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

Florencia Zanollo. Hideaki Takeda.

July 11, 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

Anime –like any other animation projects– have voice actors to play roles of each character. They usually have multiple seasons or adaptations based on original content –which can be manga, games, visual novels, etc– and it's not uncommon for a character to have always the same actor voicing it.

Since some time ago anime industry is growing bigger each year and so is seiyuu industry. Seiyuu can become very popular, with a great international fanbase and work in different areas other than voice acting, for example as singers, on theaters, etc.

1.1 Wikidata

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).

1.2 MyAnimeList

Wikidata information about seiyuu's works is really incomplete that's why 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).

[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/

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

1.3 Data retrieved

All in all we were able to retrieve the following information for 2956 seiyuu and 7614 anime.

- For Seiyuu:
  - Name
  - Debut (this was obtained from oldest work aired date)
  - Gender
  - Popularity (member_favorites information of MAL)
  - Work (anime roles with anime information plus wheter is a main role or not)
- 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).

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

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 definitions

This social network is of a particular kind called *two-mode networks* which consists of a set of actors (seiyuu) and events (anime). So there exists two ways of viewing it, one will be from seiyuu perspective, using anime in common for edges; the other being from anime perspective, using seiyuu in common for edges. We chose the former since we found more interesting they being actual people and using other information about them such as debut and gender.

So our social network consists of voice actors 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.

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 export the graph and open it on Gephi, for a more visual analysis.

And also we needed to build the edges dynamically because according to our definition they depend on the time frame we are looking at. For example, for at least 10 works in common, 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

In this section we are going to compare and analyze two definitions of relationship for our social network in order to understand more about it structure and decide on a definition:

- at least 1 work in common
- at least 10 works in common

Both of them during the time frame between the first debut registered (1960) and the year of observation.

Is easy to tell at first glance that this social network is really interconnected. With merely 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.

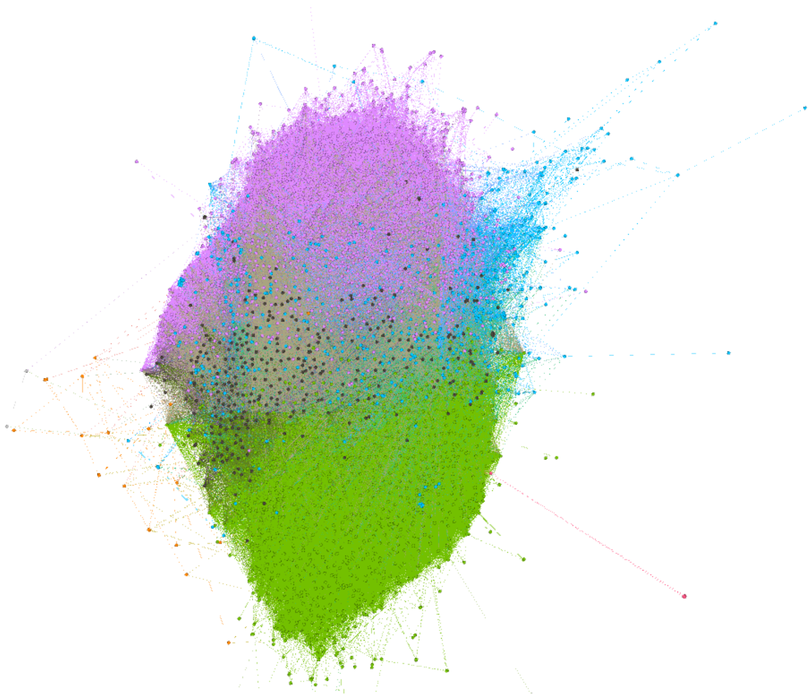


Figure 2.1: At least one work in common graph coloured by community.

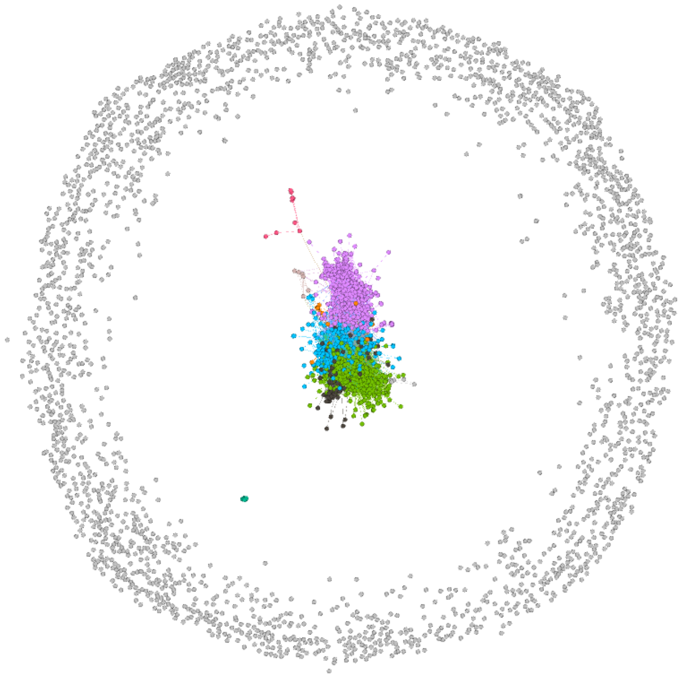


Figure 2.2: At least ten works in common graph coloured by community. Big cluster at the center, surrounded by loosely connected nodes.

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

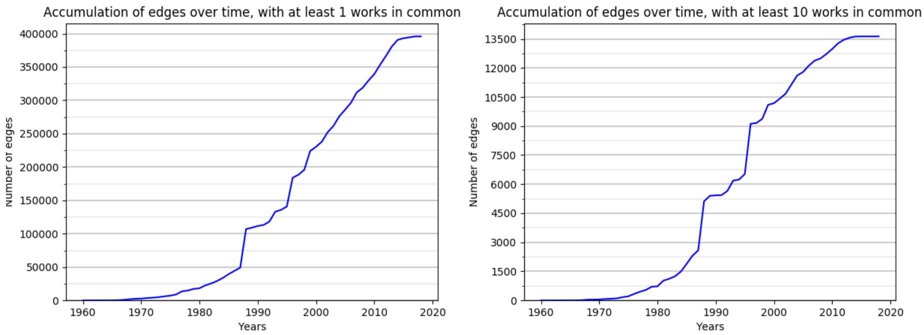
Table 2.1: Graph analysis

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

As proven by Fig. 2.3 grouwth of edges by year follows a similar 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.



(a) At least one work in common

(b) At least ten works in common

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

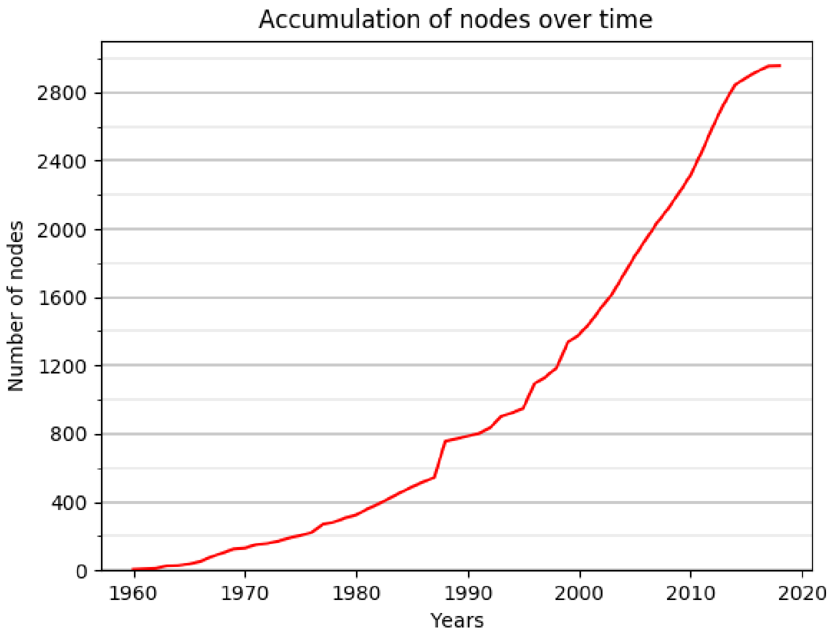


Figure 2.4: Grouwth of nodes over time.

Table 2.4 shows top 10 nodes, for degree and betweenness centrality for "at least 1 work in common" definition. And Table 2.7 does the same for "at least 10 works in common".

Name	Degree
Takehito Koyasu	1545
Akira Ishida	1488
Mamiko Noto	1422
Nobuo Tobita	1417
Daisuke Namikawa	1390
Nobuyuki Hiyama	1358
Rikiya Koyama	1331
Jrta Kosugi	1322
Keiji Fujiwara	1312
Kazuhiko Inoue	1309

Table 2.2: Top 10 degree

Name	Betweenness Centrality
Takehito Koyasu	49982.52
Akira Ishida	40221.50
Daisuke Namikawa	30448.43
Nobuo Tobita	29363.25
Mamiko Noto	29168.18
Rie Kugimiya	29122.31
Miyuki Sawashiro	28997.40
Kazuhiko Inoue	27693.88
Daisuke Ono	27034.592
Keiji Fujiwara	26802.69

Table 2.3: Top 10 Betweenness centrality

Table 2.4: At least one work in common

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 2.5: Top 10 degree

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	7413.72
Kenichi Ogata	7160.54
Harumi Sakurai	6775.76
Keiji Fujiwara	5980.58
Yoshimasa Hosoya	5607.95

Table 2.6: Top 10 Betweenness centrality

Table 2.7: At least ten works in common

2.4 Conclusion

Both networks have fairly similar top 10s so it points to them having similar structure and connections amoung their nodes, aside from actual values.

From now on our definition for edges will be: *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 so much.

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.

Table 2.8 shows a little more information about seiyuu that appear on top 10s.

Table 2.8: More information about seiyuu

Name	Popularity	Debut	Gender	Birthyear
Takehito Koyasu	7235	1988	Male	1967
Akira Ishida	7612	1989	Male	1967
Mamiko Noto	7544	1988	Female	1980
Daisuke Namikawa	8304	1988	Male	1976
Katsuyuki Konishi	3702	1996	Male	1973
Keiji Fujiwara	2778	1986	Male	1964
Junichi Suwabe	10838	1996	Male	1972
Toshiyuki Morikawa	2455	1981	Male	1967
Rie Kugimiya	31668	1996	Female	1979
Jun Fukuyama	26811	1981	Male	1978
Kenichi Ogata	52	1974	Male	1942
Harumi Sakurai	341	2005	Female	1982
Yoshimasa Hosoya	4852	2006	Male	1982
Nobuo Tobita	139	1981	Male	1959
Nobuyuki Hiyama	1723	1988	Male	1967
Rikiya Koyama	2919	1996	Male	1963
Jrta Kosugi	114	1985	Male	1957
Miyuki Sawashiro	26501	1988	Female	1985
Kazuhiko Inoue	2445	1974	Male	1954
Daisuke Ono	24080	1996	Male	1978

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 and it has a social component, seems logic to use member_favorites as a representation of popularity. We can also get popularity and score of anime 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 favorites 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.

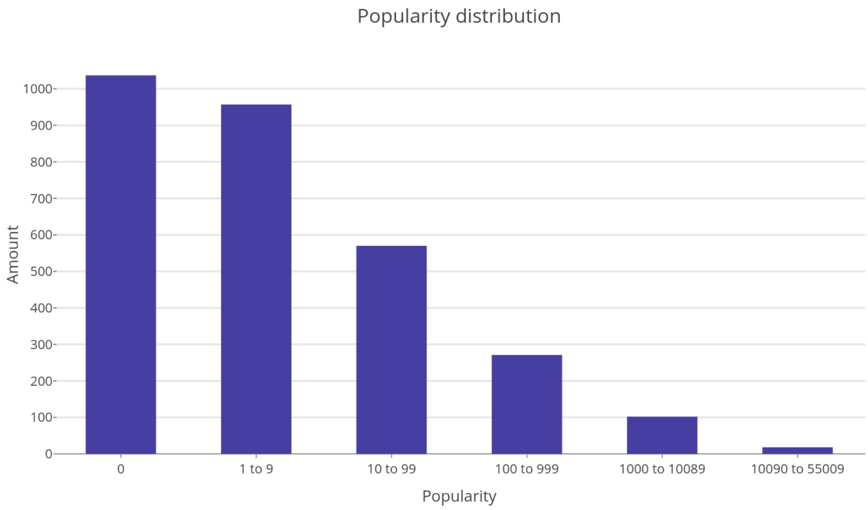


Figure 3.1: Amount of seiyuu with that popularity, divided into groups for better visualization.

Some metrics about popularity:

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

Table 3.1: Top 10 popular seiyuu

Name	Popularity
Kana Hanazawa	56637
Hiroshi Kamiya	49685
Mamoru Miyano	43942
Rie Kugimiya	31668
Jun Fukuyama	26811
Miyuki Sawashiro	26501
Tomokazu Sugita	24449
Daisuke Ono	24080
Saori Hayami	18322
Aya Hirano	18094

3.1 Correlation with only one feature

Our first approach to explaining popularity was using Pearson correlation.

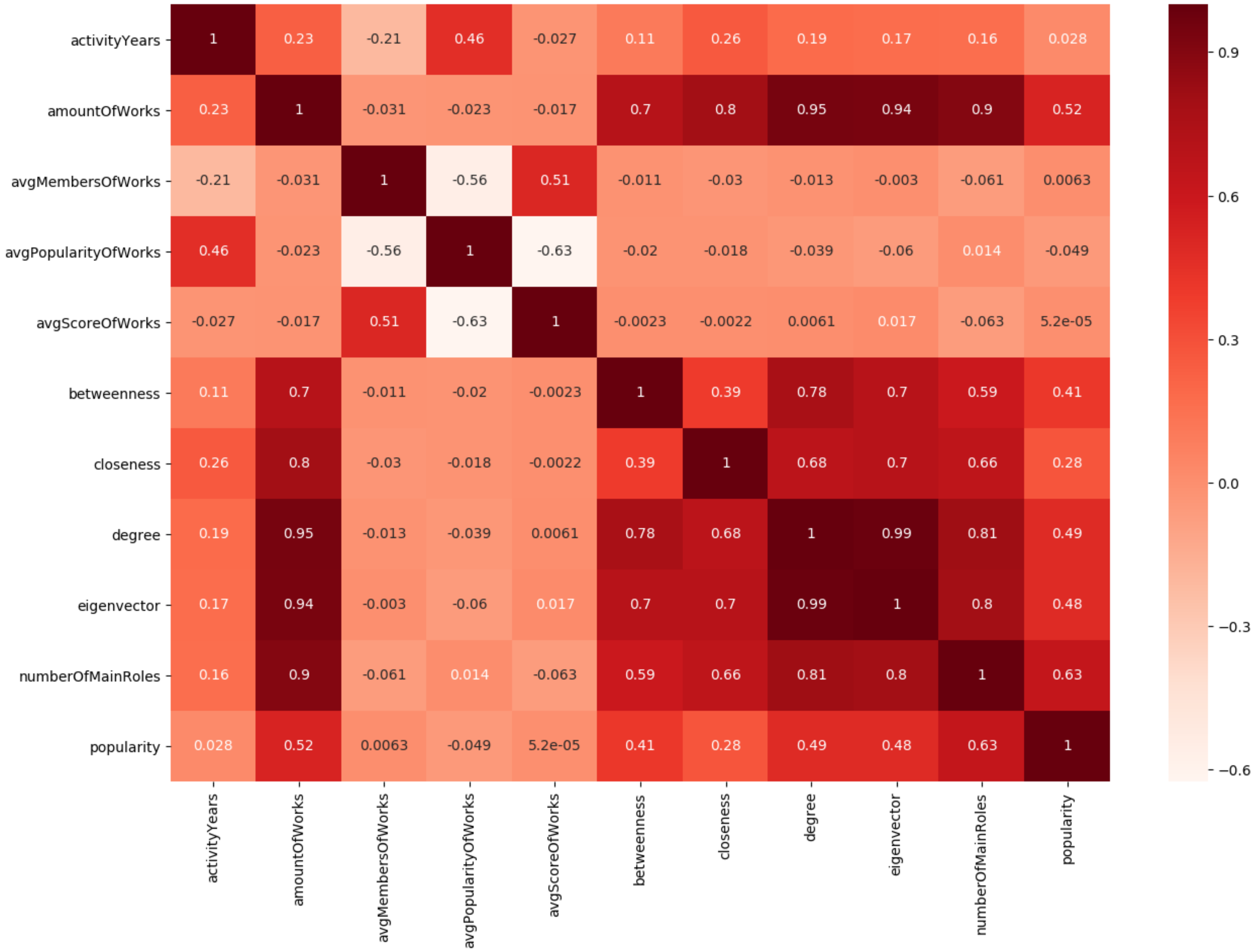


Figure 3.2: Pearson correlation between popularity and attribute of nodes (using all works).

As shown on Fig. 3.2 a fairly big correlation can be seen between popularity and amount of works. This attribute doesn't have the biggest correlation with popularity but "number of main roles" was added to the end of this investigation since we didn't had the data for doing so before. We are showing only average of values for work's attributes (ex. favorites) mostly for better visualization but for predictions we use sum, mean, median and maximum.

Since our dataset is biased in favor of more modern anime we thought of 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.

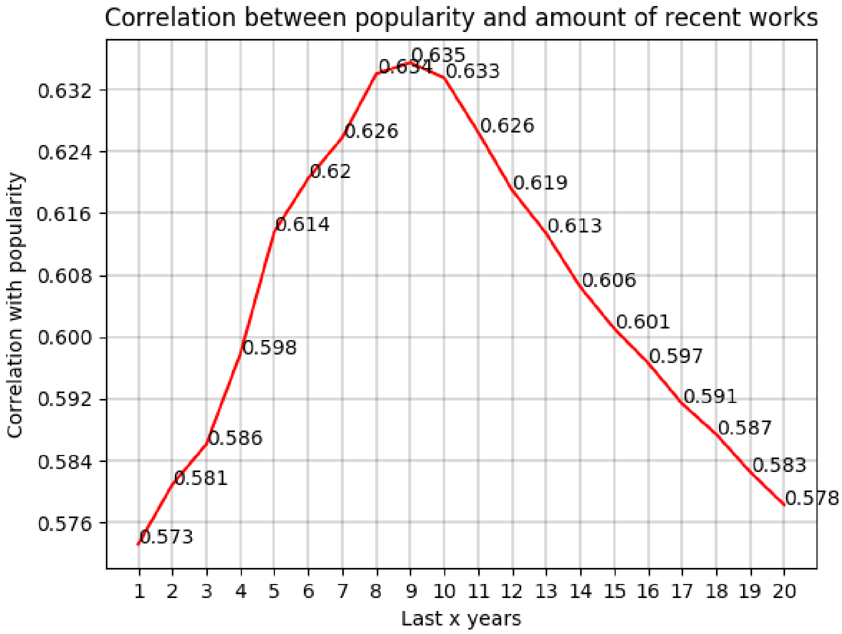


Figure 3.3: Last X years means works from 2018-X to present.

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

Fig. 3.4 shows the result of running Pearson again but using information from last 9 years of works only. It's important to clarify that we didn't build the network using only last 9 years, so betweenness centrality and degree are exactly the same as before. We left "amount of works" attribute for easy comparision against "amount of recent works".

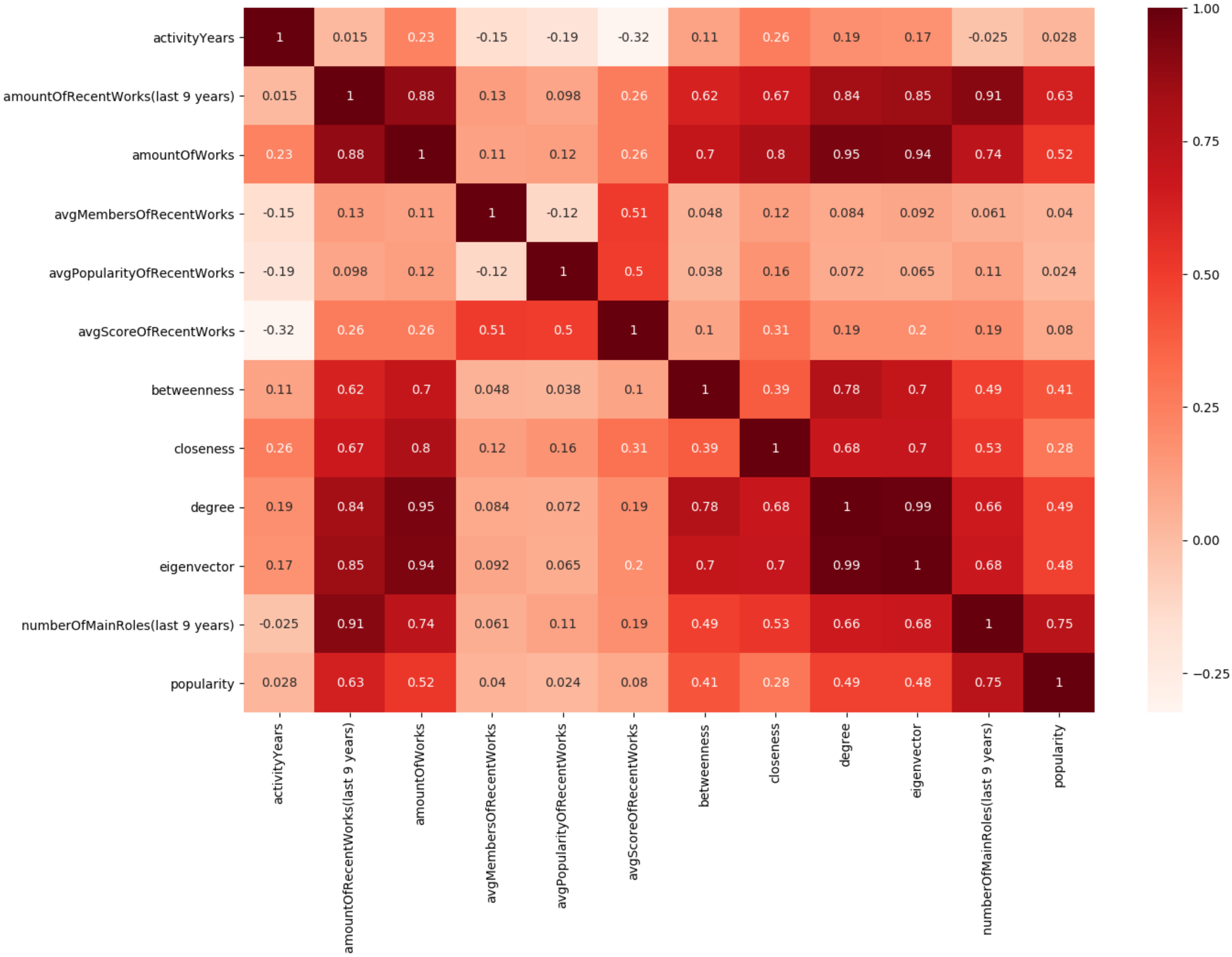
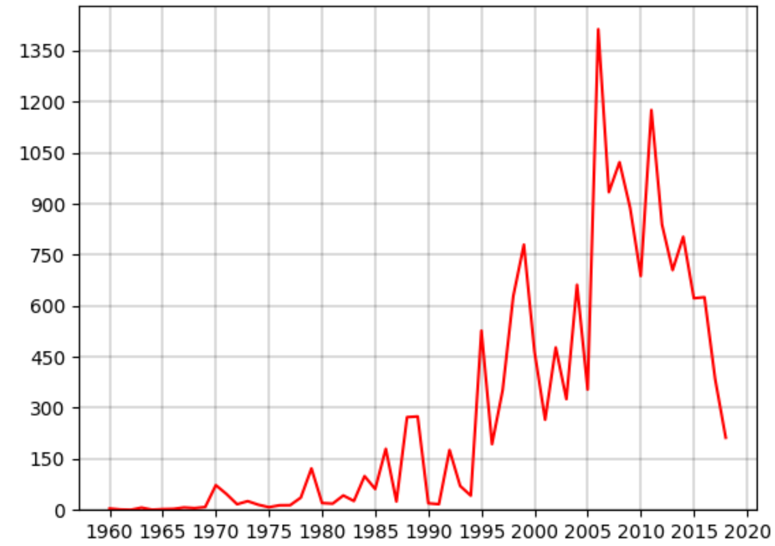


Figure 3.4: Pearson correlation between popularity and attribute of nodes (using recent works only).

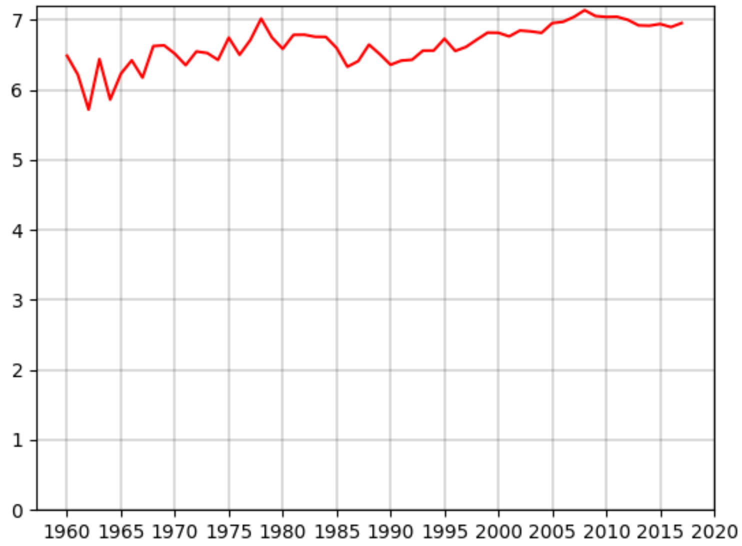
3.1.1 Why last 9 years of works has more correlation?

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".

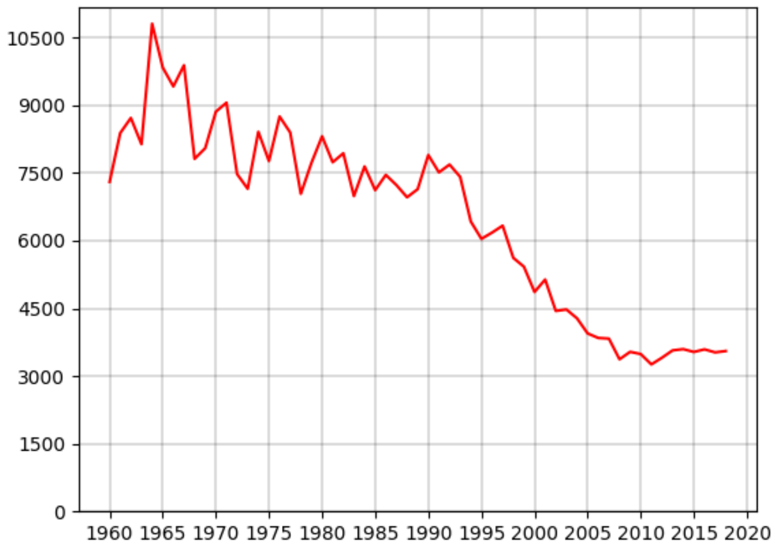
Fig. 3.5a and 3.5b shows an improvement in average of scores and favorites. Biggest peak of favorites is on 2005, this may have to do with the start of MAL (2006), see Fig. 3.5a. We suppose as users started to use MAL they favorite



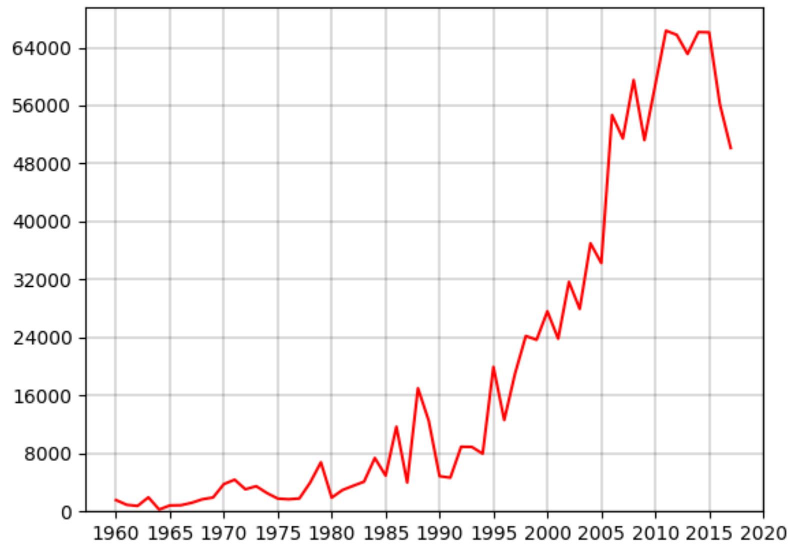
(a) Average amount of favorites per year.



(b) Average score per year.



(c) Average popularity per year.



(d) Average amount of members per year.

Figure 3.5: Averages of some atributes of anime divide by year.

anime they liked from that year and only some of the old ones; from then on they used the website often and favorite new works as they began airing.

Year 2018 was left out of Fig. 3.5b because, since 2018 is not finished yet, the average of scores was unusually small.

As Fig. 3.5c shows, popularity goes down in last years. 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.

We suppose the attribute ”members” of an anime is taken from how many users have it on any of their lists. If that is correct Fig. 3.5d tells us feature of adding anime in watched / watching / plan to watch lists is really used. As for our experience on the web and social media we can confirm this is the most used feature of MAL. So this makes it one of the best metrics to measure ”public” as another sense of popularity of anime.

As we can see on Fig. 3.6 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.

Figure 3.6: Amount of works divided by years which were aired for the first time.

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.

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
  - Activity years (2018-debut)
- Works data:
  - Amount
  - Top 5 genre
  - Favorites
  - Score
  - Popularity
  - Members
  - Number of main roles
- 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
- DecisionTreeClassifier
- LinearRegression
- KNeighborsClassifier
- LinearDiscriminantAnalysis
- GaussianNB
- SVM

We have to take into account popularity variance when trying to predict it since usual error metrics don't distinguish between small or big values. If we predict exactly a seiyuu that has >50000 popularity but we make "small" mistakes predicting seiyuu with <100 popularity then, is it a good prediction? and having into account more than $\frac{3}{4}$ of them have <100 popularity?

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. This is why we use r2_score[1] for accuracy comparation.

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

3.2.1 Only one category

Table 3.2: Only one category R2 score results

	Personal	Graph	Work	RecentWorks
DecisionTreeClassifier	-0.02	0.20	-0.35	-0.71
DecisionTreeRegressor	-0.02	-0.09	0.39	-1.05
GaussianNB	-1.01	-0.05	0.02	0.05
KNeighborsClassifier	-0.02	0.05	-0.14	0.24
LinearDiscriminantAnalysis	-0.02	-1.16	0.58	0.19
LinearRegression	-0.00	0.37	0.35	-0.03
SVM	-0.02	0.01	-0.02	-0.05

3.2.2 Groups of two categories

Table 3.3: Two categories R2 score results (R: recent works, P: personal, G: graph, W: work)

	P+G	P+W	P+R	G+W	G+R	W+R
DecisionTreeClassifier	-1.52	0.24	-1.10	0.20	0.50	-0.87
DecisionTreeRegressor	-0.71	0.19	0.33	-0.68	0.38	0.28
GaussianNB	-0.01	0.02	0.04	0.00	0.07	0.02
KNeighborsClassifier	0.05	0.09	0.30	0.24	0.36	0.37
LinearDiscriminantAnalysis	-0.01	0.35	0.53	0.20	0.54	0.48
LinearRegression	0.26	0.51	0.53	0.50	0.64	0.24
SVM	-0.02	-0.00	0.01	-0.02	0.02	-0.03

[1] `http://scikit-learn.org/stable/modules/model_evaluation.html#r2-score-the-coefficient-of-determination`

3.2.3 Groups of three categories

Table 3.4: Three categories R2 score results (R: recent works, P: personal, G: graph, W: work)

	P+G+W	P+G+R	P+W+R	G+W+R
DecisionTreeClassifier	-0.88	-1.79	0.41	-0.18
DecisionTreeRegressor	-0.78	-1.47	-0.42	-0.06
GaussianNB	-0.01	0.02	0.09	-0.01
KNeighborsClassifier	0.07	0.22	0.41	0.10
LinearDiscriminantAnalysis	0.73	0.65	0.53	0.59
LinearRegression	0.57	0.60	0.30	0.49
SVM	-0.02	-0.03	0.05	-0.02

3.2.4 All categories

Table 3.5: Only one category R2 score results

	AllFeatures
DecisionTreeClassifier	-1.51
DecisionTreeRegressor	0.04
GaussianNB	0.04
KNeighborsClassifier	0.40
LinearDiscriminantAnalysis	0.53
LinearRegression	0.53
SVM	0.01

3.3 Conclusion

Conclusion

Bibliography