

Word2Vector

Word2Vector는 단어를 벡터 공간에 표현하는 혁신적인 자연어 처리 기술입니다. 이 기술은 단어의 의미적 관계를 수학적으로 표현하여, 컴퓨터가 인간의 언어를 더 잘 이해할 수 있게 해줍니다.

예를 들어, '왕 - 남자 + 여자 = 여왕'과 같은 단어 간의 관계를 벡터 연산으로 표현할 수 있으며, 이는 자연어 처리의 다양한 분야에서 활용됩니다. 특히 문서 분류, 감성 분석, 기계 번역 등에서 핵심적인 역할을 수행합니다.

1.Word2Vector

단어를 숫자로 바꿔야 하는 이유

컴퓨터는 문자를 바로 이해하지 못합니다. 대신, 숫자나 코드로 바꿔야만 정보를 처리할 수 있습니다. 예를 들어:

- 사람은 "사랑"이라는 단어를 보면 감정을 떠올립니다.
- 하지만 컴퓨터는 "사랑"이라는 글자를 이해하지 못합니다. 대신 숫자로 변환해 "사랑"의 의미를 계산하거나 비교할 수 있도록 해야 합니다.

이 과정에서 사용하는 것이 **WordVector**입니다.

WordVector란?

WordVector는 단어를 숫자로 표현한 것입니다.

단순히 "사랑 = 1, 가족 = 2"처럼 고유 번호를 매기는 것이 아니라, **단어의 의미와 관계를 반영한 숫자 집합(벡터)**으로 표현합니다.

비유로 이해하기

단어를 위치로 표현하기

단어를 숫자로 바꾸는 과정을 지도에 비유할 수 있습니다.

- **지도 비유:** 단어 하나하나를 지도 위의 점으로 생각해보세요.
 - "사랑"과 "가족"은 서로 가깝게 위치합니다. (두 단어가 의미상 가까움)
 - "사랑"과 "컴퓨터"는 멀리 떨어져 있습니다. (두 단어가 의미상 멀음)
- 컴퓨터는 이 점들의 위치를 숫자 좌표로 저장합니다.
 - 예: "사랑" = (1.2, 3.4), "가족" = (1.1, 3.3), "컴퓨터" = (10.5, 8.2)

즉, WordVector는 단어를 숫자 좌표로 표현해 단어 간의 거리와 관계를 알 수 있게 해줍니다.

WordVector가 어떻게 만들어지나요?

컴퓨터가 단어를 숫자로 바꾸는 방법은 다음과 같습니다:

- 1. 단어를 많이 읽기:** 컴퓨터는 책이나 기사처럼 많은 문장을 읽으면서 단어들이 주로 어떤 단어들과 같이 사용되는지 배웁니다.
 - 예: "사랑"은 "가족", "우정", "행복" 같은 단어들과 자주 같이 나옵니다.
- 2. 단어 관계 계산:** 컴퓨터는 단어들 간의 관계를 학습한 뒤, 이 관계를 숫자로 표현합니다.
 - "사랑"은 "가족"과 가까운 숫자를 가집니다.
 - "사랑"은 "컴퓨터"와 먼 숫자를 가집니다.
- 3. 숫자 좌표 만들기:** 컴퓨터는 각 단어에 대해 좌표(숫자 벡터)를 생성합니다.
 - "사랑" = (1.2, 3.4)
 - "가족" = (1.1, 3.3)
 - "컴퓨터" = (10.5, 8.2)

현실에서 WordVector 사용 사례

1. 영화 추천

- "로맨스 영화"를 좋아하면 "사랑"과 비슷한 단어를 가진 영화(예: "우정", "행복")가 추천됩니다.

2. 챗봇

- WordVector를 사용해 사용자의 질문과 비슷한 의미를 가진 답변을 찾아냅니다.

3. 검색 엔진

- "애플 스마트폰"을 검색하면 WordVector를 통해 관련 있는 단어(예: "아이폰", "iOS")를 찾아 검색 결과에 포함합니다.

2. gensim 라이브러리

gensim 라이브러리

gensim 라이브러리는 Python에서 쉽게 Word2Vec을 사용할 수 있게 도와주는 도구이며, `from gensim.models.word2vec import Word2Vec`를 통해 이 기능을 불러옵니다.

CBOW와 Skip-gram

Word2Vec은 단어를 숫자로 바꾸는 두 가지 주요 방법을 제공합니다: CBOW와 Skip-gram

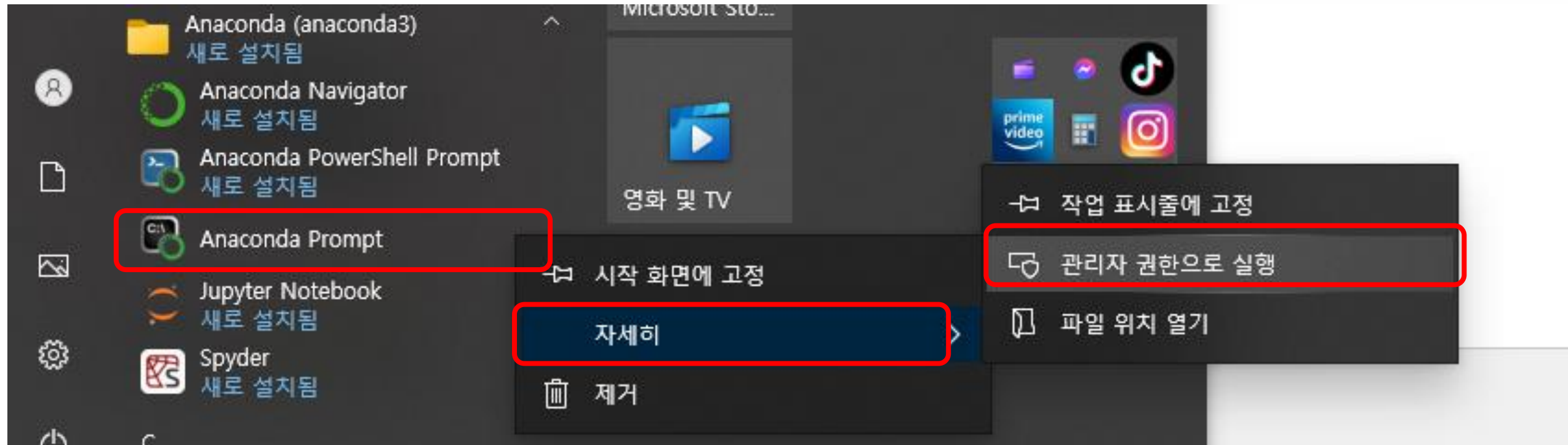
1. CBOW (Continuous Bag of Words) 의미: "주변 단어를 보고 중심 단어를 예측"합니다. 작동 방식: 문장에서 특정 단어의 앞뒤 단어(주변 단어)를 보고 중심 단어를 추측합니다. 예: "나는 ___을 좋아합니다."에서 나는, 을, 좋아합니다.를 이용해 ___(중심 단어)를 예측. 특징: 속도가 빠릅니다. 대량의 데이터를 처리하기에 적합합니다. 중심 단어가 자주 사용될수록 학습 성능이 좋아집니다.
2. Skip-gram 의미: "중심 단어를 보고 주변 단어를 예측"합니다. 작동 방식: 문장에서 특정 단어(중심 단어)를 보고 앞뒤 단어(주변 단어)를 추측합니다. 예: "나는 사랑을 좋아합니다."에서 사랑(중심 단어)을 이용해 나는, 좋아합니다.(주변 단어)를 예측. 특징: 의미가 적게 등장하는 단어(희귀 단어)도 잘 학습합니다. CBOW보다 계산 비용이 높아 학습 속도가 느릴 수 있습니다. CBOW vs Skip-gram 요약

CBOW와 Skip-gram

	CBOW	Skip-gram
목적	주변 단어로 중심 단어 예측	중심 단어로 주변 단어 예측
속도	빠름	느림
희귀 단어	잘 학습되지 않음	잘 학습함
적합한 경우	데이터가 크고 단어가 자주 등장할 때	데이터가 적고 희귀 단어가 중요한 경우

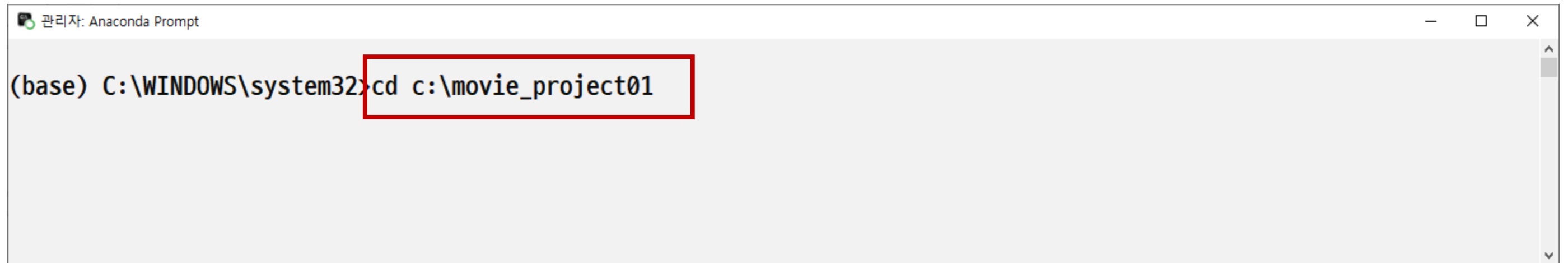
3. Word2Vector 실습

Word2Vector 실습



Word2Vector실습

- 아래를 입력 합니다

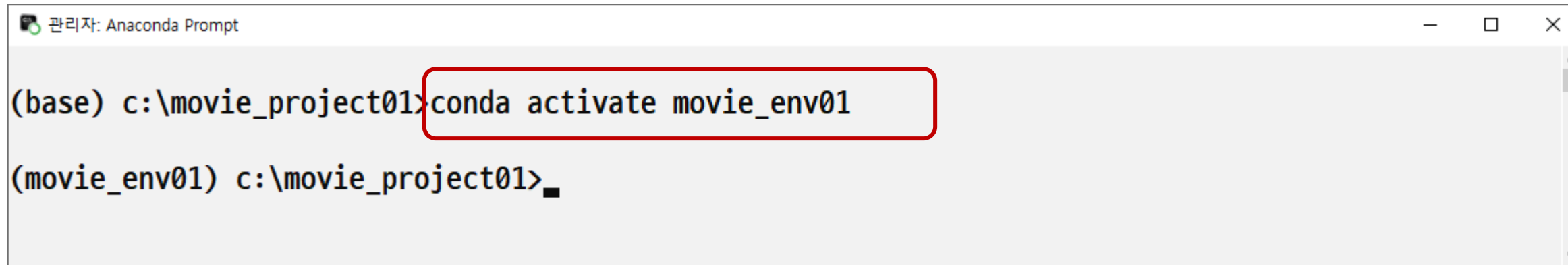


A screenshot of an Anaconda Prompt terminal window. The title bar at the top reads "관리자: Anaconda Prompt" and includes standard window control buttons (minimize, maximize, close). The terminal content shows the prompt "(base) C:\WINDOWS\system32" followed by the command "cd c:\movie_project01". The command and its arguments are enclosed in a red rectangular box, indicating where the user should input the command.

```
(base) C:\WINDOWS\system32>cd c:\movie_project01
```

Word2Vector실습

- 아래 명령을 입력 합니다



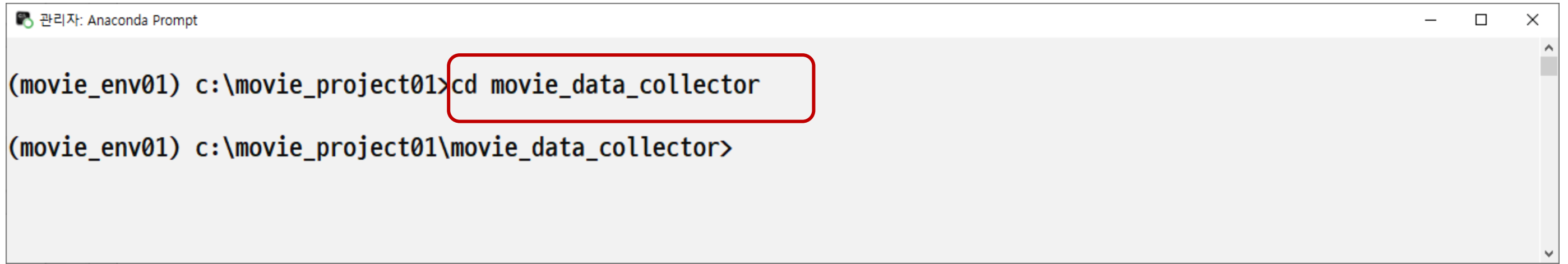
```
관리자: Anaconda Prompt
(base) c:\movie_project01>conda activate movie_env01
(movie_env01) c:\movie_project01>_
```

- 이 명령어의 의미:

- conda: Anaconda라는 프로그램에서 사용되는 명령어입니다.
- activate: 특정 가상 환경을 사용하도록 전환하는 명령입니다.
- movie_env01: 이전에 만들어둔 가상 환경의 이름입니다.
- 가상 환경을 활성화하는 것은 프로젝트 전용 작업 공간에 들어가는 것과 같습니다.movie_env01은 특정 작업(예: 영화 프로젝트)을 위한 개인 작업 공간입니다.이 공간 안에서는 다른 프로젝트와 관련 없는 도구나 설정을 사용할 걱정 없이 해당 프로젝트에 최적화된 환경에서 작업할 수 있습니다.즉, 이 작업 공간은 영화 프로젝트를 위한 "전용 책상" 같은 역할을 합니다.

Word2Vector실습

- 아래를 입력 합니다



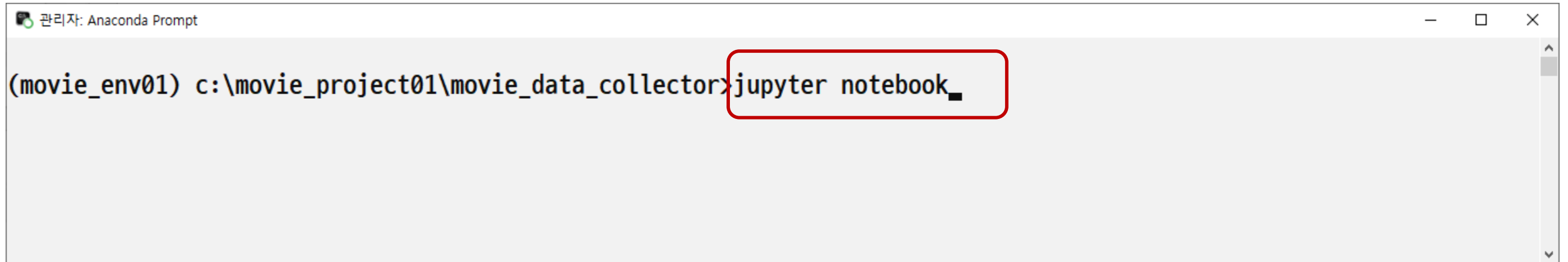
A screenshot of an Anaconda Prompt terminal window. The title bar reads '관리자: Anaconda Prompt'. The terminal shows two lines of text: the first line is '(movie_env01) c:\movie_project01>cd movie_data_collector' and the second line is '(movie_env01) c:\movie_project01\movie_data_collector>'. The command 'cd movie_data_collector' in the first line is highlighted with a red rounded rectangle.

```
(movie_env01) c:\movie_project01>cd movie_data_collector
(movie_env01) c:\movie_project01\movie_data_collector>
```

- 프로젝트가 다운로드된 movie_data_collector 디렉토리로 이동합니다

Word2Vector실습

- 아래를 입력 합니다



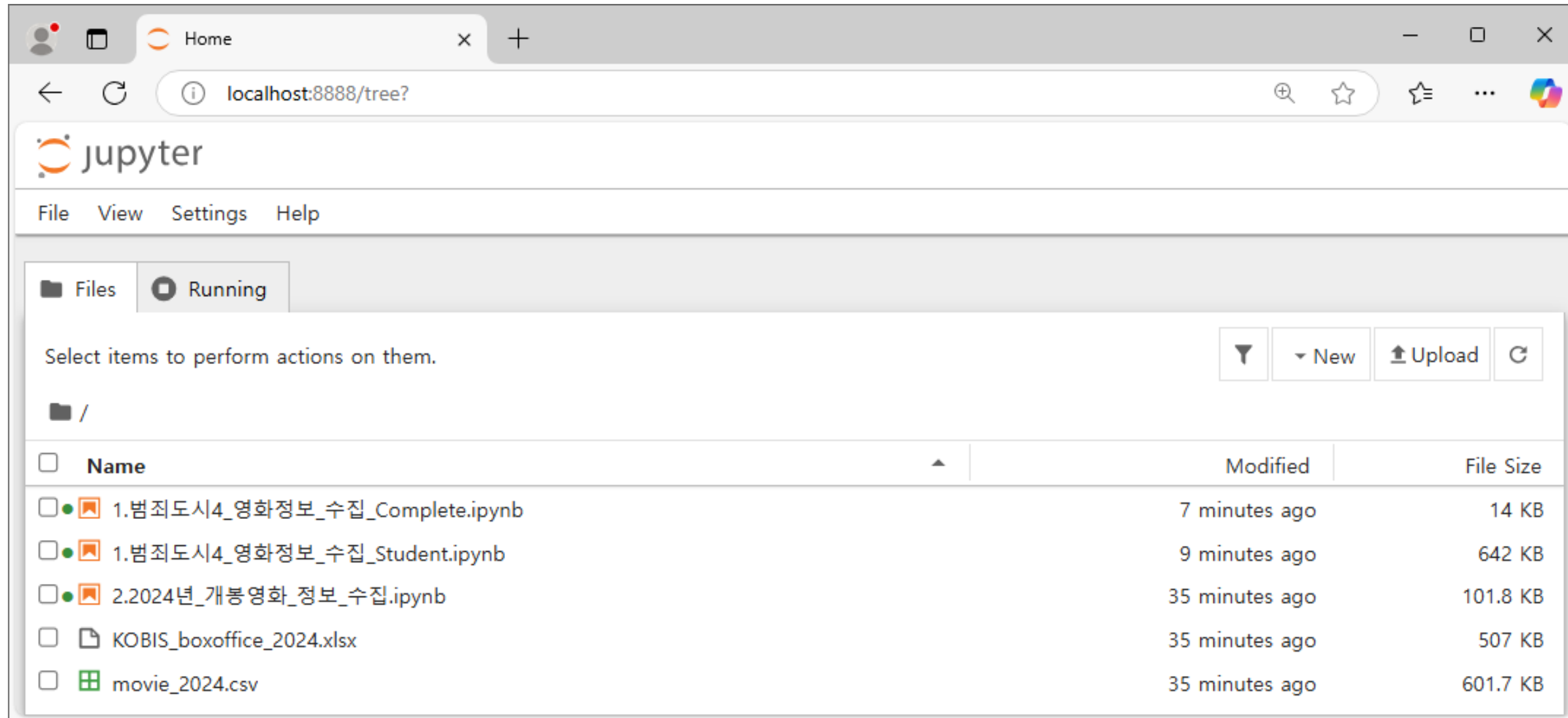
The image shows a screenshot of an Anaconda Prompt window. The title bar reads '관리자: Anaconda Prompt'. The command prompt shows the current directory as 'c:\movie_project01\movie_data_collector' and the command 'jupyter notebook' is being entered. The command 'jupyter notebook' is highlighted with a red rounded rectangle. A cursor is at the end of the command line.

```
(movie_env01) c:\movie_project01\movie_data_collector>jupyter notebook_
```

- jupyter notebook은 Jupyter Notebook을 실행하라는 명령입니다.명령을 실행하면 컴퓨터의 웹 브라우저(크롬, 엣지 등)에서 Jupyter Notebook 화면이 열립니다.

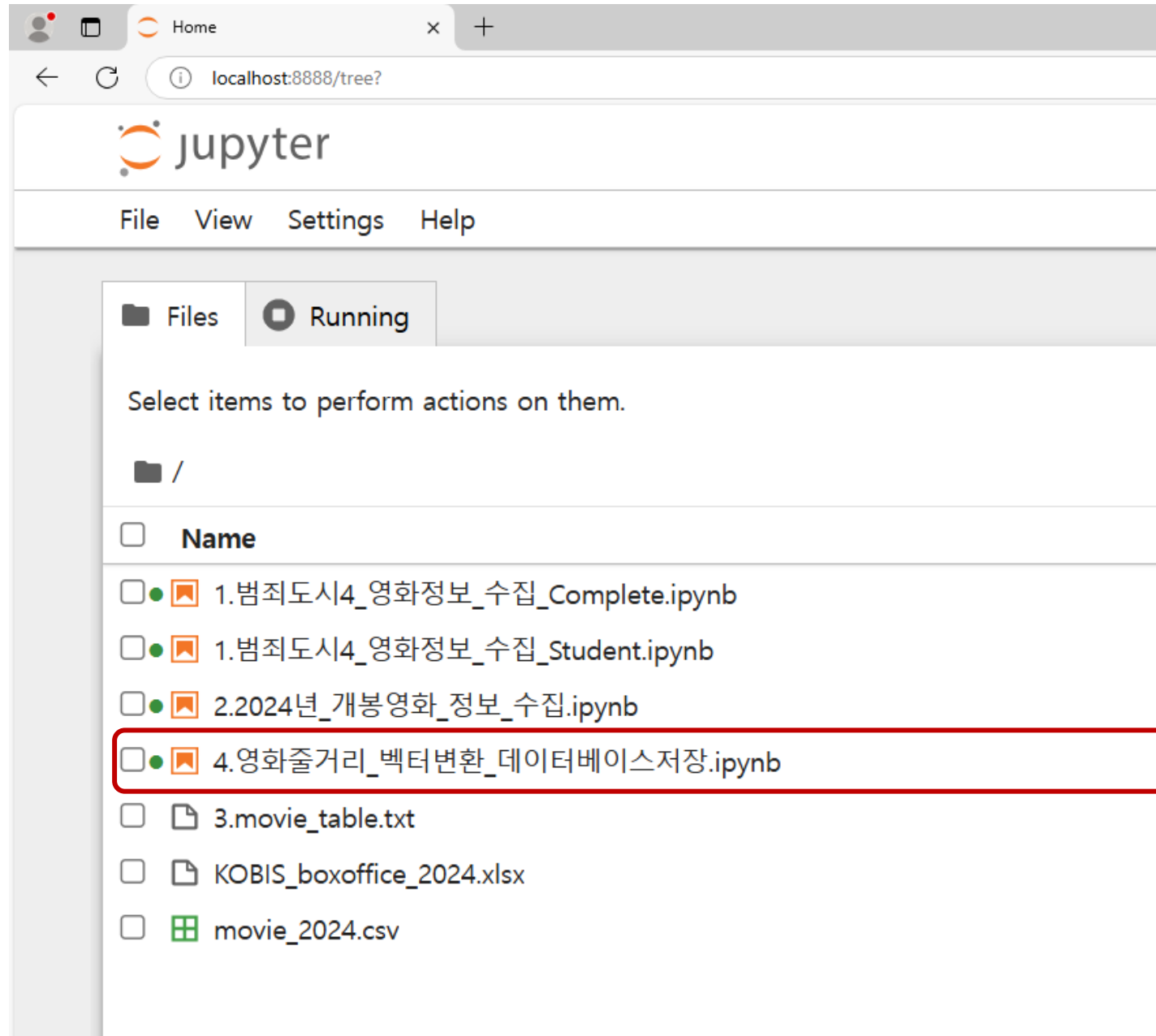
Word2Vector실습

- 웹 브라우저 주소 표시줄에 <http://localhost:8888> 을 입력 합니다



Word2Vector실습

■ 4.영화줄거리_벡터변환_데이터베이스저장.ipynb 를 클릭 합니다



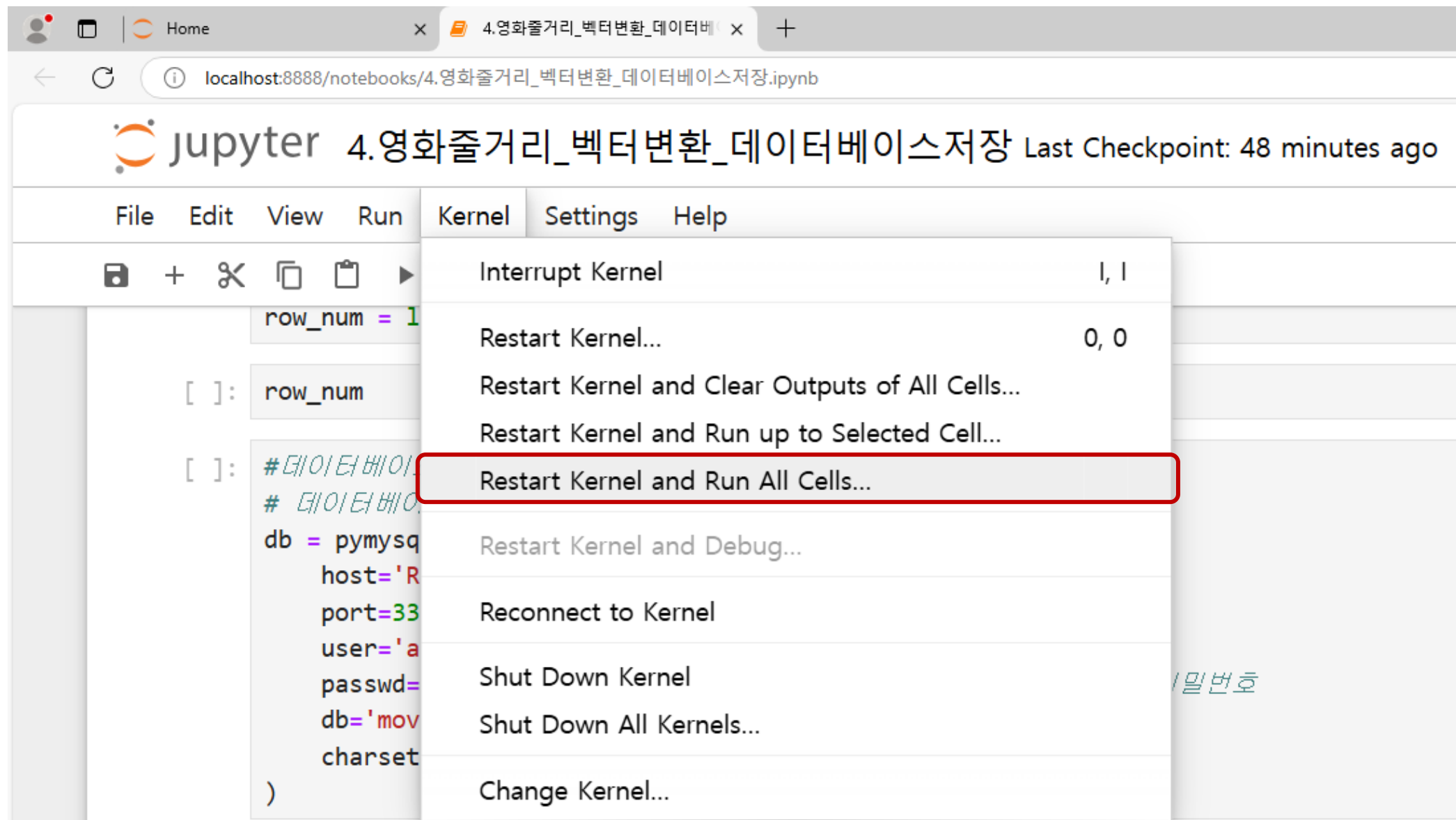
Word2Vector실습

- 화면 아래로 스크롤 한 후 다음 부분을 수정 합니다
- Ctrl+S를 눌러서 저장 합니다

```
[ ]: #데이터베이스 연결
# 데이터베이스 연결 설정
db = pymysql.connect(
    host='RDS 엔드포인트를 입력 합니다', # 데이터베이스 서버 주소
    port=3306, # 데이터베이스 연결 포트
    user='admin', # 데이터베이스 사용자 이름
    passwd='RDS 비밀번호를 입력 합니다', # 데이터베이스 사용자 비밀번호
    db='movie_db', # 데이터베이스 이름
    charset='utf8' # 데이터 인코딩 설정
)
```

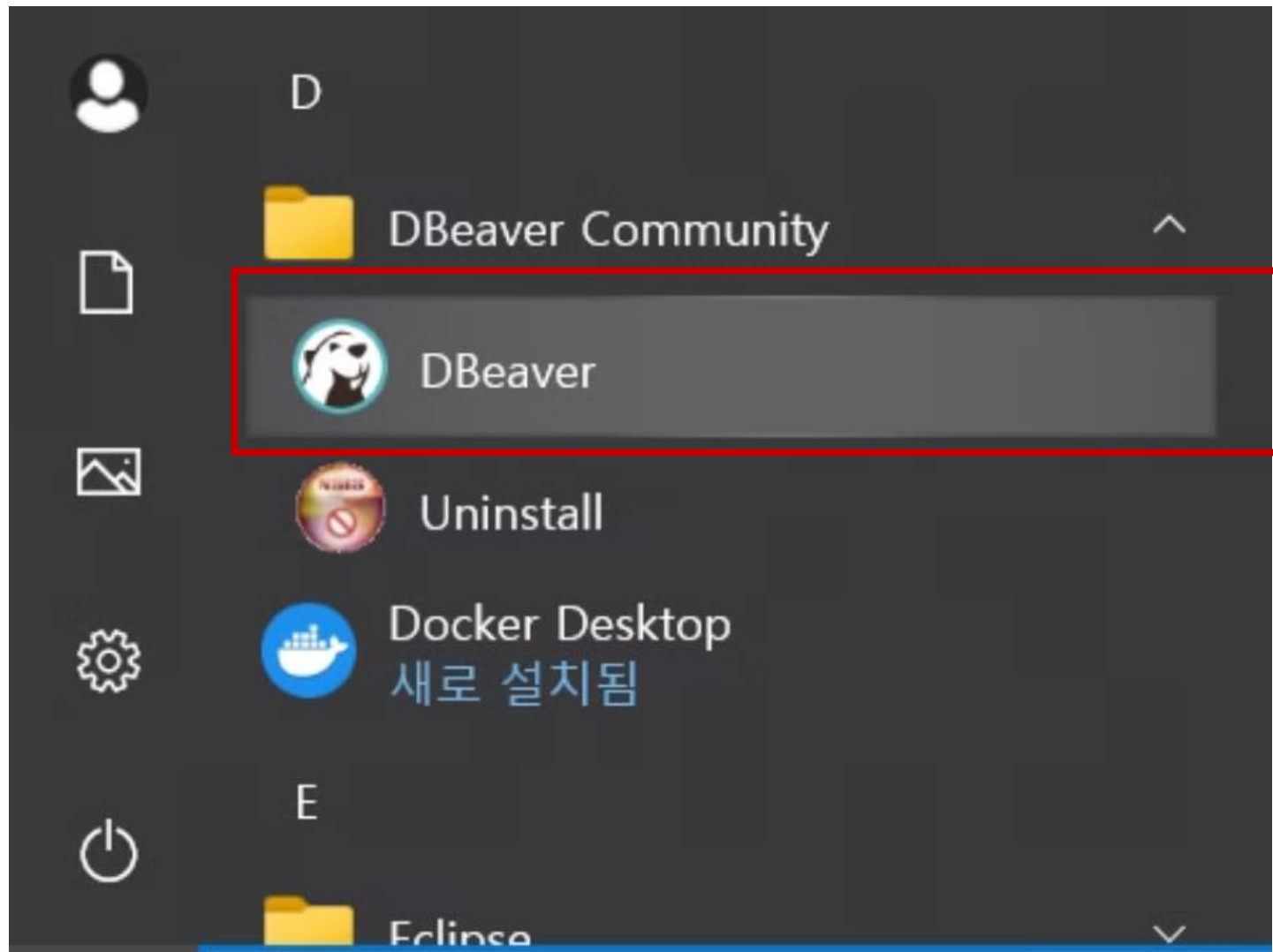
Word2Vector실습

- 모두 실행을 선택 합니다



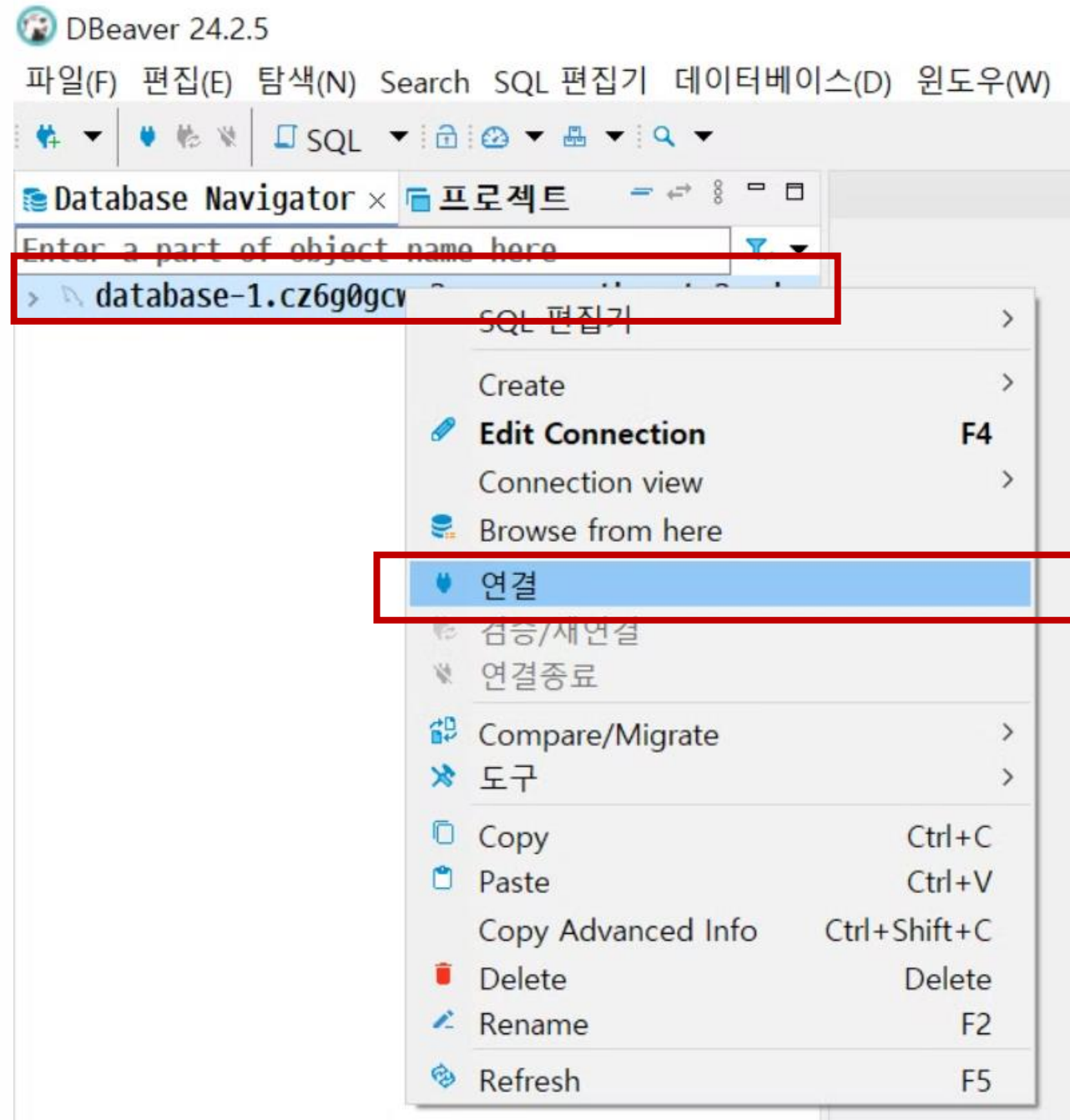
Word2Vector실습

DBeaver를 선택합니다



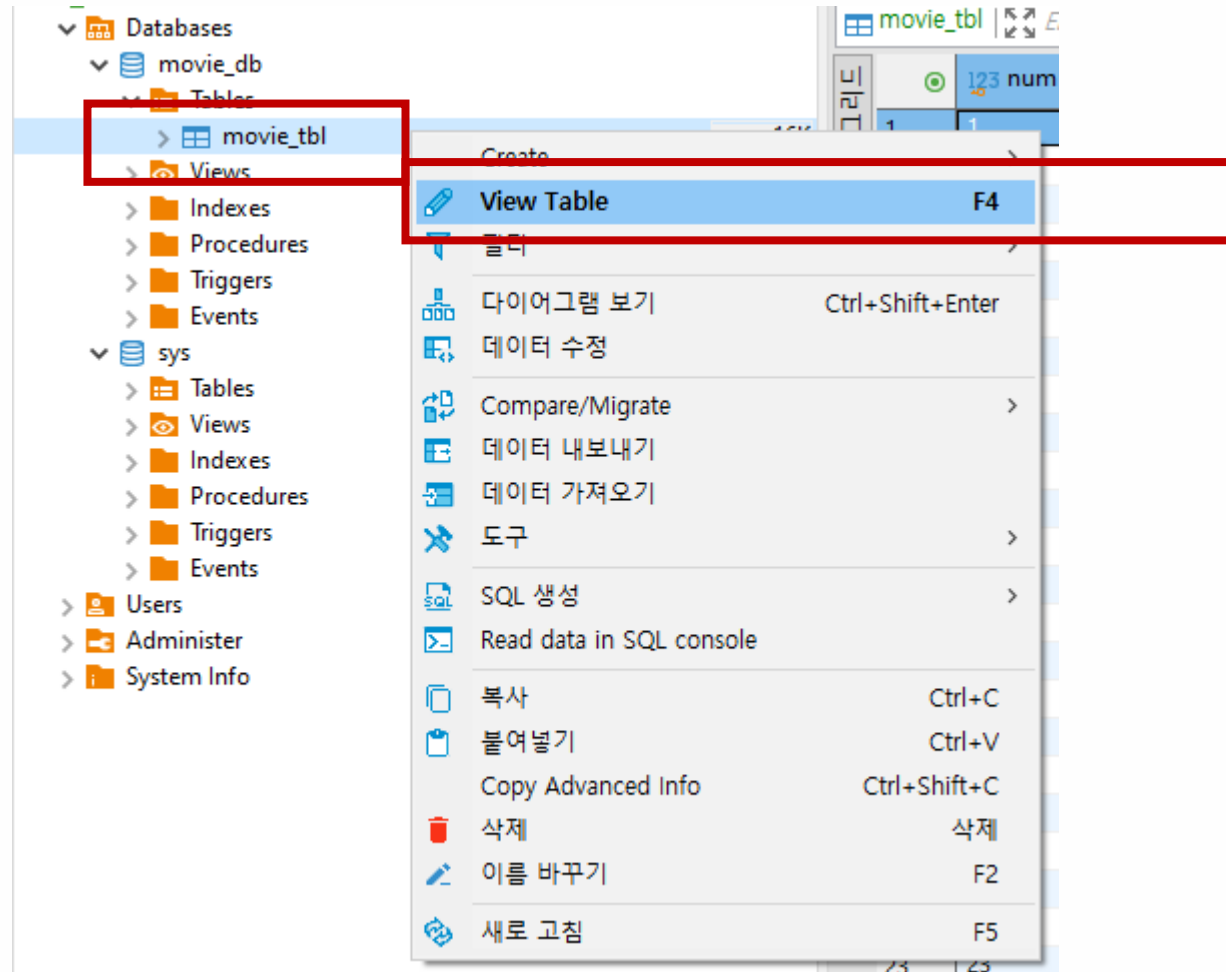
Word2Vector실습

데이터베이스 연결 정보를 선택하고 연결을 선택합니다



Word2Vector실습

테이블을 클릭하고 View Table 을 선택 합니다



Word2Vector실습

데이터 탭을 클릭해서 영화 데이터가 저장되었는지 확인 합니다

Property		Data		엔티티 관계도		movie-rds.cz6g0gcww2ra.ap-northeast-2.rds.amazonaws.com										Databases		movie_db		Tables		mov					
movie_tbl		Enter a SQL expression to filter results (use Ctrl+Space)																									
		123 num		A-Z title		A-Z director		A-Z actor		A-Z synopsis		A-Z poster		open_date		A-Z degr											
1		1		파묘		장재현		최민식,김고은,유해진,이도현,김재철,김민준,김병오,전진기,박정자,박지일,이종구		미국 LA, 거액의 의뢰를 받은 무당		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-02-22		15세 이상											
2		2		범죄도시4		허명행		김무열,박지환,이동휘,마동석,곽자형,이범수,김민재,이지훈,이주빈,김도건,김지훈		신종 마약 사건 3년 뒤, 괴물형사 '이		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-04-24		15세 이상											
3		3		인사이드 아웃 2		켈시 만		에이미 포엘러,토니 헤일		디즈니.픽사의 대표작 <인사이드		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-06-12		전체 관람											
4		4		베테랑2		류승완		황정민,정해인,장윤주,정만식,신승환,오달수,오대환,김시후,안보현,진경,권해효,변		가족들도 못 챙기고 밤낮없이 범조		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-09-13		15세 이상											
5		5		파일럿		김한결		조정석,이주영,한선화,신승호,오민애,김지현,서재희,박다운,현봉식,서현철,유재석,		하루 아침에 인생 추락한 스타 파일		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-07-31		12세 이상											
6		6		탈주		이종필		이제훈,구교환,홍사빈,서현우,이성욱,정준원,장영남,송강,이솜,이호정,배철수,이호		"내 앞 길 내가 정했습니다" 휴전선		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-07-03		12세 이상											
7		7		소방관		곽경택		주원,곽도원,유재명,이유영,김민재,오대환,이준혁,장영남		살리기 위해, 살아남기 위해 하루하		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-12-04		12세 이상											
8		8		한섬가이즈		남동협		이성민,이희준,공승연,박지환,이규형,우현		"우리가 뭐 빠지는 게 있노? 집도		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-06-26		15세 이상											
9		9		하이재킹		김성한		하정우,여진구,성동일,채수빈,문유강,임형택,전정관,정예진,문우진,홍정혜,전영,변		1971년 겨울 속초공항 여객기 조종		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-06-21		12세 이상											
10		10		시민덕희		박영주		라미란,공명,염혜란,박병은,장윤주,이무생,안은진,김을호		내 돈을 사기 친 그 놈이 구조 요청		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-01-24		15세 이상											
11		11		외계+인 2부		최동훈		류준열,김태리,김우빈,이하늬,염정아,조우진,김의성,진선규,신정근,윤경호,이시훈,		반드시 돌아가야 한다. 모두를 지켜		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-01-10		12세 이상											
12		12		서울의 봄		김성수		심가영,윤호림,황정민,이미라,정우성,이성민,박해준,김성균,김의성,정동환,안내상,		1979년 12월 12일, 수도 서울 군사		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2023-11-22		12세 이상											
13		13		사랑의 하츠피		김수훈		이지현,조경아,최낙윤,여민정,홍소영,홍범기,김하영,류승곤,조현정,김은아,장예나,		처음 본 순간, 한눈에 반해버렸어!		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-08-07		전체 관람											
14		14		그녀가 죽었다		김세휘		변요한,신혜선,이엘,전정일		"나쁜 짓은 절대 안 해요. 그냥 보기		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-05-15		15세 이상											
15		15		노랑: 죽음의 바다		김한민		김윤석,백윤식,정재영,허준호,김성규,이규형,이무생,최덕문,안보현,박명훈,박훈,문		임진왜란 발발로부터 7년이 지난 '		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2023-12-20		12세 이상											
16		16		히든페이스		김대우		송승헌,조여정,박지현,박지영,박성근,변종희,박영준,이혜령,유예린,주예린,이소영,		'갈렸다 지켜봤다 벗겨졌다' 지휘자		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-11-20		청소년 관											
17		17		뎃글부대		안국진		박상혁,손희서,손석구,김성철,김동휘,홍경,이찬유,이현,박민국,김한솔		실력 있지만 허세 가득한 사회부		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-03-27		15세 이상											
18		18		대도시의 사랑법		이연희		김고은,허지유,방정민,노상현,정휘,오동민,김채은,권영은,서벽준,박지안,이유진,정		미친X과 게이가 만났다! 바야흐로		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-10-01		15세 이상											
19		19		청설		조선희		홍경,송유라,노윤서,김민주,정용주,정혜영,현봉식,안민영,고경만,장인호,하승연,김		손으로 설렘을 말하고 가슴으로 시		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-11-06		전체 관람											
20		20		명탐정 코난: 100만 달러의 펜타그램		나가오카 치카		김선혜,강수진,신용우,최재호		홋카이도 하코다테에 있는 '오노어		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-07-17		12세 이상											
21		21		행복의 나라		추창민		조정석,이선균,유재명,우현,이원종,전배수,송영규,최원영,강말금,박훈,이현균,진기		1979년 10월 26일, 대통령 암살 사:		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-08-14		12세 이상											
22		22		탈출: 프로젝트 사일런스		김태곤		이선균,주지훈,김희원,문성근,예수정,김태우,박희본,박주현,김수안,하도권,장광,최		붕괴 위기의 공항대교, 생존자 전		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-07-12		15세 이상											
23		23		보통의 가족		허진호		설경구,장동건,김희애,수현		물질적 욕망을 우선시하며 살인자		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-10-16		15세 이상											
24		24		원더랜드		김태용		탕웨이,배수지,박보검,정유미,최우식,탕준상,성병숙,최무성,이엘,강애심,김성령,빅		언제 어디서든 다시 만날 수 있습니		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-06-05		12세 이상											
25		25		아마존 활명수		김창주		류승룡,진선규,이그르 페드로소,루안 브를,J.B. 올리베이라,염혜란,고경표		어서 와, 아마존은 처음이지 전 양:		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-10-30		12세 이상											
26		26		설계자		이요섭		강동원,이무생,이미숙,이현우,탕준상,김홍파,김신록,이동휘,정은채		"정말 우연이라고 생각해요?" 의로		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-05-29		15세 이상											
27		27		빅토리		박범수		이혜리,방정민,박세완,이정하,조아람,최지수,백하이,권유나,염지영,이한주,박효은		1999년 세기말 거제, 출만이 전부		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-08-14		12세 이상											
28		28		도그데이즈		김덕민		윤여정,유해진,김윤진,정성화,김서형,다니엘 헤니,이현우,탕준상,윤채나,김고은		깔끔한 성격의 계획형 싱글남 '민		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-02-07		12세 이상											
29		29		소풍		김용균		나문희,김영욱,박근형,류승수,이항나,공상아,임지규,최선자,이용이,한태일,곽자형,		60년 만에 찾아간 고향, 16살의 추		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-02-07		12세 이상											
30		30		임영웅 아이 히어로 더 스타디움		조우영,정현철		임영웅		10만 영웅시대와 함께 상암발을 점		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-08-28		전체 관람											
31		31		1승		신연식		송강호,박정민,장윤주		"그래도 한 번은 이기겠죠?" 지도		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-12-04		12세 이상											
32		32		브레이드이발소: 뽕스타의 탄생		정지환		엄상현,박윤희,강시현,홍범기,김현욱,윤아영,김보나,안영미,정의진		대한민국 NO.1 애니메이션 '브레드		https://search.pstatic.net/common?type=o&size=176x264&quality=85&direct		2024-09-14		전체 관람											