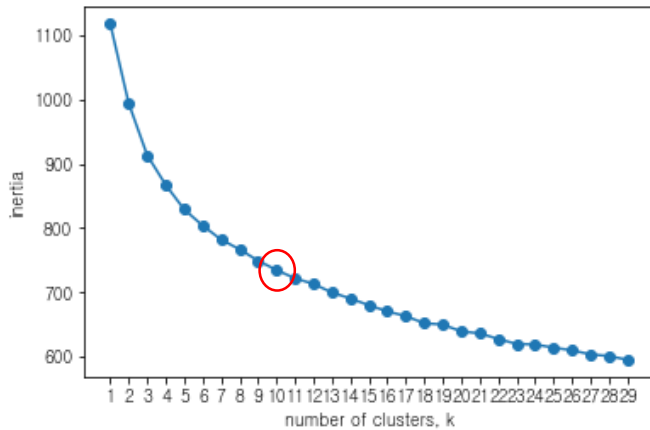
# Experiment - Results

word2vec 클러스터링 (K means)

	v1	v2	v3	...	Vn
W1	-0.1	0.3	0.12	...	0.55
W2	-0.5	0.1	0.16	..	0.2
W3	0.2	-0.2	0.2	...	0
...	...	...	...	...	...
짜	0.2	0.1	0	...	0

word2vec

K means

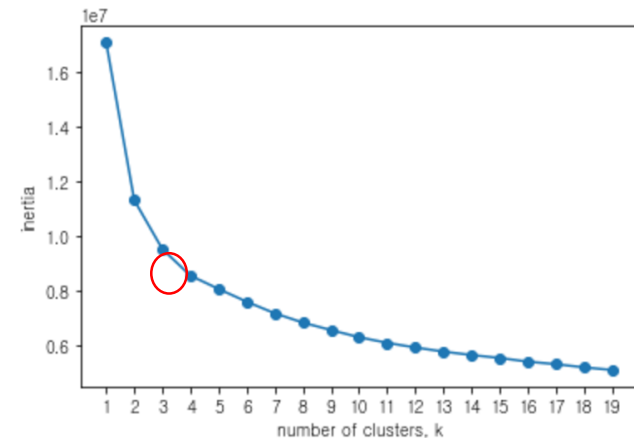


토픽 클러스터링 ( k = 10)

	T0	T1	T2	...	T9
U1	1	5	12	...	5
U2	2	1	16	..	2
U3	4	2	2	...	0
...	...	...	...	...	...
Uk	2	1	0	...	0

유저 별 토픽 단어 사용 횟수 벡터 생성

K means



유저 클러스터링 ( k = 4)

# Experiment - Results

토픽 클러스터 내 단어

Topic 0 (레이드, 항해)

항구	공대
해적	나룻배
원정	쾌속
대륙	노르드
레이드	가나다라
약탈	영지
크라켄	범선
헤엄	바다

Topic 5 (파티, 부정적)

어렵	파티원
ㅠㅠ	감사
아프	몸빵
죄송	부탁드려
파티	근딜
ㅈㅈ	힘드네
——	선빵
ㅈㅇㅇ	파장

Topic 7 (욕)

거지	느려
강아지	임마
멘붕	힘듦
18의 변형	노예
삼시세끼의 변형	Me친의 변형
바보	짜증
이상한	같다
더럽	없나

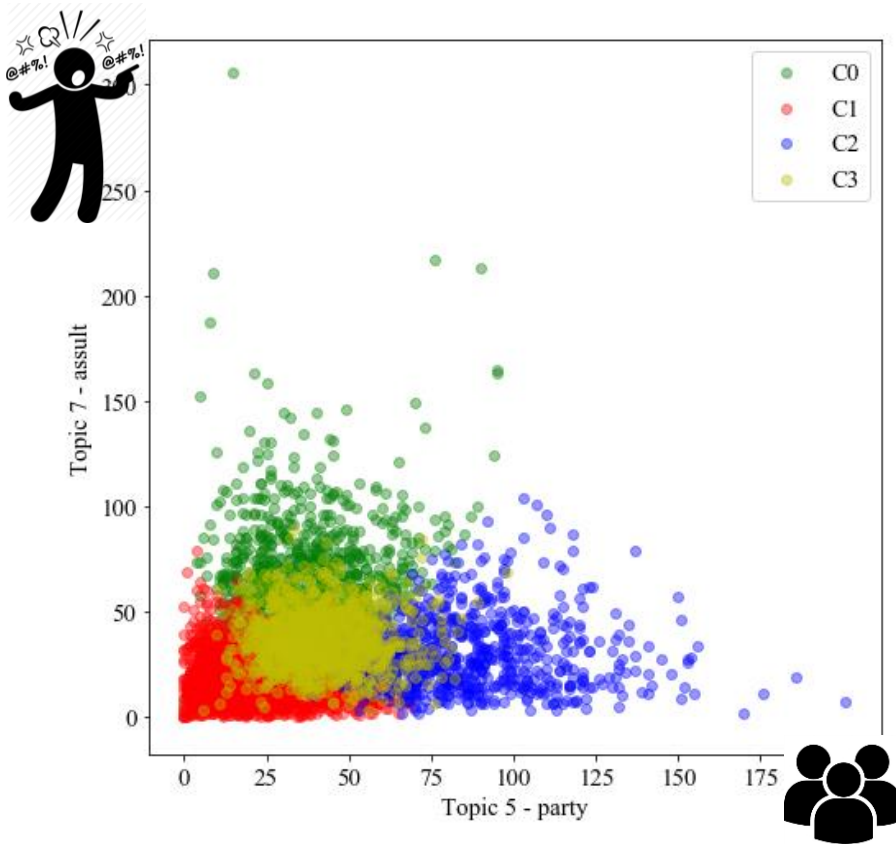
Topic 9 (거래)

얼마	거래
사고	필요해
팔고	상점
골드	통나무
쿿주세	ㅠㅠ
있으	ㅈㅈ
가격	비싸
교환	재료

	단어 개수	주제
Topic 0	137	레이드, 항해
Topic 1	109	직업, 장비
Topic 2	118	이동, 의문
Topic 3	123	방향, 위치
Topic 4	140	게임 콘텐츠, 평가
Topic 5	185	파티, 부정적
Topic 6	127	부사
Topic 7	126	욕
Topic 8	64	채집
Topic 9	112	거래
전체	1,241	

# Experiment - Results

각 유저 클러스터 이해 (K-means, k = 4)



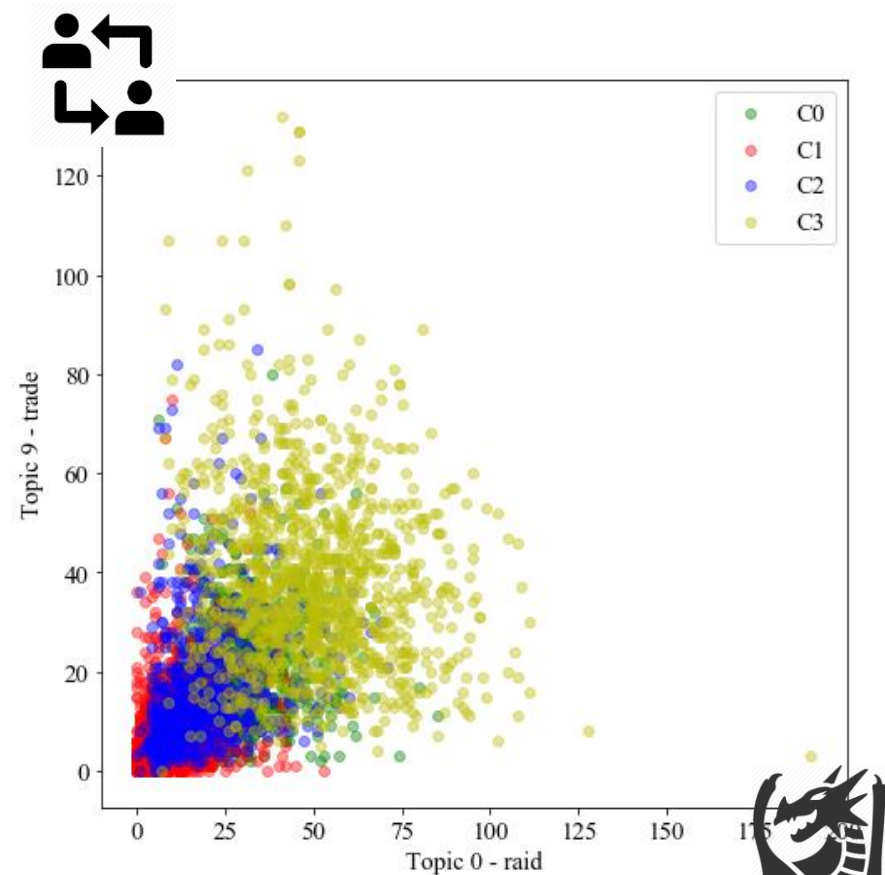
Topic

Topic 0 : 레이드, 항해

Topic 5 : 파티, 부정적

Topic 7 : 욕

Topic 9 : 거래



User cluster

C0 (초록색) : 욕설유저

C1 (빨간색) : 과묵한 유저

C2 (파란색) : 파티 + 부정적인 유저

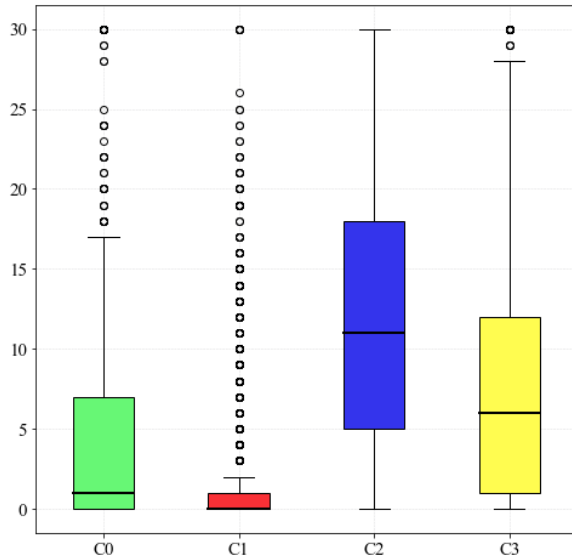
C3 (노란색) : 다양한 콘텐츠 이용 유저

# Experiment - Results

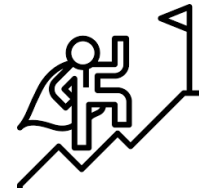
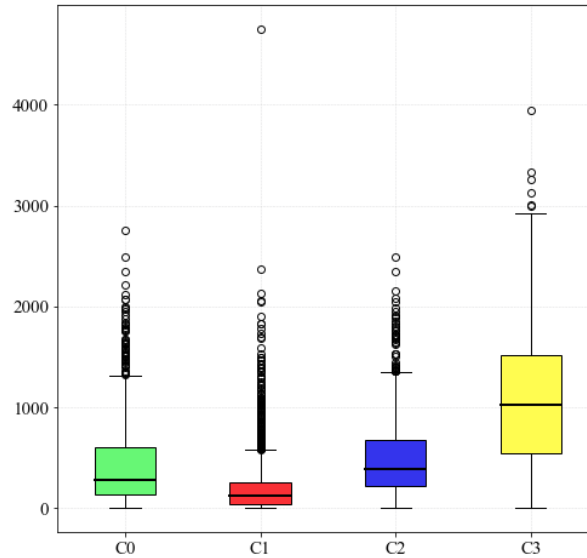
유저 클러스터 별 특징



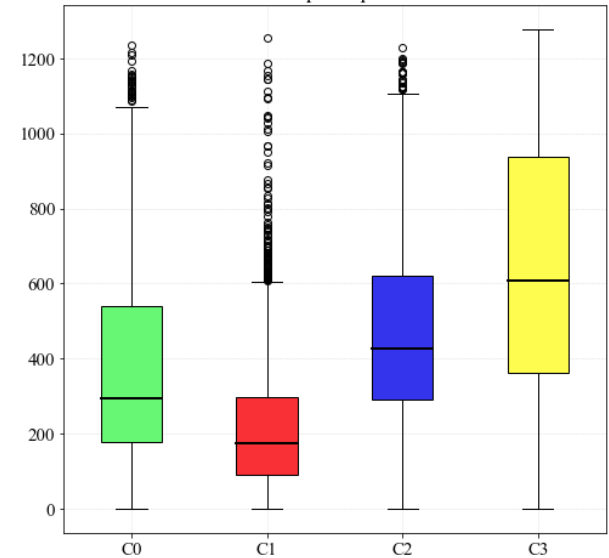
social indeon



harvest



complete quest



User cluster

C0 (초록색) : 욕설유저

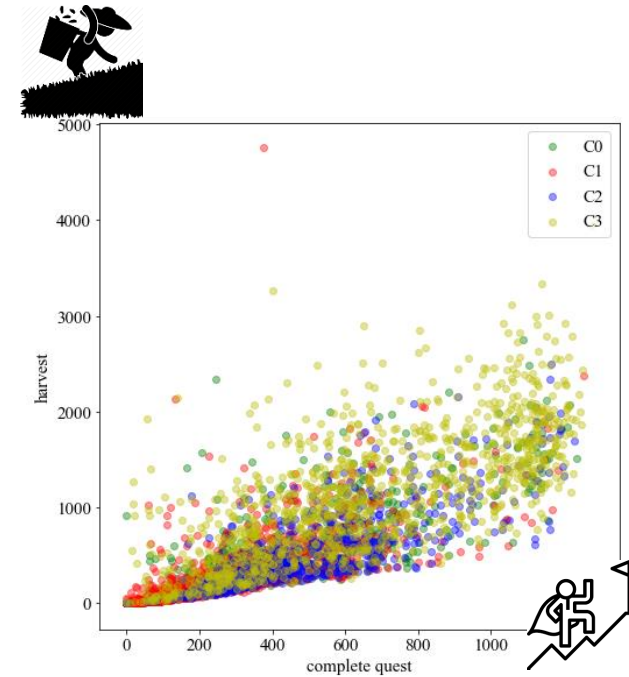
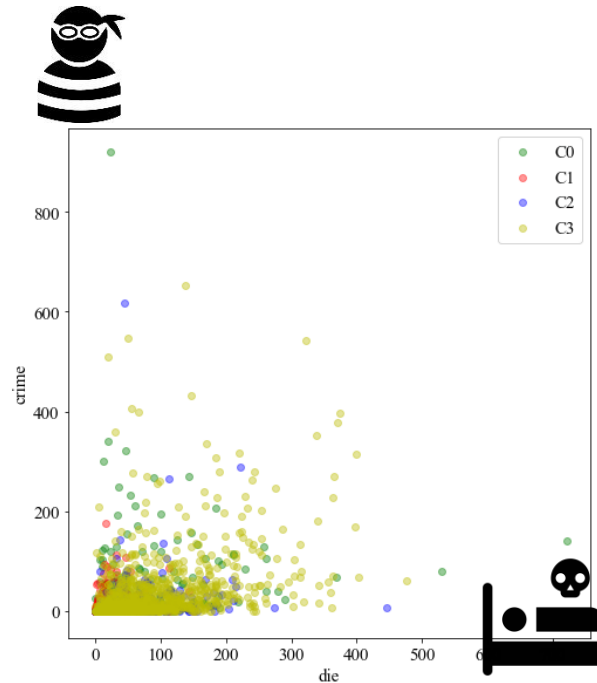
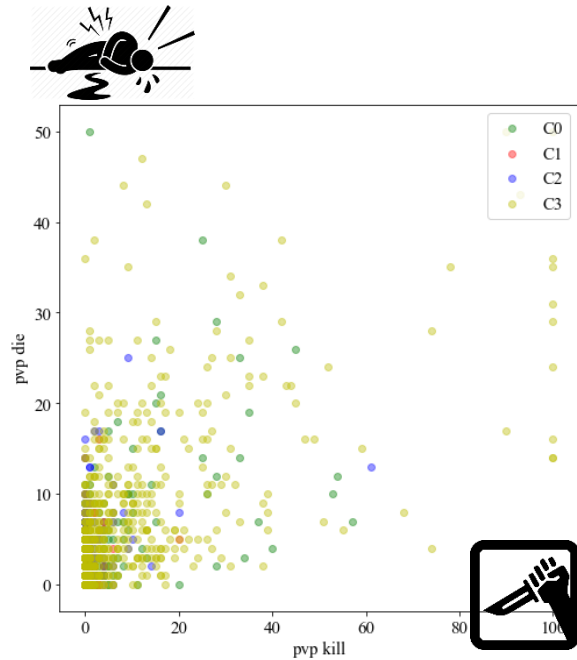
C1 (빨간색) : 과묵한 유저

C2 (파란색) : 파티 + 부정적인 유저

C3 (노란색) : 다양한 콘텐츠 이용 유저

# Experiment - Results

유저 클러스터 별 특징



User cluster

C0 (초록색) : 욕설유저

C1 (빨간색) : 과묵한 유저

C2 (파란색) : 파티 + 부정적인 유저

C3 (노란색) : 다양한 콘텐츠 이용 유저



# Conclusion

1

## 기존 clustering

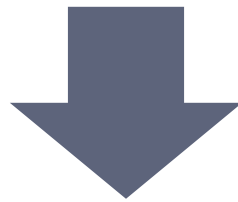
실..패..



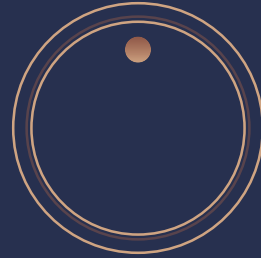
2

## Word2vec을 이용한 clustering

클러스터링된 단어쌍들이 대화 주제를 나타냄  
대화 주제에 따라 클러스터링이 가능



대화 주제에 따라 유저의 특징을 파악할 수 있다!!  
게임 전용 단어와 오타를 그룹화할 수 있는 가능성을 보여줬다!!



Thank you