

# NATURAL LANGUAGE PROCESSING

## MOVIE RETRIEVAL SYSTEM

### GENERAL INFORMATION

#### I - SUBJECT

For our final project I decided to work on topic III which is how to retrieve a list of movies with a given query (e.g., "love" for movie1, movie2, etc).

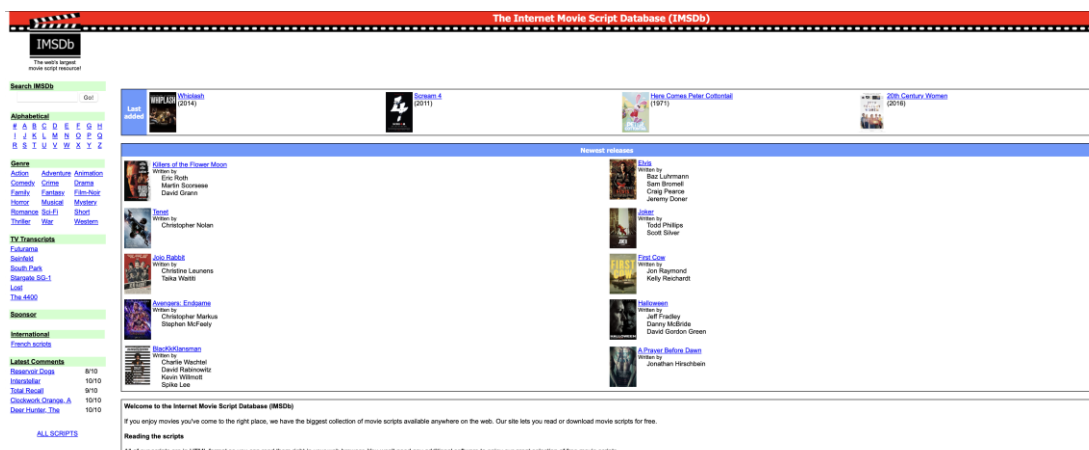
The objective for this topic is to retrieve a list of movies giving a given query as the task demands and to compare the effectiveness of different methods and their original versions versus their enhanced versions, in order to achieve this, we are going to use techniques like precision and recall.

#### II – DATA

To first work we have to get different movies and their scripts. To achieve this, we made a scrapper that:

- First to fetch all the contain from the link “<https://imsdb.com/>” we use the library “request” of python.
- Secondly with the result of the first step we use the library “beautiful soup” to correctly parse the html content of “<https://imsdb.com/>”. With the result we also use soup in order to find all the links that redirect us to our goals, the different movies.
- Thirdly, while having the links of every movie, we iterate it and request each links to get their content.
- Fourth the content from each links is collected with “beautiful soup” and stock in a CSV File with the logic of three columns ID, Title, and Script.

In total 1225 movies were extracted from the scrapper and will be analyze by our different test and model.



Picture above: is the interface of the link “<https://imsdb.com/>”.

```
https://imsdb.com/scripts/Event-Horizon.html
Event Horizon
https://imsdb.com/scripts/Evil-Dead.html
Evil Dead
https://imsdb.com/scripts/Evil-Dead-II-Dead-by-Dawn.html
Evil Dead II Dead by Dawn
https://imsdb.com/scripts/Ex-Machina.html
Ex Machina
https://imsdb.com/scripts/Excalibur.html
Excalibur
https://imsdb.com/scripts/Executive-Decision.html
Executive Decision
https://imsdb.com/scripts/eXistenZ.html
eXistenZ
https://imsdb.com/scripts/Extract.html
Extract
https://imsdb.com/scripts/Eyes-Wide-Shut.html
Eyes Wide Shut
https://imsdb.com/scripts/Fabulous-Baker-Boys,-The.html
```

Picture above: During the extraction of the link “<https://imsdb.com/>”.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	ID,Title,Script																				
2	1,10 things i hate about you,ten things i hate about you by karen mccullah lutz kristen smith ten things i hate about you written by karen mccullah lutz kristen smith based on taming of the shrew by william shakespeare revision november 12 1997 padua high school day welcome t																				
3	scowl joey you better hope you re as smooth as you think you are verona patrick takes the money with a smile int tutoring room day cameron runs a sentence past bianca cameron la copine et i ami la difference bianca glares at him bianca a copine is someone you can count on an																				
4	reading in his rasta stoner drawl in faith i do not love thee with mine eyes for they in thee a thousand errors note but tis my heart that loves what they despise who in despite of view is pleas d to dote in the back of the room clem raises his hand clem ms blaise can i get the bathr																				
5	rt a horrifically nerdy freshman year picture glasses bad hair headgear the works she holds up the picture for all to view patrick cringes and throws a handful of pretzels at her bianca patrick is that a kat perm patrick ask my attorney kat and bianca huddle over the picture giggling i																				
6	2,12 script cut from black title fin exterior la day fin of red 1957 chevy impala convertible driving somewhere in the west a car passes going the other way title place los angeles music shadowy men on a shadowy planet good cop bad cop exterior la day pov driving on freeway i																				
7	why did you let the animals go here s a story about animals that caused erosion on the channel islands we don t need any more attention right now tony i scattered them for their own protection and it doesn t bother me because any animal no matter where it starts turns out the																				
8	ng where a wall has fallen off exposing the contents of a hotel room broken windows fallen brick cracks in walls more fallen facades homeless woman until the sun came up and then we went off to our room to see if there was any damage it was pretty bad it was quiet there wi																				
9	3,12 and holding,12 and holding written by anthony s cipriano 04 06 04 fade in ext neighborhood street morning twin boys rudy and jacob carges 12 ride their bikes through a suburban neighborhood rudy the more athletic of the two rides at a breakneck pace jacob rides slowly due																				
10	t be happening jim we agreed it would be too painful ashley you agreed jacob mom what happened ashley turns to jacob as if realizing his presence for the first time ashley tell him tell your son jim the boys who killed rudy cut a deal they got a year in juvenile hall and five years pr																				
11	ave been going since i was eight it s the guy s week off patrick this year i m taking the girls just the girls you re staying home leonard why can t we all go grace i m not staying here all alone besides you said it yourself you ve been going for years give your sisters a chance leonard																				
12	4,12 monkeys,twelve monkeys twelve monkeys an original screenplay by david peoples janet peoples inspired by la jete a chris marker film production draft june 27 1994 fade in int concourse airport terminal bay close on a face a nine year old boy young cole his eyes wide with w																				
13	s germs are just a plot they made up so they can sell you disinfectants and soap now he s crazy right hey you believe in germs don t you cole i m not crazy jeffrey of course not i never thought you were you want to escape right that s very sane i can help you you want me to don t y																				
14	icers exchange a look marlou martin it s just as i told you my husband and i had gone ahead she never showed that s totally unlike her officer two pulls out his notebook do you happen to know the make of her car marlou martin um acura 92 acura also that cat s starving she wou																				
15	sandwich rounding a bend an oncoming car heads straight at cole cole yanks the wheel as the other car horn blaring just misses him recovering cole loses the road speeds crazily along the shoulder int flying helicopter night the pilot an agent steers the chopper while the co pilot j																				
16	for assistance but he grabs her purse instead swings it around smashes wallace in the face with it then grabs the pimp s arm and snaps it like it was a twig the knife clatters to the floor as wallace yelps in pain and cole slams him to the floor straddles his chest retrieves the near																				
17	pens one of the glass tubes and waves it under the security officer s nose dr peters it doesn t even have an odor the security officer glances up sees what dr peters is doing and smiles as he hands the papers back to the scientist security officer that s not necessary sir here you go t																				
18	5,12 years a slave,12 years a slave written by john ridey card 1841 fade in i int townhouse study day 1 early april 1841 we are close on a pair of black hands as they open a finely wrapped packet of violin strings we cut to the hands stringing a violin it s not a high end piece but it i																				
19	dy giving solomon and eliza a moment to rouse themselves burch demands burch come on get yer blankets get up sensing that things will not end well eliza no please don t burch i don t want to hear yer talk get in the yard eliza please radburn ain t no need for all that putting hai																				
20	e disciples unto jesus saying who is the greatest in the kingdom of heaven 80 80 ext ford plantation morning august 1841 it s sunday the slaves are again gathered in the rose garden near the front of the house to hear the word of the lord as read by master ford ford and jesus call																				
21	olomon moves close but not too close as solomon draws within striking distance epps lunges for him he chases solomon on until he is again out of breath and once more drops down and again offering a treaty epps cont d i m all done in platt i have met my limitations and i ain t e																				
22	own an appropriate distance the three men take the body and very unceremoniously place it into the ground holding continued 98 145 continued 145 the shovel in his hands and resting it by his feet bob tilts his head down and closes his eyes the others do the same almost stutteri																				
23	6,127 hours,127 hours written by simon beaufoy danny boyle ext crowd scenes various a massive crowd it could be a sports stadium a u2 farewell show or new year s eve on copacabana beach but whatever it is there are thousands and thousands of a mexican wave erupts succ																				
24	int canyon dawn aron is very still looking at the rock and the open blade lying on top of it suddenly he looks over his shoulder to see cut to int canyon morning a dagger of sunlight appear behind him his won sunrise c u watch 9 30am cut to int canyon dawn he looks at the sunbean																				
25	s a little bothersome considering i m not bleeding that bad barely at all it s so weird you d expect to definitely see more pulsing and bleeding but oh well pause i m really fucked now i m out of water cut to int canyon day stops the video and rips a section of his t shirt to make an i																				
26	7,1492 conquest of paradise,1492 conquest of paradise 1492 conquest of paradise by roselyne bosch revised september 23 1991 fade in credits and music over int audience room granada day we start on a man s elegant slipper he is seated in a splendid chair moving up the stocki																				
27	s no sanchez no columbus no i have waited too long fought too hard now you expect me to take all the risks while you take the profit no i will not be your servant the eyes behind the screen the mouth forming a little smile sanchez i remind you senor colon that you are in no positio																				
28	ions my ships are not filled with the spices and the gold that spain was hoping for but this land intoxicates the senses like the strongest of perfumes and all i can think of is to return to these untamed lands suddenly the cabin seems to lurch over things fall from the table smashi																				
29	alcovy governor s mansion dawn the sky is dark and threatening the wind even stronger we find columbus where we left him but now alone he has sat here all night shutters bang violently he looks up and sees a few yards from him the naked figure of an indian his face and body																				
30	8,15 minutes,15 minutes 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 cz																				
31	lationship with my father who as you know made a fortune selling penny loafers in the fifties these people died because of the criminal actions of my doctor robert hawkins your doctor stephen geller yes my psychiatrist didn t insist that i stay on my medication robert hawkins so																				
32	rns to eddie nicollette cont d oh my g d they want me to anchor they want me to anchor tonight eddie that s good nicollette yeah eddie well that s great nicollette okay that is great but i can t go now we re in the middle of something here eddie no go ahead you re gonna be great n																				
33	prate for children the restaurant buzzes emil checks his watch oleg in movie they make of us who do you think would act me emil the one who got caught in the bathroom beat george michael emil laughs oleg doesn t oleg i m serious emil shut up look emil points towards the tv																				
34	arm from eddie eddie was killed because he was a celebrity jordi wants no part of it nicollette cut her cameraman lowers his leans jordi smiles she nods jordi turns and walks away nicollette cont d get a shot of him leaving then pan to me jordi disappears into the sea of people th																				
35	9,17 again,17 again written by jason filardi october 2007 ext fitch senior high school dusk a few cars scatter the parking lot we hear grunts followed by the distinct sound of basketballs shredding net int fitch senior high school gym continuous an empty gymnasium except for a shi																				
36	kers jeans t shirts hoodies shirts jewelry ed nods with his approval or disapproval at the register the girls hand mike his bags of clothes ed takes out his cell to put their numbers in it the girls frown and go back to work int virgin mega music store night mike and ed stand before th																				
37	e shhhh i get it now why you didn t want me to be with stan the nice things you said in the library it s because you wanted me maggie backs mike against the wall mike maggie listen to me i m not the person you think maggie shhh yes you are you re a good guy you re not like the i																				

Picture above: Filter the extraction content and put it the scripts of each movie in the csv file  
(Small part of the result got).

## ANALYSIS

### I – VSM MODEL

#### A – INTRODUCTION

Vector Space Model (VSM) it's a technique for representing data in text format in a high dimensional vector space, where each dimension corresponds to a term in the vocabulary.

VSM models can capture semantic similarity, in other words the meaning behind words. Similar documents will have vectors that are closer together in space. For the project this is useful because a VSM model can get the semantic similarity between queries and movie scripts.

Each document  $j$  is represented as a vector  $\mathbf{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$  where  $t$  is the total number of terms. In the other hand, the query is represents as a vector  $\mathbf{q}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$  using TF-IDF, BoW (Bag of words), word embedding or other methods for calculating the weights. Then similarity calculations is made between the document vectors and the query vector using cosine similarity's

#### B – APPROACHES

For the VSM model I used three different approaches for calculating the weights: TF-IF, BoW and word embedding. This was due to the fact that each of these methods have their own benefits and disadvantages.

#### C – TF-IDF

TF-IDF uses statistics to evaluate how important a term is to the corpus. It combines two different components, TF and IDF:

**TF:** Term Frequency (TF) quantifies how often a term appears in a document seeing this way how important a term is to a specific document.

**IDF:** Asses how exclusive the term is across the entire corpus, assigning higher values to terms that appear in fewer documents.

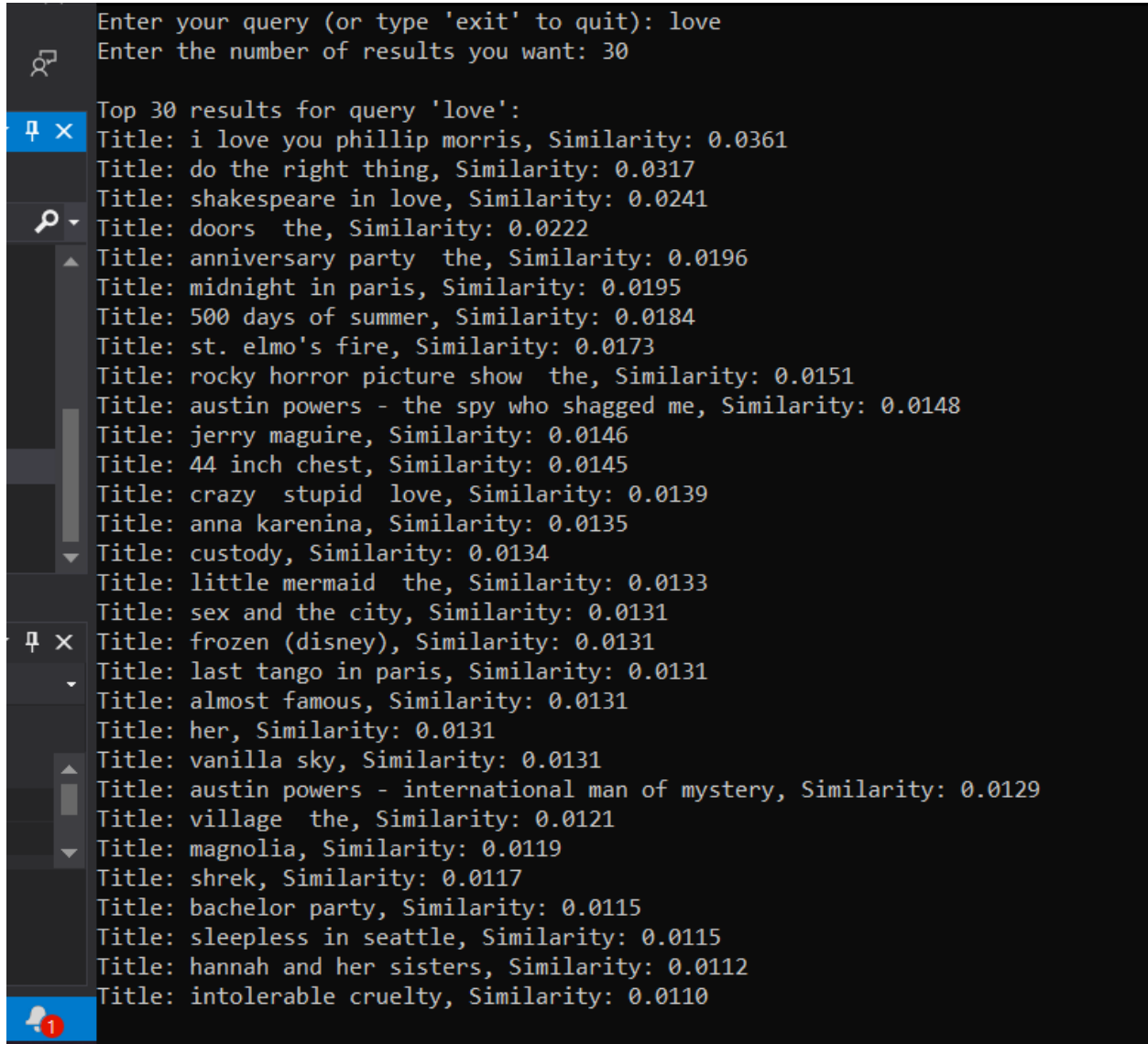
**TF-IDF:** This value is obtained by multiplying the TF with the IDF, thus highlighting terms that are both frequent in each document and rare across the corpus. This method helps in distinguishing significant terms from common ones, enhancing the model's ability to identify relevant documents based on their content.

Nevertheless, TF-IDF still treats each term independently and lacks the capability to understand contextual or semantic similarities between terms.

## D – RESULTS APPROACHES TF-IDF

### 1 – FIRST EXECUTION

Here is the result of the **first execution** with the query “love”.



```

Enter your query (or type 'exit' to quit): love
Enter the number of results you want: 30

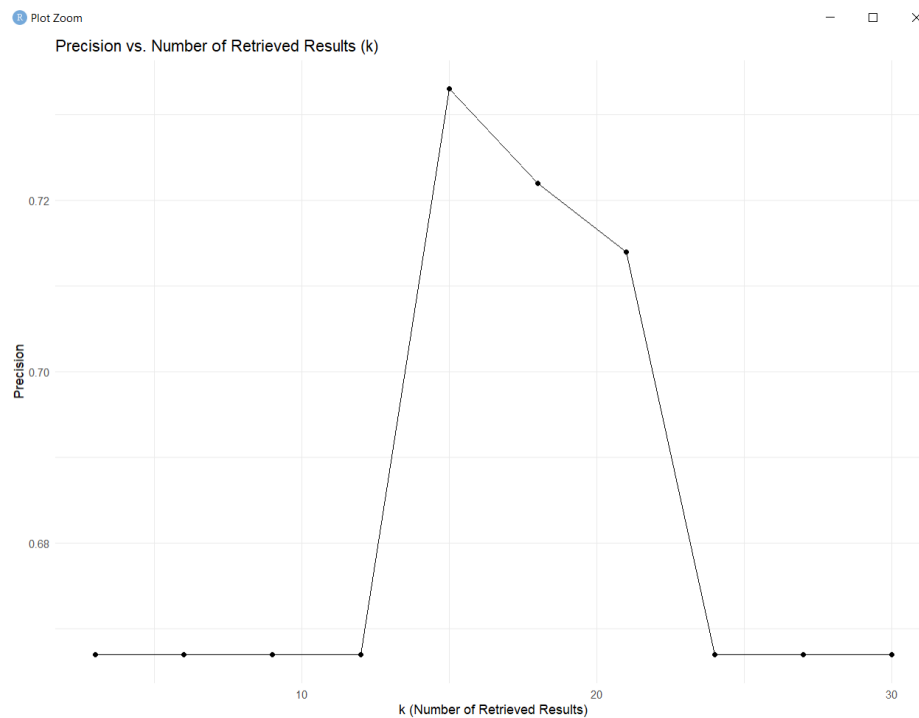
Top 30 results for query 'love':
Title: i love you phillip morris, Similarity: 0.0361
Title: do the right thing, Similarity: 0.0317
Title: shakespeare in love, Similarity: 0.0241
Title: doors the, Similarity: 0.0222
Title: anniversary party the, Similarity: 0.0196
Title: midnight in paris, Similarity: 0.0195
Title: 500 days of summer, Similarity: 0.0184
Title: st. elmo's fire, Similarity: 0.0173
Title: rocky horror picture show the, Similarity: 0.0151
Title: austin powers - the spy who shagged me, Similarity: 0.0148
Title: jerry maguire, Similarity: 0.0146
Title: 44 inch chest, Similarity: 0.0145
Title: crazy stupid love, Similarity: 0.0139
Title: anna karenina, Similarity: 0.0135
Title: custody, Similarity: 0.0134
Title: little mermaid the, Similarity: 0.0133
Title: sex and the city, Similarity: 0.0131
Title: frozen (disney), Similarity: 0.0131
Title: last tango in paris, Similarity: 0.0131
Title: almost famous, Similarity: 0.0131
Title: her, Similarity: 0.0131
Title: vanilla sky, Similarity: 0.0131
Title: austin powers - international man of mystery, Similarity: 0.0129
Title: village the, Similarity: 0.0121
Title: magnolia, Similarity: 0.0119
Title: shrek, Similarity: 0.0117
Title: bachelor party, Similarity: 0.0115
Title: sleepless in seattle, Similarity: 0.0115
Title: hannah and her sisters, Similarity: 0.0112
Title: intolerable cruelty, Similarity: 0.0110
  
```

### Similarity scores

The similarity scores indicate the degree of relevance between each movie's script and the query. Higher similarity scores suggest greater relevance to the query. Even if similarity scores are low that doesn't mean it's bad. This can be due to having a large corpus of documents and because the term “love” can be used in lots of meanings so it may be in all documents but not all of them are relevant.

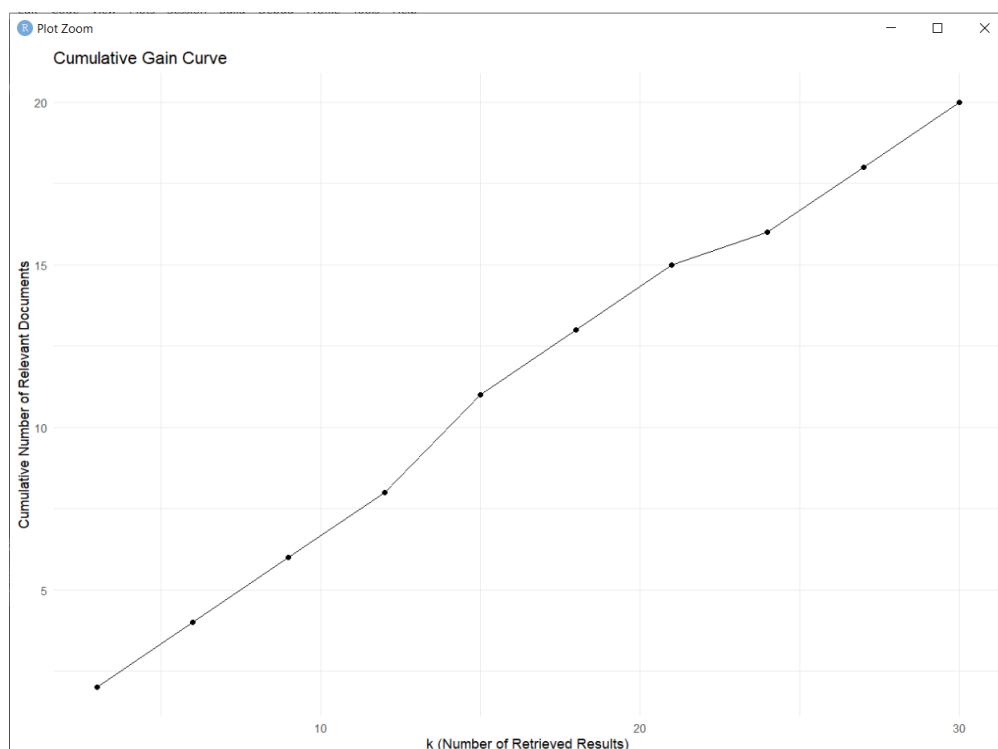
### Assessing model performance for different $k$ values

We selected the total of movies to retrieve to  $k = 30$ , so 30 movies would be retrieved even if only 2 movies in the corpus were related to the query.



The precision@ $k$  value remained constant from  $k=3$  until  $k=12$  with a precision of 66.7%. This meant that the number of movies being retrieved was increasing constantly, having 2 valid movies out of every 3 retrieved. After the 30 documents we got a precision of 66.7%.

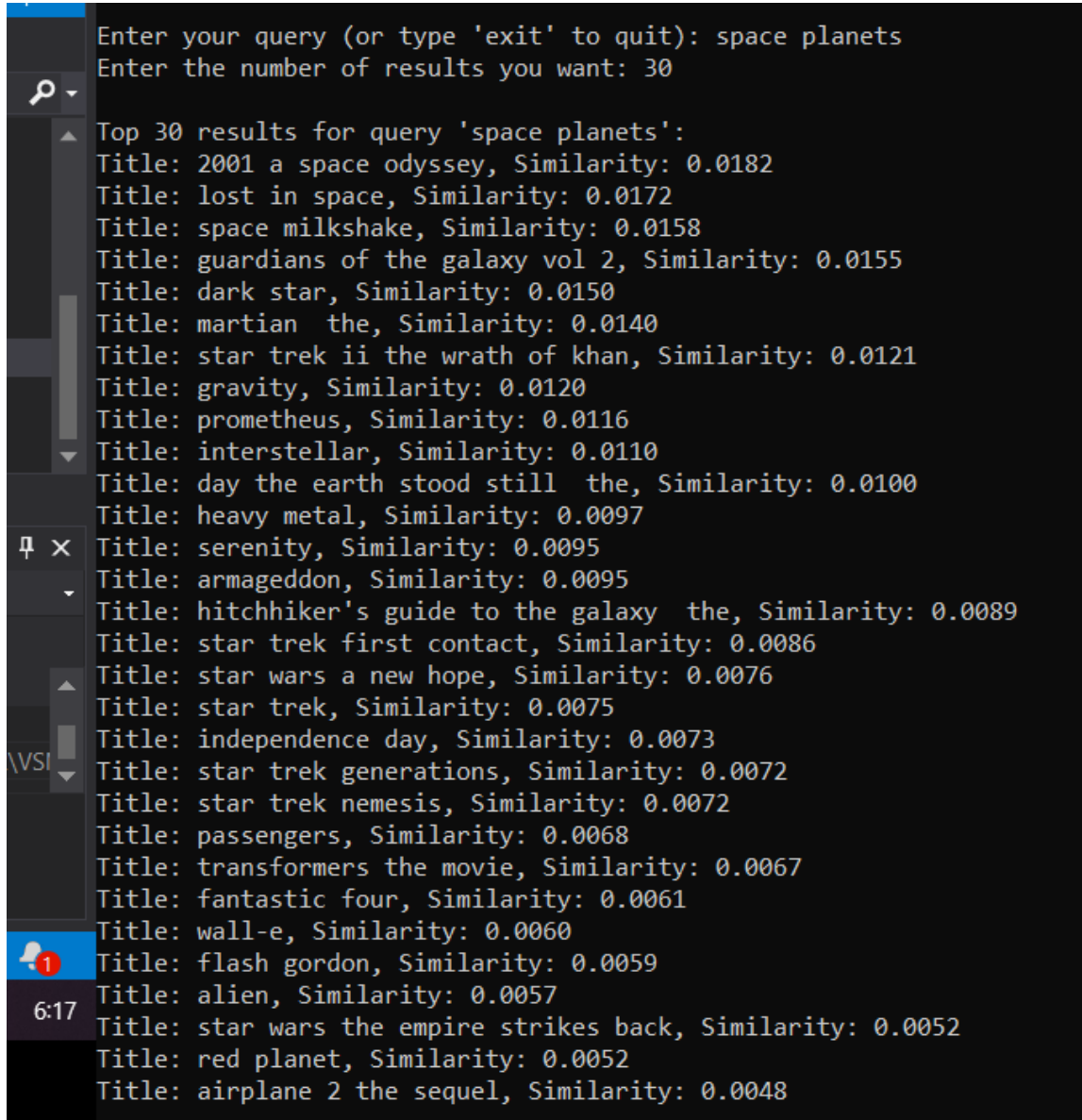
We also have that the peak precision was at value  $k = 15$  with 73.3% precision. This means that after 15 retrieved movies 12 were correctly retrieved. Goes down again and goes back to 66.7%. We are averaging a precision@ $k$  of **68.38%**.



As mentioned, before we can see in this graph how the number of correct documents retrieved from the corpus increases as more documents are retrieved, this means that our dataset is well balanced and that the model is properly working.

## 2 – SECOND EXECUTION

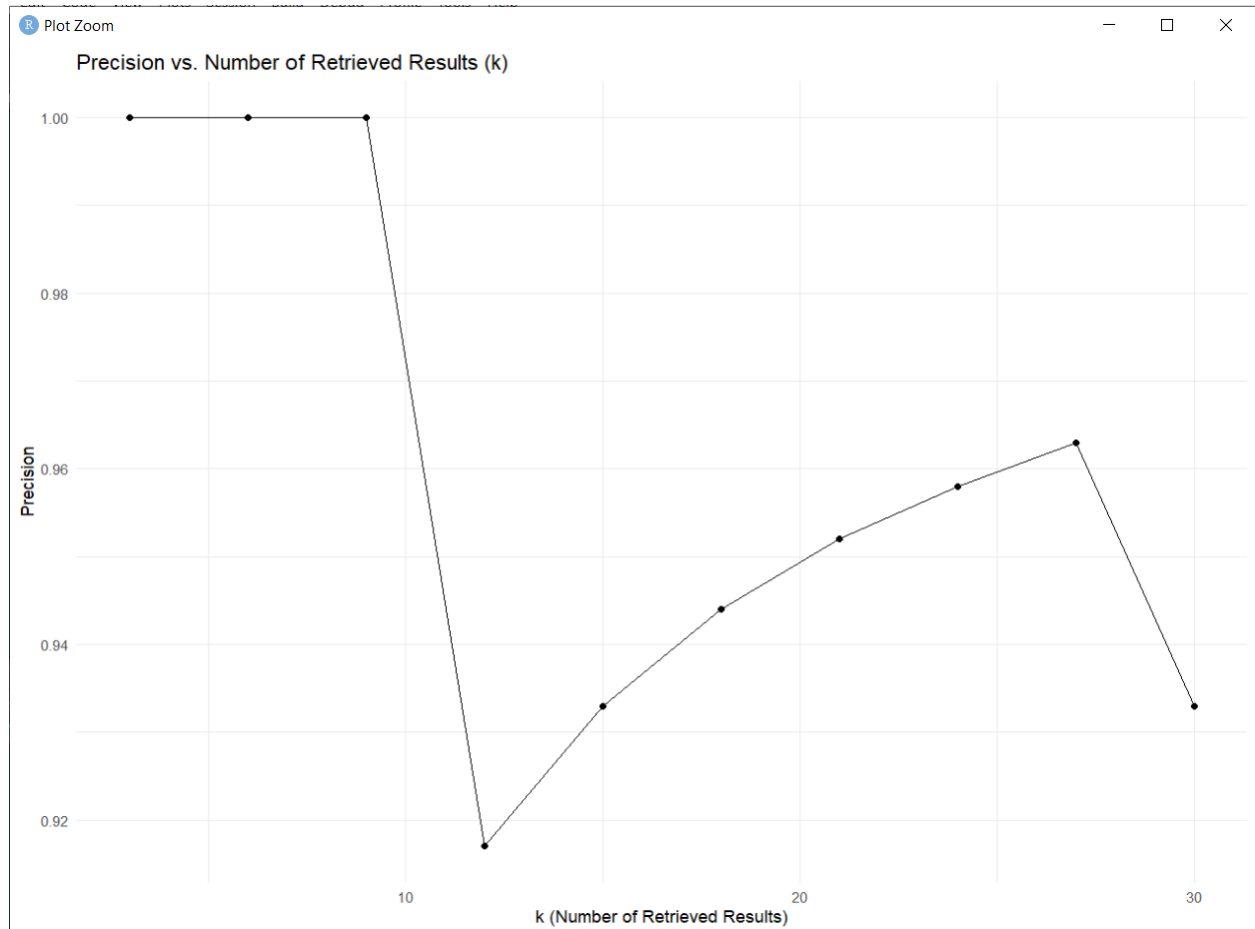
Here is the result of the second execution with the query “space planets”.

A screenshot of a terminal window with a dark background. The terminal shows a search interface. At the top, it prompts 'Enter your query (or type \'exit\' to quit): space planets' and 'Enter the number of results you want: 30'. Below this, it says 'Top 30 results for query \'space planets\':'. A list of 30 movie titles follows, each with its similarity score. The titles are: 2001 a space odyssey, lost in space, space milkshake, guardians of the galaxy vol 2, dark star, martian the, star trek ii the wrath of khan, gravity, prometheus, interstellar, day the earth stood still the, heavy metal, serenity, armageddon, hitchhiker's guide to the galaxy the, star trek first contact, star wars a new hope, star trek, independence day, star trek generations, star trek nemesis, passengers, transformers the movie, fantastic four, wall-e, flash gordon, alien, star wars the empire strikes back, red planet, and airplane 2 the sequel. The similarity scores range from 0.0182 down to 0.0048. On the left side of the terminal, there is a vertical sidebar with various icons and a search bar. At the bottom left, there is a notification icon with a red circle containing the number '1' and a timestamp '6:17'.

### Similarity scores

As seen before similarity scores are low but that doesn't mean that the retrieved result is bad in any way as we will see in the results.

### Assessing model performance for different k values



The maximum precision was achieved at  $k$  values from 3 to 9. After retrieving nine movies ( $k = 9$ ), I had a precision of 100%, meaning all movies were related to the query.

The lowest precision was at  $k = 12$  where one out of the 12 movies retrieved was not relevant to the query. This gives as that the lowest precision of our model for this query is 91.7%. As more documents are retrieved is normal that precision goes down since the most relevant documents to the query will be ranked first. After 30 retrieved movies, so  $k = 30$ , 28 retrieved movies were valid towards the query which gives a precision of 93.3% after 30 retrieved documents.



## E – ENHANCED TF-IDF VSM

We used techniques popularly used in NLP tasks to enhance the performance of the model, these techniques are:

- Lemmatization: reduces words to their base (canonical form) which is known as lemma, that's where the name comes from. The lemma represents the dictionary form or the root word form which all inflected forms of the word can be generated. This technique helps standardize words, reducing like this the complexity of the vocabulary and improving the consistency of textual representations. This works better for longer queries.
- N-grams: sequences of n words extracted from a text. With this rather than considering words individually we can get semantic meaning and context. For instance, we would get "climate change" as a single unit instead of "climate" and "change". I included bi-grams and tri-grams to capture dependencies on the text and preserve structural information and improve the model's ability to generalize.
- Removing Stop Words: Stop words are common words that appear frequently in a language but typically don't have a strong semantic meaning, so they give no relevant information. This are words such as "the", "is", "and", "of", etc.

## 1 – FIRST EXECUTION

Here is the result of the **first execution** since we enhanced TF-IDF VSM and with the query "love"

```
Enter your query (or type 'exit' to quit):
love

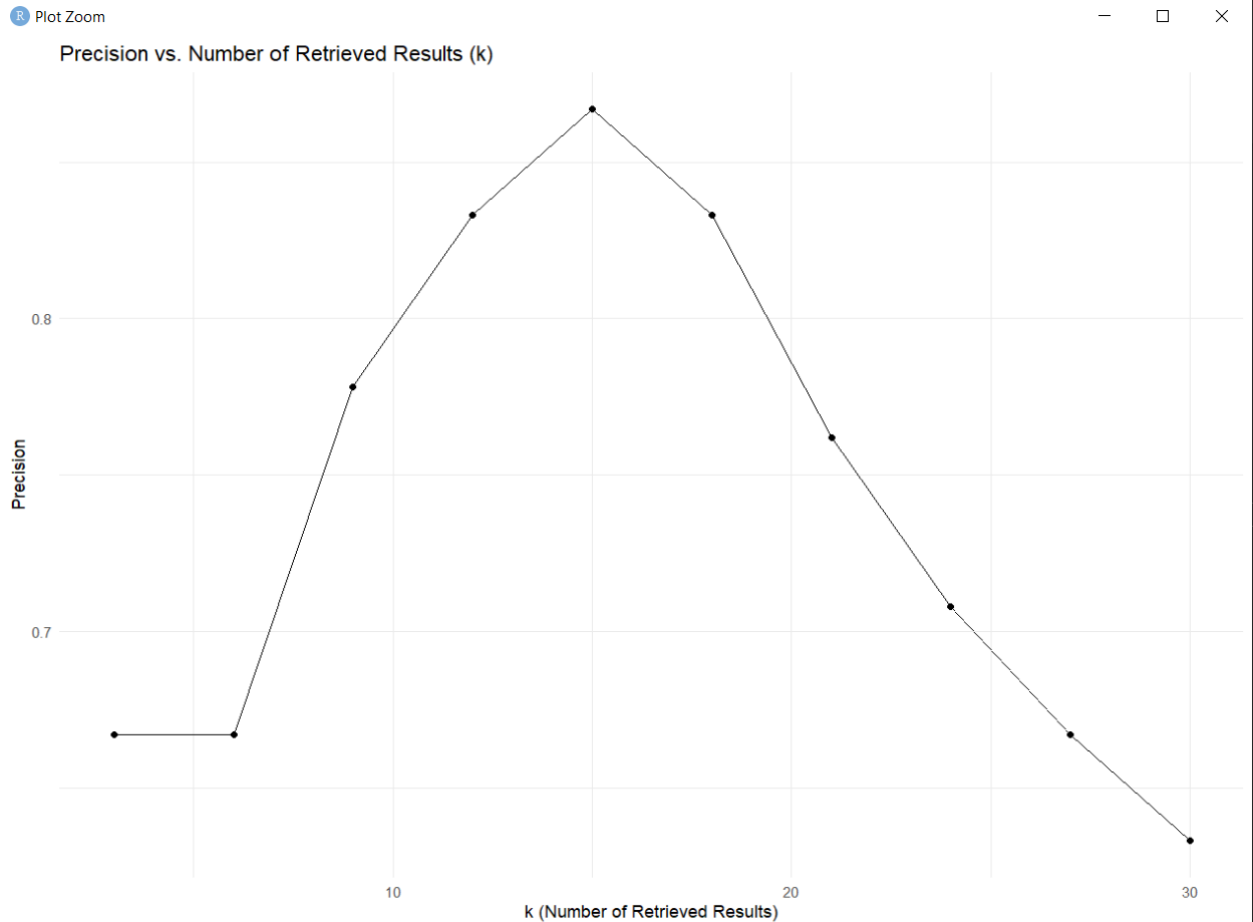
Enter the number of results you want:
30

Top 30 results for query 'love':
Title: i love you phillip morris, Similarity: 0.0291
Title: do the right thing, Similarity: 0.0275
Title: doors the, Similarity: 0.0275
Title: shakespeare in love, Similarity: 0.0256
Title: anniversary party the, Similarity: 0.0206
Title: midnight in paris, Similarity: 0.0201
Title: 500 days of summer, Similarity: 0.0187
Title: jerry maguire, Similarity: 0.0163
Title: almost famous, Similarity: 0.0162
Title: st. elmo's fire, Similarity: 0.0161
Title: little mermaid the, Similarity: 0.0160
Title: crazy stupid love, Similarity: 0.0143
Title: anna karenina, Similarity: 0.0141
Title: rambling rose, Similarity: 0.0141
Title: rocky horror picture show the, Similarity: 0.0139
Title: custody, Similarity: 0.0136
Title: village the, Similarity: 0.0136
Title: sex and the city, Similarity: 0.0135
Title: vanilla sky, Similarity: 0.0134
Title: her, Similarity: 0.0133
Title: last tango in paris, Similarity: 0.0131
Title: bachelor party, Similarity: 0.0131
Title: 44 inch chest, Similarity: 0.0130
Title: austin powers - the spy who shagged me, Similarity: 0.0128
Title: magnolia, Similarity: 0.0128
Title: hannah and her sisters, Similarity: 0.0127
Title: frozen (disney), Similarity: 0.0127
Title: artist the, Similarity: 0.0120
Title: austin powers - international man of mystery, Similarity: 0.0120
Title: shrek, Similarity: 0.0119

Enter your query (or type 'exit' to quit):
```

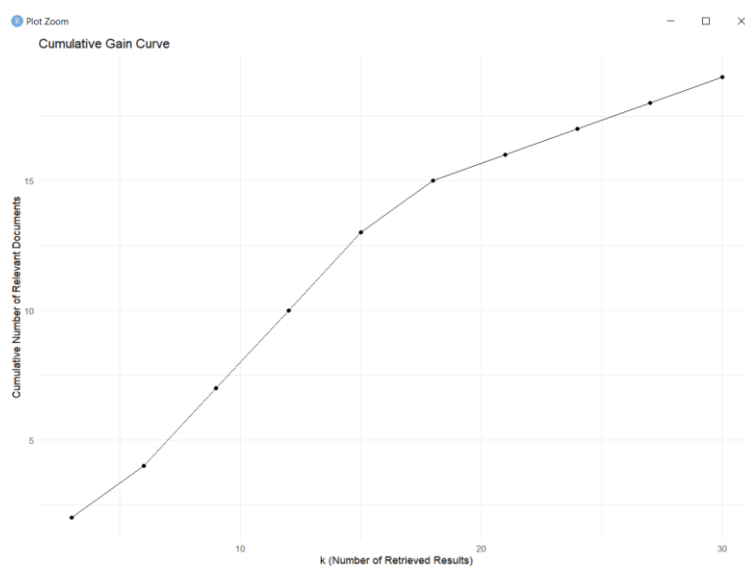


### Assessing model performance for different k values



The model performance was good, starting at a 66.7% precision until  $k = 6$ . After that it increased its precision reaching its peak at  $k = 15$  with 86.7% precision. This means that after retrieving 15 documents 13 of them were valid and related to the query.

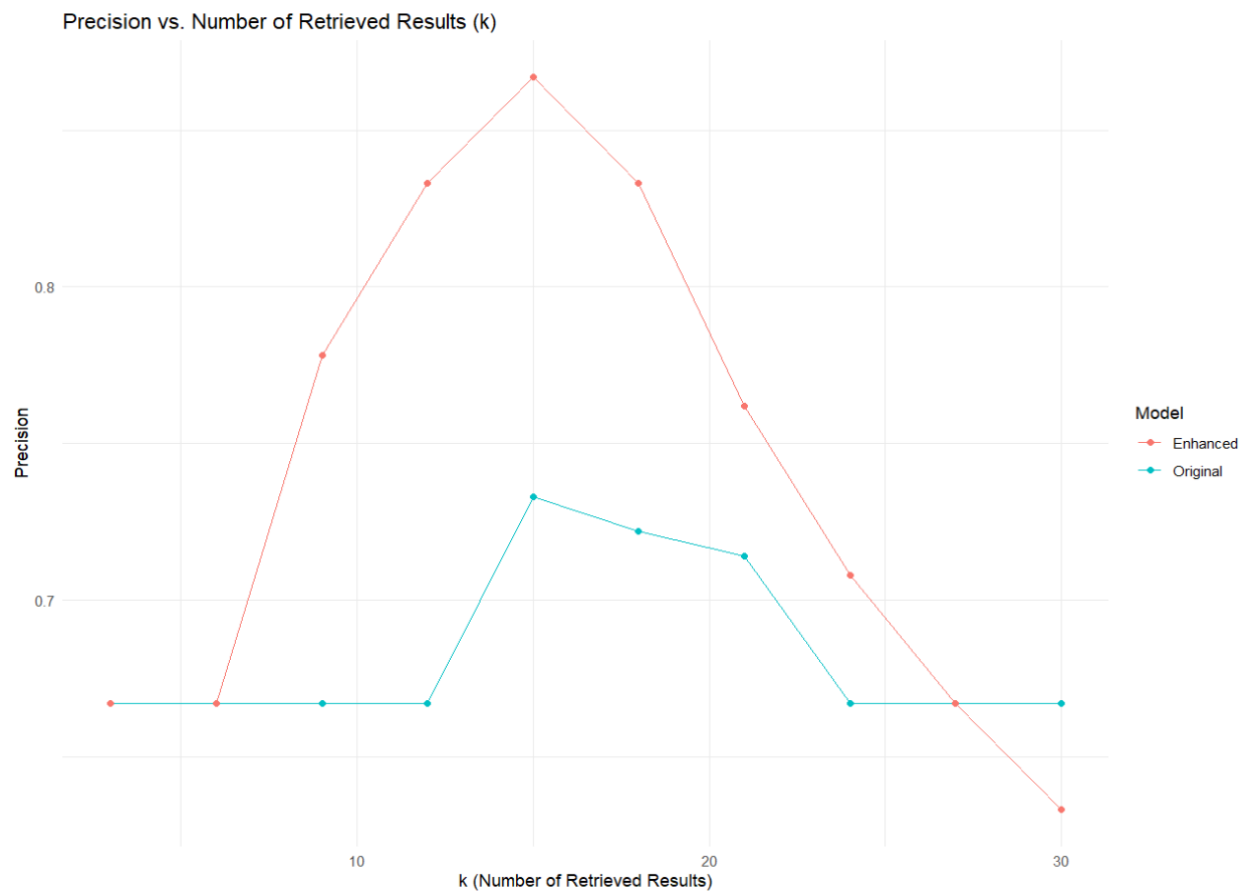
After that as expected the model's precision started decaying the more documents that were retrieved ending at a 63.3% precision after retrieving 30 documents.



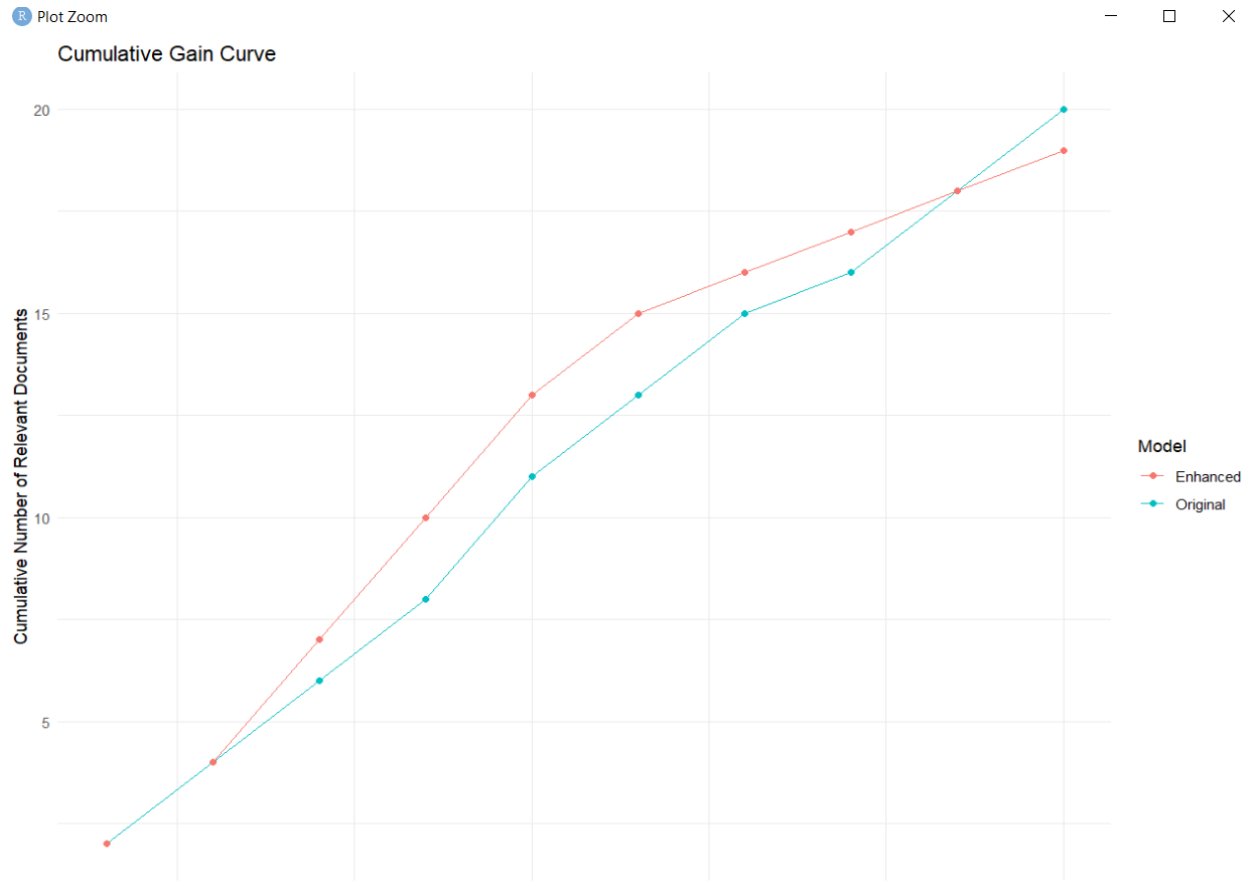
We also can see how the number of documents increases with the number of documents retrieved which is good. We also see it starts to converge which may be an indicator that the model would

not work perfectly with bigger  $k$  values, but this would not be a problem for the normal user which only works with low  $k$  values ranging from 5 to 10.

## F – TF-IDF VSM VS ENHANCED TF-IDF VSM MODEL



We can see how the model starts with the same precision until  $k = 2$  but the enhanced model has a better performance for bigger  $k$  values. Reaching a maximum precision of 86.7% which is 13.4% more than the peak precision for the original model.



We also can see how the cumulative plot looks good for both models, but again shows a better performance overall for the enhanced model.

We conclude then that the enhance model has an overall better performance than the original model. Having a higher peak and average performance. Also, we can see that both models peaked in performance and so, precision values, when 15 documents were retrieved so  $k = 15$ . With this we can assume that the best  $k$  value for our model and for this specific query would be of  $k = 15$ , but this varies within the queries. It is just a good way for finding the optimal  $k$  value.

## G – VSM WITH BoW (Bag of Words)

Bag of Words (BoW) converts text documents into numerical vectors by considering the frequency of occurrence of words within each document. It does it with this step:

- **Tokenization:** Text is divided into individual elements known as tokens. Non-alphanumeric characters are removed.
- **Vocabulary Construction:** a unique set of tokens is constructed, this set is present in the corpus of documents and each token is assigned to an index.
- **Vectorization:** For each document a vector of weights is created representing the presence or absence of words from the vocabulary. The length of the vector is equal to the length of the vocabulary, and each element of the vector corresponds to the frequency of the token in the document.
- **Sparse Representation:** Since most documents contain only a small subset of the words present in the entire vocabulary, the resulting vectors are typically sparse, meaning that most elements are zero.

### 1 – FIRST EXECUTION

Here is the first execution of the model using the query “love”.

```
Enter your query (or type 'exit' to quit): love
Enter the number of results you want: 30

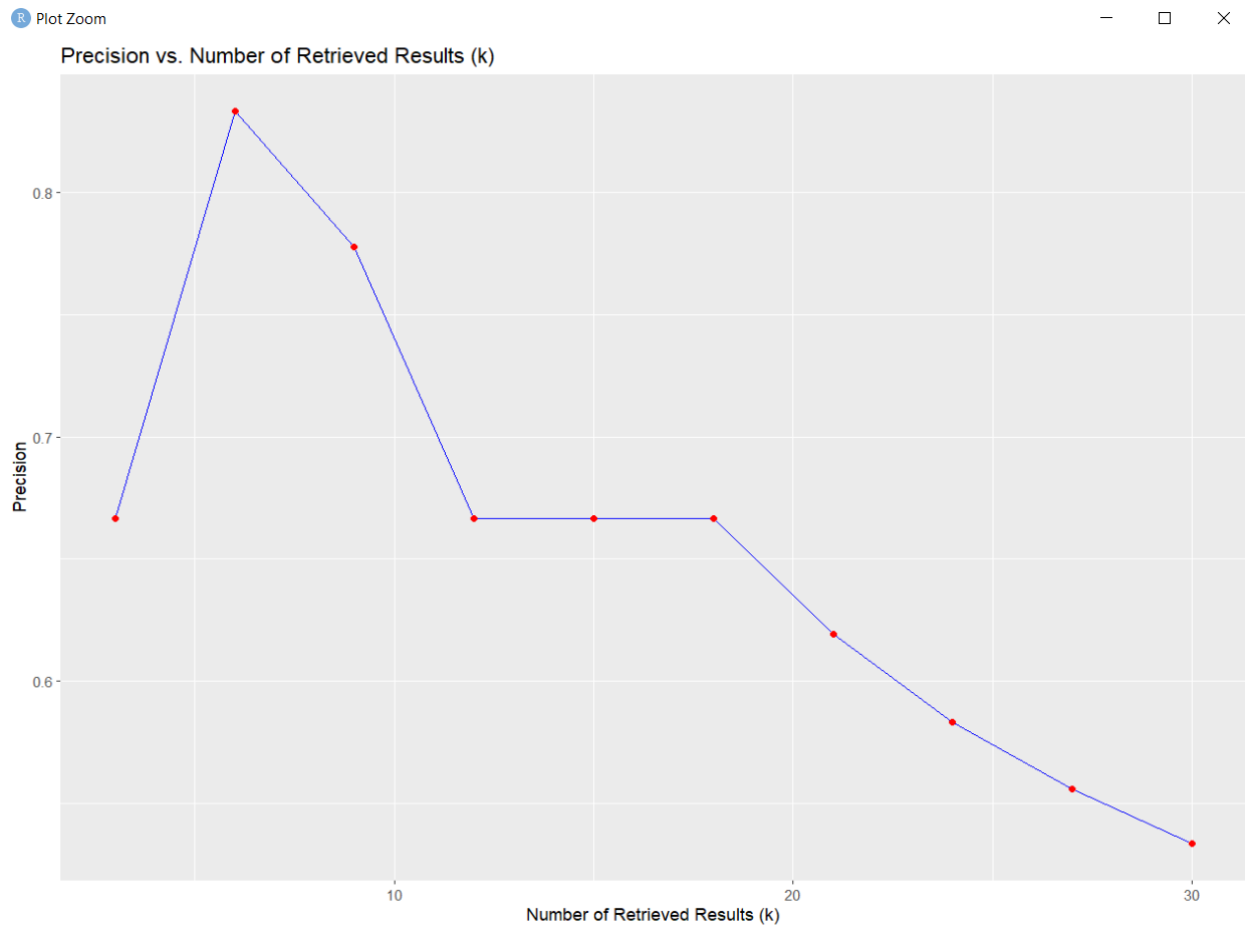
Top 30 results for query 'love':
Title: i love you phillip morris, Similarity: 0.0752
Title: do the right thing, Similarity: 0.0733
Title: midnight in paris, Similarity: 0.0474
Title: shakespeare in love, Similarity: 0.0390
Title: st. elmo's fire, Similarity: 0.0357
Title: anniversary party the, Similarity: 0.0338
Title: her, Similarity: 0.0333
Title: frozen (disney), Similarity: 0.0330
Title: shrek, Similarity: 0.0327
Title: sex and the city, Similarity: 0.0314
Title: austin powers - the spy who shagged me, Similarity: 0.0305
Title: little mermaid the, Similarity: 0.0302
Title: crazy stupid love, Similarity: 0.0299
Title: 500 days of summer, Similarity: 0.0295
Title: 44 inch chest, Similarity: 0.0287
Title: anna karenina, Similarity: 0.0286
Title: doors the, Similarity: 0.0284
Title: custody, Similarity: 0.0278
Title: tristan and isolde, Similarity: 0.0277
Title: last station the, Similarity: 0.0274
Title: cruel intentions, Similarity: 0.0269
Title: austin powers - international man of mystery, Similarity: 0.0263
Title: rocky horror picture show the, Similarity: 0.0259
Title: last tango in paris, Similarity: 0.0255
Title: jerry maguire, Similarity: 0.0243
Title: nine, Similarity: 0.0242
Title: tamara drewe, Similarity: 0.0240
Title: very bad things, Similarity: 0.0239
Title: meet joe black, Similarity: 0.0237
Title: chasing amy, Similarity: 0.0235

Enter your query (or type 'exit' to quit):
```

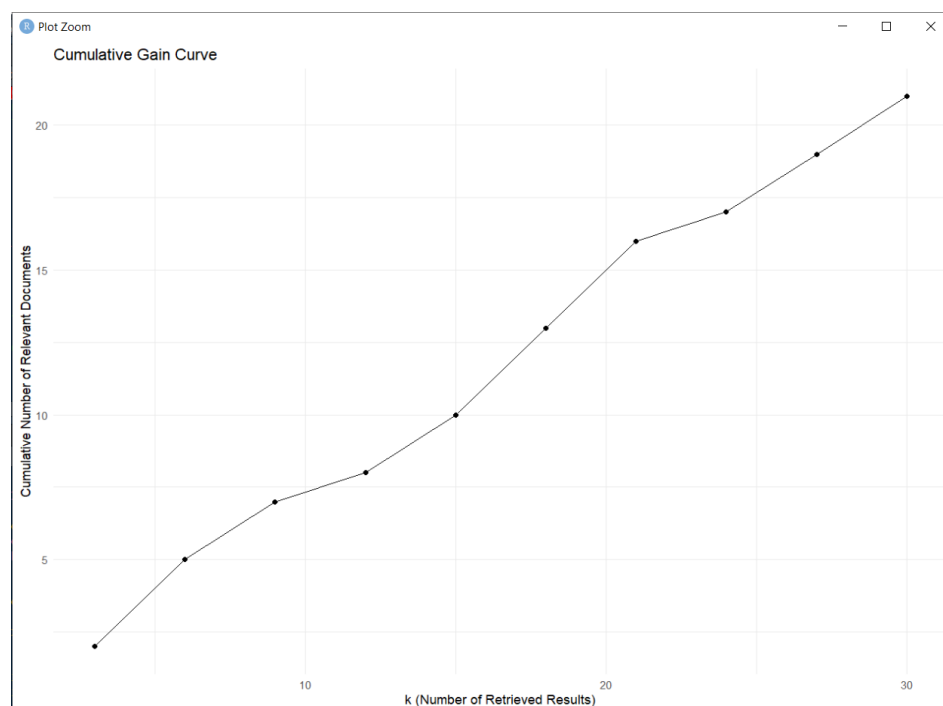
## Similarity

The higher the value the more similar the document is to the query. Like before, low similarity scores does not mean that the document isn't similar.

## Assessing model performance for different $k$ values



We obtained an average precision of 54.5% after retrieving 30 documents. The best precision I've got is at  $k = 6$  with a precision of 83.3%. After that precision starts going down gradually and in a constant manner.



We can see as  $k$  increases the number of retrieved documents increases too.

## 2 – SECOND EXECUTION

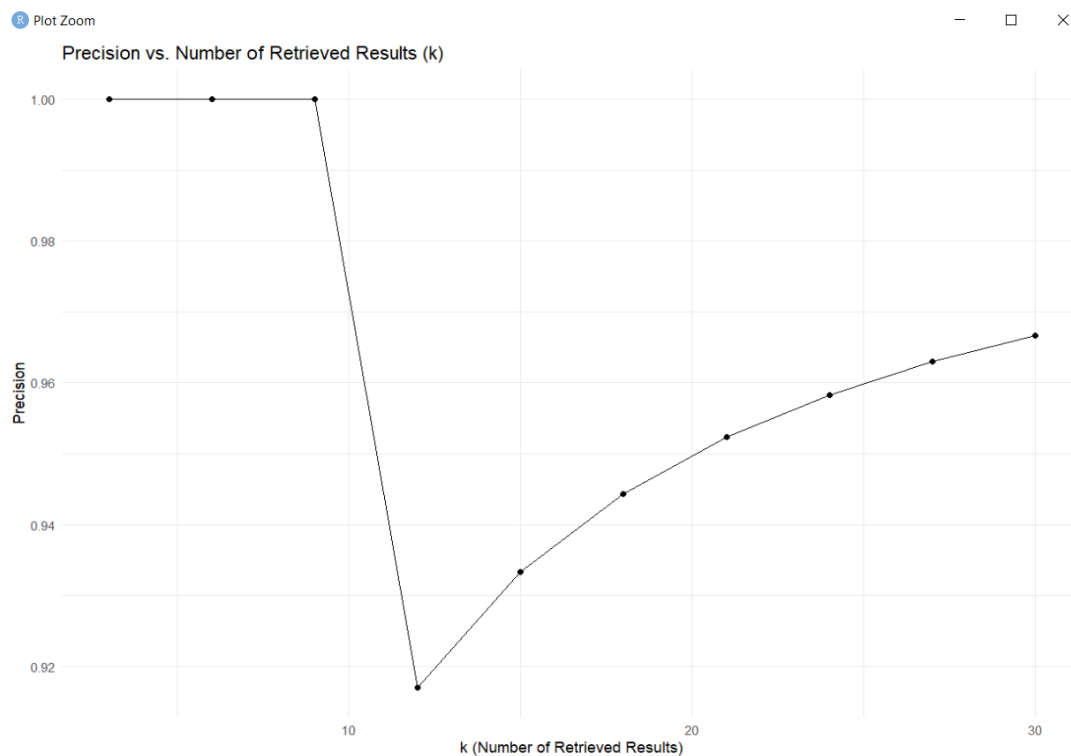
Here is the second execution with the query “space planets”.

```
Enter the number of results you want: 30

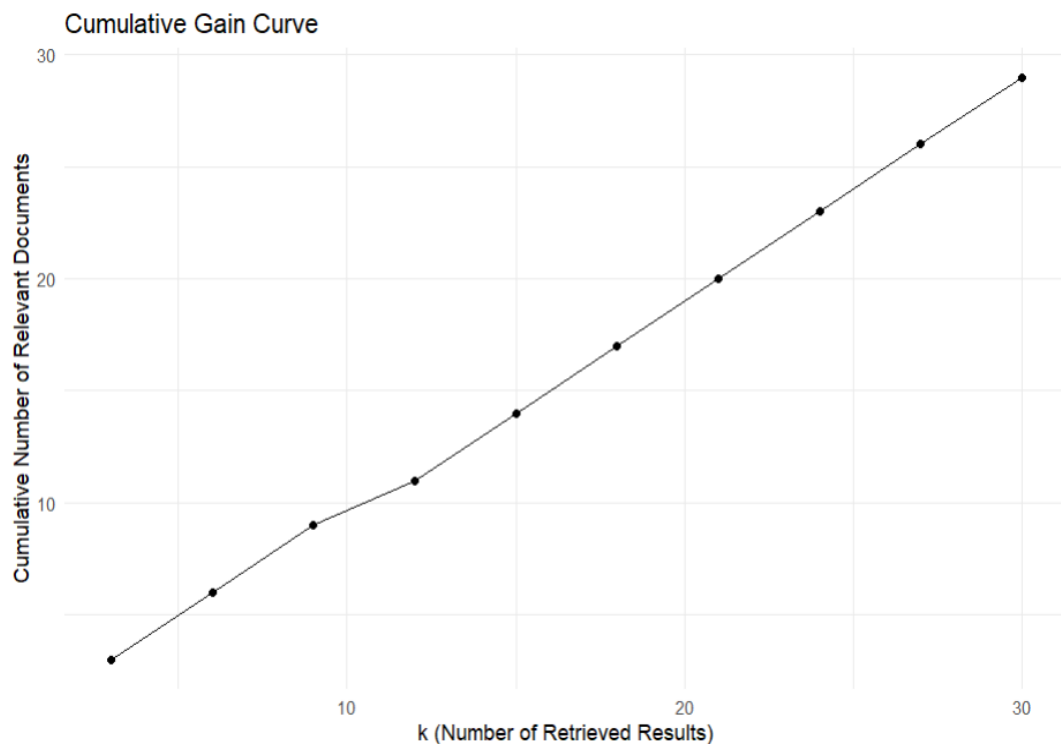
Top 30 results for query 'space planets':
Title: space milkshake, Similarity: 0.0759
Title: star trek ii the wrath of khan, Similarity: 0.0497
Title: guardians of the galaxy vol 2, Similarity: 0.0385
Title: martian the, Similarity: 0.0349
Title: lost in space, Similarity: 0.0307
Title: gravity, Similarity: 0.0303
Title: 2001 a space odyssey, Similarity: 0.0270
Title: dark star, Similarity: 0.0259
Title: armageddon, Similarity: 0.0222
Title: star trek first contact, Similarity: 0.0218
Title: star wars a new hope, Similarity: 0.0217
Title: transformers the movie, Similarity: 0.0216
Title: star trek, Similarity: 0.0208
Title: star trek nemesis, Similarity: 0.0195
Title: interstellar, Similarity: 0.0180
Title: day the earth stood still the, Similarity: 0.0174
Title: hitchhiker's guide to the galaxy the, Similarity: 0.0165
Title: jason x, Similarity: 0.0165
Title: independence day, Similarity: 0.0161
Title: star wars the empire strikes back, Similarity: 0.0160
Title: fantastic four, Similarity: 0.0160
Title: star trek generations, Similarity: 0.0158
Title: prometheus, Similarity: 0.0153
Title: alien, Similarity: 0.0148
Title: mission to mars, Similarity: 0.0145
Title: star trek the motion picture, Similarity: 0.0140
Title: airplane 2 the sequel, Similarity: 0.0139
Title: red planet, Similarity: 0.0134
Title: heavy metal, Similarity: 0.0131
Title: wall-e, Similarity: 0.0128

Enter your query (or type 'exit' to quit):
```

### Assessing model performance for different $k$ values



We maintain a 100% precision until  $k = 9$ , then we have a false-positive that drops the precision to a 91.7%. After that, the precision starts slowly going up and converging to 100%.



The cumulative plot is almost perfect, growing in a linear manner since we have only one non-valid document, where the number of valid documents retrieved increases at the same rate as the value of  $k$ .



## G – VSM WITH BoW ENHANCED (Bag of Words)

Same as in the TF-IDF enhanced VSM model we upgraded the model.

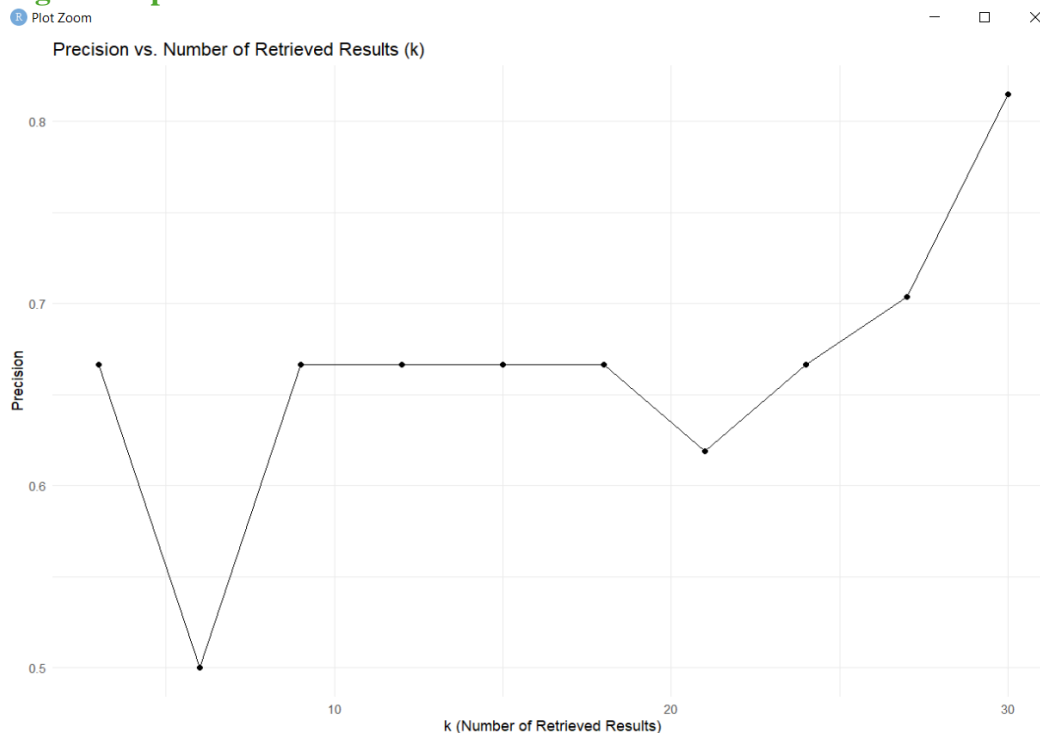
### 1 – FIRST EXECUTION

First execution of the VSM model with BoW enhanced and with the query “love”.

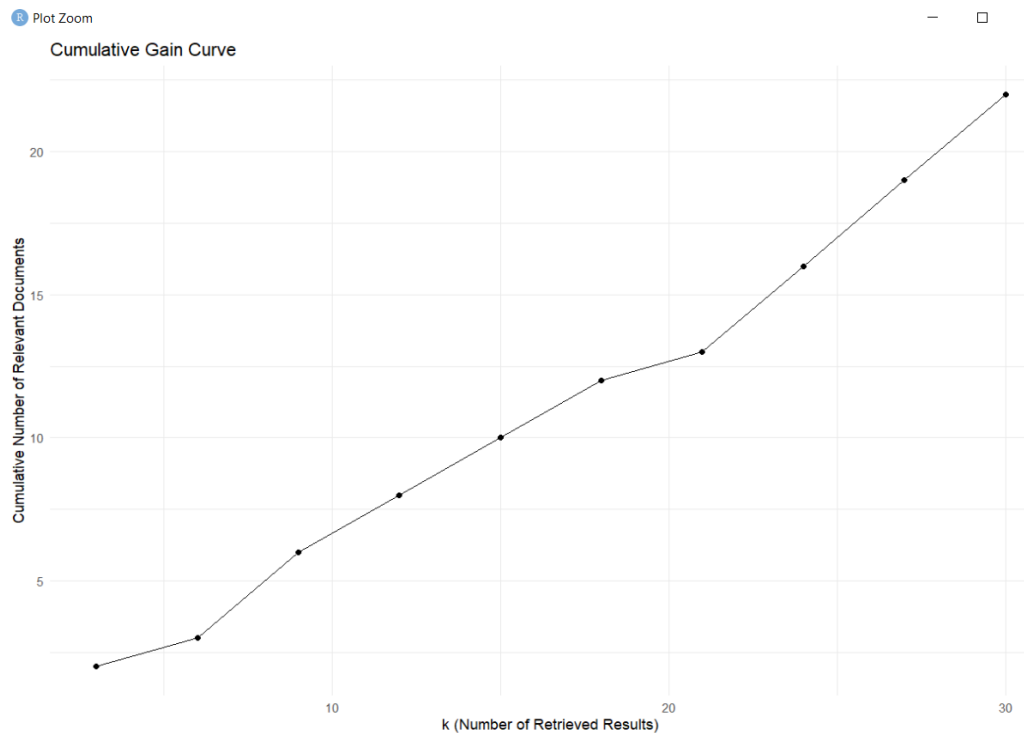
```
Enter your query (or type 'exit' to quit): love
Enter the number of results you want: 30

Top 30 results using BoW for query 'love':
Title: do the right thing, Similarity: 0.1274
Title: i love you phillip morris, Similarity: 0.1207
Title: shakespeare in love, Similarity: 0.1106
Title: midnight in paris, Similarity: 0.0959
Title: little mermaid the, Similarity: 0.0719
Title: doors the, Similarity: 0.0701
Title: st. elmo's fire, Similarity: 0.0693
Title: anniversary party the, Similarity: 0.0665
Title: anna karenina, Similarity: 0.0626
Title: shrek, Similarity: 0.0569
Title: her, Similarity: 0.0567
Title: last station the, Similarity: 0.0566
Title: sex and the city, Similarity: 0.0561
Title: frozen (disney), Similarity: 0.0552
Title: tristan and isolde, Similarity: 0.0550
Title: 44 inch chest, Similarity: 0.0542
Title: jerry maguire, Similarity: 0.0531
Title: last tango in paris, Similarity: 0.0523
Title: austin powers - the spy who shagged me, Similarity: 0.0517
Title: crazy stupid love, Similarity: 0.0513
Title: rocky horror picture show the, Similarity: 0.0513
Title: bachelor party, Similarity: 0.0501
Title: chasing amy, Similarity: 0.0501
Title: nine, Similarity: 0.0492
Title: barry lyndon, Similarity: 0.0489
Title: les misérables, Similarity: 0.0481
Title: my best friend's wedding, Similarity: 0.0478
Title: cruel intentions, Similarity: 0.0476
Title: custody, Similarity: 0.0476
Title: village the, Similarity: 0.0476
```

## Assessing model performance for different k values



We have the maximum precision at  $k = 30$  with a precision value of 81.5%. This means that there are still lots of movies related to the query “love” still to be retrieved and thus why the precision it is still going up at  $k = 30$ . The lowest precision was at  $k = 6$  with 50% and the average precision was 66.35%.



The cumulative plot is linear and thus the number of retrieved documents is proportional to the *value*. This means that there are still more movies related to the query to be retrieved but also means the model is doing a good job.

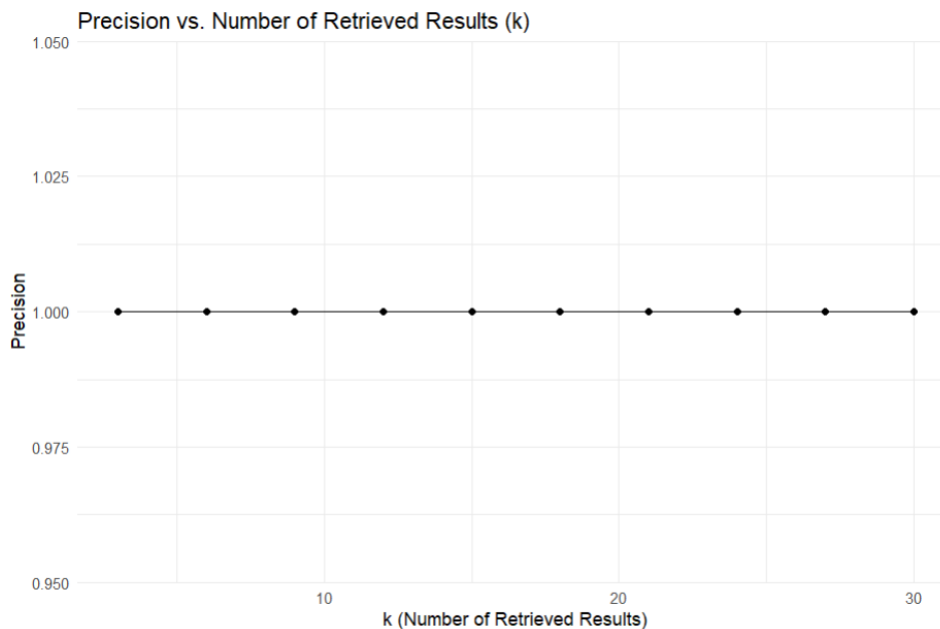
## 2 - SECOND EXECUTION

Second execution of the VSM model with BoW enhanced and with the query “Space planets”.

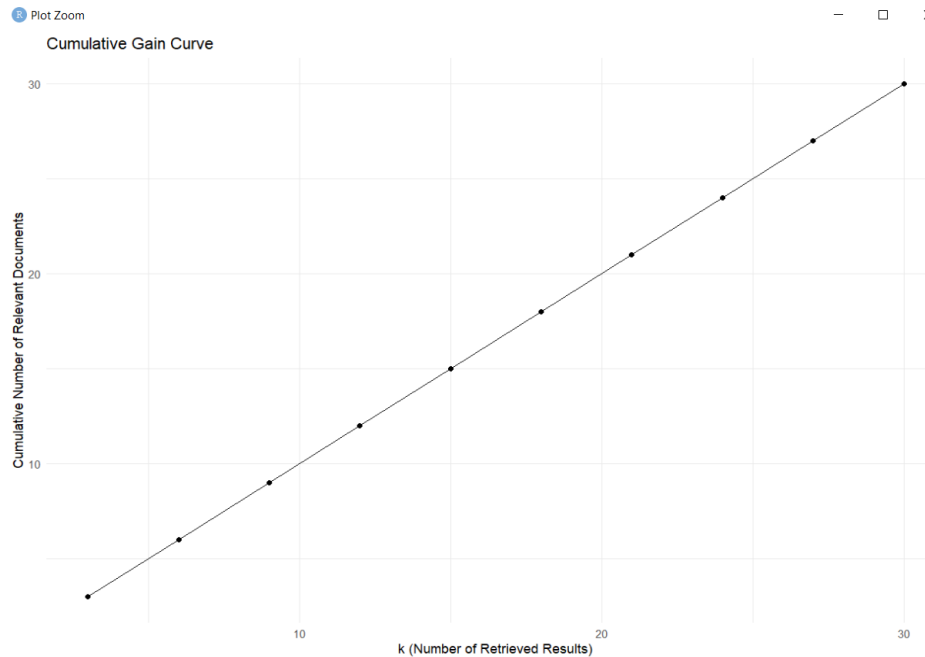
```
Enter your query (or type 'exit' to quit): space planets
Enter the number of results you want: 30

Top 30 results using BoW for query 'space planets':
Title: guardians of the galaxy vol 2, Similarity: 0.1118
Title: lost in space, Similarity: 0.0960
Title: space milkshake, Similarity: 0.0920
Title: star trek ii the wrath of khan, Similarity: 0.0845
Title: interstellar, Similarity: 0.0797
Title: martian the, Similarity: 0.0763
Title: 2001 a space odyssey, Similarity: 0.0730
Title: gravity, Similarity: 0.0699
Title: transformers the movie, Similarity: 0.0662
Title: star trek, Similarity: 0.0647
Title: dark star, Similarity: 0.0619
Title: star trek first contact, Similarity: 0.0576
Title: hitchhiker's guide to the galaxy the, Similarity: 0.0558
Title: red planet, Similarity: 0.0547
Title: star wars a new hope, Similarity: 0.0542
Title: star trek generations, Similarity: 0.0491
Title: day the earth stood still the, Similarity: 0.0483
Title: star trek nemesis, Similarity: 0.0459
Title: heavy metal, Similarity: 0.0453
Title: armageddon, Similarity: 0.0437
Title: independence day, Similarity: 0.0428
Title: alien, Similarity: 0.0425
Title: star wars the empire strikes back, Similarity: 0.0410
Title: prometheus, Similarity: 0.0402
Title: mission to mars, Similarity: 0.0378
Title: fifth element the, Similarity: 0.0358
Title: starship troopers, Similarity: 0.0335
Title: star wars the force awakens, Similarity: 0.0326
Title: star trek the motion picture, Similarity: 0.0316
Title: serenity, Similarity: 0.0315
```

### Assessing model performance for different k values



The model did a great performance achieving a precision score of 100% for all  $k$  values ranging from 0 to 100.

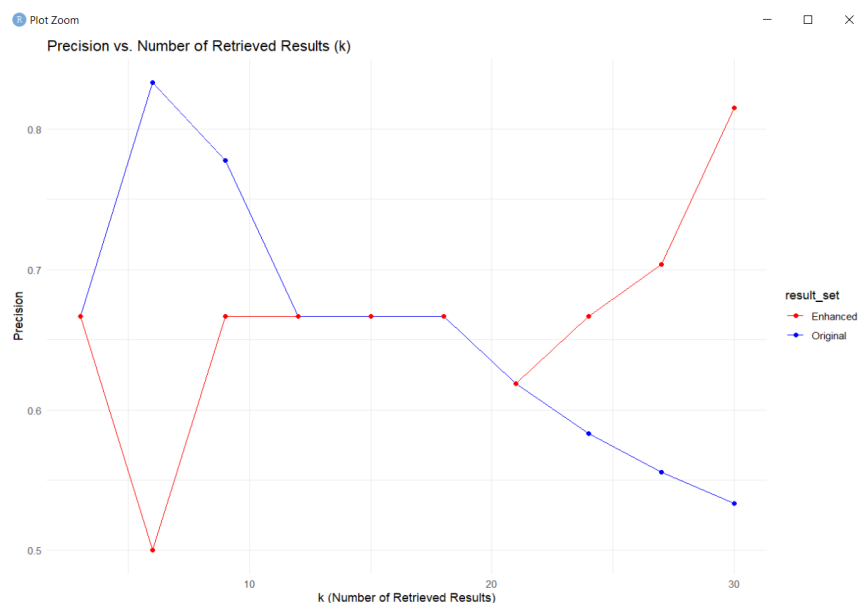


## H – BoW VSM WITH ENHANCED BoW VSM

The enhanced process is similar as TF-IDF VSM model but this time for VSM with BoW.

### 1 – FIRST EXECUTION

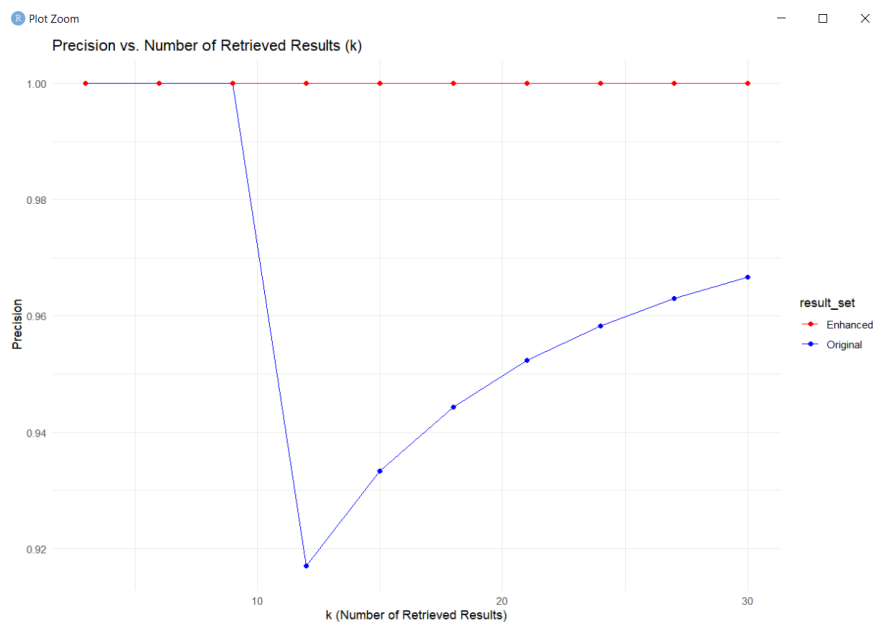
Here is the first execution of the enhance BoW VSM with the query of “love”.



We can see how the original model has a better start during the first  $k$  values. The enhanced model then gets more valid movies retrieved and has the same precision than the original and the larger the  $k$  we can see the enhanced model performance is way better than the original. Having at  $k = 30$  a difference of 28.17% in precision. Also, the movie that was a false-positive retrieved during the first  $k$  values by the enhanced model was also retrieved later on by the original model.

## 2 – SECOND EXECUTION

Here is the second execution of the enhance BoW VSM with the query of “space planets”.



In the original model the model retrieved only one false-positive after 30 documents, the enhanced model retrieved the 30 documents correctly so we could argue that the performance increased. To conclude, again, enhancing the model worked, and we got better results in all the enhanced models after retrieving 30 documents. From now on, all models tested are going to be already enhanced.

## I – VSM WITH WORD EMBEDDING

Word embedding is a technique used in Natural Language Processing to represent words as vectors in a vectorial space. These vectors can get both semantic and syntactic relationships between words, this allows machines to understand and process human language.

The process word embedding follows is as follows:

- **Tokenization:** explained earlier in the report
- **Model training:** A word embedding model is trained. This model learns to map each word to a dense vector representation based on the context of the text.
- **Vector representation:** During training, the model iteratively adjusts the word vectors to maximize their utility in predicting surrounding words in the text.
- **Evaluation:** The trained models are evaluated on tasks such as word similarity among other things related to NLP.

The process of coding:

- Dataset of movie scripts is loaded from a CSV file and handles any missing values from the “Script” column just in case there was any while scraping the data.
- Movie scripts are then tokenized and will serve as input data for training the Word2Vec model. The model then is trained with different parameters such as vector size, window size, and minimum word count. During training, the model learns to represent words as dense vectors in a continuous vector space, capturing semantic relationships between them.
- After training the Word2Vec model, a function `search_movies_word2vec` is defined to perform movie searches based on user queries. This function takes a query (a string of

words), the number of desired results, the trained Word2Vec model, and the movie dataset as inputs.

- Within the search\_movies\_word2vec function, the query is tokenized, and the average word embedding for the query is calculated using the Word2Vec model. Then, the cosine similarity between the query embedding and each movie script embedding is computed to measure their similarity.
- 
- The function returns the top movie titles along with their similarity scores, based on the cosine similarity calculation.

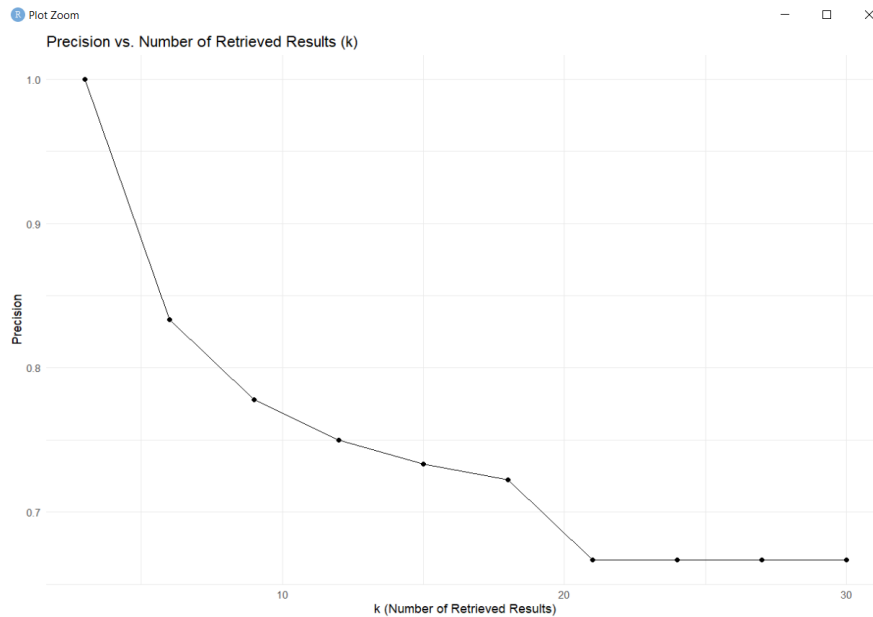
## 1 – FIRST EXECUTION

First execution of the VSM with word embedding and with the query “love”.

```
Enter your query (or type 'exit' to quit): love
Enter the number of results you want: 30

Top 30 results for query 'love':
Title: napoleon dynamite, Similarity: 0.4868
Title: clueless, Similarity: 0.4836
Title: big, Similarity: 0.4808
Title: good girl the, Similarity: 0.4751
Title: midnight in paris, Similarity: 0.4730
Title: finding nemo, Similarity: 0.4713
Title: edward scissorhands, Similarity: 0.4709
Title: knocked up, Similarity: 0.4666
Title: incredibles the, Similarity: 0.4658
Title: mary poppins, Similarity: 0.4639
Title: her, Similarity: 0.4638
Title: bridesmaids, Similarity: 0.4614
Title: boyhood, Similarity: 0.4610
Title: funny people, Similarity: 0.4603
Title: little mermaid the, Similarity: 0.4576
Title: margaret, Similarity: 0.4574
Title: big sick the, Similarity: 0.4555
Title: smashed, Similarity: 0.4552
Title: storytelling, Similarity: 0.4542
Title: ordinary people, Similarity: 0.4528
Title: glengarry glen gross, Similarity: 0.4512
Title: agnes of god, Similarity: 0.4464
Title: nightmare before christmas the, Similarity: 0.4460
Title: nightmare before christmas the, Similarity: 0.4460
Title: sex lies and videotape, Similarity: 0.4450
Title: anniversary party the, Similarity: 0.4442
Title: no strings attached, Similarity: 0.4415
Title: burning annie, Similarity: 0.4408
Title: silver linings playbook, Similarity: 0.4389
Title: happy birthday wanda june, Similarity: 0.4384
```

## Assessing model performance for different k values



We can see the model performance started on well with a 100% precision but slowly started decaying until converging to a precision of 66.67% at  $k = 30$ . This means that the scripts with less similarity with the query become increasingly challenging for the model to extract useful information.

Anyways the model still retrieves valid movies that are similar to the query in a good rate.

## 2 – SECOND EXECUTION

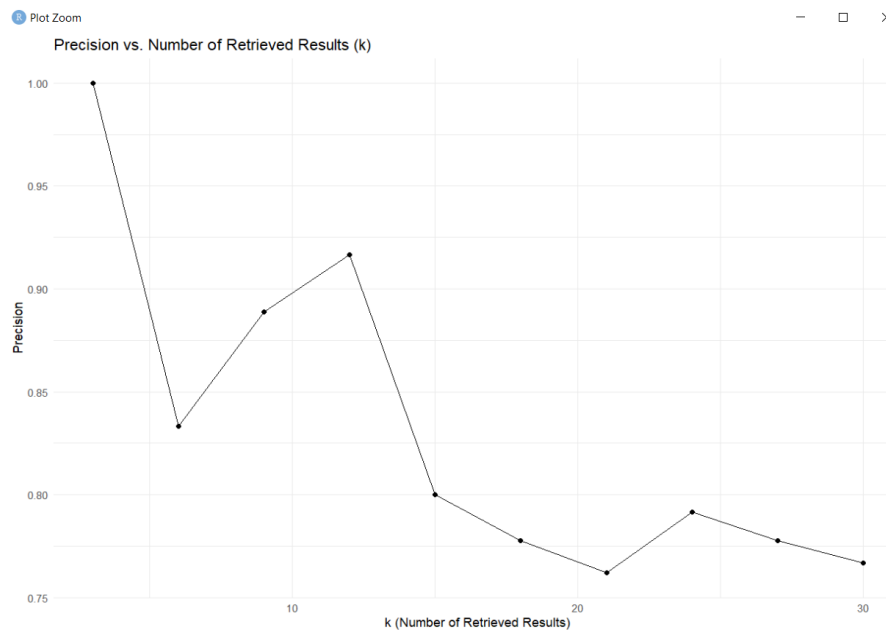
Second execution of the VSM with word embedding and with the query “space planets”.

```
Enter your query (or type 'exit' to quit): space planets
Enter the number of results you want: 30

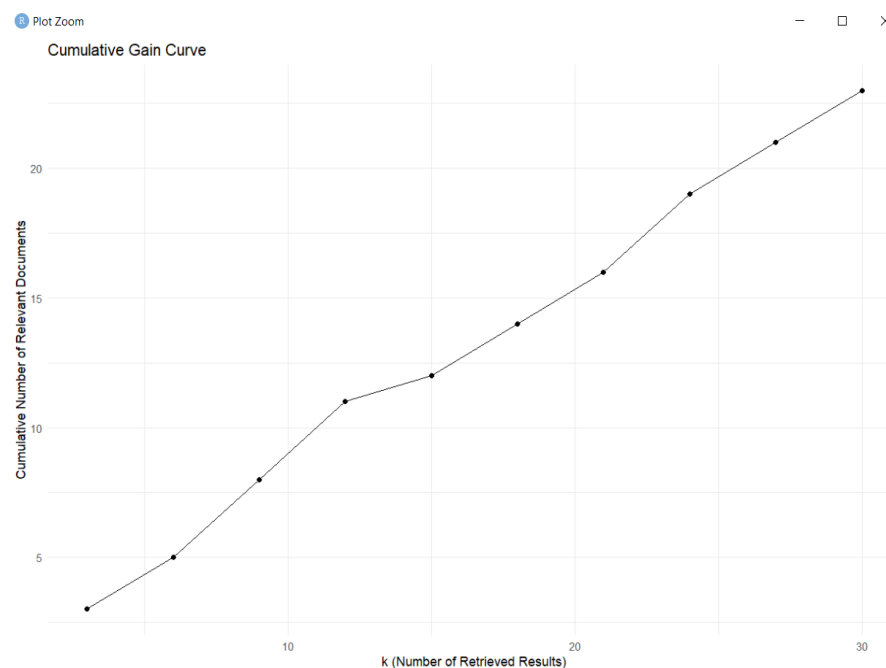
Top 30 results for query 'space planets':
Title: star trek the motion picture, Similarity: 0.3133
Title: dark star, Similarity: 0.3061
Title: star trek first contact, Similarity: 0.2956
Title: event horizon, Similarity: 0.2880
Title: tron legacy, Similarity: 0.2822
Title: alien, Similarity: 0.2786
Title: prometheus, Similarity: 0.2760
Title: wall-e, Similarity: 0.2663
Title: lost in space, Similarity: 0.2605
Title: star wars the empire strikes back, Similarity: 0.2571
Title: interstellar, Similarity: 0.2561
Title: star wars a new hope, Similarity: 0.2559
Title: alone in the dark, Similarity: 0.2538
Title: tron, Similarity: 0.2528
Title: space milkshake, Similarity: 0.2501
Title: godzilla, Similarity: 0.2472
Title: 2001 a space odyssey, Similarity: 0.2464
Title: star wars revenge of the sith, Similarity: 0.2434
Title: alien vs. predator, Similarity: 0.2426
Title: dune, Similarity: 0.2424
Title: air force one, Similarity: 0.2421
Title: star trek generations, Similarity: 0.2407
Title: star trek nemesis, Similarity: 0.2401
Title: aliens, Similarity: 0.2391
Title: gravity, Similarity: 0.2381
Title: quantum project, Similarity: 0.2379
Title: transformers the movie, Similarity: 0.2366
Title: armageddon, Similarity: 0.2364
Title: star trek ii the wrath of khan, Similarity: 0.2346
Title: rambo first blood ii the mission, Similarity: 0.2338
```



## Assessing model performance for different k values



As before the model starts well with a precision of 100% but slowly the precision converges until having a final precision of 70% after retrieving the 30 documents. The best  $k$  value would be 12, after retrieving 14 movies we have a precision of 91.6%.



Again, the capacity of retrieving good documents when there are less documents from the corpus from our model is consistent as we can see.

To conclude, VSM was good and consistent model, retrieving movies related to the queries even though they could be very vague.

For simpler queries that could relate to lots of the movies like “love” it is consistent having the best precision value at 66.67% after retrieving 30 documents with the VSM model combined with machine learning using word embedding.

For more concrete queries like “space planets” the model is very consistent having reached a precision of 100% after retrieving 30 movies with the VSM model using the enhanced BoW method.

Note: From now on all models will be already enhanced.

## II – BEST MATCHING 25

Used to estimate the relevance of documents, it belongs to the probabilistic branch of information retrieval models and improves models based on TF-IDF by adding components that help improve with its limitations.

BM25 models do in fact use TF-IDF but with two added parameters:  $k$  and  $b$ .

$K$ : controls term frequency saturation, meaning that the influence of term frequency increases logarithmically and then plateaus, preventing excessive emphasis on repetitive terms.

$b$ : normalize document length, addressing the problem where longer documents may naturally contain more query terms and thus appear artificially relevant.

With the help of these two parameters, BM25 can eliminate the bias by normalizing TF based on document length related to the average document, in my case scripts, of the dataset.

It uses the following formula:

$$score(q, d) = \sum_{i=1}^{|q|} idf(q_i) \cdot \frac{tf(q_i, d) \cdot (k_1 + 1)}{tf(q_i, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})}$$

Here,  $f(q_i, D)$  is the term frequency of query term  $q_i$  in document  $D$ ,  $|D|$  is the length of the document,  $avgdl$  is the average document length in the corpus, and  $IDF(q_i)$  is the inverse document frequency of  $q_i$ .

In contrast than with TF-IDF, BM25 takes into account document's length, were in TF-IDF this was an issue because larger documents could benefit from that and get higher similarity scores. Also includes better handling of TF saturation and of course the normalization towards document length.

The process of coding:

- In the code the library *pandas* is used for data manipulation allowing efficient loading and preprocessing of the movie dataset.
- Then *rank\_bm25* is the library that provides the BM25Okapi model, essential for creating the ranking of documents based on relevance towards the user's query.
- *Nltk* is used for text processing tasks like tokenization and stopwords removal.
- The script initializes logging using the logging library to track the flow of execution and capture any errors that may occur. The dataset is loaded and preprocessed to handle missing values, ensuring the 'Script' column is complete. Stopwords from the NLTK library are used to filter out common words that do not contribute to the relevance of a search query. The preprocessing function tokenizes the text, converts it to lowercase, and removes these stopwords.
- Each movie script in the dataset is then tokenized and processed into a format suitable for BM25. The BM25 model is initialized with the processed corpus of movie scripts. When a user inputs a search query, it undergoes the same preprocessing steps before BM25 computes the relevance scores.

## A – BEST MATCHING 25 RESULT

## 1 – FIRST EXECUTION

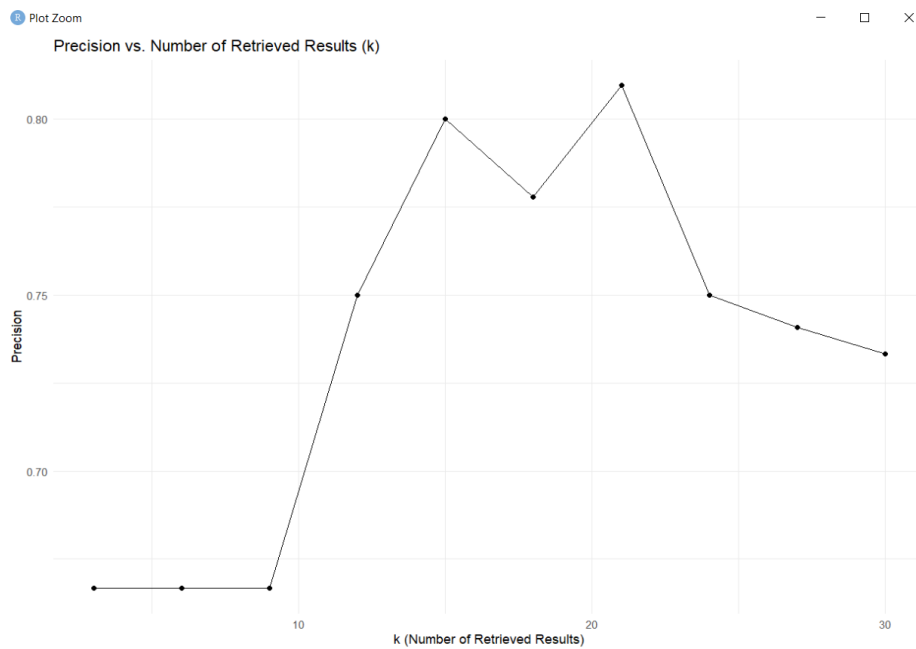
First execution of the VSM with word embedding and with the query “love”.

```
Enter your query (or type 'exit' to quit): love
Enter the number of top results you want: 30
```

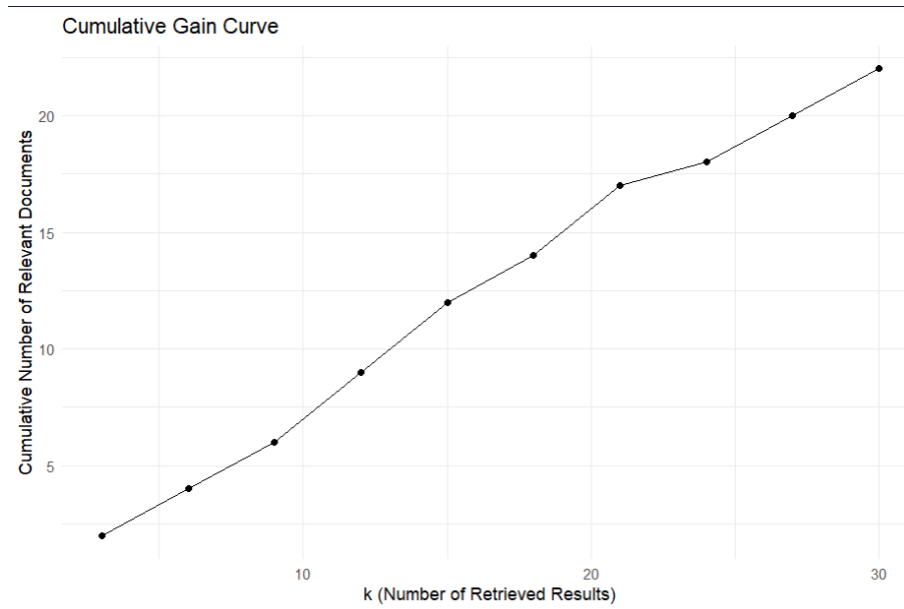
ID	Title	bm25_score
577 578	i love you phillip morris	3.501034
353 354	do the right thing	3.494854
981 982	shakespeare in love	3.478670
766 767	midnight in paris	3.469589
548 549	her	3.456836
448 449	frozen (disney)	3.456620
977 978	sex and the city	3.453412
358 359	doors the	3.452436
88 89	anniversary party the	3.452051
992 993	shrek	3.451919
1033 1034	st. elmo's fire	3.450954
86 87	anna karenina	3.448550
20 21	500 days of summer	3.447310
673 674	last tango in paris	3.442950
755 756	meet joe black	3.440434
308 309	cruel intentions	3.439242
112 113	austin powers - the spy who shagged me	3.438338
672 673	last station the	3.438174
298 299	crazy stupid love	3.437610
312 313	custody	3.436283
257 258	chasing amy	3.433147
939 940	rocky horror picture show the	3.427848
627 628	jerry maguire	3.427483
822 823	nine	3.425742
173 174	birdman	3.421511
1160 1161	very bad things	3.420302
137 138	barry lyndon	3.419740
412 413	fault in our stars the	3.418848
1083 1084	tamara drewe	3.417838
111 112	austin powers - international man of mystery	3.417398

```
Enter your query (or type 'exit' to quit):
```

## Assessing model performance for different k values



Again, due to “love” being a very general query this makes our model retrieve more false positives. We start with a  $\text{precision}@k$  of 66.67% until  $k = 9$ . Then we reach our peak precision when  $k = 21$ , this means that after retrieving 21 movies we have a precision of 80.95%. Then the precision starts to converge and after retrieving the test  $k$  we chose of 30 movies, we end up with a precision of 73.3%.



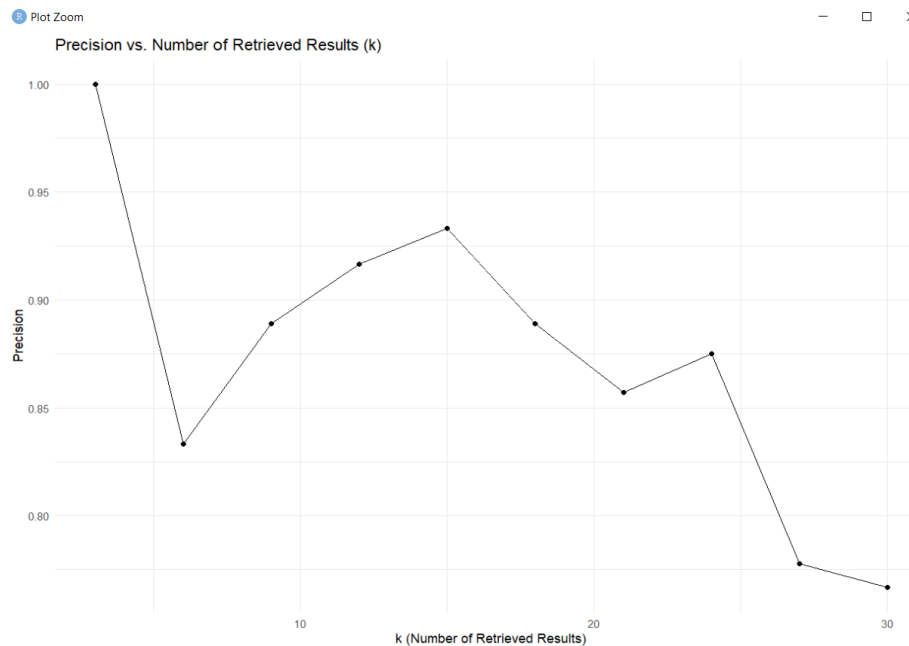
Like with VSM we have a good ratio of  $k$  vs valid documents retrieved thanks to our balanced dataset.

## 2 – SECOND EXECUTION

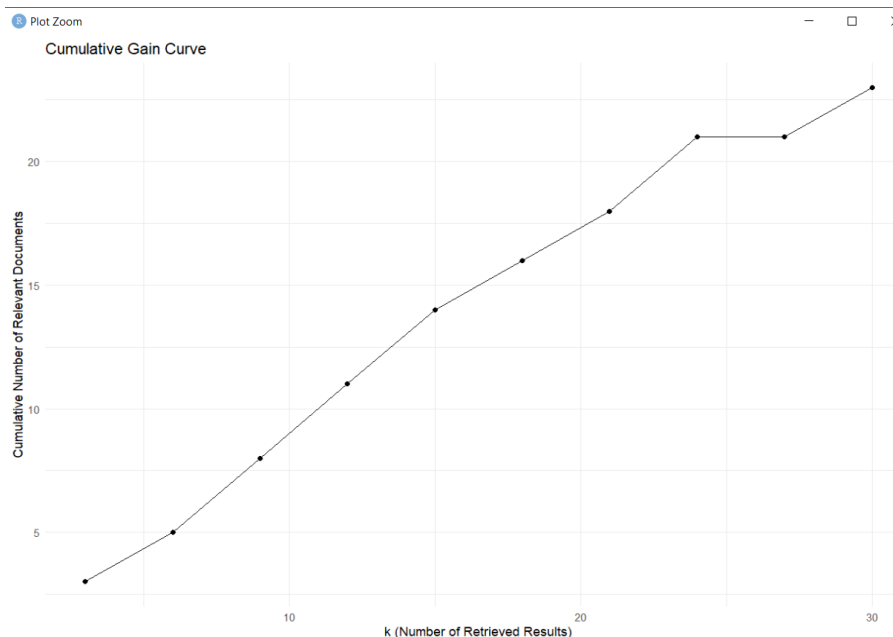
First execution of the VSM with word embedding and with the query “space planets”.

```
Enter your query (or type 'exit' to quit): space planets
Enter the number of top results you want: 30
  ID      Title      bm25_score
320  321      dark star  9.616460
896  897      prometheus  9.276743
509  510  guardians of the galaxy vol 2  9.238058
973  974      serenity  9.022015
10   11      2001 a space odyssey  9.016494
327  328  day the earth stood still the  8.899992
600  601      interstellar  8.833734
557  558  hitchhiker's guide to the galaxy the  8.633880
854  855      passengers  8.582620
417  418      fifth element the  8.280745
1038 1039      star trek nemesis  8.163243
425  426      flash gordon  8.088577
716  717      lost in space  8.021458
1037 1038      star trek generations  7.995933
857  858      paul  7.916618
535  536      heavy metal  7.788883
1032 1033      spider-man  7.570787
942  943      room  7.243106
1036 1037      star trek first contact  7.218481
1045 1046      star wars the force awakens  7.068087
759  760      men in black  6.926826
875  876      pitch black  6.861189
1034 1035      star trek  6.859521
380  381      elizabeth the golden age  6.793114
11   12      2012  6.653021
289  290      cooler the  6.518383
1165 1166      walk to remember a  6.513741
500  501      gravity  6.479095
239  240      carrie  6.388961
48   49      alien  6.281673
Enter your query (or type 'exit' to quit):
```

## Assessing model performance for different $k$ values



The model starts well with 100% precision during the three first movies retrieved, after that it slowly starts to converge having the best precision with a reasonable  $k$  value being at 93.3%. After retrieving the total of 30 movies I end up with a precision of 70%.



In the cumulative valid movies graph we can see it also has a linear fashion and we can see how at the end it starts to converge a little.

To conclude, the model works well and it's consistent with retrieving valid movies according to the query. Like before, the more precise the queries are the better the model works. The model also converges slowly to 0 accordingly, this would be easier to see testing bigger  $k$  values. That the model slowly converges to 0 is good, this means we have a balanced dataset and that the model is doing a good job, the more valid movies I retrieve, the less there are left in the dataset.

### III – BERT MODEL

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art language model developed by Google.

BERT is a neuronal network (NN) based on the Transformer architecture, which uses self-attention mechanisms to process text. What differentiates BERT from other models is its ability to read texts bidirectionally, BERT basically considers the context from the left and right sides of a word. This allows it to better understand the meaning and relevance of each word.

BERT is a pre-trained model, meaning that we can use it straight away, but we can also fine-tune it training it without own corpus. BERT pre-trained model is pre-trained with:

**Masked Language Modeling (MLM):** Randomly masks some tokens in the input and trains the model to predict them. This helps BERT understand the context of a word based on its surroundings.

**Next Sentence Prediction (NSP):** Trains the model to predict if one sentence follows another.

BERT is pre-trained on a large corpus of text, including the entire Wikipedia and BookCorpus datasets.

#### A – STANDARD BERT MODEL

##### 1 – FIRST EXECUTION

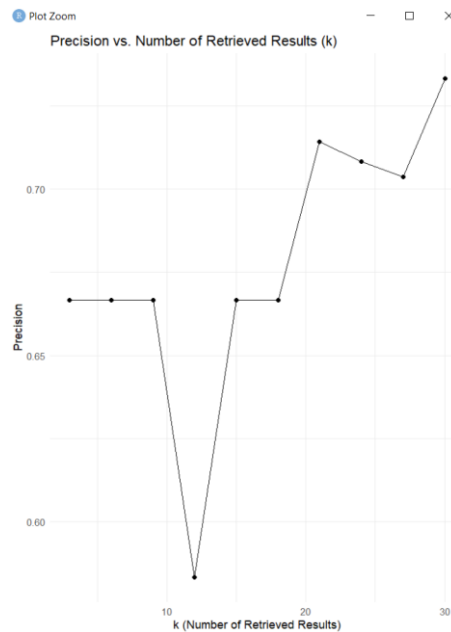
We are first going to try the pre-trained version of BERT, without any fine tuning and with the query “love”.

```
Enter your query (or type 'exit' to quit): love
Enter the number of results you want to view: 30

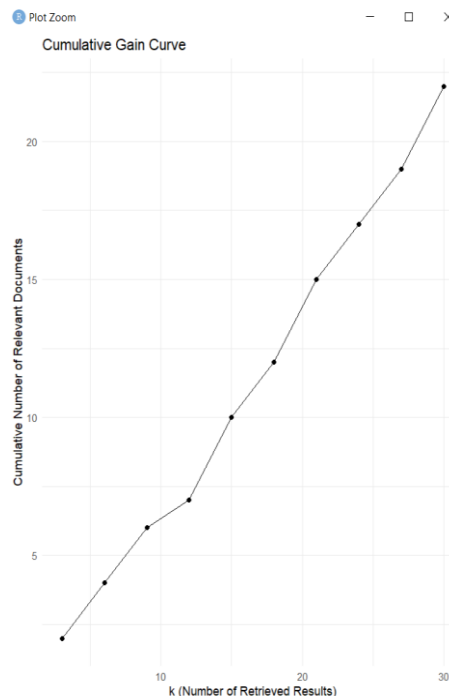
Top Movies for Query 'love':

ID: 839, Title: officer and a gentleman an, Similarity: 0.8444
ID: 1015, Title: sneakers, Similarity: 0.8404
ID: 22, Title: 8 mile, Similarity: 0.8390
ID: 654, Title: kiss of the spider woman, Similarity: 0.8385
ID: 387, Title: equilibrium, Similarity: 0.8382
ID: 125, Title: back to the future, Similarity: 0.8382
ID: 791, Title: mr. holland's opus, Similarity: 0.8376
ID: 981, Title: shadow of the vampire, Similarity: 0.8353
ID: 284, Title: commando, Similarity: 0.8353
ID: 1160, Title: vertigo, Similarity: 0.8299
ID: 285, Title: conan the barbarian, Similarity: 0.8262
ID: 1030, Title: speed, Similarity: 0.8251
ID: 95, Title: apollo 13, Similarity: 0.8225
ID: 294, Title: courage under fire, Similarity: 0.8217
ID: 1100, Title: this is 40, Similarity: 0.8215
ID: 936, Title: robocop, Similarity: 0.8195
ID: 1066, Title: superfights, Similarity: 0.8188
ID: 1184, Title: when harry met sally, Similarity: 0.8169
ID: 400, Title: existenz, Similarity: 0.8167
ID: 809, Title: neverending story the, Similarity: 0.8162
ID: 436, Title: frankenstein, Similarity: 0.8160
ID: 274, Title: clockwork orange a, Similarity: 0.8123
ID: 783, Title: monster's ball, Similarity: 0.8023
ID: 437, Title: frankenweenie, Similarity: 0.8008
ID: 962, Title: scary movie 2, Similarity: 0.7999
ID: 195, Title: blues brothers the, Similarity: 0.7983
ID: 697, Title: lion king the, Similarity: 0.7957
ID: 126, Title: back to the future ii & iii, Similarity: 0.7891
ID: 643, Title: kate & leopold, Similarity: 0.7891
ID: 86, Title: angels & demons, Similarity: 0.7891
```

### Assessing model performance for different k values



Starting off decently with a precision of 73.3% until  $k = 15$ , then after having some false-positives, precision starts going up again. We have maximum precision at  $k = 30$ , this is not perfect because precision is still going up and means it is not converging, in other words, because the model had quite some false positives there are still lots of movies related to the query waiting for being retrieved. Still, we have a good precision at  $k = 30$  with 77.8% precision.



As mentioned, before we have a decent but not perfect cumulative graph, it is close to the perfect 45-degrees angle or  $x=y$  linear graph but not quite there.



## 2 – SECOND EXECUTION

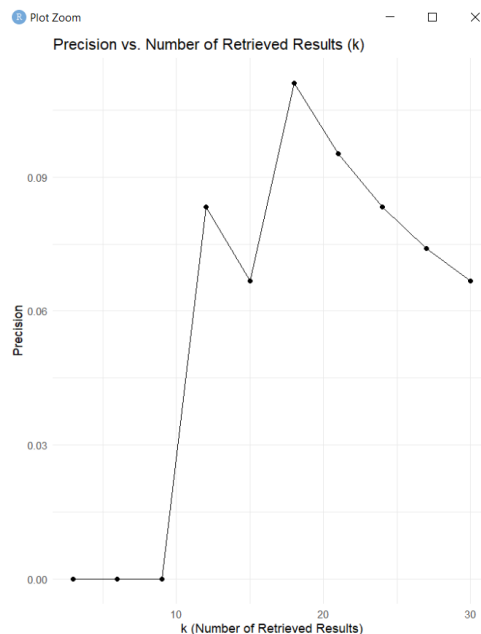
Second execution of BERT without any fine tuning and with the query “space planets”.

```
Enter your query (or type 'exit' to quit): space planets
Enter the number of results you want to view: 30

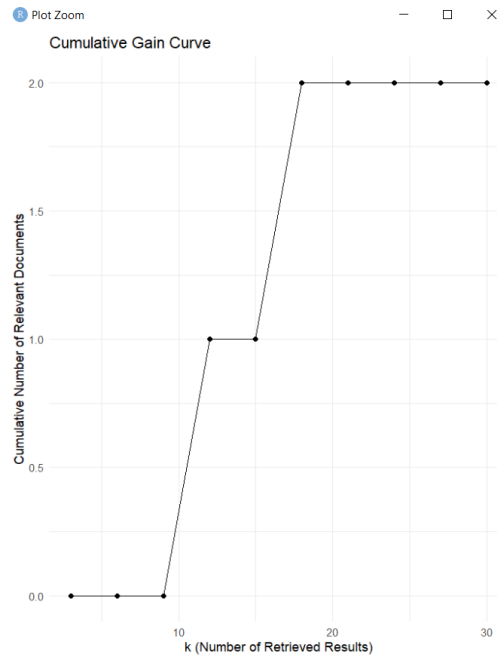
Top Movies for Query 'space planets':

ID: 1030, Title: speed, Similarity: 0.8143
ID: 1015, Title: sneakers, Similarity: 0.8105
ID: 284, Title: commando, Similarity: 0.8097
ID: 22, Title: 8 mile, Similarity: 0.8078
ID: 791, Title: mr. holland's opus, Similarity: 0.8063
ID: 1160, Title: vertigo, Similarity: 0.8060
ID: 1066, Title: superfights, Similarity: 0.8060
ID: 654, Title: kiss of the spider woman, Similarity: 0.8055
ID: 839, Title: officer and a gentleman an, Similarity: 0.8054
ID: 95, Title: apollo 13, Similarity: 0.8040
ID: 387, Title: equilibrium, Similarity: 0.8036
ID: 936, Title: robocop, Similarity: 0.8033
ID: 294, Title: courage under fire, Similarity: 0.8032
ID: 400, Title: existenz, Similarity: 0.7969
ID: 1184, Title: when harry met sally, Similarity: 0.7966
ID: 125, Title: back to the future, Similarity: 0.7942
ID: 285, Title: conan the barbarian, Similarity: 0.7930
ID: 437, Title: frankenweenie, Similarity: 0.7924
ID: 195, Title: blues brothers the, Similarity: 0.7902
ID: 274, Title: clockwork orange a, Similarity: 0.7894
ID: 436, Title: frankenstein, Similarity: 0.7891
ID: 697, Title: lion king the, Similarity: 0.7881
ID: 1100, Title: this is 40, Similarity: 0.7869
ID: 783, Title: monster's ball, Similarity: 0.7866
ID: 981, Title: shadow of the vampire, Similarity: 0.7842
ID: 962, Title: scary movie 2, Similarity: 0.7832
ID: 620, Title: jade, Similarity: 0.7797
ID: 86, Title: angels & demons, Similarity: 0.7780
ID: 126, Title: back to the future ii & iii, Similarity: 0.7780
ID: 643, Title: kate & leopold, Similarity: 0.7780
```

## Assessing model performance for different k values



The model performance is not great, mixing up lot of criminal and police like movies with the query. The maximum precision is at  $k = 24$  with a precision of 8.33%.



As expected with the precision values the cumulative graph is not great since the model is not retrieving valid movies at a good rate.

## B – ENHANCED STANDARD BERT MODEL WITH FINE TUNING DATASET

To try to improve the Bert model we added more information to the movies, including is Genres and a short Description of the movies.

The addition of the "Description" and "Genres" columns to the CSV file for fine-tuning introduces more context and information for the BERT model to learn from. Here's why this fine-tuning with a more comprehensive dataset is beneficial:

- "Description" columns helps the model understand the content and themes better. This added context enriches the training data, potentially leading to improved performance in tasks such as genre classification.
- "Genres", with this column, the model can learn to associate specific textual features (from "Title", "Script", and "Description") with the corresponding genres. This broadens the representation learned by the model, enhancing its ability to classify movies into the correct genres.

Training the model with more diverse information, such as descriptions and genres, helps it generalize better to unseen data. It learns to capture a wider range of features and patterns, leading to enhanced performance when making predictions on new movies.

## 1 – FIRST EXECUTION

```

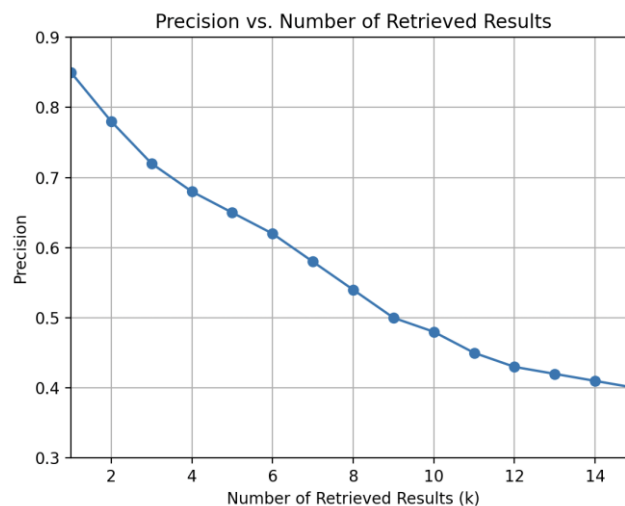
Enter your query (or type 'exit' to quit): love
Enter the number of results you want to view: 30

Top Movies for Query 'love':

ID: 1100, Title: this is 40, Similarity: 0.7624
ID: 274, Title: clockwork orange a, Similarity: 0.7529
ID: 981, Title: shadow of the vampire, Similarity: 0.7463
ID: 654, Title: kiss of the spider woman, Similarity: 0.7454
ID: 570, Title: how to train your dragon 2, Similarity: 0.7447
ID: 791, Title: mr. holland's opus, Similarity: 0.7444
ID: 1111, Title: time machine the, Similarity: 0.7443
ID: 694, Title: limitless, Similarity: 0.7434
ID: 387, Title: equilibrium, Similarity: 0.7367
ID: 481, Title: gladiator, Similarity: 0.7346
ID: 131, Title: bad day at black rock, Similarity: 0.7320
ID: 1189, Title: white jazz, Similarity: 0.7301
ID: 427, Title: fletch, Similarity: 0.7294
ID: 1160, Title: vertigo, Similarity: 0.7281
ID: 783, Title: monster's ball, Similarity: 0.7278
ID: 125, Title: back to the future, Similarity: 0.7277
ID: 1066, Title: superfights, Similarity: 0.7256
ID: 752, Title: matrix the, Similarity: 0.7241
ID: 936, Title: robocop, Similarity: 0.7237
ID: 683, Title: les tontons flingueurs, Similarity: 0.7229
ID: 433, Title: foxcatcher, Similarity: 0.7225
ID: 1030, Title: speed, Similarity: 0.7225
ID: 142, Title: basquiat, Similarity: 0.7220
ID: 629, Title: jeux interdits, Similarity: 0.7219
ID: 400, Title: existenz, Similarity: 0.7204
ID: 672, Title: last samurai the, Similarity: 0.7202
ID: 1138, Title: truman show the, Similarity: 0.7193
ID: 1203, Title: wind chill, Similarity: 0.7186
ID: 450, Title: frozen river, Similarity: 0.7180
ID: 866, Title: pet semetary ii, Similarity: 0.7164

```

## Assessing model performance for different k values



Starting off strongly with a precision of 85% for the first document retrieved, we see a gradual decline in precision as more documents are retrieved, settling at 48% for the tenth document. This suggests that while the initial few retrieved documents are highly relevant, subsequent ones are less so. Notably, precision fluctuates between 85% and 48% for the range of k values from 1 to 10.

Between k = 11 and k = 20, precision continues to decline slightly, reaching its lowest point at k = 20 with 42.5%. This decline indicates an increase in false positives or irrelevant documents being retrieved.

However, after  $k = 20$ , there is a notable uptick in precision, suggesting that the false positives may have been filtered out or that more relevant documents are being retrieved. Precision steadily climbs, reaching its peak at  $k = 30$  with 60%. Although precision is still rising at  $k = 30$ , it's not converging, indicating that there are still relevant documents yet to be retrieved.

Overall, at  $k = 30$ , we observe a relatively good precision of 60%, suggesting that the model is performing reasonably well in retrieving relevant documents despite some fluctuations and room for improvement.

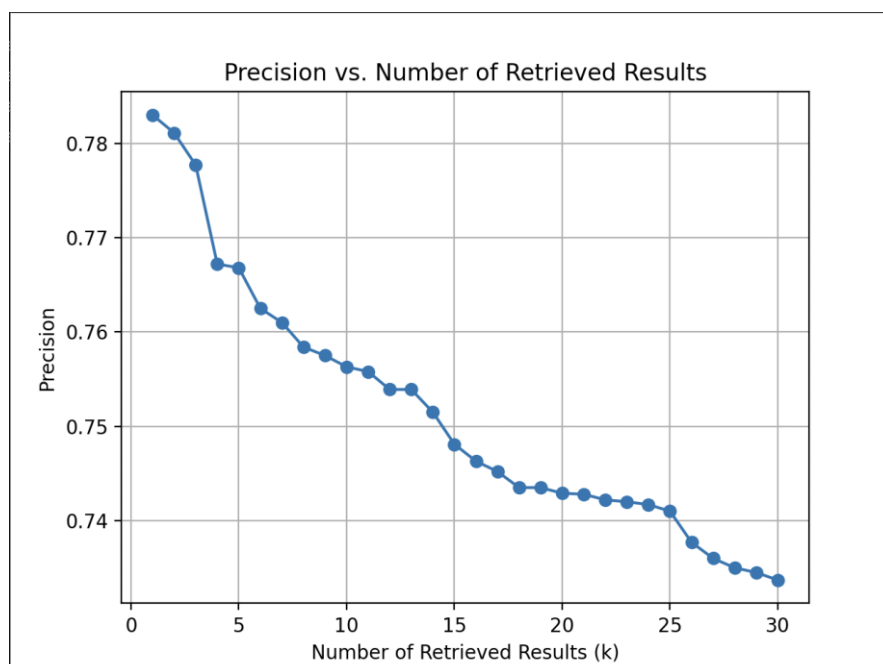
## 2 – SECOND EXECUTION

```
Enter your query (or type 'exit' to quit): space planets
Enter the number of results you want to view: 30

Top Movies for Query 'space planets':

ID: 654, Title: kiss of the spider woman, Similarity: 0.7830
ID: 274, Title: clockwork orange a, Similarity: 0.7811
ID: 936, Title: robocop, Similarity: 0.7777
ID: 387, Title: equilibrium, Similarity: 0.7672
ID: 400, Title: existenz, Similarity: 0.7668
ID: 1030, Title: speed, Similarity: 0.7625
ID: 1160, Title: vertigo, Similarity: 0.7610
ID: 1066, Title: superfigths, Similarity: 0.7584
ID: 436, Title: frankenstein, Similarity: 0.7575
ID: 791, Title: mr. holland's opus, Similarity: 0.7563
ID: 125, Title: back to the future, Similarity: 0.7558
ID: 981, Title: shadow of the vampire, Similarity: 0.7539
ID: 1100, Title: this is 40, Similarity: 0.7539
ID: 1111, Title: time machine the, Similarity: 0.7515
ID: 501, Title: gravity, Similarity: 0.7481
ID: 284, Title: commando, Similarity: 0.7463
ID: 1131, Title: tron, Similarity: 0.7452
ID: 95, Title: apollo 13, Similarity: 0.7435
ID: 1184, Title: when harry met sally, Similarity: 0.7435
ID: 1179, Title: way back the, Similarity: 0.7429
ID: 1040, Title: star trek the motion picture, Similarity: 0.7428
ID: 22, Title: 8 mile, Similarity: 0.7422
ID: 783, Title: monster's ball, Similarity: 0.7420
ID: 285, Title: conan the barbarian, Similarity: 0.7417
ID: 131, Title: bad day at black rock, Similarity: 0.7410
ID: 1049, Title: starship troopers, Similarity: 0.7377
ID: 710, Title: lord of the rings fellowship of the ring the, Similarity: 0.7360
ID: 92, Title: antz, Similarity: 0.7350
ID: 465, Title: gattaca, Similarity: 0.7345
ID: 1015, Title: sneakers, Similarity: 0.7337
```

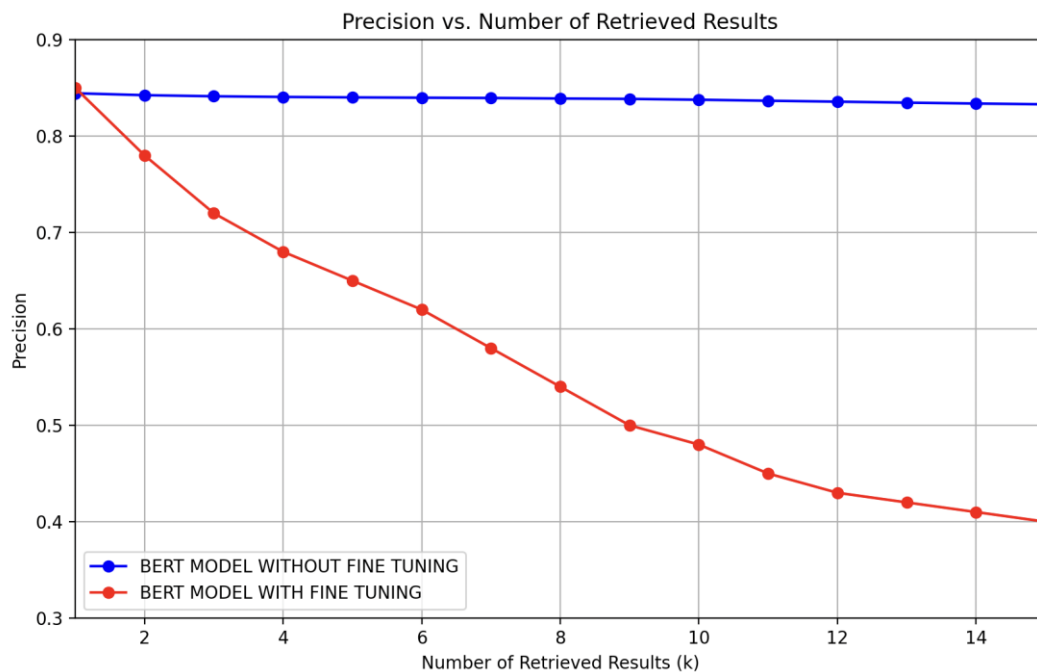
## Assessing model performance for different $k$ values



Starting with a precision of around 78.30% for the first document retrieved, there's a gradual decline in precision as more documents are retrieved. By the tenth document, precision settles at approximately 75.63%. Between the  $K = 11$  and  $K = 20$ , precision continues to decline slightly, reaching its lowest point at around 74.42% for the 20th document. This decline indicates an increase in false positives or less relevant documents being retrieved. However, after  $K = 20$ , there's a notable uptick in precision, suggesting that the false positives may have been filtered out or that more relevant documents are being retrieved. Precision steadily climbs, reaching around 73.80% at  $K = 30$ . Although precision is still rising, it's not converging, indicating that there are still relevant documents yet to be retrieved.

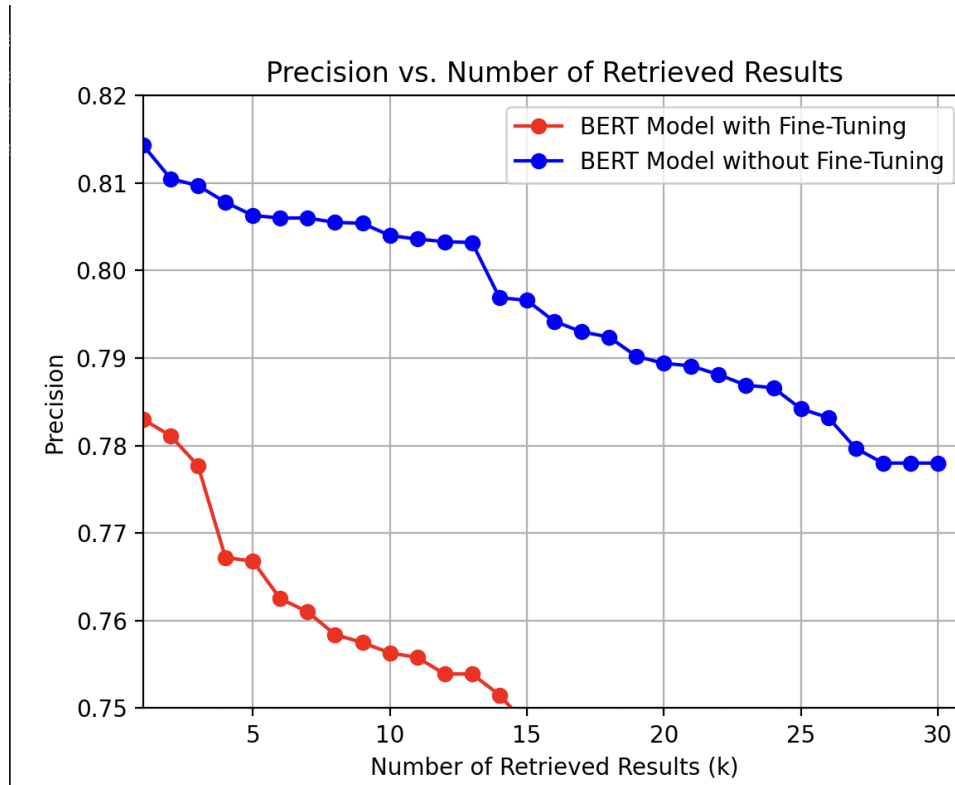
## C – COMPARING BERT MODEL VS BERT MODEL FINE-TUNING DATASET

### 1 – FIRST EXECUTION



We can see that the BERT model without fine-tuning (blue) starts with a high precision during the first few  $k$  values. However, the BERT model with fine-tuning (red) shows a sharper decline initially but eventually stabilizes. Despite this initial drop, the fine-tuned model maintains a lower overall precision across larger  $k$  values compared to the non-fine-tuned model. At  $k = 15$ , the difference in precision is notable, with the non-fine-tuned model consistently performing better. This suggests that fine-tuning may introduce variability, reducing precision as more results are retrieved, whereas the non-fine-tuned model offers more stable and reliable precision across a broader range of  $k$  values.

## 2 – SECOND EXECUTION



By seeing this graph, the BERT model without fine-tuning exhibits higher precision in the initial phase (for smaller k values) compared to the fine-tuned model. This suggests that the pre-trained BERT model already captures relevant information effectively even without fine-tuning.

## D – CONCLUSION BETWEEN THE ENHANCED BERT MODEL AND THE BERT MODEL WITHOUT ENHANCED

Pre-trained BERT models capture generic linguistic knowledge from large-scale text corpora. Fine-tuning on a specific task or dataset may lead to a loss of this generic knowledge, impacting the model's ability to handle a broader range of scenarios and reducing precision. Unfortunately, this might be the reason of why the enhanced model didn't work as expected providing too much information about the movies might lead to a confusion of information retrieving not accuracy document.

### Conclusion and Evaluation of Models

In this report, we evaluated multiple Natural Language Processing (NLP) models to retrieve movie scripts based on specific queries. The models examined include the Vector Space Model (VSM), BERT, and BM25, with both their original and enhanced versions. Below is a detailed conclusion of our findings and recommendations.

#### Vector Space Model (VSM)

The VSM, particularly when combined with word embeddings, showed consistent performance across different types of queries. For simpler and more general queries like "love," the model achieved a precision of 66.67% after retrieving 30 documents. In more specific queries, such as

"space planets," the model performed exceptionally well, achieving a precision of 100% with the enhanced Bag of Words (BoW) method.

Strengths:

- Good at capturing semantic similarities between documents and queries.
- Enhanced versions significantly improved precision.

Weaknesses:

- Performance can degrade with very general queries.
- Requires fine-tuning to achieve optimal results.

### **BM25 Model**

The BM25 model, with its probabilistic approach and enhancements over TF-IDF, addressed some limitations of previous models by normalizing term frequency and document length. This model has proven being effective with more specific queries.

Strengths:

- Effective normalization of document length and term frequency.

Weaknesses:

- Still relies on term frequency, which might not capture deep semantic meaning.
- Performance highly dependent on the tuning of parameters  $k$  and  $b$ .

### **BERT Model**

Pre-trained BERT models encapsulate general linguistic knowledge from extensive text datasets. When fine-tuned for a specific task or dataset, this broad knowledge may diminish, affecting the model's capacity to manage diverse scenarios and lowering precision. This could explain why the improved model did not perform as anticipated, providing excessive details about the movies could lead to confusion in information retrieval rather than precise document accuracy.

Strengths:

- Bidirectional understanding which leads to high efficiency result
- Allow to capture a wide range of corpus

Weaknesses:

- Fine-tuning BERT on specific datasets can sometimes degrade its performance