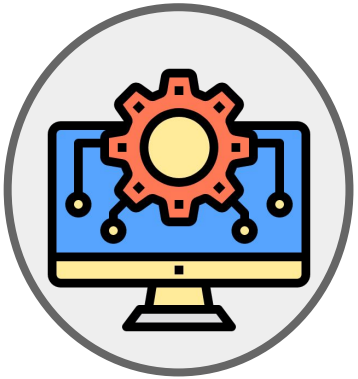
컨텐츠 기반의 NAVER 영화 추천시스템

김희아
이원호
이상민

Content

Data Scraping

Data Preprocessing



Modeling

Visualization

Data Scraping

```
title = movie.select_one('.h_movie > a')
genre = movie.select_one('.info_spec> dd:nth-of-type(1) > p > span')
director = movie.select_one('.info_spec> dd:nth-of-type(2) > p')
```

네이버 영화 [제목] ,[장르], [감독] 스크래핑

- https://movie.naver.com/movie/bi/mi/basic.nhn?code={i}
- Query Parameter 를 변경해가면서 총 178474개의 데이터 수집

Data PreProcessing

```
if re.search('genre',str(genre)) is not None:
    movie_genre = re.sub(',', ' ', genre.text).split()
else:
    movie_genre = ''

if re.search('director',str(director)) is not None:
    movie_director = director.text
else:
    movie_director = ''
```

[장르] or [감독] 누락 -> [국가] or [상영시간]이 대체

- 국가의 a 태그 제거
- 상영시간은 href 태그 존재하지 않음
  : href 태그 & 국가에 대한 코드가 아닐 경우 장르 추가

최종적으로 **총 130062개**의 데이터 수집

Modeling

기존 모델과의 차이점

- 정규 표현식 사용을 통해 제목이 정확하게 일치하지 않아도 되도록 설계

```
re.sub('₩W+', '', input_movie)
```

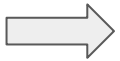
특수문자 + 공백 제거

ex) 왕의 남자 **vs** 왕의남자

```
왕의남자
                title genre
0               왕의 남자   드라마
```

ex) @골뱅이 **vs** 골뱅이

```
골뱅이
                title genre
51          낚시 바보 일지   코미디
180     위드 프렌즈 라이크 디즈   코미디
115          @골뱅이   코미디
```



띄어쓰기나 특수문자가 정확히 일치하지 않아도 같은 결과 출력

Visualization(WordCloud)

	title	genre	director
0	왕의 남자	드라마	이준익
262	야망의 세월	드라마	이종수
129	로망스	드라마	이대영
117	북의 영년	드라마	유키사다 이사오
114	에로스	드라마 왕가위	스티븐 소더버그 미켈란젤로 안토니오니
110	사랑밖엔 난 몰라	드라마	박종 윤재문
109	하나뿐인 당신	드라마	정운현
107	루키	드라마	고흥식
105	미나	드라마	김재순 최석원 이주형
103	사춘기	드라마	이창한
90	세리 누아르	드라마	알랭 코르노
78	아카메 48 폭포 자살미수	드라마	아라토 겐지로
73	아씨	드라마	고성원
72	7월 32일	드라마	진승현
70	진주탑	드라마	김묵
45	레옹 모랭 신부	드라마	장 피에르 멜빌
140	운수 좋은 날	드라마	이한종
44	무서운 아이들	드라마	장 피에르 멜빌
149	천국에 있는 것처럼	드라마	케이 폴락
161	반올림 시즌 2	드라마	최세경 박기현 김진환
256	하와이, 오슬로	드라마	에릭 포페
250	왕과 함께 한 하룻밤 - 에스더 이야기	드라마	마이클 O. 사이벨
248	달라스 362	드라마	스콧 칸
245	발라드 오브 잭 앤 로즈	드라마	레베카 밀러
243	스키조	드라마	굴삿 오마로바
240	오프 더 맵	드라마	캠벨 스콧
206	마추카	드라마	안드레스 우드
204	천국의 책방 - 연화	드라마	시노하라 테츠오
200	용봉투	드라마	두기봉
190	더티 댄싱 2	드라마	리사 니에미

Output with WordCloud

Thank You!


세리 누아르
북의 영년
야망의 세월
진주탑
시튼
레옹 모랭 신부
하나뿐인 당신
왕의 남자
사춘기
로망스
운수 좋은 날
사랑밖엔 난 몰라
전국에 있는 것처럼
루키