



# 영화 장르 예측

TEAM T 2 M

NLP TEAM PROJECT

2019 / 06 / 15

박  
강  
구  
김  
심  
이

나  
제  
영  
민  
대  
설

윤  
순  
서  
지  
범  
희

# INDEX

영 화 장 르 예 측

- ■ 개요
- ■ 개발 과정
- ■ 시행 착오
- ■ 데모

## 프로젝트 주제

영화 및 영상의 스크립트를 분석하여  
해당하는 장르를 예측해 낸다

### 예상 활용 분야

- 영상, 영화 콘텐츠 시장에서의 서비스 기능
  - 영상 콘텐츠 추천, 영화 데이터 관리 등

## 데이터 수집



<https://www.imsdb.com/>

## 데이터 수집



### Selenium

Selenium is a portable  
software testing framework  
for web applications.

크롤링 도구

# 데이터 수집

Genre	
Action	<a href="#">Adventure</a>
Comedy	<a href="#">Crime</a>
Family	<a href="#">Fantasy</a>
Horror	<a href="#">Musical</a>
Romance	<a href="#">Sci-Fi</a>
Thriller	<a href="#">War</a>

## Action Movie Scripts

[15 Minutes](#) (Undated Draft)

*Written by John Hertzfield*

[2012](#) (2008-02 Second draft)

*Written by Roland Emmerich, Harald Kloser*

[30 Minutes or Less](#) (2009-12 Draft)

*Written by Michael Diliberti, Matthew Sullivan*

[48 Hrs.](#) (Undated Draft)

*Written by Steven E. De Souza, Walter Hill, Roger Spottiswoode, Larry Gross, Jeb Stuart*

[A Most Violent Year](#) (Undated Draft)

*Written by J.C. Chandor*

[Above the Law](#) (1987-04 Draft)

*Written by Steven Pressfield, Ronald Shusett, Andrew Davis, Steven Seagal*

[Abyss, The](#) (1988-08 Draft)

*Written by James Cameron*

[Air Force One](#) (Undated Draft)

*Written by Andrew W. Marlowe*

# 데이터 수집

## Action Movie Scripts

[15 Minutes](#) (Undated Draft)  
Written by John Hertzfield

[2012](#) (2008-02 Second draft)  
Written by Roland Emmerich, Harald Kloser

[30 Minutes or Less](#) (2009-12 Draft)  
Written by Michael Diliberti, Matthew Sullivan

[48 Hrs.](#) (Undated Draft)  
Written by Steven E. De Souza, Walter Hill, Roger

[A Most Violent Year](#) (Undated Draft)  
Written by J.C. Chandor

[Above the Law](#) (1987-04 Draft)  
Written by Steven Pressfield, Ronald Shusett, An

[Abyss, The](#) (1988-08 Draft)  
Written by James Cameron

[Air Force One](#) (Undated Draft)  
Written by Andrew W. Marlowe

No Poster  
No Poster  
No Poster  
No Poster  
No Poster  
No Poster

**IMDb opinion**  
None available

**IMDb rating**  
Not available  
**Average user rating**  
 (5.75 out of 10)

**Writers**  
[John Hertzfield](#)

**Genres**  
[Action](#)  
[Crime](#)  
[Thriller](#)

[Read "15 Minutes" Script](#)

# 데이터 수집

**FADE IN**

on the words CZECH AIRLINE. We are panning across the words on the side of the plane.

**INT. AIRPLANE**

**ANGLE DOWN**

on a tray table. Crumpled Czech bills and coins are on it. Hands are counting the money. The airline hostess announces the arrival at JFK - in CZECH. A hand reaches into a breast pocket - pulling out two passports. One is opened. Belongs to EMIL SLOVAK. The next passport belongs to OLEG RAZGUL. The hand passes the Oleg Razgul passport to the man next to him. We notice several empty airline bottles of vodka and a small disposable camera on Oleg's tray table. The passport is set down. Oleg picks it up. We hear Emil's voice in CZECH. The scene is subtitled in ENGLISH.

**EMIL (V.O.)**

Just do what I do. Say the same thing I say. Don't open your mouth.

**OLEG (V.O.)**

Okay.

**INT. PASSPORT CONTROL - KENNEDY AIRPORT - DAY**

CAMERA DOLLIES down a long line of passengers. They are

# 데이터 처리



사용한 Python Library

# 데이터 처리 - preprocess



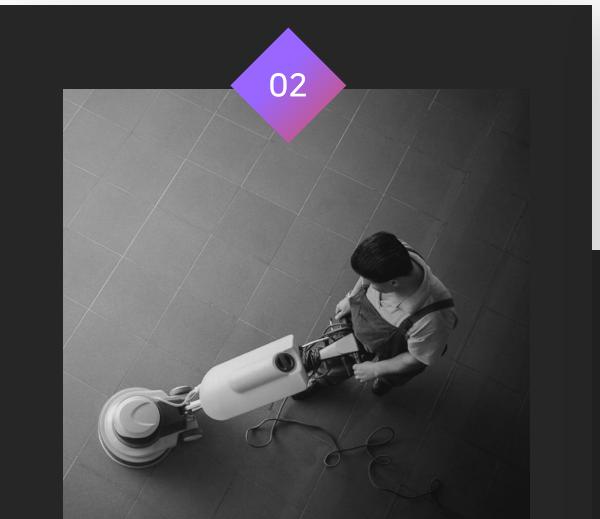
## Tokenization

토큰화

문서를 단어로 분리하는 단계

- `nltk.sent_tokenize (문서)`
- `nltk.word_tokenize (문장)`

# 데이터 처리 - preprocess



Cleaning

클리닝

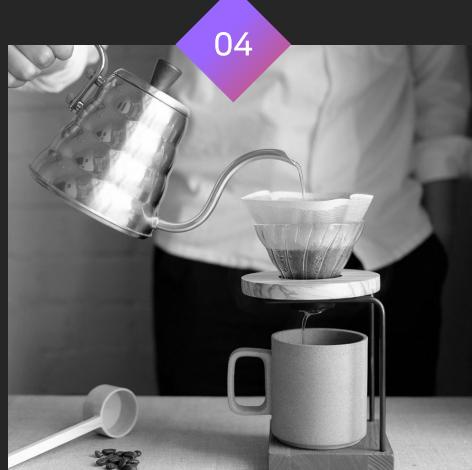
- 세글자 미만 단어 제외
- 소문자로 변경
- 숫자 제거

# 데이터 처리 - preprocess



- 불필요한 단어 제거  
(전치사, 관사, etc.)
- `nltk.corpus.stopwords.words('stopword')`
- 실험적으로 stopword 추가

# 데이터 처리 - preprocess



## Lemmatization

표제어 추출

- 단어의 기본 형태(표제어) 추출, 통합
- Stemming이 포함
- `nltk.stem.WordNetLemmatizer()`
- `nltk.stem.lematize()`

# 데이터 처리 - tagging

## 품사 정보 태깅

- nltk.tag.pos\_tag

## 단어 추출

- 태깅된 품사 정보 이용
- 명사, 동사만 추출

# 데이터 처리 - indexing

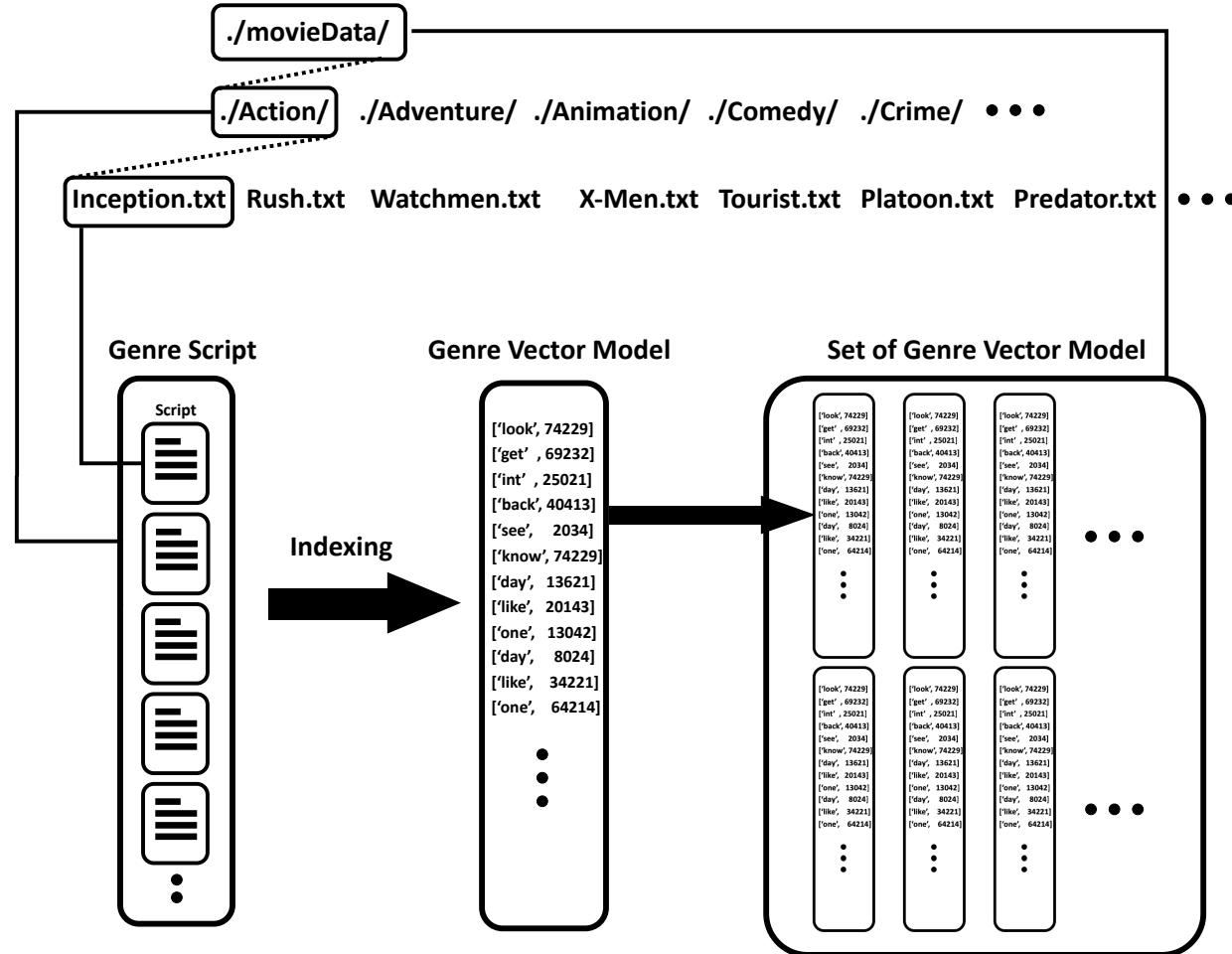
## 목록 읽어오기

- 장르 폴더 > 장르 목록
- 각 장르 > [스크립트 목록](#)

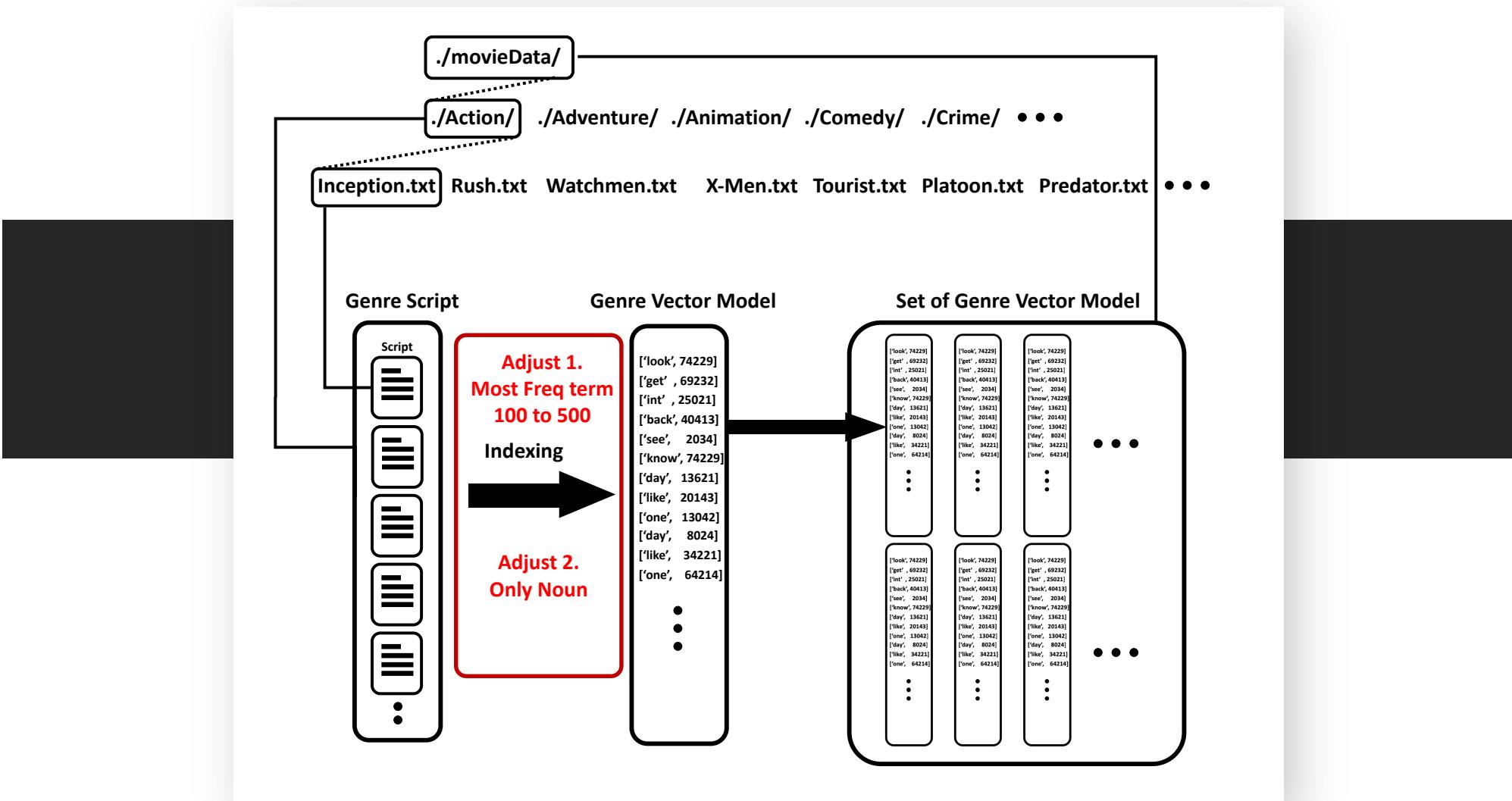
## Inverted Indexing

- 각 장르 단어 빈도로
- nltk.FreqDist

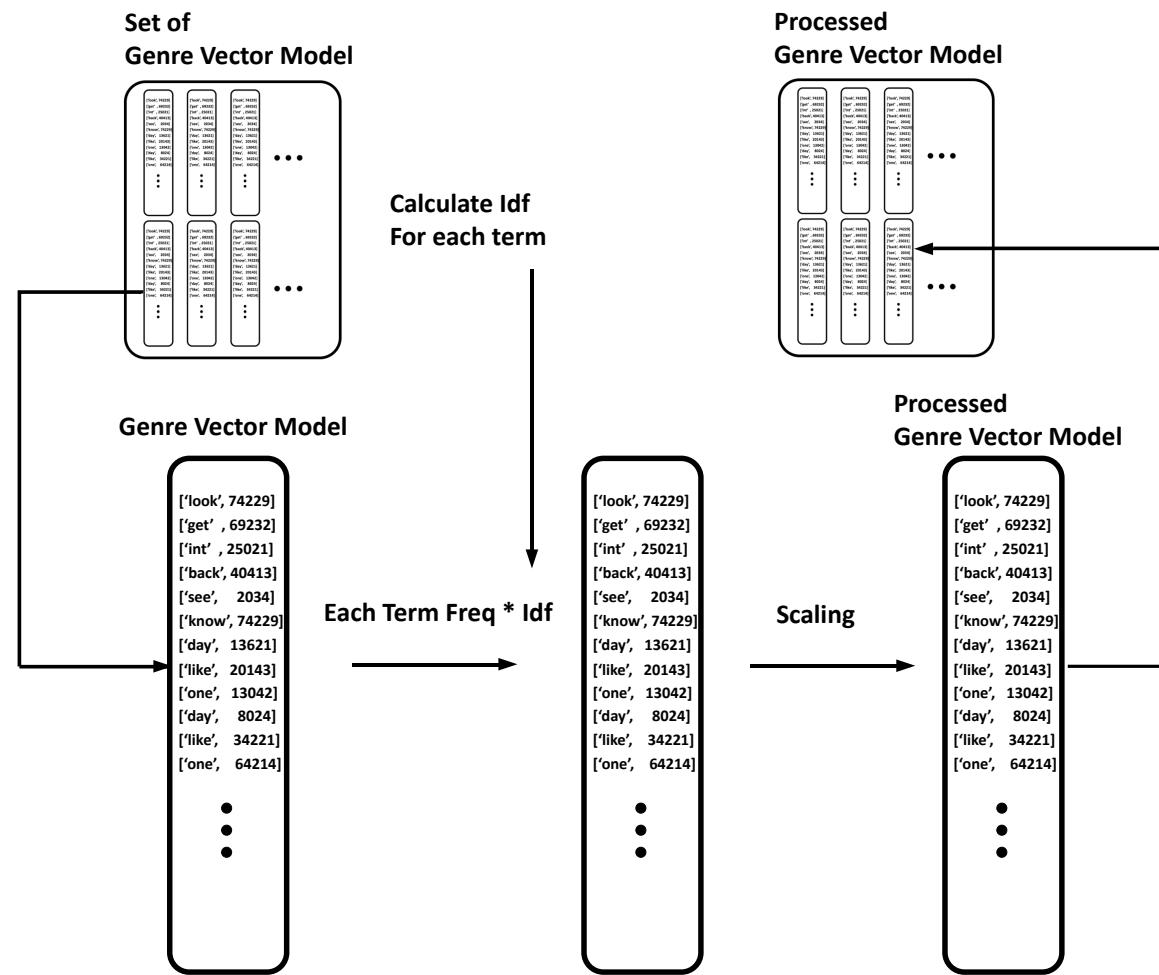
# 벡터 모델링 - tf·idf



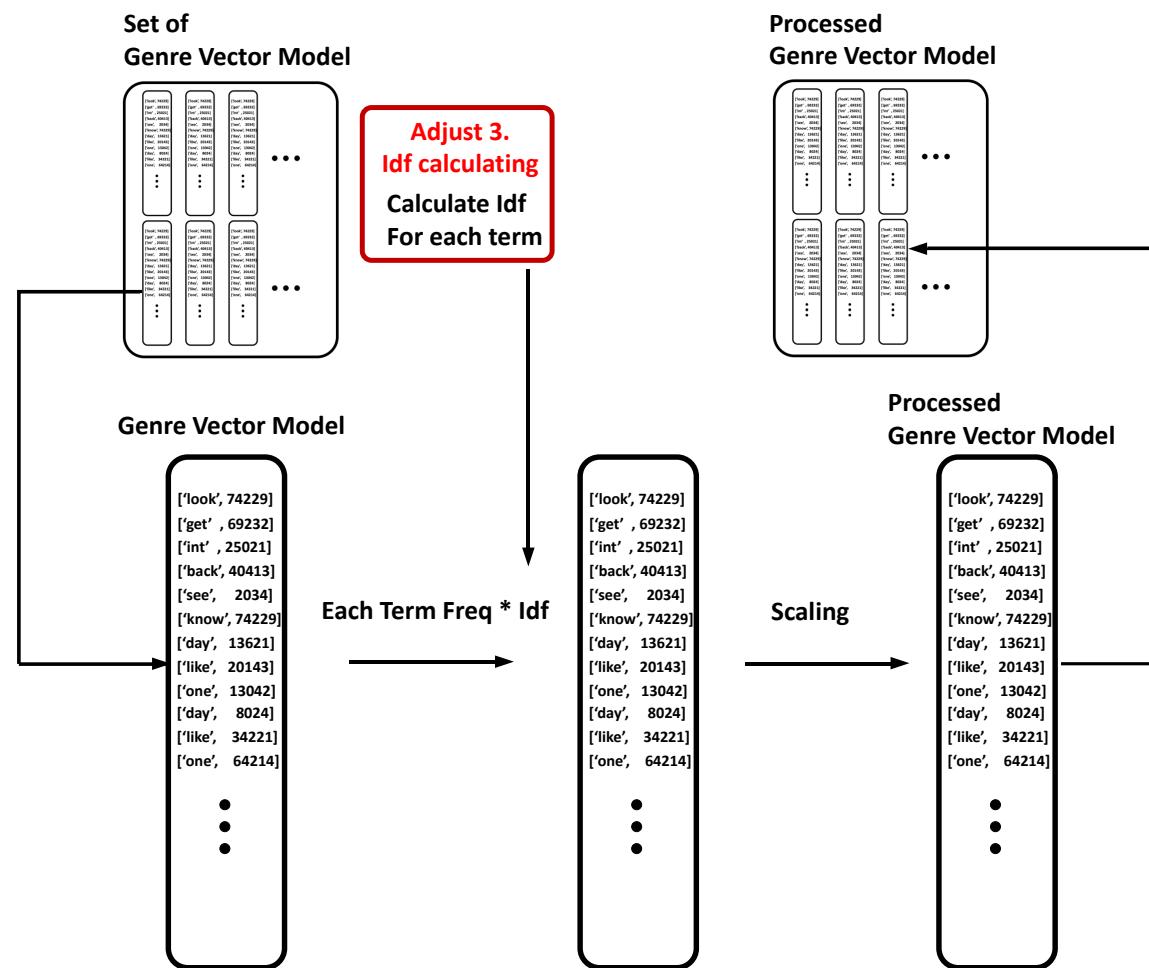
# 벡터 모델링 - tf·idf



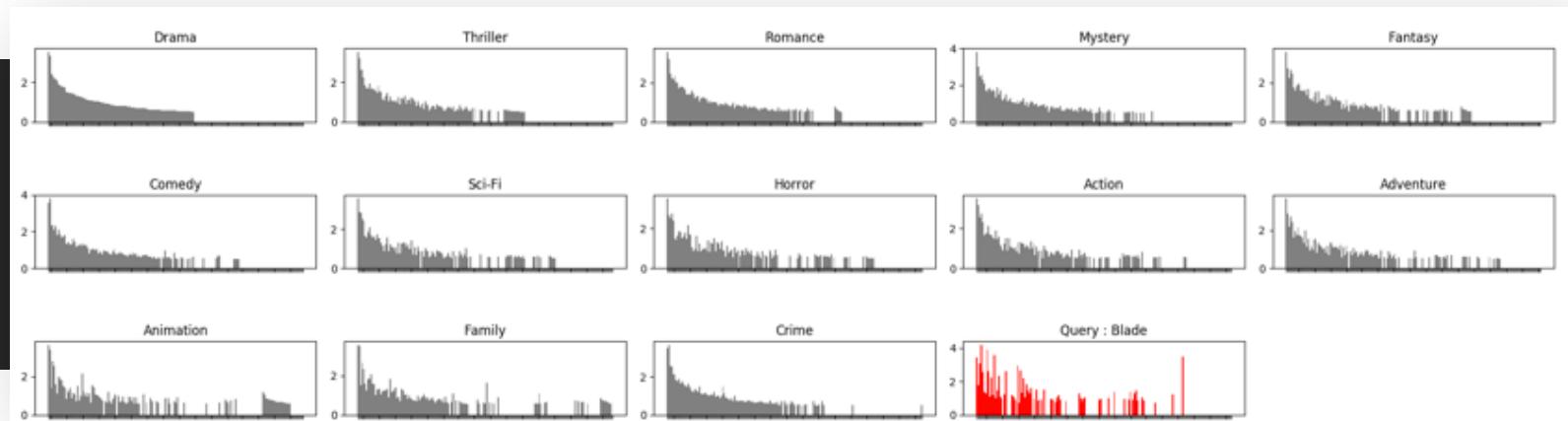
# 벡터 모델링 - tf·idf



# 벡터 모델링 - tf·idf



# 벡터 모델링 - tf·idf



## 벡터 모델링 - LSA

영화 별 장르 스크립트의 폴더에서 .txt 파일을 가져온다.

영화 스크립트 텍스트를 전처리한다.

영화 스크립트 텍스트를 벡터화한다.

## 벡터 모델링 - LSA

벡터화된 영화 스크립트의 차원 감소를 수행한다.

input을 받아 LSA 구성 요소를 사용하여 문서 유사성을 계산한다.

가장 유사한 장르 상위 3개를 가려낸다.

# 벡터 모델링 - LSA

영화 제목		장르 1	장르 2	장르 3
Thor	실제 장르	Adventure	Action	Fantasy
	분석한 장르	Adventure	Action	Sci-fi
Alien	실제 장르	Action	Sci-Fi	Horror
	분석한 장르	Action	Sci-Fi	Adventure
Godfather	실제 장르	Drama	Crime	
	분석한 장르	Drama	Comedy	Romance
Hacker	실제 장르	Drama	Crime	Action
	분석한 장르	Drama	Comedy	Romance
Hostage	실제 장르	Crime	Action	Drama
	분석한 장르	Crime	Mystery	Family
Inglourious-Basterds	실제 장르	Action	Adventure	
	분석한 장르	Drama	Crime	Family
Jurassic-Park-The-Los t-World	실제 장르	Action	Adventure	Sci-fi
	분석한 장르	Action	Adventure	Sci-fi
Lord-of-the-Rings-The -Two-Towers	실제 장르	Action	Adventure	Fantasy
	분석한 장르	Action	Adventure	Sci-fi
Mad-Max-2-The-Road-Warrior	실제 장르	Action	Adventure	Sci-fi
	분석한 장르	Action	Adventure	Sci-fi
Titanic	실제 장르	Drama	Romance	
	분석한 장르	Action	Adventure	Sci-fi
Total-Recall	실제 장르	Thriller	Adventure	Sci-fi
	분석한 장르	Action	Adventure	Sci-fi
War-Horse	실제 장르	Drama		
	분석한 장르	Animation	Thriller	Mystery
Yes-Man	실제 장르	Comedy	Romance	
	분석한 장르	Comedy	Romance	Drama

Dimensionality: 2

정확도: 약 57%

# 벡터 모델링 - LSA

영화 제목		장르 1	장르 2	장르 3
Thor	실제 장르	Adventure	Action	Fantasy
	분석한 장르	Adventure	Action	Sci-fi
Alien	실제 장르	Horror	Sci-Fi	Action
	분석한 장르	Horror	Sci-Fi	Adventure
God Fother	실제 장르	Drama	Crime	
	분석한 장르	Drama	Crime	Romance
Hacker	실제 장르	Drama	Crime	Action
	분석한 장르	Drama	Crime	Drama
Hostage	실제 장르	Crime	Action	Drama
	분석한 장르	Crime	Mystery	Thriller
Inglourious-Basterds	실제 장르	Action	Adventure	
	분석한 장르	Drama	Comedy	Romance
Jurassic-Park-The-Los t-World	실제 장르	Action	Adventure	Sci-fi
	분석한 장르	Action	Adventure	Sci-fi
Lord-of-the-Rings-The -Two-Towers	실제 장르	Action	Adventure	Fantasy
	분석한 장르	Action	Adventure	Fantasy
Mad-Max-2-The-Road- Warrior	실제 장르	Action	Adventure	Sci-fi
	분석한 장르	Action	Adventure	Sci-fi
Titanic	실제 장르	Drama	Romance	
	분석한 장르	Action	Adventure	Family
Total-Recall	실제 장르	Adventure	Thriller	Sci-fi
	분석한 장르	Action	Thriller	Sci-fi
War-Horse	실제 장르	Drama		
	분석한 장르	Adventure	Action	Fantasy
Yes-Man	실제 장르	Comedy	Romance	
	분석한 장르	Comedy	Family	Animation

Dimensionality: 13

정확도: 약 63%

# 평가

# 데이터 수집

## 데이터 부족

장르별로 스크립트 개수가 불균형적  
→ 스크립트 개수가 너무 적은 장르는 제외

# 데이터 처리

## 고유 명사

장르 판단과 무관한 인명, 지명과 같은 고유 명사의 다빈도 출현

→ 실험적으로 제거

# 벡터 모델링

## Dimensionality

Dimension이 많아서 오류가 생겼을 때  
정확한 원인을 파악하기 힘듦

# 데모 시연



demonstration of team project

---

## Q & A

---

감사합니다