Natural Language Processing

Final Report

Team Project – Movie Genre Prediction

학 과	컴퓨터공학부	
팀 명	T2M	
팀 원	박 나 윤	김 민 지
	강 제 순	심 대 범
	구 영 서	이 설 희
제 출 일	2019. 6. 15 (토)	

목차

1. 프로젝트 주제 및 설계

- 주제

  개인 과제였던 영화 스크립트 분석에서 한 발 더 나아가, 장르별로 특징을 구분 짓고 새로운 영화의 스크립트를 입력 받았을 때 해당하는 한 개 이상의 장르를 예측해낸다.

  영화 및 영상 컨텐츠 시장에서의 서비스로 확장될 수 있다. 예를 들어, 사용 패턴을 분석하여 개인에 알맞는 영상 컨텐츠를 추천하거나 영화 데이터를 더욱 효율적으로 관리할 수 있을 것이다.

- 설계

  - 데이터 소스: IMSDb (https://www.imsdb.com/)

  - 장르 분류

Genre		
Action	Adventure	Animation
Comedy	Crime	Drama
Family	Fantasy	Film-Noir
Horror	Musical	Mystery
Romance	Sci-Fi	Short
Thriller	War	Western

  → 데이터가 상대적으로 매우 부족한 장르를 제외하고 최종적으로 아래 13 가지의 장르에 대해서만 프로젝트를 진행하기로 하였다.

Action	Adventure	Animation	Comedy
Crime	Drama	Family	Fantasy
Horror	Mystery	Romance	Sci-Fi
Thriller			

2. 개발 과정

- **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,Ro

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

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

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

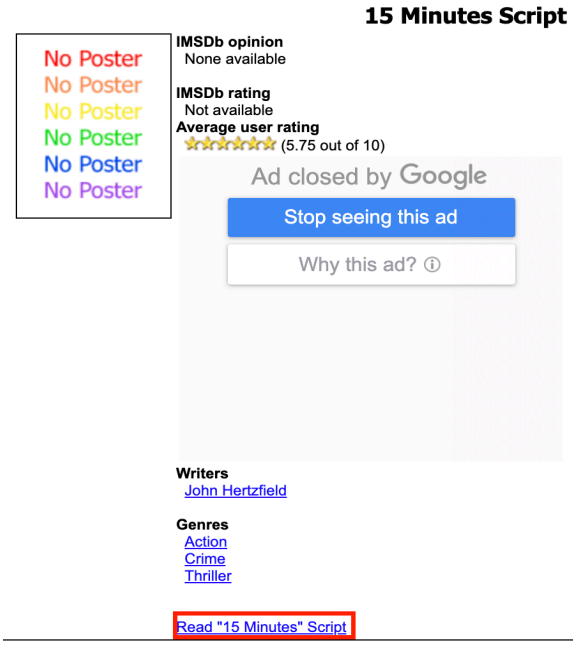
Air Force One (Undated Draft)
Written by Andrew W. Marlowe

Alien (1978-06 Draft)
Written by Walter Hill,David Giler

Alien 3 (1991-01 Draft)

<그림 1>

장르를 선택하면 우측에 영화 스크립트 리스트(그림 1)가 출력된다. 이 장르에 해당하는 각 영화 스크립트를 수집하는 것이 이 단계의 목표이다.



15 Minutes Script

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

IMSDb opinion
None available

IMSDb rating
Not available
Average user rating
(5.75 out of 10)

Ad closed by Google
Stop seeing this ad
Why this ad?

Writers
John Hertzfield

Genres
Action
Crime
Thriller

Read "15 Minutes" Script

<그림 2>

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 split into two lines - one for Americans, the other for

<그림 3>

스크립트를 클릭하면 중간에 < Read '영화제목' script >가 존재한다(그림 2). 이 링크에는 해당 영화의 스크립트 페이지(그림 3)가 연결되어있다.

크롤링의 프로세스를 간단하게 정리하면 아래와 같다.

① https://www.imsdb.com/ 에서 좌측의 장르 리스트를 가져온다.

② 각 장르의 스크립트 리스트를 가져온다.

③ 각각의 스크립트를 가져와 txt 형태로 저장한다.

크롤링은 Selenium 을 통해서 진행하였다. Selenium 은 주로 웹앱을 테스트하는데 이용하는 프레임워크이다. 이것은 webdriver 라는 API 를 통해 운영체제에 설치된 Chrome 등의 브라우저를 제어한다.

브라우저를 직접 동작시킨다는 것은 JavaScript 를 이용해 비동기적으로 혹은 뒤늦게 불러와지는 컨텐츠들을 가져올 수 있다는 것이다. 즉, '눈에 보이는' 컨텐츠라면 모두 가져올 수 있음을 뜻한다. 우리가 requests 에서 사용했던 .text 의 경우 브라우저에서 '소스보기'를 한 것과 같이 동작하여,

JS 등을 통해 동적으로 DOM 이 변화한 이후의 HTML 을 보여주지 않는 반면, Selenium 은 실제 웹 브라우저가 동작하기 때문에 JS 로 렌더링이 완료된 후의 DOM 결과물에 접근이 가능하게 된다.

selenium 으로 css selector 를 이용하여 html 의 tag 즉, element 를 찾고 직접 컨트롤한다. 그리고 이어서 앞서 설명한 프로세스를 진행한다. 각 장르별로 폴더를 만들고, 해당 장르의 영화 스크립트를 하나씩 접근하여 txt 로 출력하여 저장하게 된다.

여기서 생기는 예외는 1) pdf 파일의 데이터를 가져오는 것과 2) script 가 없는 영화 스크립트가 있다는 것, 이 두가지였습니다. 이를 해결하기 위해 각 pdf 파일은 따로 수동으로 저장하고, script 가 없는 영화는 무시하고 건너뛰는 것으로 해결하였다.

- **데이터 처리**

- 사용한 Python 라이브러리: NLTK (Natural Language Toolkit)

- **데이터 전처리**

① Tokenization : 문서를 단어로 분리시키는 단계

문서를 nltk.sent_tokenize(문서)와 nltk.word_tokenize(문장)을 사용하여 토큰화한다.

② Cleaning : 세 글자 미만의 단어 제외, 소문자로 변경, 숫자 제거

③ Remove stopwords : 전치사, 관사 등 문서의 특징을 표현하는데 불필요한 단어를 삭제한다.

from nltk.corpus import stopwords → stopwords.words('english') → 토큰에서 stopwords 를 제거한다.

결과를 보고 불필요한 단어들은 stopwords 에 추가하여 삭제

④ 표제어 추출 (Lemmatization) : 단어의 기본 형태(표제어) 추출한다. 단어를 기본형으로 변형시켜 다른 형태로 표현된 단어들을 통합해준다. 단순한 Stemming 보다는 Lemmatization 이 더 단어의 형태가 보존되는 양상을 보이기 때문에 Lemmatization 을 선택하였다.

from nltk.stem import WordNetLemmatizer →WordNetLemmatizer(), lemmatize()를 이용하였다.

- **태깅**

① nltk.tag의 pos_tag를 이용하여 단어마다 품사 정보를 태깅한다.

② 태깅된 품사 정보를 이용하여, 명사/동사 단어만 추출한다.

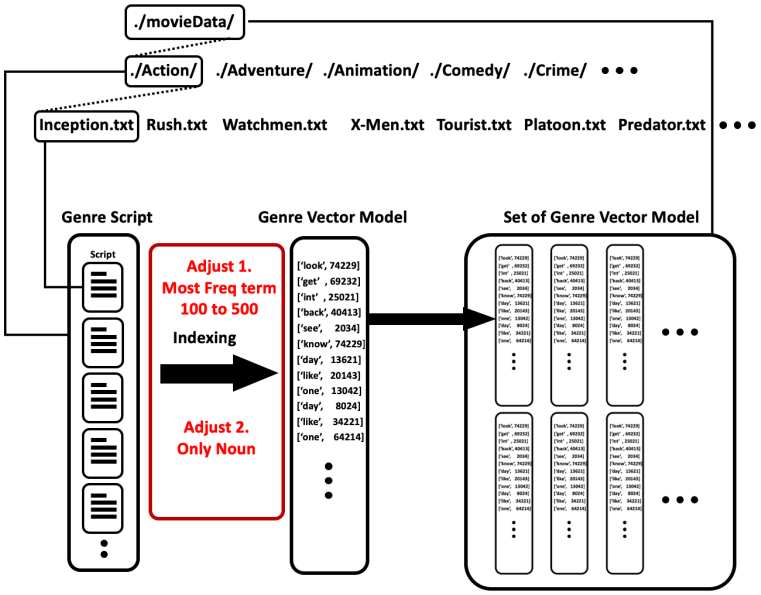
- **인덱싱**

① 장르 폴더 내의 장르들 목록을 읽어 오고, 각 장르들 내의 스크립트 목록을 읽어 온다.

② 한 장르 안의 문서들의 단어를 빈도로 inverted indexing → from nltk import FreqDist 를 사용한다.

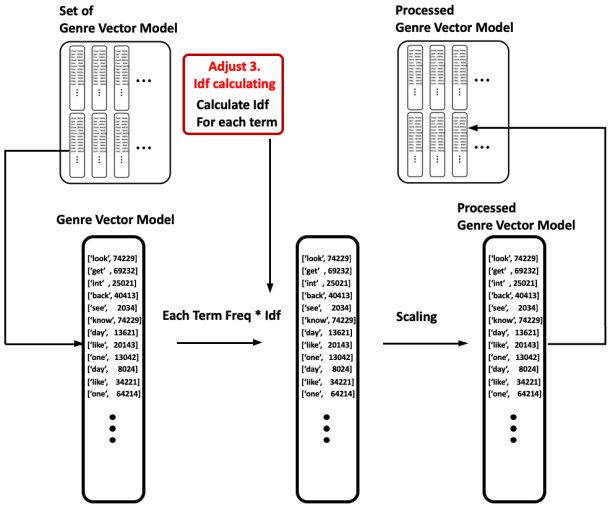
• 벡터 모델링

수집한 데이터들은 ./movieData/ 디렉토리 안에 각 장르별 디렉토리에 나뉘어 저장되어 있다. 장르가 여러 개 라벨링 되어 있는 영화의 경우에는, 각 장르 폴더 마다 스크립트가 존재한다. 각 장르별로 수집된 데이터 양은 차이가 있으며, 40 개 에서 600 개까지 범위로 수집되었다. 이러한 데이터 양의 편차는 개선이 필요한 부분이다. 총 1092 개의 개별 영화 스크립트가 존재하며, 장르 별로 중복되어 존재하는 파일 갯수는 총 2999 개 이다. 가장 적은 데이터는 Animation 장르로, 43 개의 데이터가 존재한다.



< 수집한 데이터에 대응하는 벡터 모델 구조 >

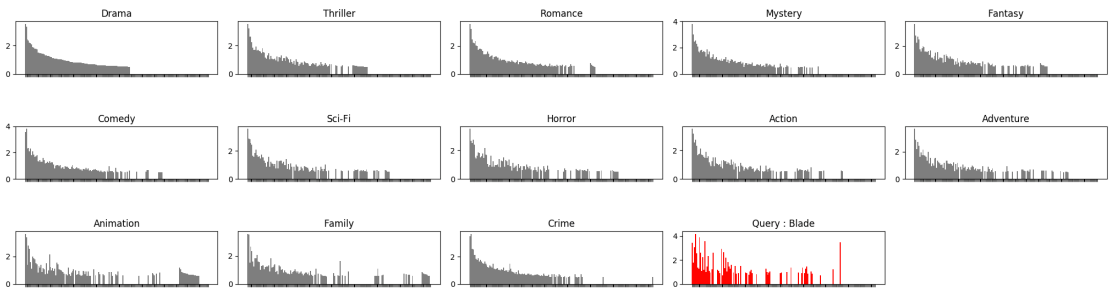
위의 그림은 수집한 데이터들이 프로그램 내에서 어떻게 벡터 모델이 되는지 대략적으로 나타낸다. 어떠한 장르에 속한 모든 스크립트는, 하나의 장르 스크립트로 통합된다. 따라서 본 프로젝트에서 다루는 13 개의 장르는 각각 하나의 스크립트로 나타내어지고, 이후 인덱싱 되어 장르 당 하나의 벡터 모델이 생성된다. 인덱싱 된 장르 벡터 모델은 각자가 가진 Term 과, Term Frequency 값을 갖는다. 이 때, 인덱싱 된 단어들 중, 각 장르 스크립트 내에서 가장 많은 빈도를 가지는 단어 N 개를 선별하여 장르 벡터 모델을 생성하였다. 이 때에, 각 장르 별로 N 개 만큼 추린 이후, 여기서 나온 모든 선별된 단어들을 합하여, 하나의 벡터 차원을 만들어 진 후에, 이에 따라 각각의 장르 벡터 모델들이 만들어 진다. 따라서 모든 장르 벡터 모델은 같은 차원을 갖는다. 장르별로 100 개를 추렸을 때에는 175 개의 Dimension 이 생성되었으며, N 값이 커지게 되면 차원은 증가한다.



< 장르 벡터 모델 프로세싱 >

생성된 장르 벡터 모델은, Weight 적용과 Scaling 과정을 거쳐 사용되게 된다. 생성된 장르 벡터 모델들을 모아 각 Term 에 대한 DF 값을 계산한 후, 이를 이용해 IDF 수식에 따라 계산 후 각 장르 벡터 모델에 적용된다. 그리고, 각 벡터들을 같은 크기로 Scaling 하여 쿼리 스크립트와 장르 벡터 간의 계산이 공평히 이루어 질 수 있도록 하였다.

프로세싱 과정을 거쳐 최종적으로 완성된 장르 벡터 모델들은 GenreVectorModel 객체 내에 저장되어 있게 되며, 쿼리 스크립트가 들어오게 되면, 쿼리 벡터와 Similarity 계산을 하여 장르 추론을 수행한다. 이 때, 쿼리 스크립트는 장르 벡터 모델의 Dimension 과 같은 차원의 벡터로 변환되어 수행된다.



< 시각화 된 장르 벡터 모델과 쿼리 벡터 모델 >

변환 된 쿼리 벡터 모델은, 장르 벡터 모델이 거친 프로세싱 과정을 거치게 되며, 이때 idf 적용 및 스케일링이 수행된다. 쿼리 벡터 모델이 처리 과정을 마치면, 장르와 자신의 각 요소의 차들을 합하여 Distance 정도를 계산하게 된다.

- 분석 및 성능 개선을 위한 시도

본 팀이 해당 프로그램에서 성능 향상을 위해 조정한 부분은 크게 3 부분이다.

1. 장르 벡터 모델을 만드는 과정에서, 사용되는 단어의 수를 조정해 보았다. 인덱싱 과정에서 가장 빈도 높은 단어 N 개를 추출하여 만들게 되는데, N 값을 100 에서 700 까지 조정하며 테스트해 보았다. N 값이 늘어날 수록 장르 벡터 모델의 차원은 늘어나게 되며, 쿼리 벡터와의 Distance 비교에 사용되는 차원이 더 많아지게 된다. 실험적으로, 100~700 의 범위 내에서는 N 값이 늘어날 수록 성능이 개선되는 경향을 보였다.

2. 인덱싱 된 값에서, 특정 품사만 추출하여 장르 벡터 모델을 생성해 테스트 해 보았다.

	Drama	Thriller	Romance	Mystery	Fantasy	Comedy	Sci-Fi
1.	*Look*	*Look*	*Look*	*Look*	*Look*	*Get*	*Look*
2.	*Get*	*Get*	*Get*	*Get*	*Get*	*Look*	*Get*
3.	*Int*	*Int*	*Int*	*Back*	*Back*	*Back*	*Int*
4.	*Back*	*Back*	*Know*	*Int*	*See*	*Int*	*Back*
5.	*See*	*See*	*Back*	*See*	*Int*	*Know*	*See*
6.	*Know*	*One*	*See*	*Know*	*One*	*See*	*One*
7.	*Day*	*Know*	*Day*	*Door*	*Come*	*Like*	*Like*
8.	*Like*	*Door*	*Like*	*Come*	*Like*	*One*	*Ext*

9.	*One*	*Take*	*Come*	*One*	*Know*	*Go*	*Come*
10.	*come*	*like*	*take*	*go*	*hand*	*day*	*know*

< 각 장르 별 가장 많이 인덱싱 되는 Term 순위>

품사를 제한하지 않고 인덱싱 된 Term 들을 Freq 에 따라 Rank 를 매겨보면, 상위권의 많은 Term 들이 동사를 품사로 가진다. 특히 'Look'과 'Get'이 장르를 불문하고 최상위에 랭크 되었다. 초기 버전 테스트에서 R-precision 평가가 30% 대의 저조한 성능이 나와, 특정 품사를 제거해 보는 방법을 사용하여 테스트해 보았다.

3. Idf 수식을 다양하게 변화시켜 테스트 하였다.

Idf Cases	Expression
Default :	$\log_2 2 + \frac{N}{(1+df)}$
Idf Case 1	$\frac{N}{(1+df)}$
Idf Case 2	$\log_{10} 10 + \frac{N}{(1+df)}$
Idf Case 3 :	$\frac{N}{1+\log_2(1+df)}$

Default 값은 원래 $\log_2 \frac{N}{(1+df)}$ 값 이었으나, 위 부분이 2 미만일 경우 음수로 나오는 문제가 발생하여, 양수 값만 나올 수 있도록 2 를 더하여 식을 세웠다. Case 1 의 경우에는, Default 식에서의 결과 값의 차를 더 크게 해주기 위한 의도로, log 를 씌우지 않은 식을 세워 테스트 해 보았다. Case 2 의 식에서는, Term 마다 나타나는 df 의 값의 차이로 인한 idf 값의 편차를 줄여주기 위한 의도로 식을 작성하여 테스트 해 보았다. Case 3 수식은, 편차 값을 더 적절히 할 수 없을까 고민하여 만들어 낸 식이다.

- **평가**

성능 평가는 모델링할 때 사용되지 않은 14 개의 영화 스크립트 데이터에 대해 장르 예측 결과를 판단하는 것으로 진행되었다. 평가 척도로는 hamming loss, precision / recall, R-precision 을 사용하였다.

사용된 스크립트는 다음과 같다.

Alien	Lord-of-the-Rings-The-Two-Towers
Blade	Mad-Max-2-The-Road-Warrior
Godfather	Thor
Hackers	Titanic
Hostage	Total-Recall
Inglorious-Basterds	War-Horse
Jurassic-Park-The-Lost-World	Yes-Man

- **Hamming Loss**

하나의 영화가 하나의 장르 만을 갖지 않고 여러 개의 장르로 분류될 수 있다. 따라서 영화 장르 분석 문제를 다중 레이블 분류 문제로 바라보았다. 다중 레이블 분류에 대한 성능 평가 척도로 hamming loss 라는 것이 사용된다. Hamming loss 는 총 레이블 수에 대한 잘못 예측된 레이블의 비율을 계산한 것으로 다음과 같은 loss function 식으로 계산된다

$$\frac{1}{|N| \cdot |L|} \sum_{i=1}^{|N|} \sum_{j=1}^{|L|} xor(y_{i,j}, z_{i,j})$$

$y_{i,j}$는j번째 영화의 i번째 장르에 대한 실제 Boolean 값이고, $z_{i,j}$는 j번째 영화의 i번째 장르에 대한 예측 Boolean 값이다. $|N|$은 14, $|L|$은 13으로 각각 영화 스크립트 수와 장르 수를 뜻한다.

→ Result: 결과는 각 영화 별 hamming loss 값, 실제 장르, 예측 장르 그리고 평균 총 hamming loss 값을 엑셀 파일로 저장하였다.

- **Precision / Recall / F-measure**

  Precision 과 Recall 은 이진 분류에 대한 성능 평가 척도로 사용된다. 이를 다중 레이블 분류 문제에 적용하기 위해 각 레이블 별 단일 분류 문제로 전환하여 바라보았다. 13 개의 장르에 대해서 각 장르를 포함하고있는 영화들의 집합을 true set 으로 두고, 실제 해당 장르로 예측한 수를 계산함으로써 장르별 precision 과 recall, f-measure 값을 구하였다.

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{TRUE\ predictions}{whole\ predictions}$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TRUE\ predictions}{total\ existing\ TRUE}$$

$$\text{F measure} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

  → Result: 결과는 각 영화 별 Precision, Recall, F-measure 값, 그리고 평균 Precision, Recall, F-measure 값을 엑셀 파일로 저장하였다.

- **R-precision**

  R-precision 은 R 개의 관련 문서를 가진 쿼리에 대해 R 번째 랭크까지의 결과에 대한 precision 을 계산하는 방식이다. 해당 프로젝트에서 하나의 영화는 R 개의 장르를 갖고, 장르 분석을 수행했을 때 13 개의 각 장르에 대한 유사도가 랭크로 나타난다. 상위 R 개까지 예측된 장르에 대해서 실제 영화가 갖는 R 개의 장르와 비교하여 precision 을 계산하였다.

  → Result: 결과는 각 영화 별 R-precision 값, 실제 장르, 예측 장르 그리고 평균 총 R-precision 값을 엑셀 파일로 저장하였다.

3. 실행 결과

	idf: Case 0 $\frac{N}{1+\log_2(1+df)}$					idf: Default $\log_2 2+\frac{N}{(1+df)}$				
	#########	TESTIFY				########	TESTIFY			
OnlyIndex Noun	Title	Hamming	Actual Ge	Predict Genre		Title	Hamming	Actual Ge	Predict Genre	
GenreVectorSize : 100	Titanic	0.307692	['Romanc	['Action', 'Adventure']		Titanic	0.307692	['Romanc	['Action', 'Adventure']	
	Total-Recall	0.153846	['Sci-Fi', 'A	['Action', 'Horror', 'Sci-Fi', 'Thriller']		Total-Rec	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Mystery', 'Thriller']	
	Godfather	0.153846	['Crime', 'I	['Action', 'Drama']		Godfather	0	['Crime', 'I	['Crime', 'Drama']	
	Inglourious-Basterds	0.307692	['Adventur	['Drama', 'Thriller']		Inglouriou	0.307692	['Adventur	['Mystery', 'Romance']	
	Mad-Max-2-The-Road-V	0.153846	['Sci-Fi', 'A	['Action', 'Comedy', 'Sci-Fi']		Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Drama']	
	Hackers	0.461538	['Action', '	['Comedy', 'Drama', 'Romance', 'Sci-Fi']		Hackers	0.461538	['Action', '	['Comedy', 'Crime', 'Mystery', 'Romance']	
	War-Horse	0.153846	['Drama']	['Adventure']		War-Hors	0.153846	['Drama']	['Adventure']	
	Yes-Man	0.307692	['Romanc	['Animation', 'Family']		Yes-Man	0.307692	['Romanc	['Animation', 'Family']	
	Lord-of-the-Rings-The-T	0.153846	['Adventur	['Adventure', 'Fantasy', 'Sci-Fi']		Lord-of-th	0.153846	['Adventur	['Adventure', 'Fantasy', 'Sci-Fi']	
	Hostage	0.307692	['Action', '	['Crime', 'Mystery', 'Romance', 'Thriller']		Hostage	0.307692	['Action', '	['Comedy', 'Crime', 'Mystery', 'Thriller']	
	Blade	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Horror']		Blade	0.153846	['Sci-Fi', 'H	['Fantasy', 'Horror', 'Sci-Fi']	
	Jurassic-Park-The-Lost-V	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Family', 'Horror', 'Thriller']		Jurassic-P	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Fantasy', 'Horror', '	
	Alien	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Horror', 'Sci-Fi']		Alien	0.307692	['Sci-Fi', 'H	['Adventure', 'Horror', 'Mystery', 'Sci-Fi']	
	Thor	0.153846	['Adventur	['Action', 'Adventure', 'Romance']		Thor	0.153846	['Adventur	['Action', 'Adventure', 'Romance']	
	average hl	0.21978				average h	0.230769			
	#########	#########				########	#########			
	#########	TESTIFY				########	TESTIFY			
Indexing Default	Title	Hamming	Actual Ge	Predict Genre		Title	Hamming	Actual Ge	Predict Genre	
GenreVectorSize : 100	Titanic	0.307692	['Romanc	['Action', 'Adventure']		Titanic	0.307692	['Romanc	['Adventure', 'Family']	
	Total-Recall	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Mystery', 'Thriller']		Total-Rec	0.153846	['Sci-Fi', 'A	['Action', 'Mystery', 'Sci-Fi', 'Thriller']	
	Godfather	0	['Crime', 'I	['Crime', 'Drama']		Godfather	0	['Crime', 'I	['Crime', 'Drama']	
	Inglourious-Basterds	0.307692	['Adventur	['Comedy', 'Drama']		Inglouriou	0.307692	['Adventur	['Comedy', 'Romance']	
	Mad-Max-2-The-Road-V	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Comedy']		Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Thriller']	
	Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']		Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']	
	War-Horse	0.153846	['Drama']	['Adventure']		War-Hors	0.153846	['Drama']	['Adventure']	
	Yes-Man	0.307692	['Romanc	['Animation', 'Family']		Yes-Man	0.307692	['Romanc	['Animation', 'Family']	
	Lord-of-the-Rings-The-T	0.153846	['Adventur	['Adventure', 'Fantasy', 'Mystery']		Lord-of-th	0	['Adventur	['Action', 'Adventure', 'Fantasy']	
	Hostage	0.153846	['Action', '	['Crime', 'Drama', 'Mystery', 'Thriller']		Hostage	0.307692	['Action', '	['Comedy', 'Crime', 'Mystery', 'Thriller']	
	Blade	0.153846	['Sci-Fi', 'H	['Horror', 'Sci-Fi', 'Thriller']		Blade	0	['Sci-Fi', 'H	['Action', 'Horror', 'Sci-Fi']	
	Jurassic-Park-The-Lost-V	0.307692	['Sci-Fi', 'A	['Action', 'Adventure', 'Family', 'Fantasy', 'Thriller']		Jurassic-P	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Fantasy', 'Horror', '	
	Alien	0.307692	['Sci-Fi', 'H	['Action', 'Adventure', 'Mystery', 'Sci-Fi']		Alien	0.307692	['Sci-Fi', 'H	['Adventure', 'Horror', 'Mystery', 'Sci-Fi']	
	Thor	0.153846	['Adventur	['Adventure', 'Fantasy', 'Mystery']		Thor	0.153846	['Adventur	['Action', 'Adventure', 'Sci-Fi']	
	average hl	0.208791				average h	0.186813			
	#########	#########				########	#########			
	#########	TESTIFY				########	TESTIFY			
Indexing Default	Title	Hamming	Actual Ge	Predict Genre		Title	Hamming	Actual Ge	Predict Genre	
GenreVectorSize : 300	Titanic	0.307692	['Romanc	['Action', 'Adventure']		Titanic	0.307692	['Romanc	['Action', 'Adventure']	
	Total-Recall	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Drama', 'Thriller']		Total-Rec	0.153846	['Sci-Fi', 'A	['Action', 'Crime', 'Sci-Fi', 'Thriller']	
	Godfather	0.153846	['Crime', 'I	['Drama', 'Romance']		Godfather	0	['Crime', 'I	['Crime', 'Drama']	
	Inglourious-Basterds	0.307692	['Adventur	['Drama', 'Thriller']		Inglouriou	0.307692	['Adventur	['Comedy', 'Drama']	
	Mad-Max-2-The-Road-V	0.153846	['Sci-Fi', 'A	['Action', 'Crime', 'Sci-Fi']		Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Crime']	
	Hackers	0.307692	['Action', '	['Action', 'Comedy', 'Crime', 'Romance']		Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']	
	War-Horse	0.153846	['Drama']	['Adventure']		War-Hors	0.153846	['Drama']	['Adventure']	
	Yes-Man	0.307692	['Romanc	['Animation', 'Family']		Yes-Man	0.307692	['Romanc	['Animation', 'Family']	
	Lord-of-the-Rings-The-T	0	['Adventur	['Action', 'Adventure', 'Fantasy']		Lord-of-th	0	['Adventur	['Action', 'Adventure', 'Fantasy']	
	Hostage	0.153846	['Action', '	['Crime', 'Drama', 'Horror', 'Thriller']		Hostage	0.307692	['Action', '	['Comedy', 'Crime', 'Mystery', 'Thriller']	
	Blade	0.307692	['Sci-Fi', 'H	['Fantasy', 'Horror', 'Thriller']		Blade	0.153846	['Sci-Fi', 'H	['Action', 'Fantasy', 'Horror']	
	Jurassic-Park-The-Lost-V	0	['Sci-Fi', 'A	['Action', 'Adventure', 'Horror', 'Sci-Fi', 'Thriller']		Jurassic-P	0	['Sci-Fi', 'A	['Action', 'Adventure', 'Horror', 'Sci-Fi', 'Th	
	Alien	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Sci-Fi', 'Thriller']		Alien	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Horror', 'Sci-Fi']	
	Thor	0.307692	['Adventur	['Action', 'Crime', 'Romance']		Thor	0.153846	['Adventur	['Action', 'Adventure', 'Sci-Fi']	
	average hl	0.208791				average h	0.175824			
	#########	#########				########	#########			
	#########	TESTIFY				########	TESTIFY			
Indexing Default	Title	Hamming	Actual Ge	Predict Genre		Title	Hamming	Actual Ge	Predict Genre	
GenreVectorSize : 500	Titanic	0.307692	['Romanc	['Action', 'Thriller']		Titanic	0.307692	['Romanc	['Action', 'Adventure']	
	Total-Recall	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Drama', 'Thriller']		Total-Rec	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Mystery', 'Thriller']	
	Godfather	0	['Crime', 'I	['Crime', 'Drama']		Godfather	0	['Crime', 'I	['Crime', 'Drama']	
	Inglourious-Basterds	0.307692	['Adventur	['Drama', 'Thriller']		Inglouriou	0.307692	['Adventur	['Comedy', 'Drama']	
	Mad-Max-2-The-Road-V	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Thriller']		Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Crime']	
	Hackers	0.153846	['Action', '	['Comedy', 'Crime', 'Drama', 'Thriller']		Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']	
	War-Horse	0	['Drama']	['Drama']		War-Hors	0.153846	['Drama']	['Adventure']	
	Yes-Man	0.307692	['Romanc	['Animation', 'Family']		Yes-Man	0.153846	['Romanc	['Comedy', 'Family']	
	Lord-of-the-Rings-The-T	0	['Adventur	['Action', 'Adventure', 'Fantasy']		Lord-of-th	0	['Adventur	['Action', 'Adventure', 'Fantasy']	
	Hostage	0.153846	['Action', '	['Crime', 'Drama', 'Horror', 'Thriller']		Hostage	0.307692	['Action', '	['Comedy', 'Crime', 'Mystery', 'Thriller']	
	Blade	0.153846	['Sci-Fi', 'H	['Action', 'Horror', 'Thriller']		Blade	0.153846	['Sci-Fi', 'H	['Action', 'Horror', 'Thriller']	
	Jurassic-Park-The-Lost-V	0.153846	['Sci-Fi', 'A	['Action', 'Drama', 'Horror', 'Sci-Fi', 'Thriller']		Jurassic-P	0	['Sci-Fi', 'A	['Action', 'Adventure', 'Horror', 'Sci-Fi', 'Th	
	Alien	0.153846	['Sci-Fi', 'H	['Action', 'Drama', 'Sci-Fi', 'Thriller']		Alien	0	['Sci-Fi', 'H	['Action', 'Horror', 'Sci-Fi', 'Thriller']	
	Thor	0.307692	['Adventur	['Adventure', 'Drama', 'Romance']		Thor	0.153846	['Adventur	['Action', 'Adventure', 'Drama']	
	average hl	0.186813				average h	0.164835			
	#########	#########				########	#########			
	#########	TESTIFY				########	TESTIFY			
Indexing Default	Title	Hamming	Actual Ge	Predict Genre		Title	Hamming	Actual Ge	Predict Genre	
GenreVectorSize : 700	Titanic	0.307692	['Romanc	['Action', 'Thriller']		Titanic	0.307692	['Romanc	['Action', 'Adventure']	
	Total-Recall	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Drama', 'Thriller']		Total-Rec	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Mystery', 'Thriller']	
	Godfather	0.153846	['Crime', 'I	['Drama', 'Thriller']		Godfather	0	['Crime', 'I	['Crime', 'Drama']	
	Inglourious-Basterds	0.307692	['Adventur	['Comedy', 'Drama']		Inglouriou	0.307692	['Adventur	['Comedy', 'Drama']	
	Mad-Max-2-The-Road-V	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Thriller']		Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Crime']	
	Hackers	0.153846	['Action', '	['Comedy', 'Crime', 'Drama', 'Thriller']		Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']	
	War-Horse	0.153846	['Drama']	['Fantasy']		War-Hors	0.153846	['Drama']	['Adventure']	
	Yes-Man	0.307692	['Romanc	['Animation', 'Family']		Yes-Man	0.153846	['Romanc	['Comedy', 'Family']	
	Lord-of-the-Rings-The-T	0	['Adventur	['Action', 'Adventure', 'Fantasy']		Lord-of-th	0	['Adventur	['Action', 'Adventure', 'Fantasy']	
	Hostage	0.153846	['Action', '	['Crime', 'Drama', 'Horror', 'Thriller']		Hostage	0.307692	['Action', '	['Comedy', 'Crime', 'Mystery', 'Thriller']	
	Blade	0.153846	['Sci-Fi', 'H	['Action', 'Horror', 'Thriller']		Blade	0.153846	['Sci-Fi', 'H	['Action', 'Horror', 'Thriller']	
	Jurassic-Park-The-Lost-V	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Drama', 'Horror', 'Thriller']		Jurassic-P	0	['Sci-Fi', 'A	['Action', 'Adventure', 'Horror', 'Sci-Fi', 'Th	
	Alien	0.153846	['Sci-Fi', 'H	['Action', 'Drama', 'Sci-Fi', 'Thriller']		Alien	0	['Sci-Fi', 'H	['Action', 'Horror', 'Sci-Fi', 'Thriller']	
	Thor	0.307692	['Adventur	['Adventure', 'Drama', 'Thriller']		Thor	0.153846	['Adventur	['Action', 'Adventure', 'Drama']	
	average hl	0.197802				average h	0.164835			
	#########	#########				########	#########			

idf: case 2 $\frac{N}{(1+df)}$

#######	TESTIFY		
Title	Hamming	Actual Ge	Predict Genre
Titanic	0.307692	['Romanc	['Action', 'Adventure']
Total-Rec	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Horror', 'Sci-Fi']
Godfather	0.153846	['Crime', 'I	['Action', 'Drama']
Inglouriou	0.307692	['Adventur	['Sci-Fi', 'Thriller']
Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Comedy', 'Sci-Fi']
Hackers	0.307692	['Action', '	['Action', 'Comedy', 'Drama', 'Sci-Fi']
War-Hors	0.153846	['Drama']	['Adventure']
Yes-Man	0.307692	['Romanc	['Animation', 'Family']
Lord-of-th	0.153846	['Adventur	['Adventure', 'Fantasy', 'Mystery']
Hostage	0.307692	['Action', '	['Crime', 'Mystery', 'Romance', 'Thriller']
Blade	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Horror']
Jurassic-P	0.307692	['Sci-Fi', 'A	['Action', 'Adventure', 'Drama', 'Family', 'Thriller']
Alien	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Horror', 'Sci-Fi']
Thor	0.153846	['Adventur	['Action', 'Adventure', 'Romance']
average h	0.21978		
#######	#########		
#######	TESTIFY		
Title	Hamming	Actual Ge	Predict Genre
Titanic	0.307692	['Romanc	['Action', 'Adventure']
Total-Rec	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Mystery', 'Thriller']
Godfather	0.153846	['Crime', 'I	['Action', 'Drama']
Inglouriou	0.307692	['Adventur	['Comedy', 'Drama']
Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Comedy']
Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']
War-Hors	0.153846	['Drama']	['Adventure']
Yes-Man	0.307692	['Romanc	['Animation', 'Family']
Lord-of-th	0.153846	['Adventur	['Adventure', 'Fantasy', 'Mystery']
Hostage	0.153846	['Action', '	['Crime', 'Drama', 'Mystery', 'Thriller']
Blade	0.153846	['Sci-Fi', 'H	['Horror', 'Sci-Fi', 'Thriller']
Jurassic-P	0.307692	['Sci-Fi', 'A	['Action', 'Adventure', 'Family', 'Fantasy', 'Thriller']
Alien	0.307692	['Sci-Fi', 'H	['Action', 'Adventure', 'Mystery', 'Sci-Fi']
Thor	0.307692	['Adventur	['Adventure', 'Mystery', 'Thriller']
average h	0.230769		
#######	#########		
#######	TESTIFY		
Title	Hamming	Actual Ge	Predict Genre
Titanic	0.307692	['Romanc	['Adventure', 'Thriller']
Total-Rec	0.153846	['Sci-Fi', 'A	['Action', 'Crime', 'Sci-Fi', 'Thriller']
Godfather	0.307692	['Crime', 'I	['Family', 'Romance']
Inglouriou	0.153846	['Adventur	['Action', 'Thriller']
Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Crime', 'Sci-Fi']
Hackers	0.461538	['Action', '	['Action', 'Comedy', 'Romance', 'Sci-Fi']
War-Hors	0.153846	['Drama']	['Adventure']
Yes-Man	0.307692	['Romanc	['Animation', 'Family']
Lord-of-th	0	['Adventur	['Action', 'Adventure', 'Fantasy']
Hostage	0.153846	['Action', '	['Crime', 'Drama', 'Horror', 'Thriller']
Blade	0.307692	['Sci-Fi', 'H	['Fantasy', 'Horror', 'Thriller']
Jurassic-P	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Drama', 'Sci-Fi', 'Thriller']
Alien	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Sci-Fi', 'Thriller']
Thor	0.307692	['Adventur	['Action', 'Crime', 'Romance']
average h	0.21978		
#######	#########		
#######	TESTIFY		
Title	Hamming	Actual Ge	Predict Genre
Titanic	0.307692	['Romanc	['Action', 'Thriller']
Total-Rec	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Drama', 'Thriller']
Godfather	0.153846	['Crime', 'I	['Drama', 'Thriller']
Inglouriou	0.307692	['Adventur	['Drama', 'Thriller']
Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Thriller']
Hackers	0.307692	['Action', '	['Comedy', 'Drama', 'Romance', 'Thriller']
War-Hors	0	['Drama']	['Drama']
Yes-Man	0.307692	['Romanc	['Animation', 'Family']
Lord-of-th	0	['Adventur	['Action', 'Adventure', 'Fantasy']
Hostage	0.153846	['Action', '	['Crime', 'Drama', 'Horror', 'Thriller']
Blade	0.153846	['Sci-Fi', 'H	['Action', 'Horror', 'Thriller']
Jurassic-P	0.153846	['Sci-Fi', 'A	['Action', 'Horror', 'Mystery', 'Sci-Fi', 'Thriller']
Alien	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Sci-Fi', 'Thriller']
Thor	0.307692	['Adventur	['Adventure', 'Drama', 'Romance']
average h	0.197802		
#######	#########		
#######	TESTIFY		
Title	Hamming	Actual Ge	Predict Genre
Titanic	0.307692	['Romanc	['Adventure', 'Family']
Total-Rec	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Mystery', 'Thriller']
Godfather	0	['Crime', 'I	['Crime', 'Drama']
Inglouriou	0.307692	['Adventur	['Comedy', 'Drama']
Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Crime']
Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']
War-Hors	0.153846	['Drama']	['Adventure']
Yes-Man	0	['Romanc	['Comedy', 'Romance']
Lord-of-th	0	['Adventur	['Action', 'Adventure', 'Fantasy']
Hostage	0.307692	['Action', '	['Comedy', 'Crime', 'Mystery', 'Thriller']
Blade	0.153846	['Sci-Fi', 'H	['Action', 'Horror', 'Thriller']
Jurassic-P	0	['Sci-Fi', 'A	['Action', 'Adventure', 'Horror', 'Sci-Fi', 'Thriller']
Alien	0.153846	['Sci-Fi', 'H	['Horror', 'Mystery', 'Sci-Fi', 'Thriller']
Thor	0.153846	['Adventur	['Action', 'Adventure', 'Crime']
average h	0.164835		
#######	#########		

idf: case 3 $\log_{10} 10 + \frac{N}{(1+df)}$

#######	TESTIFY		
Title	Hamming	Actual Ge	Predict Genre
Titanic	0.307692	['Romanc	['Adventure', 'Family']
Total-Rec	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Mystery', 'Thriller']
Godfather	0	['Crime', 'I	['Crime', 'Drama']
Inglouriou	0.307692	['Adventur	['Mystery', 'Romance']
Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Comedy']
Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']
War-Hors	0.153846	['Drama']	['Adventure']
Yes-Man	0.307692	['Romanc	['Animation', 'Family']
Lord-of-th	0.153846	['Adventur	['Adventure', 'Fantasy', 'Sci-Fi']
Hostage	0.307692	['Action', '	['Comedy', 'Crime', 'Mystery', 'Thriller']
Blade	0.153846	['Sci-Fi', 'H	['Fantasy', 'Horror', 'Sci-Fi']
Jurassic-P	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Fantasy', 'Horror', 'Sci-Fi']
Alien	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Horror', 'Sci-Fi']
Thor	0.153846	['Adventur	['Action', 'Adventure', 'Romance']
average h	0.208791		
#######	#########		
#######	TESTIFY		
Title	Hamming	Actual Ge	Predict Genre
Titanic	0.307692	['Romanc	['Adventure', 'Family']
Total-Rec	0.153846	['Sci-Fi', 'A	['Action', 'Mystery', 'Sci-Fi', 'Thriller']
Godfather	0	['Crime', 'I	['Crime', 'Drama']
Inglouriou	0.307692	['Adventur	['Comedy', 'Romance']
Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Fantasy']
Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']
War-Hors	0.153846	['Drama']	['Adventure']
Yes-Man	0.307692	['Romanc	['Animation', 'Family']
Lord-of-th	0.153846	['Adventur	['Adventure', 'Fantasy', 'Sci-Fi']
Hostage	0.307692	['Action', '	['Comedy', 'Crime', 'Mystery', 'Thriller']
Blade	0	['Sci-Fi', 'H	['Action', 'Horror', 'Sci-Fi']
Jurassic-P	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Fantasy', 'Horror', 'Sci-Fi']
Alien	0.307692	['Sci-Fi', 'H	['Adventure', 'Horror', 'Mystery', 'Sci-Fi']
Thor	0.153846	['Adventur	['Action', 'Adventure', 'Sci-Fi']
average h	0.197802		
#######	#########		
#######	TESTIFY		
Title	Hamming	Actual Ge	Predict Genre
Titanic	0.307692	['Romanc	['Action', 'Adventure']
Total-Rec	0.153846	['Sci-Fi', 'A	['Action', 'Crime', 'Sci-Fi', 'Thriller']
Godfather	0.153846	['Crime', 'I	['Crime', 'Romance']
Inglouriou	0.307692	['Adventur	['Comedy', 'Drama']
Mad-Max	0	['Sci-Fi', 'A	['Action', 'Adventure', 'Sci-Fi']
Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']
War-Hors	0.153846	['Drama']	['Adventure']
Yes-Man	0.307692	['Romanc	['Animation', 'Family']
Lord-of-th	0	['Adventur	['Action', 'Adventure', 'Fantasy']
Hostage	0.307692	['Action', '	['Comedy', 'Crime', 'Mystery', 'Thriller']
Blade	0.153846	['Sci-Fi', 'H	['Action', 'Fantasy', 'Horror']
Jurassic-P	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Fantasy', 'Horror', 'Sci-Fi']
Alien	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Horror', 'Sci-Fi']
Thor	0.153846	['Adventur	['Action', 'Adventure', 'Sci-Fi']
average h	0.186813		
#######	#########		
#######	TESTIFY		
Title	Hamming	Actual Ge	Predict Genre
Titanic	0.307692	['Romanc	['Action', 'Adventure']
Total-Rec	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Mystery', 'Thriller']
Godfather	0	['Crime', 'I	['Crime', 'Drama']
Inglouriou	0.307692	['Adventur	['Comedy', 'Drama']
Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Crime']
Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']
War-Hors	0.153846	['Drama']	['Adventure']
Yes-Man	0.307692	['Romanc	['Animation', 'Family']
Lord-of-th	0	['Adventur	['Action', 'Adventure', 'Fantasy']
Hostage	0.153846	['Action', '	['Crime', 'Drama', 'Mystery', 'Thriller']
Blade	0.153846	['Sci-Fi', 'H	['Action', 'Horror', 'Thriller']
Jurassic-P	0	['Sci-Fi', 'A	['Action', 'Adventure', 'Horror', 'Sci-Fi', 'Thriller']
Alien	0.153846	['Sci-Fi', 'H	['Action', 'Adventure', 'Horror', 'Sci-Fi']
Thor	0.153846	['Adventur	['Action', 'Adventure', 'Sci-Fi']
average h	0.175824		
#######	#########		
#######	TESTIFY		
Title	Hamming	Actual Ge	Predict Genre
Titanic	0.307692	['Romanc	['Action', 'Adventure']
Total-Rec	0.307692	['Sci-Fi', 'A	['Action', 'Crime', 'Mystery', 'Thriller']
Godfather	0	['Crime', 'I	['Crime', 'Drama']
Inglouriou	0.307692	['Adventur	['Comedy', 'Drama']
Mad-Max	0.153846	['Sci-Fi', 'A	['Action', 'Adventure', 'Crime']
Hackers	0.307692	['Action', '	['Comedy', 'Crime', 'Drama', 'Romance']
War-Hors	0.153846	['Drama']	['Adventure']
Yes-Man	0.153846	['Romanc	['Comedy', 'Family']
Lord-of-th	0	['Adventur	['Action', 'Adventure', 'Fantasy']
Hostage	0.307692	['Action', '	['Comedy', 'Crime', 'Mystery', 'Thriller']
Blade	0.153846	['Sci-Fi', 'H	['Action', 'Horror', 'Thriller']
Jurassic-P	0	['Sci-Fi', 'A	['Action', 'Adventure', 'Horror', 'Sci-Fi', 'Thriller']
Alien	0	['Sci-Fi', 'H	['Action', 'Horror', 'Sci-Fi', 'Thriller']
Thor	0.153846	['Adventur	['Action', 'Adventure', 'Drama']
average h	0.164835		
#######	#########		

위는 Hamming Loss 에 대한 값이 엑셀로 저장된 것이다. 이와 같이 Precision/Recall, R-precision 에 대한 값 또한 같은 형태로 저장된다.