Overview of anime voice actor's social network and popularity.

Florencia Zanollo. Hideaki Takeda.

July 4, 2018

Contents

Abstract

Although Social Network of actors is a relatively common object of investigation that has been addressed many times, we can say that seiyuu (anime voice actors) come from a very different industry with a distinct way to relate to each other. In this research we use Wikidata and MyAnimeList to collect seiyuu information and build a Social Network. Topics explored:

- Structure and characteristics of seiyuu Social Network.
- Understanding what properties have a main role describing and predicting popularity of seiyuu.
- Compare prediction performances between different machine learning algorithms (and different models).

Introduction

TODO

Chapter 1

Anime/Seiyuu Dataset

TODO DIVIDIR UN POCO MAS ESTA PARTE EN SUBSECTIONS Y AGREGAR MAS INFO O MUESTRAS DE DATOS

Wikidata[1] is a collaboratively edited knowledge base intended to provide a common source of data which can be used by Wikimedia projects such as Wikipedia. The information is stored in RDF format, and can be retrieved in multiple ways, one of them being through a SPARQL endpoint.

Using Wikidata's SPARQL endpoint we retrieved a list of seiyuu. This list contains all persons that have seiyuu as occupation, a total of 6472 entities were obtained[2]. Gender, birthday and birthplace information was also fetched (last two were not used in the end because it was lacking in the majority of entities).

Since Wikidata information about seiyuu's works is really incomplete, MyAnimeList (MAL)[3] was used to retrieve voice acting roles and anime information. MAL is a social networking and social cataloging application website with a large database on anime and manga that started in April 6, 2006. Users can make a list of currently watching, watched and/or favorite anime; score, review, comment and recommend similar ones. They can also comment about and favorite people working on the industry (voice actors, directors, editors, etc).

Since only 59 of Wikidata's seiyuu entities had MyAnimeList ID (MALID) property, a matching between Wikidata and MyAnimeList was done using seiyuu's complete name to retrieve the ID for those who was missing. Successfully restoring 3033 MALIDs, giving a total of 3092 seiyuus with that property; 2956 of them having at least one work according to MAL so we are using this subset for our experiments.

Using Jikan API[4] and MALID, seiyuu data, voice acting roles and more information about each anime was retrieved.

An issue to take into account is whether we unify all anime adaptations of the same intellectual property as one or take a single adaptation as a independent work. We chose the later because each adaptation has its own producer, score, popularity, among other information; it would be incorrect to say a seiyuu

[1] http://wikidata.org/

[2] There's actually 7030 seiyuu in Wikidata but only 6472 of them have an English label (name)

[3] https://myanimelist.net/

[4] https://jikan.docs.apiary.io/#

worked in a popular work when that adaptation didn't have enough fame.

Information used:

- For Seiyuu:
  - Name
  - Debut (this was obtained from oldest work aired date)
  - Gender
  - Popularity (member_favorites information of MAL)
  - Work (anime roles)
- For Works (Anime):
  - Year that began airing
  - Favorites
  - Score (from 0 to 10, MAL user based)
  - Popularity (ranking over all MAL animes)
  - Members (how many MAL users have it on their list)
  - Genres

It's important to notice that data such as popularity and scores are retrieved from MAL, which is user review based only; it may differ with actual awards winning or professional reviewing of works.

Further, this dataset is biased in favor of more recent anime and seiyuu, since it accounts for more complete data and with better quality. Oldest anime in this dataset is from 1960 having no record about previous ones. Majority of seiyuu's debut are from 1988 which leads us to think information from thereon is more complete.

The data was stored using Virtuoso server to create a local SPARQL endpoint, mongodb was also used as an intermediate storage (before formatting data as RDF).

Chapter 2

Seiyuu Social Network

Social networks consist of a finite set of actors and the relations between them. Usually represented as a graph; with actors or organizations as set of nodes and a defined relation between them as set of edges. This structures are useful to analyze complex social interactions and communities.

2.1 Node and edge definition

Our social network consists of voice actors (seiyuu) as nodes and co-workership between them as edges. It's important to notice that this social network is time dependant since each seiyuu has a debut year and each anime has an aired time; giving us freedom to choose different time frames to observe.

Aside from being time dependant there exists different possible definitions of relationship or co-workership between seiyuu. One could say two actors know each other if they have worked in at least one job together, or maybe it requires more than one. There's also a time frame to define, relationship could take into account all works of both of them or only from certain years.

TODO EXPLICAR EL HECHO DE QUE ES UNA TWO MODE NETWORK Y DESPUES PONER QUE ELEGIMOS USAR SEIYUU COMO NODOS PERO PODRIA SER AL REVES

TODO MOVER EL SIGUIENTE PARRAFO AL FINAL DE ESTE CAPITULO After observing graphs built with different interpretations of relationship, the criteria for connecting two nodes became: *at least 10 works in common, during the time frame between the first debut registered (1960) and the year of observation.* Because requiring more jobs in common means less amount of edges, this leaves a more understandable graph and we verified it does without changing its structure.

There's also other interesting definitions of relationship, for example we can use only common works from the last x years or from all time. This options weren't explored; having into account our limited time we opted to decide on TODO PARAFRASEAR ESTO UN POCO MAS one that was more useful or logic and put more effort in analyzing the data and social network at hand.

2.2 Construction

As a first approach Gephi was used to build the network. Since the graph was big enough to bring performance problems and we needed to build the edges dynamically (which couldn't be done in Gephi) NetworkX was used instead.

NetworkX was chosen because it's an easy yet powerful Python library, it doesn't get along with massive graphs but ours was not big enough to present a problem. One can also export the graph and open it on Gephi, for a more visual analysis.

We needed to build the edges dynamically because they depend on the time frame we are looking at. For example if two actors worked together in 9 jobs between 1960 and 1970 we shouldn't see an edge between them; but if they worked together again in 1971 then looking at 1960-1971 they should be connected.

2.3 Analysis

Is easy to tell at first glance that this social network is really interconnected. With only 2956 nodes it has 395887 edges when only one work in common is required and 13629 edges when asking for 10 or more. It shows a thightly interconnected cluster surrounded by poorly or not connected nodes. This cluster represents 99% of the nodes of one work in common graph and 23% of 10 works in common. In terms of modularity we can see at least four clear communities in each graph, Fig 2.1 and Fig 2.2.

Table 2.1 shows some metrics about each graph. Requiring more works in common decreases average degree circumstantially but doesn't change a lot modularity or network diameter.

Table 2.1: Graph analysis

Graph	Avg Degree	Graph Density	Modularity
One work in common	267	0.09	0.2
Ten works in common	9	0.003	0.29

Graph	Network Diameter	Connected Components
One work in common	6	18
Ten works in common	7	2261

TODO ACOMODAR TODO ESTO, PRIMERO COSAS EN COMUN/ QUE QUIERO TENER LADO A LADO SOBRE ESTOS GRAFOS Y DESPUS PARTICULARIDADES, TOP10, UN POCO DE EXPLICACIN DE CADA UNO?

As proven by Fig. 2.3 grouwth of edges by year follows the same distribution regardless of how many works in common are used to build the social network.

Fig. 2.4 shows that more than half of the nodes are from last 18 years (2000 to 2018), giving us an idea of how much seiyuu industry is growing.

TODO HACER TOP 10 Y DEMS DATOS PARA CADA GRAFO (1 Y 10 TRABAJOS EN COMN) Y DIVIDIRLO EN SECCIONES, CAMBIAR LA EXPLICACIN DE MS ARRIBA TAMBIEN (AHORA YA NO NOS QUEDAMOS CON UNO SOLO PARA ESTA SECCIN, PERO S PARA EL RESTO). PONER

TAMBIN QUE AC USO LAS DOS DEFINICIONES PARA MOSTRAR LO PARECIDAS QUE SON EN ESTRUCTURA Y QUE PARA EL RESTO USO LA DE AL MENOS 10 TRABAJOS

2.3.1 At least 1 work in common

2.3.2 At least 10 works in common

Tables 2.2 and 2.3 show top 10 nodes, for degree and betweenness centrality.

Table 2.2: Top 10 degree

Name	Degree
Takehito Koyasu	311
Akira Ishida	273
Mamiko Noto	258
Daisuke Namikawa	232
Katsuyuki Konishi	229
Keiji Fujiwara	220
Junichi Suwabe	216
Toshiyuki Morikawa	215
Rie Kugimiya	213
Nobuyuki Hiyama	201

Table 2.3: Top 10 Betweenness centrality

Name	Betweenness Centrality
Takehito Koyasu	18489.44
Mamiko Noto	10988.96
Daisuke Namikawa	9570.48
Akira Ishida	8299.19
Rie Kugimiya	7560.16
Katsuyuki Konishi	7413.71
Kenichi Ogata	7160.54
Harumi Sakurai	6775.76
Keiji Fujiwara	5980.58
Yoshimasa Hosoya	5607.95

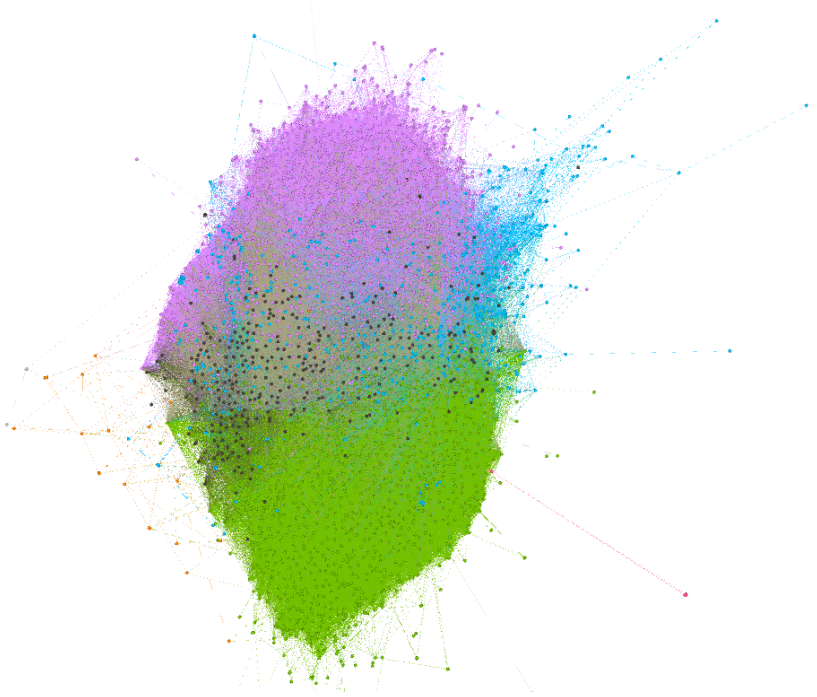

Figure 2.1: At least one work in common graph coloured by community.

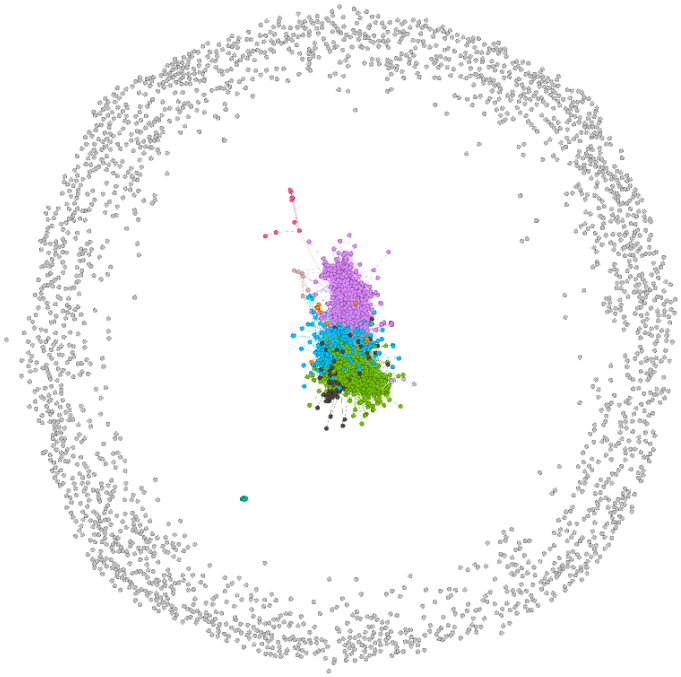

Figure 2.2: At least ten works in common graph coloured by community. Big cluster at the center, surrounded by loosely connected nodes.

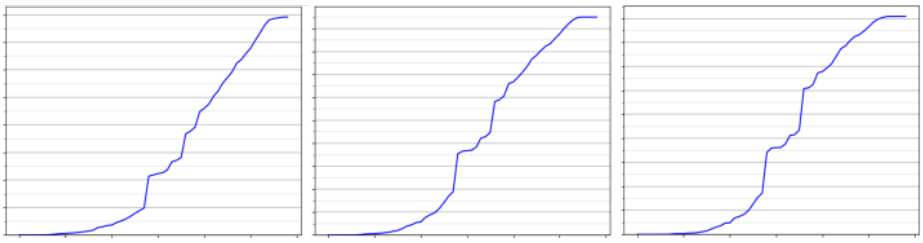

Figure 2.3: Grouwth of edges over time. For 1, 5 and 10 works in common graphs

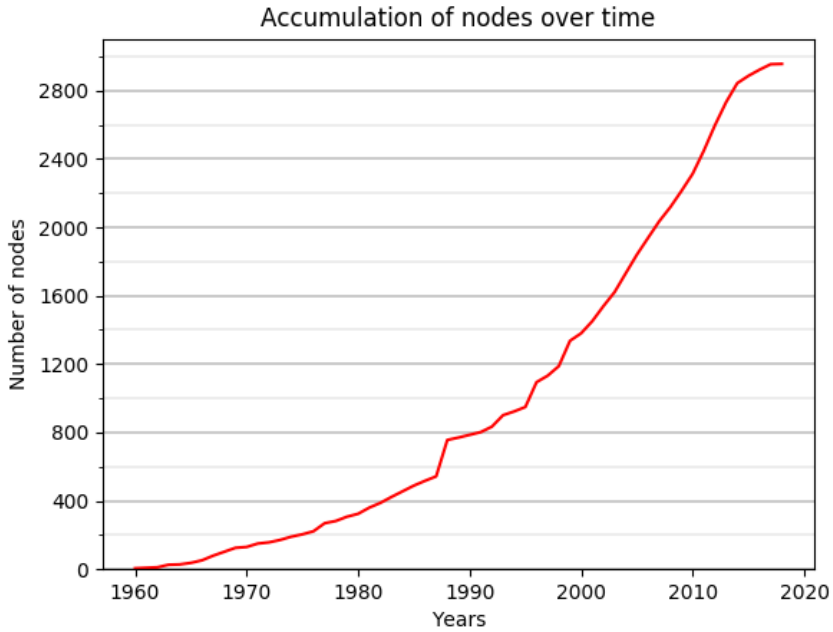

Figure 2.4: Grouwth of nodes over time.

Chapter 3

Analysis and prediction of seiyuu popularity

Popularity is an abstract criterion that must be defined as a numerical metric in order to be used for analysis and prediction. Since we are using MAL database for seiyuu and anime information and it has a social component; seems logic to use member_favorites as a representation of popularity. We can also get popularity and score for works from opinions of the same set of users.

In terms of distribution *popularity* is highly unequal –as we can observe in Fig. 3.1– having a lot of seiyuu which are no member favourites and only a few who are favorite of more than 10000 members. It's good to keep in mind that users can favorite multiple seiyuu.

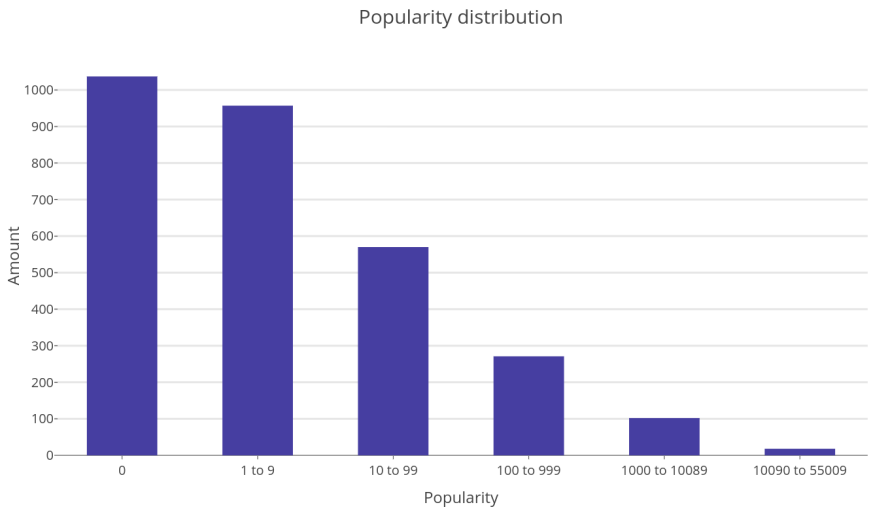

Figure 3.1: Amount of seiyuu with that popularity, divided into groups for better visualization.

EXPLICAR LO QUE SIGNIFICA QUE SEAN TAN DIFERENTES EN VALORES, MOSTRAR LOS DATOS DEL ITEMIZE This is something to

take into account when trying to predict popularity of actors or explain it using other features. TODO ADD WHY

- Mean: 289.55
- Median: 2.0
- Max: 55018
- Min: 0 (1037 values equal to zero)
- Only 120 values bigger than 1000

3.1 Correlation with only one feature

As shown in Fig. 3.2 our first approach to explaining popularity was using Pearson correlation. TODO PONER UN GRAFO DE PEARSON CON CASI TODOS LOS PARAMETROS POSIBLES

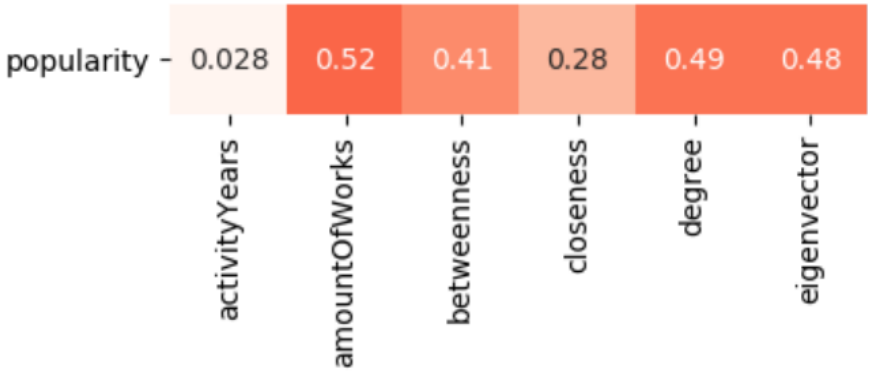

Figure 3.2: Pearson correlation between popularity and some features of nodes.

A fairly big correlation can be seen between popularity and amount of works. Since this data is biased to more modern anime we thought of trying to correlate with more recent works only. But, how recent? Last 5, 10 or 20 years? Thus correlation between popularity and works from different data frames was analyzed, Fig. 3.3.

The best result was given by recent works from last 9 years. Therefore, this definition of recent works was used from there on.

SUBSECTION PARA ESTO DE "TRATAR DE BUSCAR EXPLICACION DETRAS DE 'ULTIMOS 9 ES MEJOR'" Graphics of some characteristics of works divided by years were made, trying to shed some light over why works from last 9 years were more "important". Fig. **??** shows an improvement in average of scores and favorites, with the biggest peak in 2005. Popularity goes down but we need to consider that this metrics are from MAL and it could mean, for example, the parameter is in disuse, instead of representing how people feel about new anime.

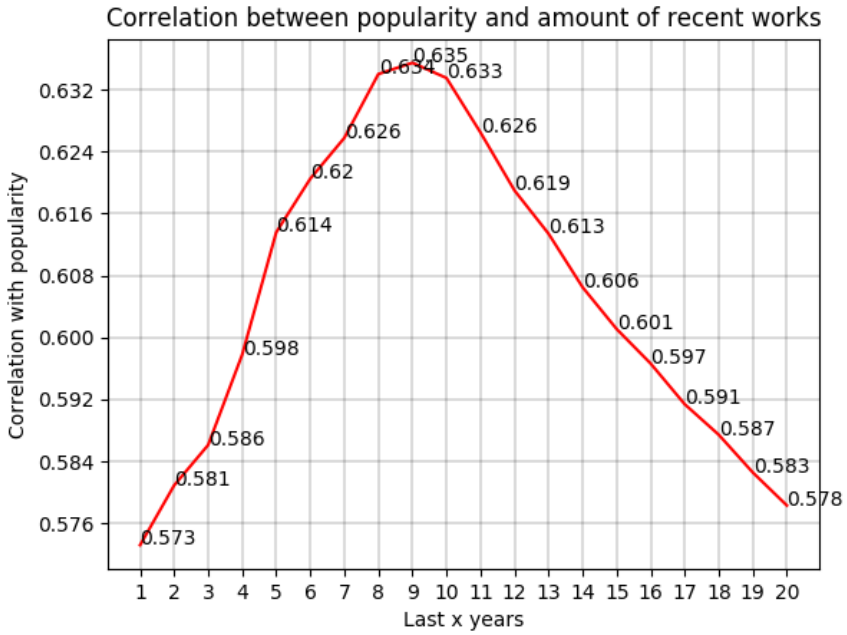

Figure 3.3: Last X years means works from 2018-X to present.

TODO ADD SCORE, MEMBERS, ETC GRAPHS AND EXPLANATION OF EACH OF THEM

As we can see on Fig. 3.5 anime industry is growing bigger each year, of course this is biased by the fact MAL will sure have every adaptation of last year but maybe not for anime from 1980.

The majority of works are from 1990 to 2018 and half of them are distributed over the last 14 years (2014 to 2018) but as far as we can tell there isnt anything particular over the last 9 years nor on year 2009. Judging by amount of favorites per year it appears that users have been more active in recent years so this could be one of the reasons.

Some interesting enough correlations are shown next TODO ADD SCATTER PLOTS AND EXPLAIN MORE

3.2 Correlation with multiple features

For this section Scikit-learn, a free software machine learning Python library, was used. The node attributes were divided into categories, leaving four distinct types:

- Personal data:
  - Debut
  - Gender

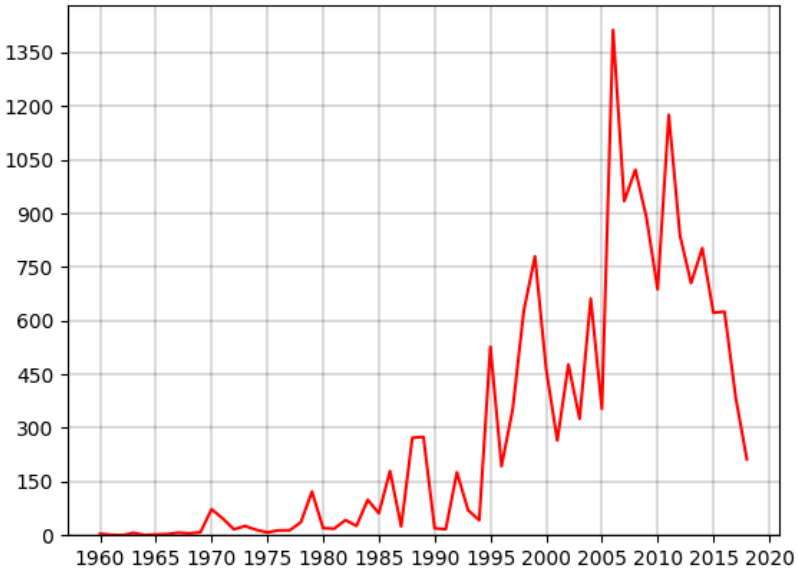

Figure 3.4: Average amount of favorites per year.

  - Activity years (2018-debut)

- Works data:
  - Amount
  - Top 5 genre
  - Favorites
  - Score
  - Popularity
- Recent works data:
  - Same as works but for only last 9 years
- Graph data:
  - Degree
  - Betweenness centrality
  - Closeness

Fitting and prediction experiments were run for each category, each combination of 2, 3 and all of them together; using 80% of seiyuu as train data and the rest as test. This was done for all following models:

- DecisionTreeRegressor

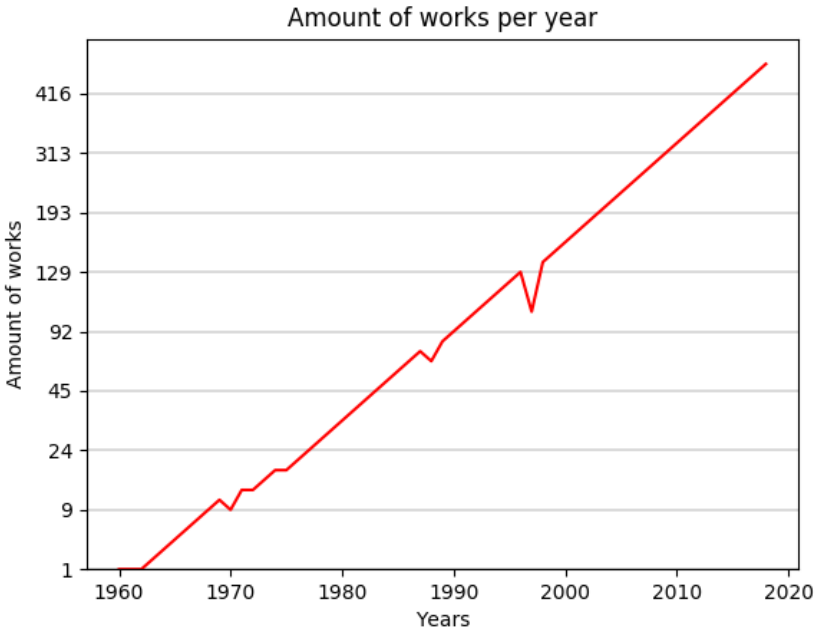

Figure 3.5: Amount of works divided by years which were aired for the first time.

- DecisionTreeClassifier
- LinearRegression
- KNeighborsClassifier
- LinearDiscriminantAnalysis
- GaussianNB
- SVM

TODO, WRITE THIS AGAIN: To compare prediction performance mean and median absolute error were used. Unfortunately since popularity variance is really high we observed good results in terms of absolute error but particular predictions were aloof. That's why we ended up using r2_score for accuracy comparation.

TODO SHOW TABLE WITH R2 SCORE RESULTS FOR EACH CATEGORY / MODEL AND GROUP OF CATEGORIES

TO ASK: SHOULD I DIVIDE INTO SECTIONS OF GROUPS OF CATEGORIES? yes

Table 3.1: Only one category R2 score results

Model / Data	Recent works	Personal	Graph	Work
DecisionTreeClassifier	0.11	-0.04	0.06	-1.3
DecisionTreeRegressor	0.33	-0.29	-0.05	0.11
GaussianNB	0	-4.7	-0.03	0
KNeighborsClassifier	0.22	-0.04	0.06	-0.05
LinearDiscriminantAnalysis	0.47	-0.04	0.32	-1.04
LinearRegression	0.57	0	0.28	0.42
SVM	-0.02	-0.04	-0.01	-0.02

3.2.1 Only one category

3.2.2 Groups of two categories

3.2.3 Groups of three categories

3.2.4 All categories

TODO ADD SOME OF THE GRAPHICS ABOUT FEATURE IMPORTANCE FOR DTC (ONLY BEST ONES) AND EXPLAIN (FOR EACH SUBSECTION)

ADD THE NEW CATEGORIZATION OF SEIYUU AND PREDICTION RESULTS (lo tengo que hacer primero)

Conclusion

Bibliography