CS6200 Information Retrieval
Assignment 4
Nov 19, 2018

Information about the search engine: Netflix

- Kinshuk Juneja

Introduction

Netflix is an amazing digital success story. It is the world's leading entertainment service that provides on-demand thousands of movies, tv shows and documentaries that subscribed users can watch anywhere and anytime on their website or app, all without commercials. The ease of service is probably what makes it so popular. With thousands of contents uploaded and many updated daily, it provides a powerful inbuilt search to filter out media according to for your own interest. Some of the categories that you can filter out movies and tv shows include: Action, Anime, Classics, Comedies, Cult, Documentaries, Horror, Sci-Fi, Sports, International, etc. The category of search I will be providing will be on these movies and tv shows. I am particularly keen on trying out queries that are not directly movie or tv show names, but something related or about them and then analyze the results to see if they match my information need.

I am a big movie buff and as a user, I like to watch movies online rather than downloading them with patience. With any website, what really intrigued me about Netflix is how do they scale and keep up with so much content and so many users. I thought to analyze a little more about the company and found some really amazing stats about it:

Netflix subscribers: 130 million (Last updated 7/16/18).
Netflix daily searches: 3 million
Estimate number of people Netflix reaches: 300 million (Last updated 19/26/17).
The number of international Netflix subscribers: 68.29 million (Last updated 4/16/18).
Percentage of Netflix users that come from outside the US: 54.6% (Last updated 4/16/18).
The number of Netflix subscribers in the U.S.: 56.71 million (Last updated 4/16/18).
Netflix revenue in 2017: $11 billion
Netflix revenue expected for 2018: $15 billion
Percentage of US adults that stream Netflix daily: 23% (Last updated 3/13/17).
An average amount of video users watches on Netflix per week: 1 billion hours (Last updated 4/19/17).
An estimated amount of ads Netflix users avoid by watching it: 160 hours of ads per year (Last updated 5/11/16).
Amount Netflix plans to spend on original content in 2018: $13 billion (Last updated 7/8/18).
Netflix estimated market cap: $140 billion (Last updated 4/16/18).
The number of countries that Netflix reaches: 200 (Last updated 3/10/17).
Netflix estimated company value: $130 billion (Last updated 3/2/18).
Percentage of US streaming service viewers Netflix reaches: 75% (Last updated 4/10/17).

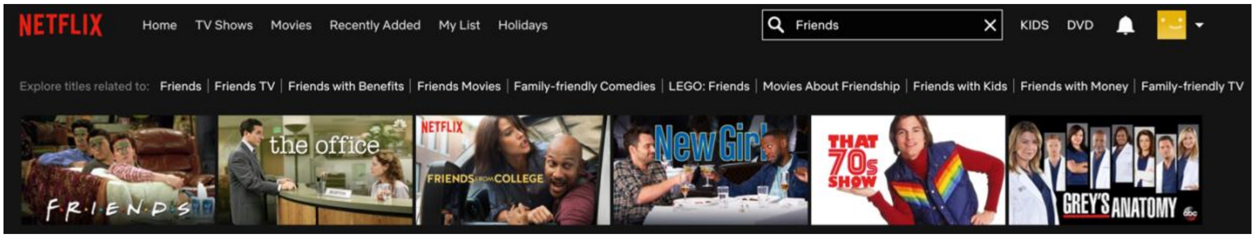
The above stats not only say a lot about the success of the company, but also gave me a lot to think about how they manage their storage cost on AWS or other cloud services, give real-time search results that are relevant, keep up with the increasing demand for new media content and various other factors to keep the company running and popular.

Analyzing Netflix's Search Results

Exact movie/tv show search:

Like any classic big search engine such as Google or Bing, Netflix also has to implement many algorithms in its backend processing to provide relevant search results. The only difference is that it does so for its dedicated domain related to movies and tv shows. For an exact query match of a movie or a tv show name, it will return a direct result containing the first result for that title (assuming it still features it). The other results are usually similar movies or tv shows based on the same cast (or at least one actor or actress), same genre, near and about same timeline for its release, similar rating, similar popularity (again based on rating or number of searches/clicks), matching some keywords if it is a two or more word query, etc.

Here is an example of a simple search query "Friends", which is a very popular tv show:

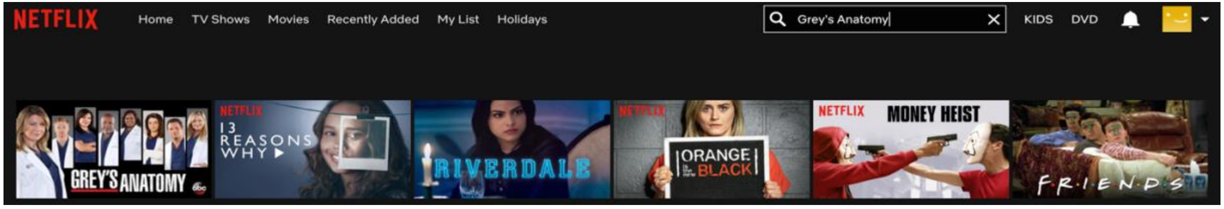

NETFLIX
Home
TV Shows
Movies
Recently Added
My List
Holidays
Friends
KIDS
DVD
Explore titles related to: Friends | Friends TV | Friends with Benefits | Friends Movies | Family-friendly Comedies | LEGO: Friends | Movies About Friendship | Friends with Kids | Friends with Money | Family-friendly TV
FRIENDS
the office
NETFLIX
FRIENDS FROM COLLEGE
New Girl
THAT 70s SHOW
GREY'S ANATOMY

The first result is expected, which is a direct name of the tv show that matches the search query. Now the next top 5 results are interesting. The second one is another popular tv show "The office" based on same genre i.e. comedy and is also a very long running tv show since 2005. I searched for IMDB rating out of curiosity to compare the ratings of the two and I found that it's not so different. "Friends" is rated 8.9/10 and "The office" is rated 8.8/10. Perhaps, the ranking is a combination of both, same genre and the least absolute difference in ranking from a popular rating site such as IMDB.

The third result is another tv show "Friends From College". The first thing that is obvious in this result is the keyword "Friends" that matches with the search query. It too falls in the same genre as the comedy but has a lower IMDB rating of 6.6/10. In comparison with the fourth result "New Girl" tv show, I found that "New Girl" has a better ranking on IMDB which is 7.7/10. The reason I can think of promoting "Friends From College" over "New Girl" could be the matching keyword "Friends" which is given more weight in relevance.

The fifth tv show is again a comedy tv series, but the last one i.e. "Grey's Anatomy" is an interesting result here. The reason I say interesting is, it is the only result in top 6 which does not fall in the comedy genre. One similarity that I can think of "Friends" with "Grey's Anatomy" is the long cast and popularity of the tv show which may have been considered to give a diverse result other than same genre parameter.

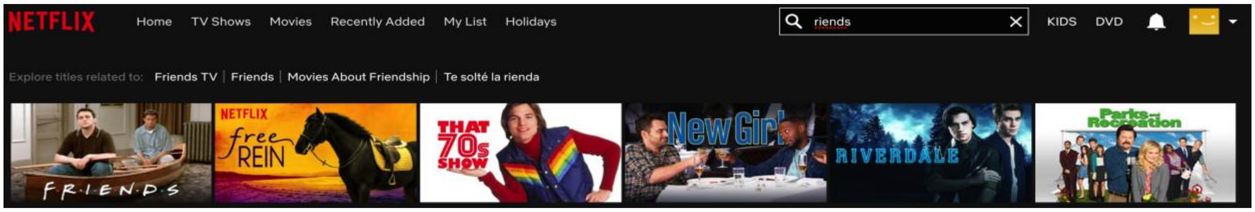
Note: I also tried out a separate search query "Grey's Anatomy" out of curiosity to check if "Friends" is listed over there as a result and it shows exactly at the same sixth position.



Search based on missing prefix, suffix and misspell queries:

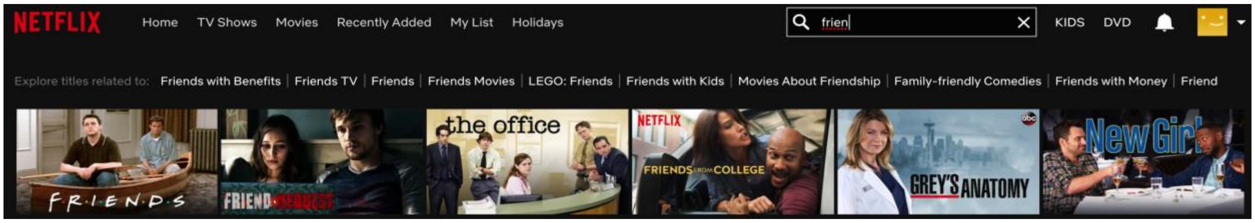
Now let’s try and search the above query example of “Friends”, but this time we will be evaluating search responds based on missing prefix, suffix and misspell and then analyze the search results.

Here is an example for a searched query with the missing prefix for original word "Friends": "riends"



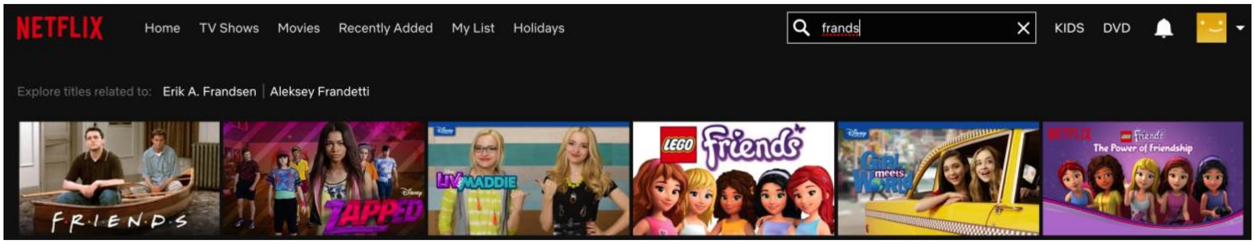
For the above query, the first result returned is still “Friends” tv show for which we had the missing prefix character ‘F’. The remaining results also fall into the same genre i.e. comedy as analyzed before also.
The second result, in particular, is, however, more interesting. If we notice the keyword "Rein" in the second result, it is very similar to our query "riends". My educated guess would be that Netflix uses Soundex algorithm and evaluated both to be the same and hence, returned the result based on the same Soundex code produced.

Here is an example for a searched query with the missing suffix for original word "Friends": "frien"
Notice, this time I omitted two characters instead of one to check the efficiency of the underlying algorithm that Netflix uses to correct an incorrect query.


NETFLIX
Home
TV Shows
Movies
Recently Added
My List
Holidays
frien
KIDS
DVD
Explore titles related to: Friends with Benefits | Friends TV | Friends | Friends Movies | LEGO: Friends | Friends with Kids | Movies About Friendship | Family-friendly Comedies | Friends with Money | Friend
FRIENDS
FRIEND REQUEST
the office
NETFLIX FRIENDS FROM COLLEGE
GREY'S ANATOMY
New Girl

For the above query, Netflix was able to detect the error for missing suffix characters i.e. ‘d' and ‘s' and produced the first result as "Friends" which was meant to type. The second and fourth result also contains the keyword "Friend/Friends” in the title of the show and it must have used term frequency to rank these based on the title after correcting the incorrect query. The other results remain different popular comedy shows except for "Grey's Anatomy" which clearly has some connection with "Friends" tv show in Netflix's data analysis.

Here is an example for a searched query with spelling error for original word “Friends”: “Frands”

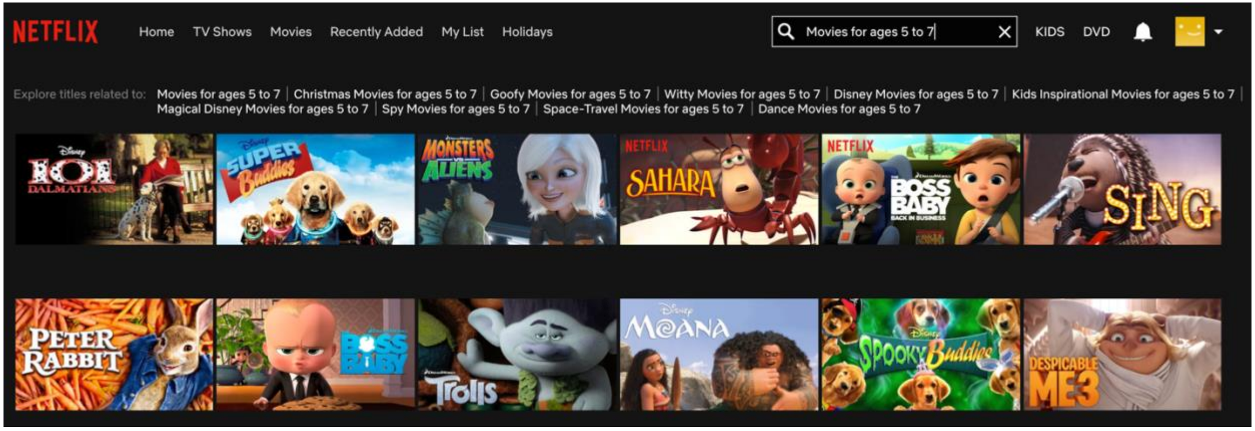


For the above query, Netflix again produced the intended result i.e. “Friends” in the first result. This hints that it is using Soundex algorithm as both come out to be the same Soundex code. The other results are different than usual this time inclined more towards children shows, but few of them still have “friends” keyword in their title which is good.

Searches based on age parameter:

Now let’s try out a more complicated search query, which is not a direct match for a movie or a tv show name.

Here is an example for a search query “Movies for age 5 to 7”:



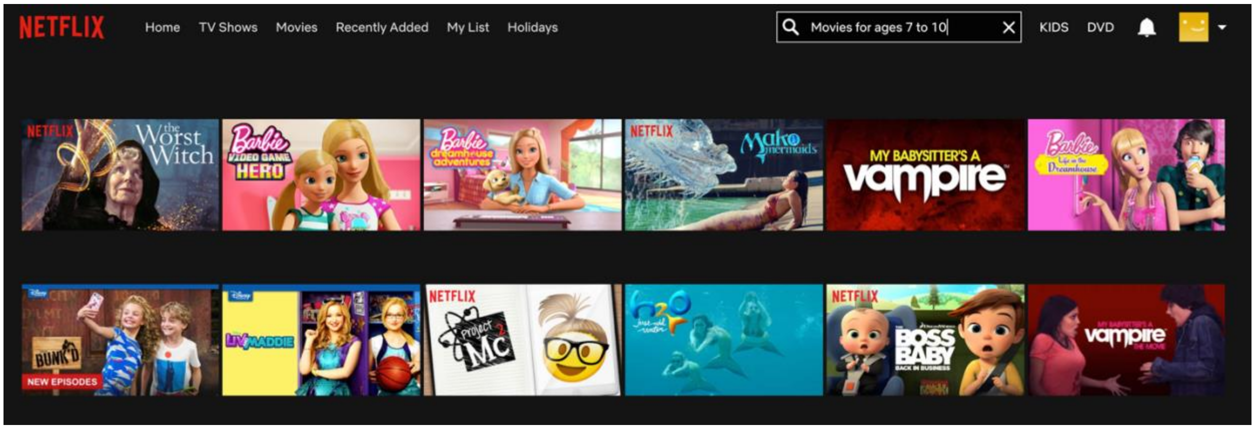
Movies for ages 5 to 7

Explore titles related to: Movies for ages 5 to 7 | Christmas Movies for ages 5 to 7 | Goofy Movies for ages 5 to 7 | Witty Movies for ages 5 to 7 | Disney Movies for ages 5 to 7 | Kids Inspirational Movies for ages 5 to 7 | Magical Disney Movies for ages 5 to 7 | Spy Movies for ages 5 to 7 | Space-Travel Movies for ages 5 to 7 | Dance Movies for ages 5 to 7

For the above query, the good thing about the result is that it displayed all the results for movies which are anime. Another good thing that I liked for this query was, there was no mention for any particular movie's name or any keyword in the search query, yet it picked up my information need to display movies that are appropriate for children. Hence, we can be sure that for these results, relevance is not based on term frequency as none of the results has any keyword matching any word in the search query. It probably picked up keyword "movies" to display only movies and not tv shows or documentaries and the keyword “age” to filter out movies based on the age specified afterward. Many of the movies listed out are by the Production house "Disney" which is known for producing movies more for children and that is anime. Lastly, needless to say, but Netflix does not show biased results by promoting Netflix's production or signed movies first over other or more popular movies as we can see in the results above (top three are "Disney" house production movies and the fourth one is "Netflix" original movie).

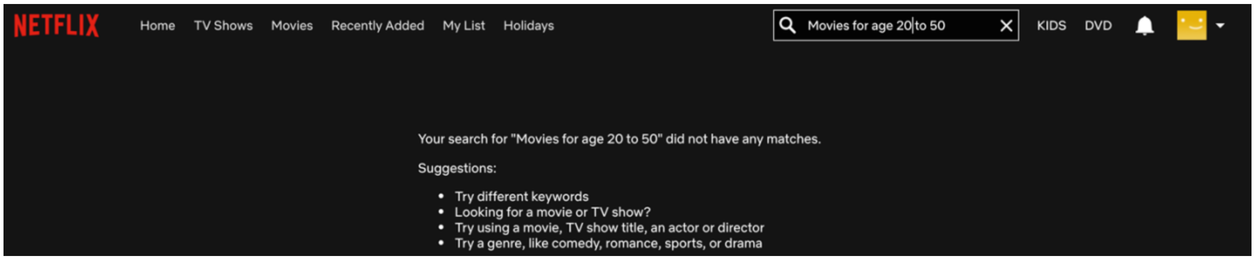
This brings me to the question; how different results will be if I change the age parameter from 7 to 16.

So, I tried my next query as: “Movies for age 7 to 10” as shown below.



Movies for ages 7 to 10

I still see a lot of anime movies, which is expected, but I also see some that are not like ("Mako Mermaids", "My Babysitter's a Vampire", "Bunk'd", "Liv And Maddie" and a few others).
I am mostly satisfied with the above results that have a mix of both anime and non-anime movies (but still suited to children). One or two results are not really movies, but tv shows such as "Bunk'd", but it is a tv show suited to children, so I am not majorly disappointed to see it there (1 tv show out of top 12).
For my next query, unfortunately, Netflix did not show any search results for specific age parameter over 20:



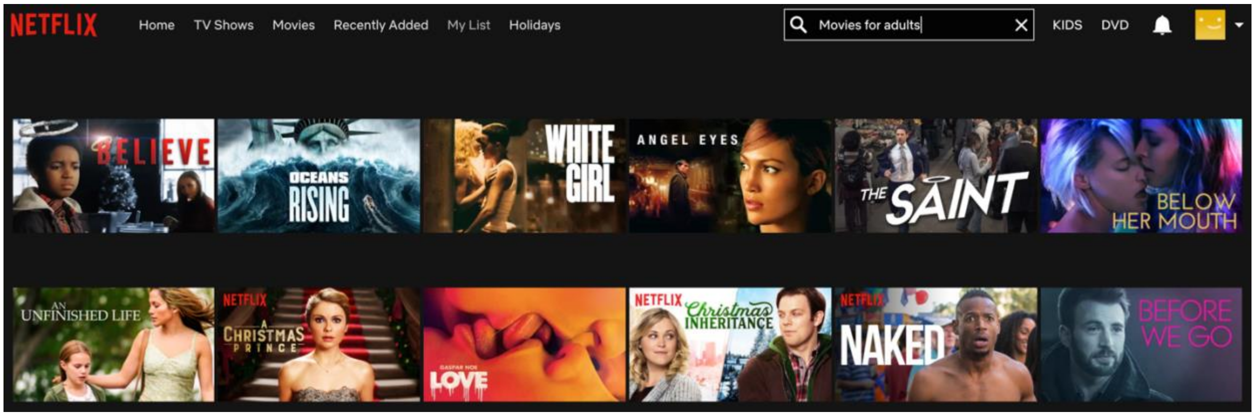
Movies for age 20 to 50

Your search for "Movies for age 20 to 50" did not have any matches.

Suggestions:

- Try different keywords
- Looking for a movie or TV show?
- Try using a movie, TV show title, an actor or director
- Try a genre, like comedy, romance, sports, or drama

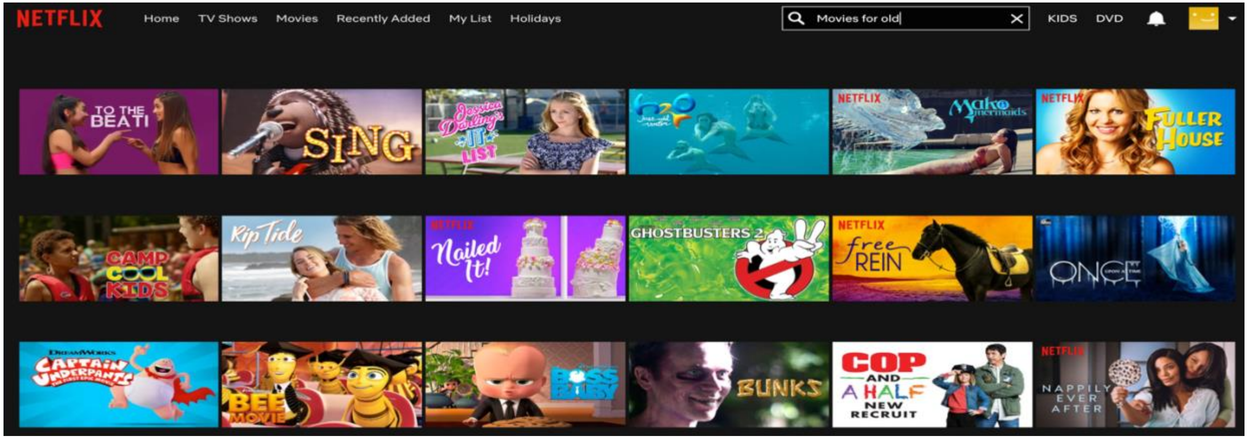
So, I tweaked my query to be a little more general such as "Movies for adults" and that worked:



Movies for adults

This query was interesting to analyze since the keyword "Adult" could mean two things: Adults who may like to watch thriller, horror or drama movies. It could also mean 'R' rated movies that only adults who are 17 and above in age can watch. I do see some 'R' rated movies and some that are not.

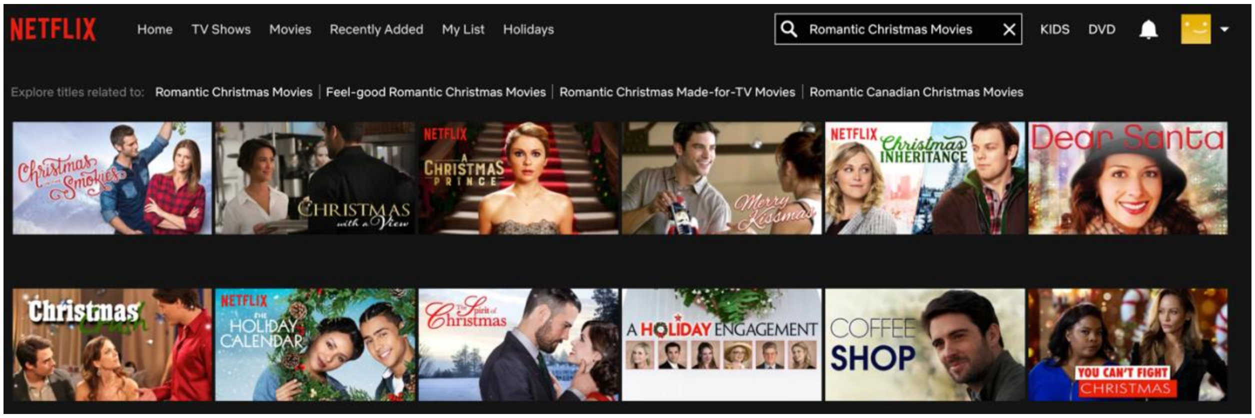
My next query was to see movies that elderly people like to watch, so I searched “Movies for old”:



This gave me some unusual results with movies that are mostly anime and suited to children. I guess, Netflix really believes in the saying that "With age, old people become more like children" (pun intended).

Search based on genre and cast:

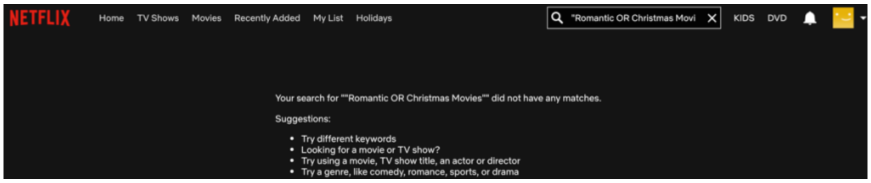
For my next query, I wanted to try a mix of movies based on Christmas festival and which are romantic. So, I tried “Romantic Christmas Movies” as a query:



Explore titles related to: Romantic Christmas Movies | Feel-good Romantic Christmas Movies | Romantic Christmas Made-for-TV Movies | Romantic Canadian Christmas Movies

All the top results are of romantic movies that are based on holiday festival Christmas. It probably picked up the keyword Christmas and matched that with movies containing that term in its title (more preference) or in the synopsis (less preference in comparison). It also combined the term "romantic" with source information of that movie listed in the genre, which is easy to obtain from that production company or rating websites such as IMDB.

Note: When I tried out a Boolean query “Romantic OR Christmas Movies”, it did not return me any result, which was surprising to me as it was a transformed query which search engines usually accept.

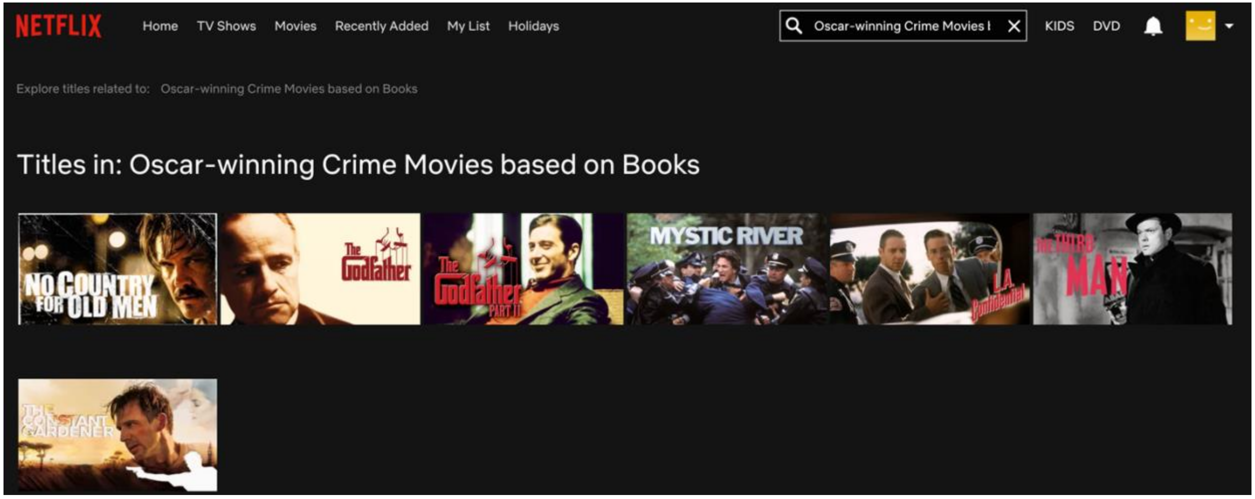


Your search for ""Romantic OR Christmas Movies"" did not have any matches.

Suggestions:

- Try different keywords
- Looking for a movie or TV show?
- Try using a movie, TV show title, an actor or director
- Try a genre, like comedy, romance, sports, or drama

For my next query, I really wanted to make it long and very specific to test the search engine capabilities. So, I searched “Oscar-winning Crime Movies based on Books”:

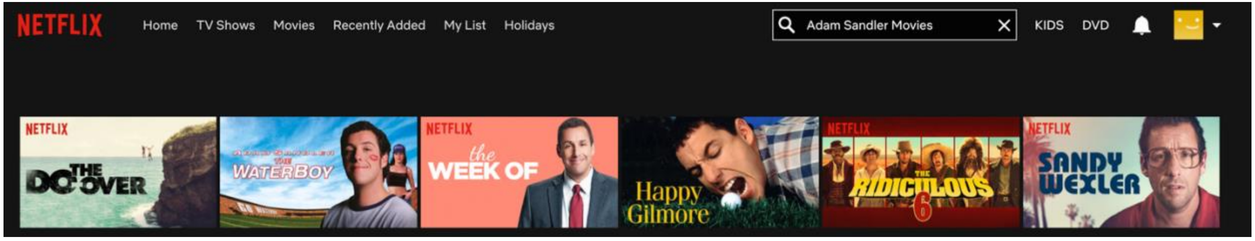


Explore titles related to: Oscar-winning Crime Movies based on Books

Titles in: Oscar-winning Crime Movies based on Books

For the previous query and to my surprise, it showed really relevant results which I verified separately on Google and IMDB. All these movies are in fact Oscar-winning and based on crime genre which was made from the stories picked up from novels written before. Because this is a much more specific and longer query, it did not display more than seven results. I am still happy to see the relevance of those seven results instead of returning me no match found. It probably picked out its source from the web about the movies that have won Oscar and which were made out of books published before. Having retrieved that list, it must have then matched its Database for such movie names that they feature.

Next, I wanted to search movies based on cast. So, the query I selected was “Adam Sandler Movies”:



All of them had Adam Sandler in their cast which was expected and good. I wanted to really analyze more about how the search results give top results already filtered out on the cast, so I looked at the release date and rating. The first movie is released in 2016 which is a recent result than the second one that was released in 1998. But, the IMDB rating of the first one which is 5.7/10 is lesser than the IMDB rating of the second one which is 6.1/10. This means that preference in the first two is obviously given to the most recent release.

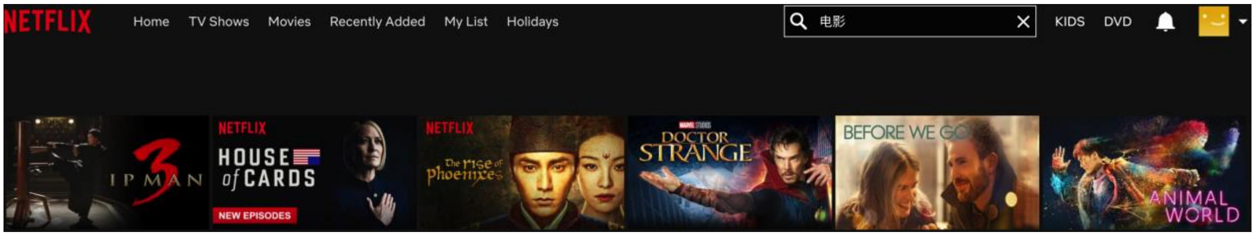
The third result "The week of" is actually the most recent release which is 2018, but it falls short on IMDB which is just 5.1/10, hence the third position. The other results are also mix of both recent and old releases and the pattern that I observed is that whenever a result contains a more recent release with not much IMDB rating difference, it is promoted first compared to an older release and slightly more (not significantly a lot) IMDB rating.

Note: My motivation to try out this query was also because I came across a fact that overall Netflix users have spent more than half a billion hours watching Adam Sandler movies, which is a lot. It might also be because of this fact that this query had maximum results out of all the queries I tried out earlier.

Search based on different language queries:

Here, I wanted to evaluate how Netflix processes queries that are not written in English.

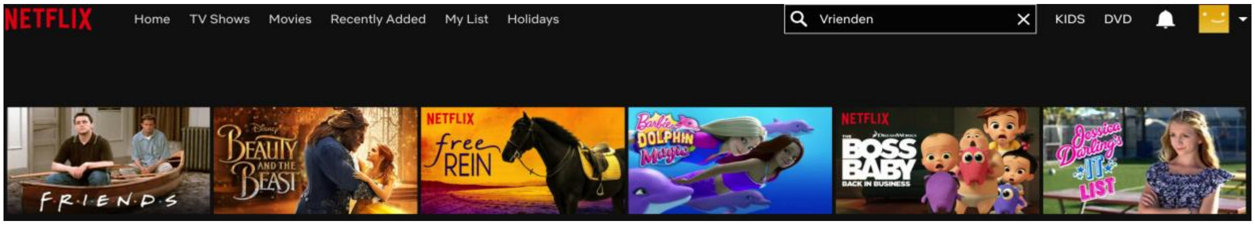
Below is a query result that I tried out for the English query “Movies” translated to Chinese: “电影”



电影

For the above query result, I found that it not only accepted the query and returned results, but it also returned some of the results that were Chinese movies in particular. I was really impressed with its capability to not only be able to process different language query other than English, but also show relevant results (even though I did not ask for Chinese movies in particular).

Next, I wanted to search Dutch translation of the word “Friends” which is a popular tv show and I have been using it in many of my queries: “Vrienden”

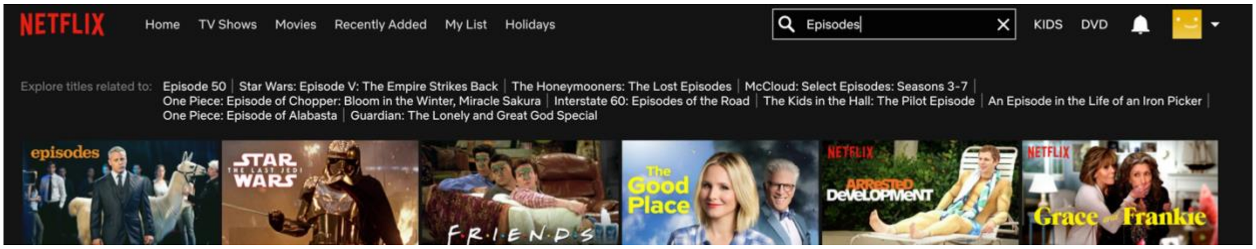


Vrienden

For the above query result, it not only showed its capability of processing different language queries other than English but also showed me an exact match of the English tv show "Friends" which was the expectation.

Search based on vague queries:

The next query I wanted to try out was to search for a tv show called "Episodes":

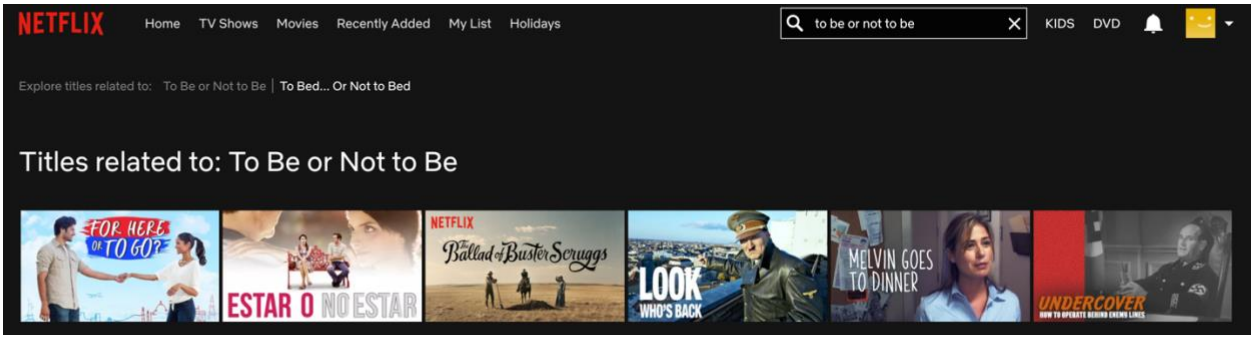


Episodes

Explore titles related to: Episode 50 | Star Wars: Episode V: The Empire Strikes Back | The Honeymooners: The Lost Episodes | McCloud: Select Episodes: Seasons 3-7 | One Piece: Episode of Chopper: Bloom in the Winter, Miracle Sakura | Interstate 60: Episodes of the Road | The Kids in the Hall: The Pilot Episode | An Episode in the Life of an Iron Picker | One Piece: Episode of Alabasta | Guardian: The Lonely and Great God Special

Needless to say, why would a tv show creator name a tv show called “Episodes” which is so confusing. When I tried out this query, “Episodes” could mean two things: either I am interested in the tv show titled “Episodes” or I am looking for episodes of any tv show. Netflix responded with both and was really good with its diverse results. The first result was in direct relation with term “Episodes” that must have matched a tv show in their database that they still feature. The rest of the results were all highly rated and popular tv shows. The second result was technically not a tv show, but it was a continuation of an earlier movie part which also makes sense.

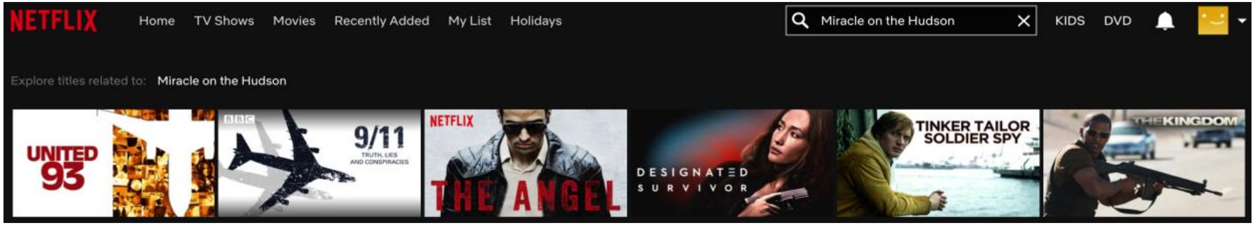
The next query I searched was not related to any movie or tv show title or the description, it was, in fact, a sentence containing all stop words i.e. "to be or not to be":



On analyzing the results, I was first surprised that it returned something considering it’s limitations to movies and tv shows domain and not so much information. Usually, a search engine ignores the stop words in processing query and gives more weight to other important terms (most importance to very rare terms). Netflix may or may not be doing that. The first result is ranked number one because it has two of the stop words i.e. "to" and "or" from the query compared to others which have one i.e. "to" or some other stop words that were not in a query like "the".
What really impressed me is the second result “Estar O No Estar” which is a Mexican film that translates to “Being or Not Being” containing two of the stop words from the original query and one similar term i.e. “being”.

Search based on informative and date queries:

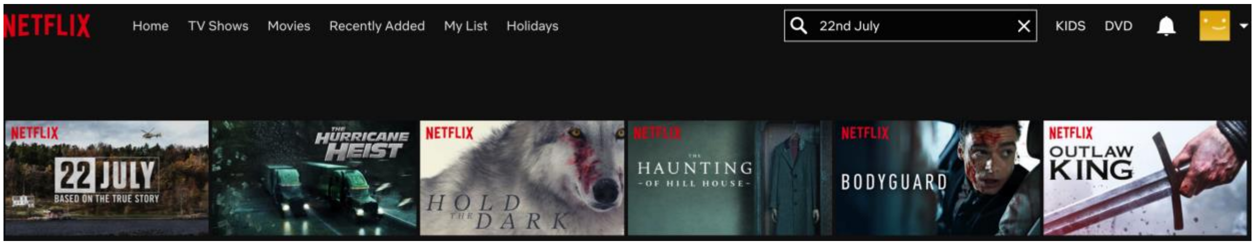
First, I wanted to test whether given the description of an incident and not a direct documentary or movie title, if Netflix is able to produce some results.
So, I chose my next query to be: “Miracle on the Hudson”



For the above query, Netflix probably could not find any documentary or movie title related to the successful landing of an aircraft on the Hudson Bay River that happened on January 15, 2009 (although I remember watching a movie titled “Sully” that used to feature on Netflix a long time back). I still liked that Netflix was able to identify the information being related to an aircraft accident and showed the top two results about 9/11.

For my next query, I wanted to give a date and analyze whether it shows some documentary/movie related to that date, or movies/shows that were released around that date.

So, I tried out with a query: “22nd July”

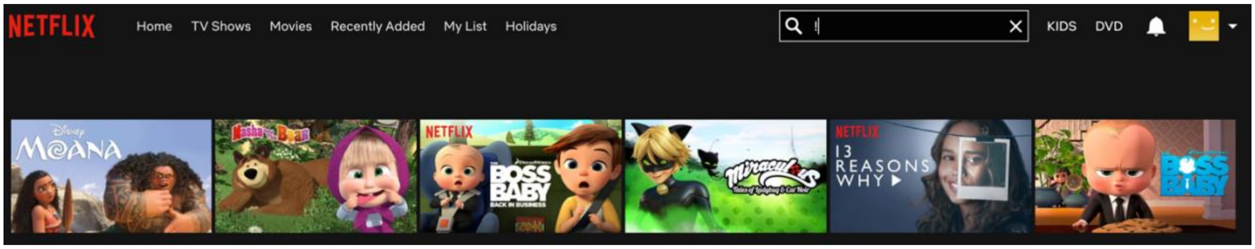


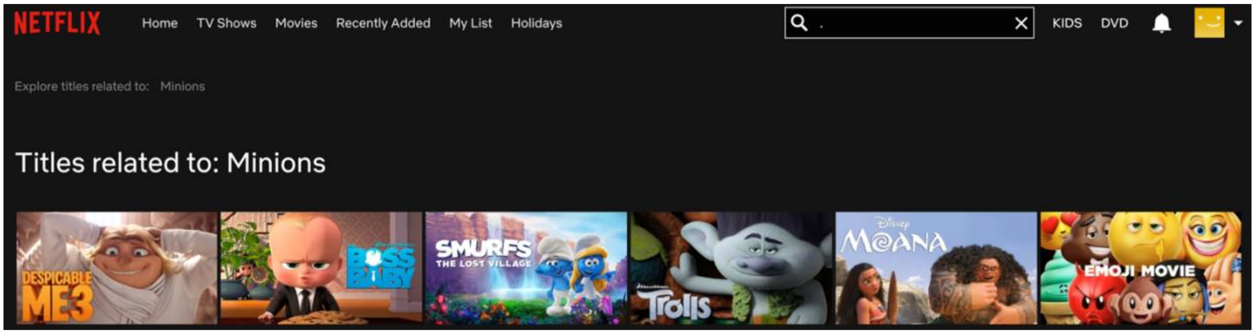
From the above query results, coincidently the first top result was a movie that was in fact made on the 22nd July incident on which there were two series of terrorist attacks in Norway. I checked the other top 5 results to verify if they were somehow related or released during the July month, but it did not appear so to be the case.

Search based on weird/gibberish queries:

Users can come up with all sorts of weird queries that makes no sense. For example: “.”, “!”, “…..”, “?”, “ffndsfjdjfsdnk". Other examples to test the edge cases could be an empty string if accepted, just space, a very long gibberish query to test if it overflow or a SQL statement.

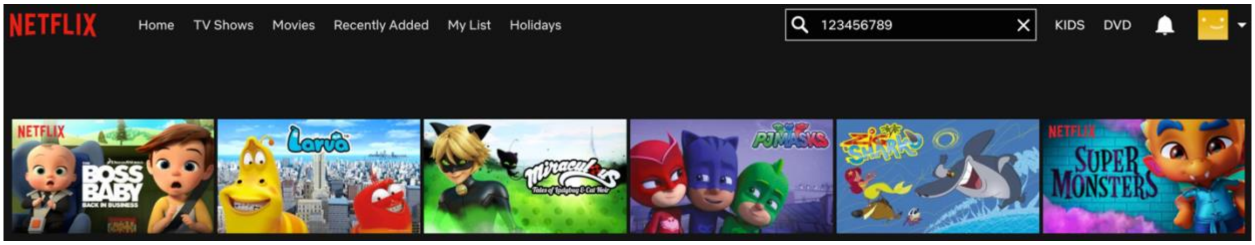
I tried them all and many other weird queries and found a pattern in results.





Explore titles related to: Minions

Titles related to: Minions



While many other queries, did not return any results, but some that did like "!", "." and "123456789" showed me all anime movies or movies suited to children. My educated guess from observing this pattern is that Netflix processes such weird queries that may have resulted from pressing any button or the ones that do not match any movie/tv show, is to draw a conclusion that such queries must have been entered by a child. Hence, they just return a list of results that are movies suited for children.

Data source and analysis

Over the course of various types of queries I tested on Netflix search engine, I learned some of its pattern and underlying ranking algorithms as discussed above.
For example, given the age parameter, it displays out results giving maximum weight to the age range specified for the query. If the query has an exact movie or tv show match, it will give first try and retrieve an exact movie or tv show titled as in query and then display rest of the results based on the same genre, cast, rating, release, similar or same keywords for a long query, and popularity. The priority for these parameters depends on the availability and it is usually a combination of two or three. The good thing is, one out of every five result stands out different from the usual pattern to give a more diverse result assuming user may have been looking for something else (no system is perfect to know what is the mind of a user).

Netflix has it’s own rating method in the form of thumbs up or thumbs down to help make a better suggestion to users. That does not mean it can rely only on this since there might be many users who do not utilize this feature. In this scenario, Netflix can look for displaying relevant results based on the rating given on IMDB and using that means to promote a result for a particular user. Other data that can be looked out could be title and synopsis to rank according to term frequency (usually term occurring in the title has been given more preference as observed in patterns above). Other data for categorizing a movie or a tv show in some genre, displaying the whole cast, giving a snippet or synopsis as mentioned will have to be from the original source i.e. film or video production companies.

Netflix is a data-driven company and it relies on various factors in deciding which show or a movie is good and should be recommended for a user. Not only that, but it also uses those factors to decide which movie or tv show should continue to be streamed and stay available on their website or app and which are the ones that are not doing so well and should be removed.

Here are some of the “events” Netflix tracks to learn more about the user and give a recommendation in searches and uses some parameters to promote ranking:

- When does a user pause, rewind or fast forward.
- What day you watch content (Netflix has found that people watch TV shows during the week and movies during the weekend).
- What time you watch content and which shows or movies do you complete.
- At which location (zip code) you watch content.

- What devices you use to watch (Do you like to use your tablet for TV shows and your Roku for movies? Do people access the Just for Kids feature more on their iPads, etc.?)
- When you pause and leave content and if you ever come back.
- The ratings are given (about 4 million per day) along with searches (about 3 million per day).
- Browsing and scrolling behavior.
- Netflix also looks at data within movies. They take various screenshots to look at "in the moment" characteristics. Netflix has confirmed they know when the credit starts rolling, but there's far more to it than just that. Some have figured these characteristics may be the volume, colors, and scenery that help Netflix find out what users like.

These above factors really play out in favor of Netflix's strength and why it is doing so well in giving appropriate recommendations to engage user's interest, support more relevance in results and deciding which movies or tv shows to feature next and which ones to keep featuring.

Technical hurdles, limitations, and scope of improvement

Netflix has a dedicated domain of movies and tv shows that it captures. All of its work is limited to this domain and sometimes working in a specific domain has its own challenges. With Netflix available to so many countries, what might be liked and highly popular in some may not be true for others. Thus, utilizing location and implementing different algorithms based on factors discussed above in displaying results and making recommendations cannot be generalized. For example, commercial movies tend to do really well in India, while that may not be the case in the US. On top of that, things are constantly changing and thus, Netflix has to learn and evolve along with users. New movies and tv shows are launched so frequently, thus making a decision which show or movie to feature on Netflix is again a learning that they have to keep doing.

In terms of search queries, I found that Netflix cannot support information or big queries related to a movie or tv show description. If the keywords or some relevance is present in title or synopsis, it will work really well. Otherwise, it can only take out keywords and search against the list present in that keyword's category such as kids if it exists.

The deciding factor of what works well for a particular user may also vary and cannot be generalized. This is a big problem not just for Netflix, but any search engine. One solution is to give diverse results along with what it classifies as relevant to address such a situation and I did notice some of them in my queries in the previous section.

A solution for big and complicated queries would be to at least give some results (which it failed for some of the queries I tried) regardless of relevance. By doing so, a user will not think something is broken and in fact may try to refine the query by seeing something rather than giving up and going away on seeing nothing. Not every user is an expert in making good queries and displaying something is better than displaying nothing. This could work out really well if one or two results from the list displayed are clicked by a user, but it was not initially what user came for. Usually displaying popular movies or tv shows in such cases would be a good strategy.

In terms of recommendations, Netflix can also utilize data sources from popular movie rating websites like IMDB (if they don’t already do so) to tally their rating against IMDB for producing better results. I noticed many times a movie that is highly popular on Netflix, is not so highly rated on IMDB and other websites. IMDB is a much more credible resource in doing so since it is a dedicated website specific to only movies and tv shows rating. Using a combination of their own data with other credible resources would be a good idea.

References

https://www.netflix.com
https://en.wikipedia.org/wiki/Netflix
https://www.imdb.com
https://gigaom.com/2012/06/14/netflix-analyzes-a-lot-of-data-about-your-viewing-habits/
https://expandedramblings.com/index.php/netflix_statistics-facts/