

# 5TT

넷플릭스 분석

# INDEX

01. 주제 선정 이유

---

02. 데이터 수집 방법

---

03. 넷플릭스와 디즈니 작품 키워드 비교

---

04. 넷플릭스 작품 분석

---

05. 개발 후기 및 느낀 점

---

# N 01. 주제 선정 이유

[글로벌 OTT] 한층 치열해지는 美 OTT 시장 경쟁 …디즈니 삼총사 약진

“이러다 다 죽어” 치킨 게임으로 치닫는 OTT 전쟁

[WEEKLY BIZ] ‘오징어 게임’처럼… 성공 노린 OTT 산업, 167조원 시장서 모두 출혈경쟁

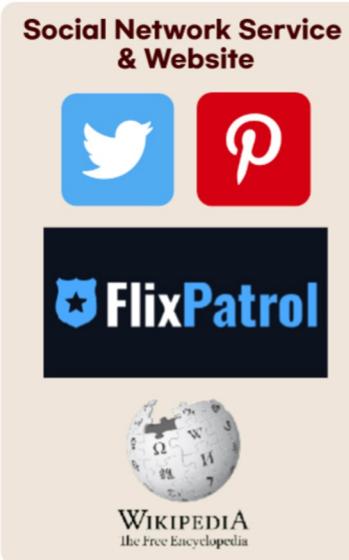
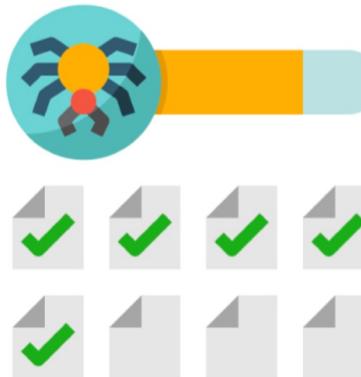


넷플릭스의 계획에 차질이 빚어진 것은 예상치 못한 막강한 경쟁자들이 시작  
어들었기 때문이다. 디즈니·픽사·루카스필름·마블 등 초호화 콘텐츠를 보..  
즈니가 2019년 11월 디즈니플러스라는 이름으로 스트리밍 사업에 뛰어들었고, 애  
플도 비슷한 시기 애플TV플러스를 출시했다. 유료 회원(아마존프라임)용 부가서  
비스 중 하나로 스트리밍을 제공해온 아마존 역시 2015년 이후 본격적으로 자체  
콘텐츠 제작에 뛰어들었다. 크고 작은 회사를 다 합치면 현재 미국 내 OTT 업체  
200개가 넘는다.

# N 02. 데이터 수집, 저장 및 정제

## 데이터 수집

### Web Crawling



- 트위터: 넷플릭스 관련 키워드 (2019.03 - 2022.03)
- 핀터레스트: 넷플릭스, 디즈니플러스 작품 포스터
- FlixPatrol: 월별 Top10 영화, TV show 목록 (날짜, 제목, 순위, 점수)
- Wikipedia: 컨텐츠 검색결과

## 데이터 저장 및 전처리

### DBMS - MySQL



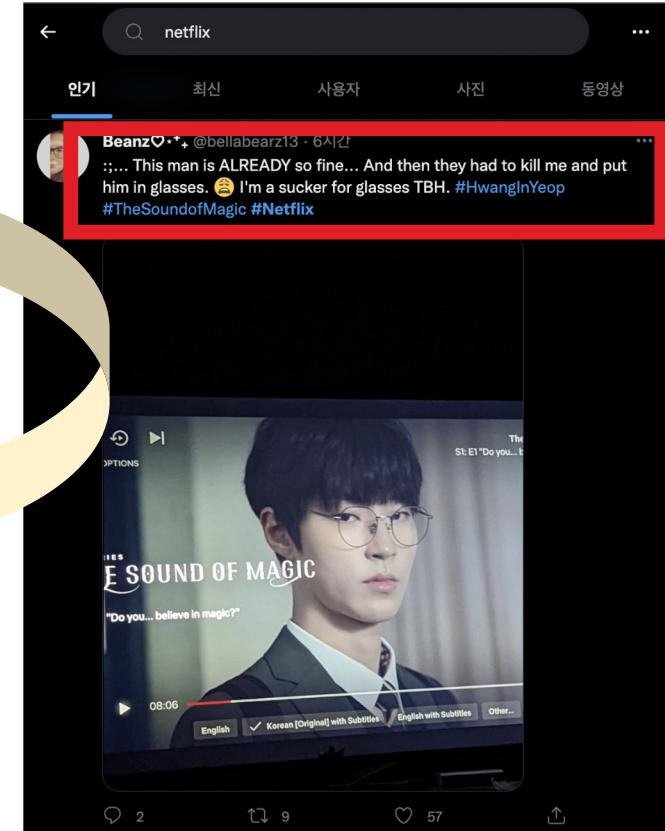
# N 02. 데이터 수집 방법 - 트위터 크롤링

## 넷플릭스 키워드로 트윗 크롤링

```
!pip install snscreape
import snscreape.modules.twitter as sntwitter

tweets_list = []
for i, tweet in enumerate(sntwitter.TwitterSearchScrapers('Netflix since:2020-12-12 until:2020-12-13').get_items()):
    if i>1000:
        break
    tweets_list.append([tweet.date, tweet.content])
```

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
12	2019-09-0 Go check out Carol and Tuesday on @netflix you'll love it I promise https://t.co/Y9nrDo38Dq													
13	2019-09-0 Netflix Netherlands pf adc a 5 temp de How to get away with murder pleaseeeee													
14	2019-09-0 They should just release all the episodes like a Netflix series													
15	2019-09-0 I shouldn't be allowed to have Netflix I don't understand moderation. Someone log on and change my password ?螺旋?螺旋?螺旋													
16	2019-09-0 @offstage_news @netflix @spookperson When you cancel your membership, it should give you a comment box to comment on why you're leaving.													
17	2019-09-0 @ruth_elena @FTPForAlways @ABCNetwork @shondarhimes @netflix Would really like Shonda to do that													
18	2019-09-0 @iTheuy @_soamii Netflix?													
19	2019-09-0 Netflix & Sunday Dinner vibes #alone #singlelife #qualitytimewithself ?螺旋?螺旋													
20	2019-09-0 @Alyssa_Milano Can't wait to see you on Netflix. Alyssa Milano at Epstein's Island.													
21	2019-09-0 @machacon_ So essentially you?螺旋 wanna Netflix and chill right?													
22	2019-09-0 I hope Louie CK has a new Netflix special in the near future. The reaction from SJW Twitter would be gold. https://t.co/zv7xqCffWE													



## N 02. 데이터 수집 방법 - 트위터 크롤링

```
# 배열에서는 split을 사용할 수 없기 때문에 문자열로 바꿔준다
string = ' '.join(tw_202003_txt['Text'].values)

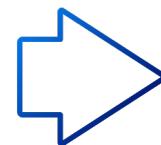
# 문자열로 바꾼 후 split 적용
string_token = string.split()

## 정규식으로 영어만 추출
import re
string_token_re = re.sub("[^a-zA-Z]", " ", str(string_token))

## (단어, 품사)로 나오게 하는 코드
from nltk.tag import pos_tag
nltk.download('averaged_perceptron_tagger')
x = word_tokenize(string_token_re)

# nn, np만 추출
df_list1_NN = df_list1[(df_list1['form'] == 'NN') | (df_list1['form'] == 'NNP')]
```

1. Tokenize 라이브러리 사용하여 트윗 토큰화
2. 정규식으로 영어만 추출
3. pos\_tag 라이브러리 사용하여 명사(Noun)만 추출

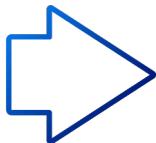


	word	count
0	list	160
1	halloween	227
2	netflix	27229
3	october	154
4	marifujarra	1
...	...	...
38412	filmlocation	1
38413	pzx	1
38414	wagc	1
38415	cleon	1
38416	hqh	1
38417	rows × 2 columns	

## N 02. 데이터 수집 방법 - 트위터 크롤링

〈월별 TOP20 중 일부〉

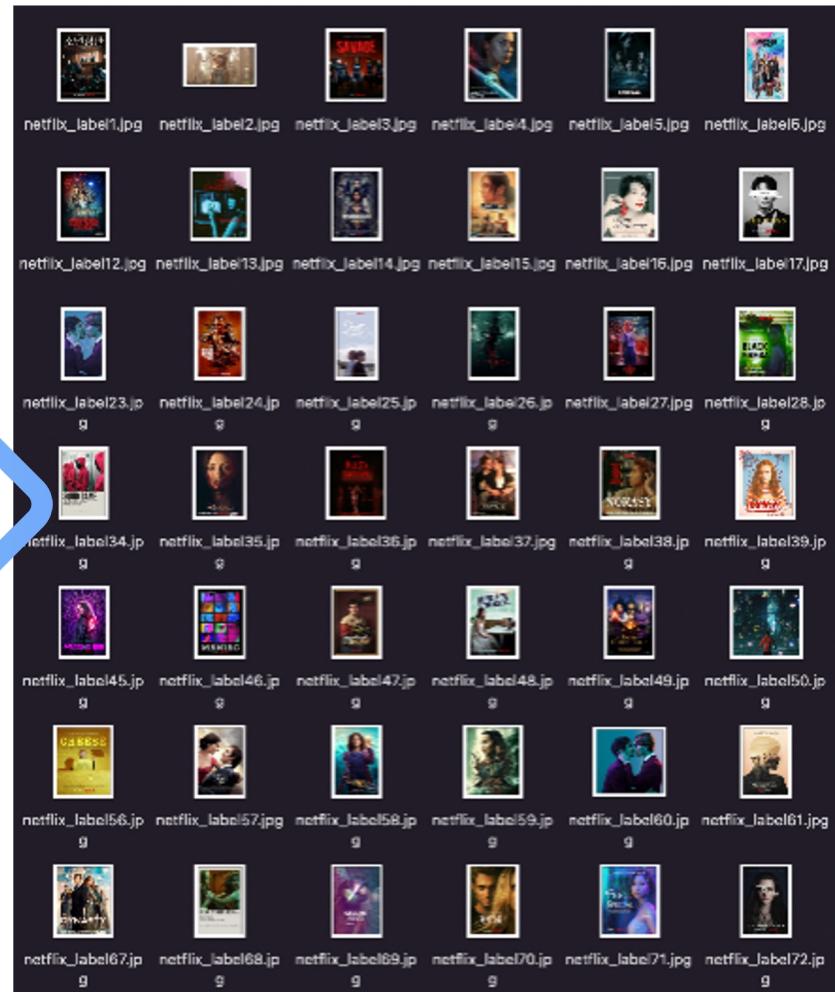
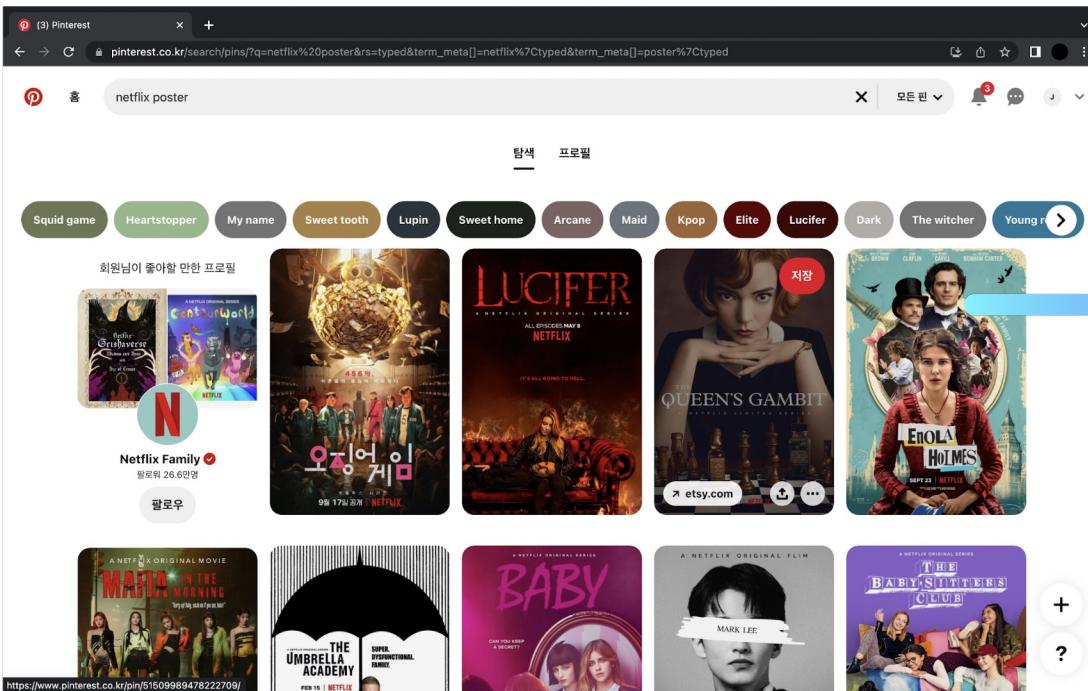
word	count	word	count
netflix	25590	netflix	26817
co	7345	co	8892
https	5364	https	6428
show	2036	show	1955
movie	1739	series	1740
series	1694	season	1698
season	1302	movie	1520
time	996	time	1111
xad	872	xad	1009
watch	793	watch	810
bebop	770	tv	775
youtube	731	disney	692
cowboy	695		
action	662		
tv	649		
amp	596		
day	561		
disney	551		



# N 03. 넷플릭스와 디즈니 작품 분위기 비교

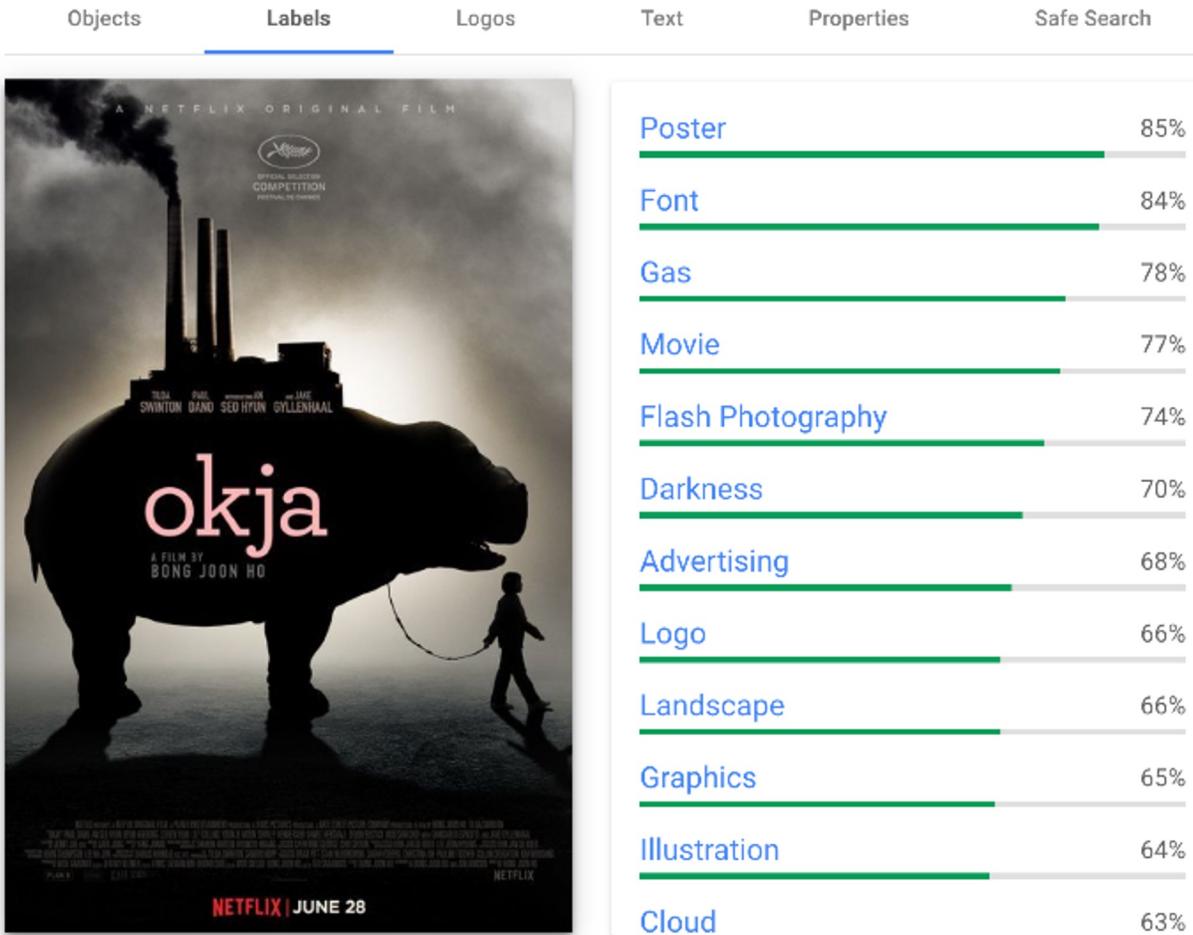
## 넷플릭스 & 디즈니플러스 포스터 이미지 크롤링

```
from src import PinterestScraper, PinterestConfig
```



# N 03. 넷플릭스와 디즈니 작품 분위기 비교

## 이미지 라벨 추출 – Google Cloud Vision

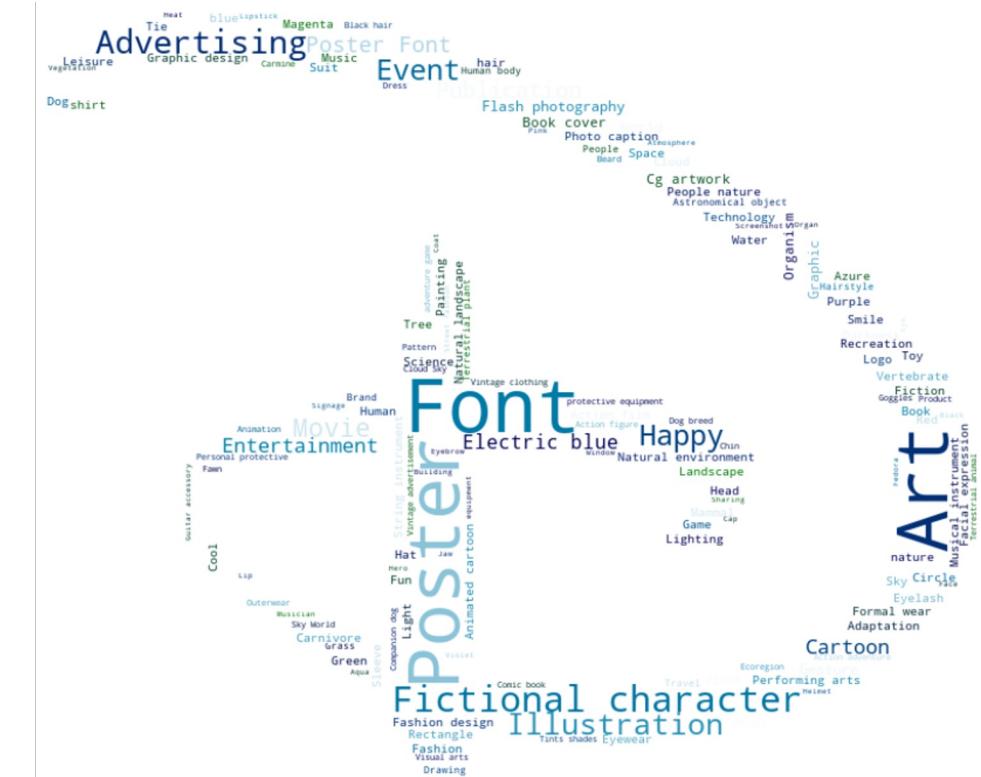
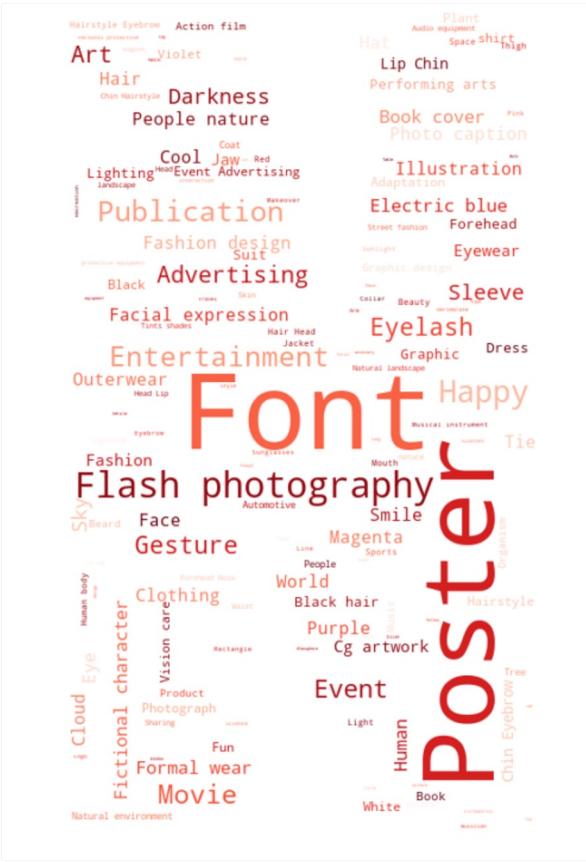


## 데이터프레임 생성

label	counts	
0	Font	171
1	Poster	151
2	Flash photography	112
3	Happy	78
4	Publication	68
...	...	...
312	Door	1
313	Daytime	1
314	Necklace	1
315	News	1
316	Kimono	1

# N 03. 시각화 - 워드클라우드

## 한계:



- 한번에 크롤링할 수 있는 이미지 개수가 (500장) 제한되어 있기에 충분한 데이터 확보 불가.

N



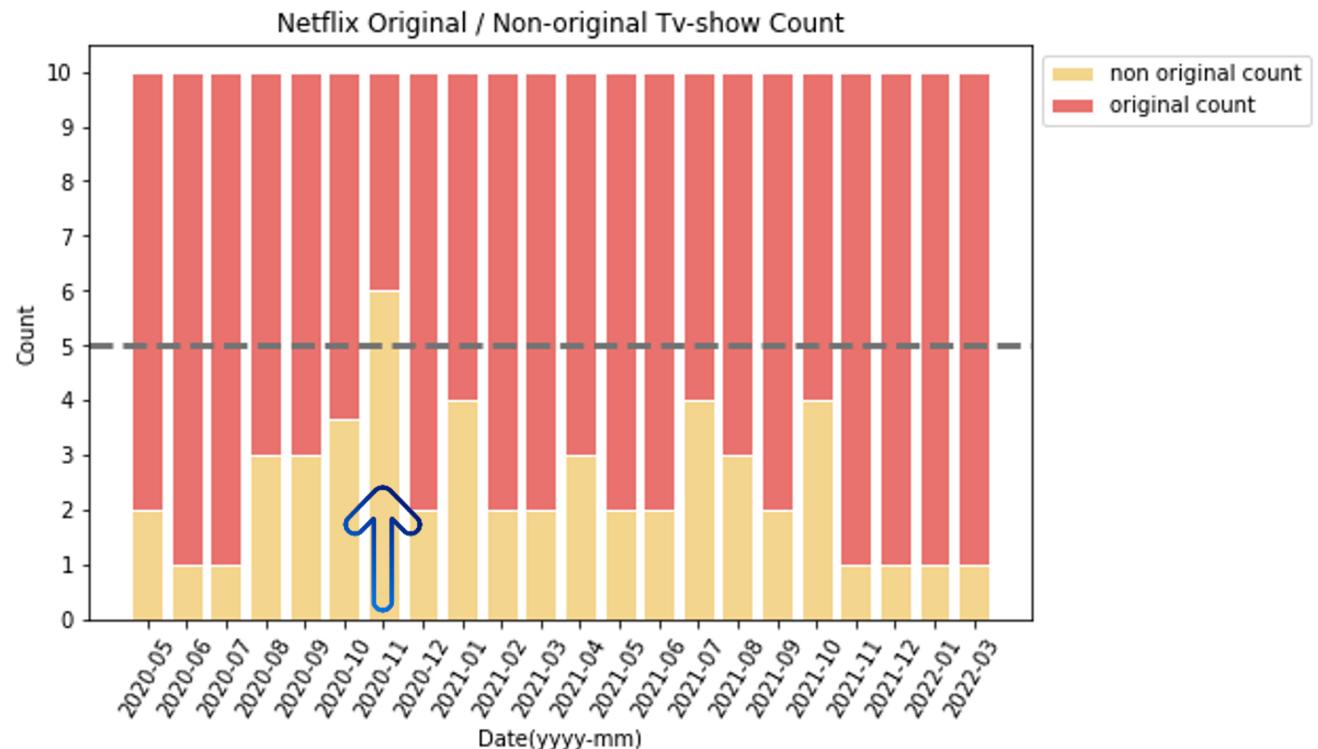
NETFLIX

넷플릭스 컨텐츠가 경쟁력을 갖추려면?

# N 04. 오리지널과 수급작 비교

## TV Show

Date	Title	Rank	Point	original
0 2020.05	Outer Banks	1	209	y
1 2020.05	Dead to Me	2	193	y
2 2020.05	Riverdale	3	123	n
3 2020.05	Avatar: The Last Airbender	4	116	n
4 2020.05	Ozark	5	111	y
...	...	...	...	...
216 2022.03	Worst Roommate Ever	6	107	y
217 2022.03	Bad Vegan: Fame. Fraud. Fugitives.	7	100	y
218 2022.03	Vikings: Valhalla	8	84	y
219 2022.03	Bridgerton	9	66	y
220 2022.03	CoComelon	10	60	n

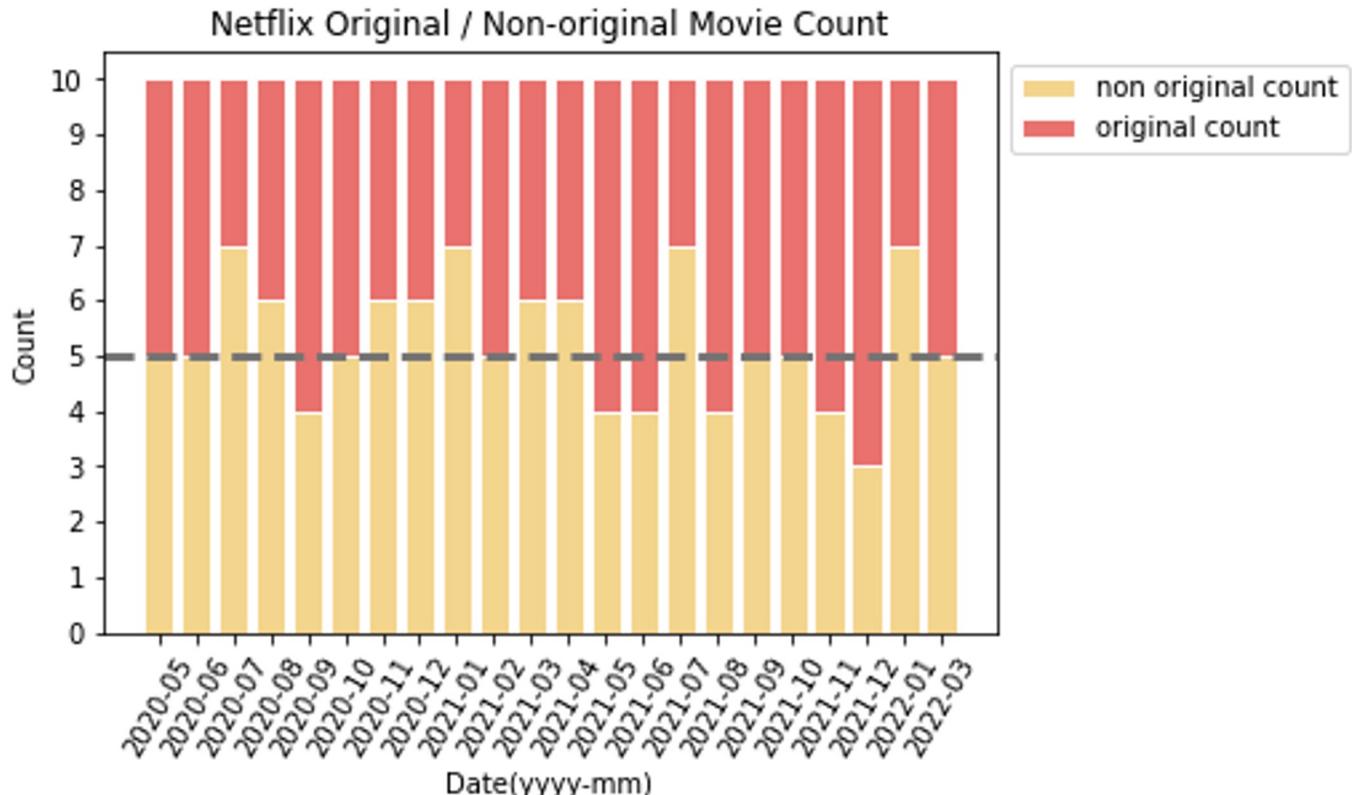


# N 04. 오리지널과 수급작 비교

## Movies

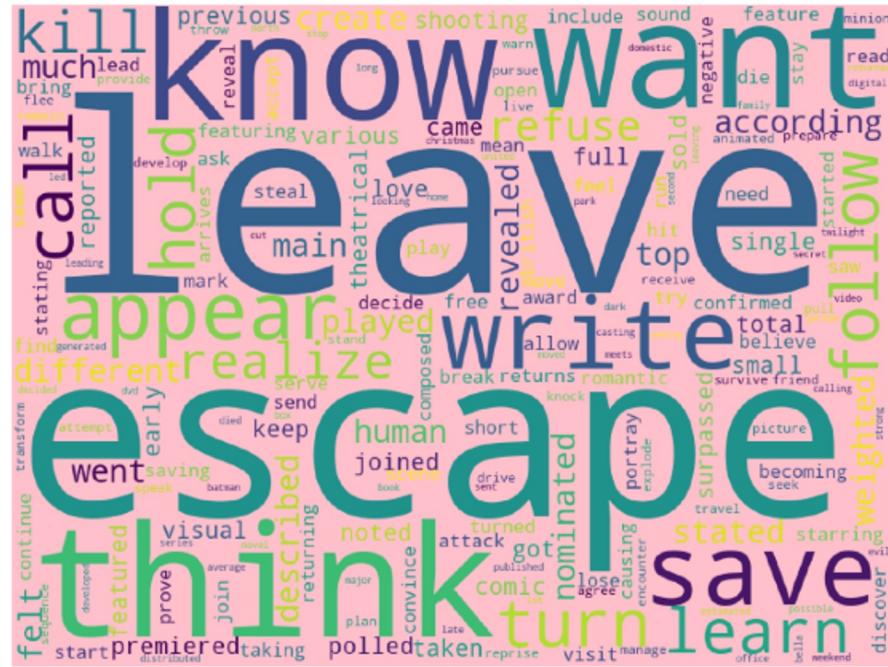
	Date	Title	Rank	Point	original
0	2020.05	The Wrong Missy	1	161	y
1	2020.05	Despicable Me	2	131	n
2	2020.05	Extraction	3	110	y
3	2020.05	The Willoughbys	4	105	y
4	2020.05	John Henry	5	93	n
...	...	...	...	...	...
215	2022.03	Black Crab	6	92	y
216	2022.03	Despicable Me 2	7	85	n
217	2022.03	The Weekend Away	8	81	y
218	2022.03	Rescued by Ruby	9	80	y
219	2022.03	Shooter	10	75	n

수급작 > 오리지널	
20년	7월 8월 11월 12월
21년	1월 3월 4월 7월
22년	1월



## N 04. 시각화 - 워드클라우드

# Movies



**〈Netflix Original〉**



Non-Original

**두 그룹 간 유사한 키워드 결과 = kill, escape, help**

# N TOP Movies on Netflix in March, 2021

1.		Yes Day	N
2.		Bigfoot Family	Animated / For Older Kids
3.		The Dark Knight	Superhero
4.		Deadly Illusions	— Thriller/ Abduction
5.		Operation Varsity Blues: The College Admissions Scandal	
6.		Parker	— Crime/ Action Crime
7.		Savages	— Thriller/ Drugs
8.		I Care a Lot	N
9.		Training Day	— Action/ Action Crime
10.		Moxie	N

## N TOP Movies on Netflix in December, 2020

1.		How the Grinch Stole Christmas	Fantasy/ Christmas
2.		The Christmas Chronicles: Part Two	N
3.		Ava	— Thriller/ Spies
4.		Peppermint	— Action/ Revenge
5.		The Christmas Chronicles	N
6.		A California Christmas	N
7.		Marauders	— Action/ Action Crime
8.		The Midnight Sky	N
9.		After We Collided	Romance/ Teens
10.		Triple 9	— Crime/ Thriller

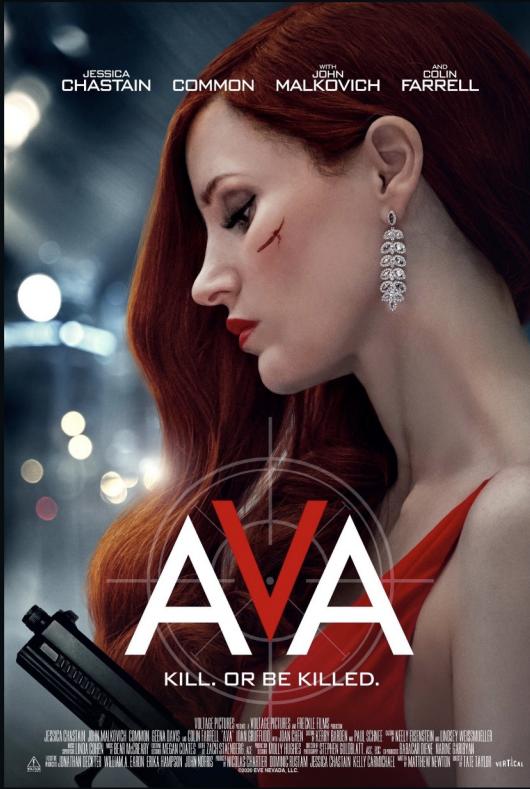
## N TOP Movies on Netflix in January, 2022

1.		Don't Look Up	N
2.		Just Go with It	Romance/ Cry movie
3.		The Royal Treatment	N
4.		Journey 2: The Mysterious Island	Adventure/ Dinosaurs
5.		Brazen	— Crime/ Crime Drama
6.		The Longest Yard	Comedy/ Football
7.		Munich: The Edge of War	N
8.		Charlie and the Chocolate Factory	Fantasy/ Fantasy Adventure
9.		Annabelle: Creation	— Horror/ Demons
10.		Tides	Movie

# N 앞으로 넷플릭스 오리지널은 어떤 장르를 제작하는 것이 좋을까?



Action



Crime



Thriller

# N 04. 분산 분석 -데이터셋 생성

## 월별 단어 & TOP10 작품 제목 교차 추출

**TOP10 제목 토큰화**

	word	date
0	me	202005
1	riverdale	202005
2	avatar	202005
3	airbender	202005
4	ozark	202005
...	...	...
390	fugitives	202203
391	vikings	202203
392	valhalla	202203
393	bridgerton	202203
394	cocomelon	202203

**월별 토큰화**

	word	count	date
0	house	191	202005
523	miss	36	202005
524	any	85	202005
525	dr	30	202005
526	stranger	79	202005
527	things	64	202005
528	stream	123	202005
529	am	50	202005
530	till	51	202005
531	wife	65	202005

**작품별 트위터 언급수**

	date	word
0	202005	wrong
1	202005	missy
2	202005	me
3	202005	extraction
4	202005	john
...	...	...
257	202203	black
258	202203	me
259	202203	weekend
260	202203	away
261	202203	ruby



# N 04. 분산 분석 -데이터셋 생성

## 월별 단어 & TOP10 작품 제목 교차 추출

	Date	Title	Count	Movie_or_tv	Country	Original	Genre	IMDB
0	202005	Riverdale	95	TV Show	United States	0	Superhero	6.9
1	202005	Avatar: The Last Airbender	696	TV Show	United States	0	Fantasy	9.2
2	202005	Ozark	128	TV Show	United States	1	Drama	8.4
3	202005	Hollywood	533	TV Show	United States	1	Drama	6.5
4	202005	Sweet Magnolias	41	TV Show	United States	1	Drama	6.9
...	...	...	...	...	...	...	...	...
286	202203	Shrek	123	Movie	United States	0	Animated	7.9
287	202203	Shrek 2	123	Movie	United States	0	Animated	7.2
288	202203	Tyler Perry's A Madea Homecoming	53	Movie	United States	1	Comedy	5.8
289	202203	The Weekend Away	49	Movie	United States	1	Thriller	5.6
290	202203	Rescued by Ruby	42	Movie	Canada	1	Drama	6.8

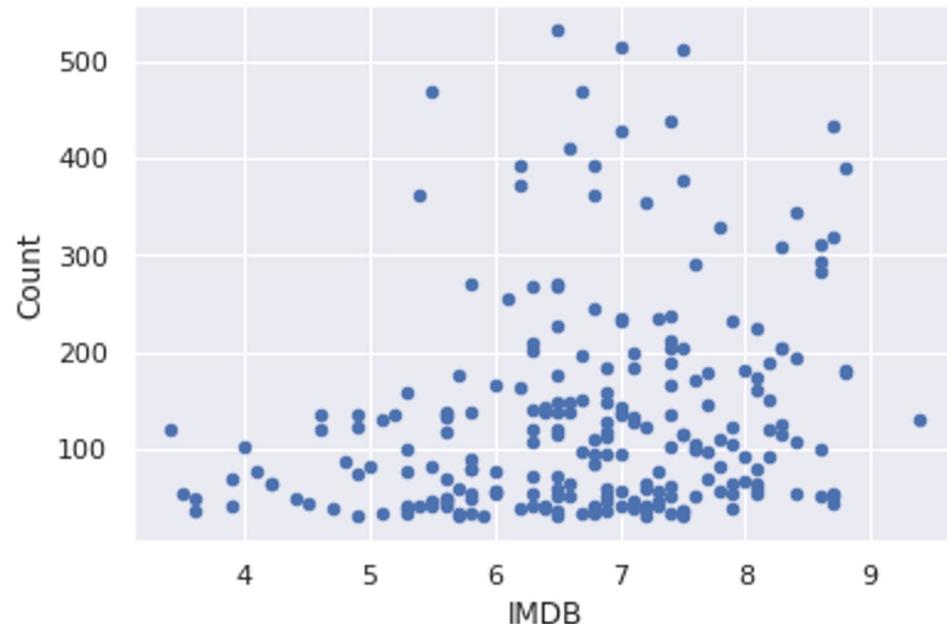
291 rows × 8 columns

# N 04. 데이터 분석 - 상관분석

## 피어슨 상관 분석

〈호응지수(Count) & IMDB〉

Pearson Correlation Coefficient



```
np.corrcoef(data_DF4["IMDB"], data_DF4["Count"])[0, 1]
```

```
0.23106370582195113
```

COR

= **0.23106370582195113**

=> 상관 없음

# N 04. 데이터 분석 - 상관분석

## 점 이분 상관 분석

Movie_or_tv	Movie_or_Tv	Original
TV Show	1	0
TV Show	1	0
TV Show	1	1
TV Show	1	0
TV Show	1	1
...	...	...
TV Show	1	1
Movie	0	0



```
stats.pointbiserialr(data_DF1["Movie_or_Tv"], data_DF1["Count"])
PointbiserialrResult(correlation=0.25200731139365035, pvalue=0.00010791038581386573)
```

### 〈호응지수(Count) & Movie or TV Series〉

COR

= 0.25200731139365035

P-value

= 0.00010791038581386573

=>상관 없음

### 〈호응지수(Count) & Original〉

COR

= 0.2715045140052942

P-value

= 2.8772882151697564e-05

=>상관 없음

# N 04. 결과

# 넷플릭스와 경쟁사(디즈니플러스)의 작품포스터에서 유의미한 특징과 차이를 발견할 수 있을까?

**한국화 수 허용까?**

**넷플릭스는 디즈니플러스에 비해  
포스터를 통한 뚜렷한 특징 파악이  
어려움**

**따라서 자사의 플랫폼을 타사와  
차별화할 수 있는 포스터로 개선 필요**

**넷플릭스 오리지널 작품과 수급  
작품 중 무엇이 더 인기있고 그  
인기작들은 어떤 특징이 있을까?**

**오리지널작과 수급작은 장르적(Action, Crime, Thriller ) 공통점이 있음**

**수익성이 보장되는 작품의 특징이  
나타남**

**인기작의 언급수에 따라 어떤 변수가  
유의미한 상관관계를 낼 수 있을까?**

양의미학 유연연체를 끝나 허통까  
도전과 도전과 도전과 도전과

IMDB 지수, tv-show, movie, 넷플릭스 오리지널 여부와 인기작과의 상관관계는 없음



# 넷플릭스의 작품선택 기준:

**IMDB 지수, TV Show, Movie, 넷플릭스 오리지널 여부와 상관없이 Action, Crime, Thriller의 장르를 가진 작품을 선택한다.**

# N 셋탑박스는 물려가고 플랫폼만 남을 미래 OTT(Over-The-Top)

TVING

hulu

NETFLIX

WATCHA

prime video

seezn

HBO  
max

tv+

Disney+

# N 05. 개발 후기 및 느낀 점

김희원: 데이터 전처리를 주로 맡아 진행하였는데, 작은 문법 하나하나도 절대 쉽지 않다는 것을 새삼 깨닫게 되었습니다. 강사님이면 삼분이면 짤 문장을 두시간씩 잡다보니 현라가 오기도 했지만, 그만큼 음 해결하는 순간이 재밌기도 하고, 더욱 공부하고 잘하고 싶은 마음이 들기도 해 마냥 맞지 않는 길은 아니구나 라고 생각했습니다. 사실 우리조가 대단하게 프로젝트를 마무리 했다고 생각하면 양심이 없지만 그래도 모두 의 상하지 않고, 오히려 으하고 힘내주는 모습에 너무 여려가지 얻은게 많은 감사한 프로젝트였습니다.

김보성: 본인의 이해력 부족으로 개발 프로젝트의 전반적인 프로세스를 잘 이해하지 못한 상태로 프로젝트 작업을 진행하는 과정에서, 의견 도출과 같은 방법이 매우 미숙하였다. 실제로 어떻게 데이터 전처리를 하며, 검증 가능한 가설세우기를 토대로 제대로 된 인사이트를 도출할 수 있는지, 더 나아가 도출한 것을 가지고 시각화 분석해서 상대방을 설득력있게 전달할 수 있을지를 고민하였다. 강사님이 알려주신 워드클라우드를 가지고 시각화를 해보았는데, 생각했던 것 만큼 결과가 잘 나오지 않았고, 오류코드를 수정하는데에 많은 시간을 할애했다. 처음엔 헛된 시간처럼 느꼈지만 전세계적인 화가 '다 빙치'가 처음 그림을 그릴 때 달걀 하나를 그리는 데에만 몇 년의 시간을 쏟았던 것처럼 내가 오류코드를 수정하는 데에 매달린 그 많은 시간들이 미래에 좋은 분석능력을 갖춘 데이터사이언스의 구성원이 되기 위한 헛되지 않은 과정이었으면 좋겠다는 생각을 하게 되었다.

민윤지: OTT서비스 주제를 제안했을 땐 데이터의 유무를 생각지도 않고 단순하게 OTT산업에 대한 흥미로움 때문이었다. 프로젝트를 시작하고 나니 너무나 당연하게도 데이터가 부족하여 데이터수집에 많은 시간을 할애했다. 거의 모든 데이터를 크롤링을 통해서 가져왔고, 그 과정은 힘들었다. 하지만 트위터 크롤링, 이미지 크롤링을 하면서 크롤링과정과 크롤링한 데이터를 정제하고 분석하는 법을 공부하고 배웠다. 무엇보다 적극적으로 의사소통하고 모르는 건 서로 알려주며 도와줬던 팀원분들에게 감사하다.

# N 05. 개발 후기 및 느낀 점

이은평: 제 스스로에 대해 많이 반성하고 깨우치는 기회가 되었습니다. 당찬 포부만으로는 세상을 해결하기 어렵고 그건 객기였습니다.... 하지만 이후 심화과정을 통해 조금 더 기술적으로 성장한 자신이 되어 다음 프로젝트 부터는 객기가 아닌 자신 있는 사람이 되고 싶다는 생각을 하게 해준 프로젝트였습니다.

신혜빈: 내가 프로젝트라는 걸 새로운 사람들과 과연 할 수 있을까? 이런 상태로 시작을 했고! 우려했던 것처럼 나에게는 너무 어렵고 힘들었고! 민폐가 안되고 싶고 뭐라도 하고 싶은데 막힐때마다 스트레스!!! 하지만 이번 프로젝트를 하면서 역시 공부도 많이 했고 좋은 조원들을 만나서 함께 해쳐나가는 소동한 경험을 했다:)

이지원: 데이터 수집이 어려운 주제라 '과연 이게 될까?'라는 의문이 들었지만 해보고 싶었던 주제라 무작정 크롤링과 팀원들을 믿고 이 주제를 선택했던 것 같습니다. 역시나 데이터 수집부터 전처리과정까지 쉬운 게 하나도 없어서 고생은 했지만 오류를 수정하고, 배운것들을 많이 접하면서 공부가 많이 됐건 것 같습니다. 무엇보다도 정말 하고 싶었던 주제와 분석방법을 시도해 볼 수 있었다는 것만으로도 너무나 만족스러웠고, 재미있는 팀원분들을 만나 즐겁게 프로젝트를 할 수 있어서 좋았습니다! 여러분들 수고 많으셨어요!!!