

Anime clustering for automatic classification and configuration of demographics

Clusterización de animes para clasificación automática y configuración de demografías

Clusterização de animes para classificação automática e configuração de demografias

Júlio César Valente Ferreira, Universidade Federal Fluminense, Niterói, Brasil e Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Rio de Janeiro, Brasil (jcvferreira@hotmail.com)

Thiago Ribeiro Furtado, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Rio de Janeiro, Brasil (151thiagoribeiro@gmail.com)

Rafael Dirques David Regis, Universidade Federal do Estado do Rio de Janeiro (UNIRIO), Rio de Janeiro, Brasil (rafaeldregis@yahoo.com.br)

Gabriela Rodrigues Diniz, Universidade Estácio de Sá, Resende, Brasil (gabirdiniz@gmail.com)

Paula Gonçalves, Universidade Federal Fluminense, Niterói, Brasil, (paulagoncalves@id.uff.br)

Vitor Pedro da Silva Castelo Tavares, Universidade Federal Rural do Rio de Janeiro, Nova Iguaçu, Brasil (vitorpdsilva@gmail.com)

ABSTRACT | The cultural industry assumed greater relevance as a productive system and expanded its market share with different forms of reception, transmission, and communication with the public, increasingly using the so-called classification and recommendation algorithms and manipulation of mass processed data, which do not require cyber-physical systems for cataloging and constant feedback from all parties involved for cataloging. In this regard, this paper proposes a methodology to support the classification and creation of corresponding groups, automatically, of cultural productions of certain segments through Robot Process Automation (RPA) techniques, to first extract public data created by fans of certain cultural segments, and Latent Dirichlet Allocation (LDA), for the clustering of these productions based on the data of the terms extracted by RPA. As a case study for this proposal,

FORMA DE CITAR

Ferreira, J. C. V., Furtado, T. R., Regis, R. D. D., Diniz, G. R., Gonçalves, P. & Tavares, V. P. da S. C. (2023). Anime clustering for automatic classification and configuration of demographics. *Cuadernos.info*, (54), 67-94. <https://doi.org/10.7764/cdi.54.53193>

we specifically observed the anime market, defined as an originally Japanese cultural product with high fan engagement and high annual production scale, supported by data obtained from two public databases data: *MyAnimeList* and *AniDB*, built collaboratively by fans. The application of the methodology allowed the automatic classification of anime, grouping them into topics that allow the proposal of a new demography of products of this genre in relation to the current one, providing a greater level of detail and allowing to contemplate the expansion of new themes.

KEYWORDS: Latent Dirichlet Allocation; Robot Process Automation; anime; topic modeling; clustering; methodology; culture industry; Japanese pop culture.

RESUMEN | *La industria cultural asumió mayor relevancia como sistema productivo y amplió su cuota de mercado con distintas formas de recepción, transmisión y comunicación con el público, con un creciente uso de los llamados algoritmos de clasificación, recomendación y manipulación de datos procesados en masa, que no requieren sistemas ciber-físicos para la catalogación ni una retroalimentación constante de todas las partes involucradas. En este sentido, este trabajo propone una metodología para apoyar la clasificación y creación de grupos correspondientes de forma automática de obras y producciones culturales de determinados segmentos mediante técnicas de Robot Process Automation (RPA) para extraer, primeramente, datos públicos creados por fans de determinados segmentos culturales, y Latent Dirichlet Allocation (LDA), para la agrupación de estos trabajos a partir de los datos de los términos extraídos por RPA. Como caso de estudio para se observó específicamente el mercado de animes, definido como un producto cultural originariamente japonés con un fuerte compromiso y apoyo de los fans y alta escala de producción anual, sustentado en datos obtenidos de dos bases de datos de carácter público construidos en colaboración por fans: MyAnimeList y AniDB. La aplicación de la metodología propuesta permitió la clasificación automática de animes, agrupándolos en temas que permiten proponer una nueva demografía de obras de este género diferente de la actual, proporcionando un mayor nivel de detalle y permitiendo contemplar la expansión de temas nuevos.*

PALABRAS-CLAVES: Latent Dirichlet Allocation; Robot Process Automation; anime; modelaje de tópicos; clusterización; metodología; industria cultural; cultura pop japonesa.

RESUMO | *A indústria cultural adquiriu maior relevância como sistema produtivo e ampliou sua participação no mercado com diversas formas de recepção, transmissão e comunicação com o público, com utilização cada vez maior dos chamados algoritmos de classificação, recomendação e manipulação de dados processados em massa, os quais prescindem de sistemas ciber-físicos para catalogação e retroalimentação constante por todas as partes envolvidas. Neste sentido, o presente trabalho propõe uma metodologia de apoio à classificação e criação de grupos correspondentes, de*

forma automática, de obras e produções culturais de determinados segmentos através de técnicas de Robot Process Automation (RPA), para primeiramente extrair dados públicos criados por fãs de certos segmentos culturais, e de Latent Dirichlet Allocation (LDA), para a agrupação destas obras com base nos dados dos termos extraídos por RPA. Como caso de estudo, observou-se em específico o mercado de anime, sendo este definido como um produto cultural originalmente japonês com alto engajamento e apoio dos fãs e alta escala de produção anual, tendo como suporte de dados aqueles obtidos em duas bases de dados públicas, o MyAnimeList e o AniDB, construídos de forma colaborativa por fãs. A aplicação da metodologia proposta permitiu a classificação automática de animes, agrupando-os em tópicos que possibilitam a proposição de uma nova demografia de obras deste gênero diferente em relação à atual, provendo um nível de detalhamento maior e permitindo contemplar a expansão de temáticas novas.

PALAVRAS-CHAVE: Latent Dirichlet Allocation; Robot Process Automation; anime; modelagem de tópicos; clusterização; metodologia; indústria cultural; cultura pop japonesa.

INITIAL CONSIDERATIONS

Digital communication has become, in an increasingly decisive way, a privileged locus of mediation with the public in different areas; whether in terms of the processes of access, reception, consumption, and appropriation, for example, of cultural products. In line with Scolari's (2018) proposition, it can be said that a new media ecosystem is being configured, based on the intensive use of technology.

A pause is needed at this point to establish better-defined boundary conditions on the dimensions of technology, without, however, sticking to its concept, whose discussion is complex (Mitcham, 1994). Thus, systematizing the contribution of the aforementioned authors, technology can be approached through four dimensions: (i) objects/artifacts, (ii) mode of knowledge, (iii) specific form of activity, and (iv) a certain human attitude towards reality.

However, Hui (2016) draws attention to the excessive prominence given to the object/artifact dimension of technology, neglecting the others, not contemplating the social and cultural perspectives. Thus, the concept of convergence disappears due to a biased view of technology, not considering the other dimensions, which we will see below as essential for a comprehensive methodology of digital communication phenomena, as it will point to the scope of appropriation of cultural products by consumers and the potential of this appropriation for the establishment of solid flows between the nodes of this ecosystem.

This new ecosystem demands a radical change from cultural production in terms of interactivity and content fragmentation. In addition to the massive use of technology, the planning and development of these products go through challenges such as attention, adherence, approval, and audience volume mediated by the insertion of the public with more robust demand dynamics (Jenkins et al., 2013).

Therefore, this paper aims to establish a methodological proposal for the use of Robot Process Automation (RPA) for the data scraping of virtual public databases of cultural works, and the adoption of clustering techniques by Latent Dirichlet Allocation (LDA) for logical segmentation and consociation automatic use of these works to create more refined analytical demographic tables with reduced cognitive demand for analysis that can be performed by practically all members of the production chain of a cultural product.

As a case study, this methodological proposal was adopted for use in anime databases (Galbraith & Schodt, 2009) to verify its feasibility possibilities.

METHODICAL CONSIDERATIONS

To achieve the objective set out in this paper, it is necessary to have access to databases that allow the collection and subsequent work of clustering.

Thus, from the research on anime, we investigated databases on this niche, as well as the level of structuring of their presentation. This information is essential for analyzing the feasibility of using RPA tools for data scraping, being a basic criterion for the choice of the bases to be adopted. To be able to analyze a large volume of textual data, it is necessary to adopt an automatic method for the formation of typologies composed by groupings of keywords (Blei et al., 2003). Therefore, it was possible to cluster anime for the establishment of possible correlations of subgenres, allowing, then, the proposition of an anime demography that can contemplate the perspectives of the members of this production chain, as producers and consumer fans, according to availability and quality of the databases.

After automatic typing, we conducted an analysis of its results, attesting or not to the feasibility of proposing a certain demography. This phase of the research proved to be important for the verification and validation that the proposed typologies had adherence from the opposition of their keywords with the anime included in this demography.

Finally, the inclusion of a more in-depth description of each topic provided as a final result the proposal of a new demographic framework for the classification of anime, which contains, for each demography: base title, robust description of its characteristics, and examples of construction.

ANIME

Anime is a type of animation created in Japan from the 1960s and whose striking aesthetic feature is that developed by artists such as Osamu Tezuka (Condry, 2013) in the first half of the 20th century. The growth of its export and consumption by Western audiences occur from the 1990s, and it is consolidated as a cultural product of global consumption with a locus of Japanese identity (Iwabuchi, 2002). Anime is characteristic of contemporary media in its interconnected networks of commercial and cultural activities that cross industries and national borders (Condry, 2013).

Anime demographics

Japanese anime is generally produced from a pattern historically established by publishers in a roll of five target audience demographics (Bryce & Davis, 2010; Galbraith & Schodt, 2009; Katsuno & Maret, 2004), these being: (i) *Kodomuke*, (ii) *Shoujo*, (iii) *Shounen*, (iv) *Josei*, and (v) *Seinen*. The concept of demographics observed by the aforementioned authors assumes that readers are able to choose mutually exclusive products among the myriad of manga genres. However, it should be noted that this typology is a research construct, which does not consider the specificities of the products, which cannot necessarily be framed into a single

demographic, according to the definitions described below, compiled from the works mentioned earlier in this section.

- i. *Kodomomuke* (子供向け): anime for children under 12, mostly starring stories about family, friends, or animals considered lovely (also understood as cute), usually with a comic slant, such as Hidenori Kusaka and Satoshi Mato Yamamoto's *Pokémon Adventures*, or coming from classic children's stories (Butler, 2019).
- ii. *Shoujo* (少女): anime for female teenagers (Galbraith & Schodt, 2009; Butler, 2019), whose age range is variable and falls between seven to 20 years old (Galbraith & Schodt, 2009), or from 10 to 18 years old (Butler, 2019), spanning a variety of genres, including slice of life stories (everyday life), sports, and novels focused on affective tensions between girls and boys. This demographic also includes the *Mahou Shoujo* genre (or magical girls genre), with the anime *Bishojo Senshi Sailor Moon* as an example (Butler, 2019). The *Shoujo* demographic is also the locus of origin of the Japanese genre *Sentou Bishoujo*, which has become one of the most widely viewed, translated as beautiful girls fighters (Saitou & Azuma, 2011), with the anime *Bishojo Senshi Sailor Moon* being its most representative example (Galbraith & Schodt, 2009).
- iii. *Shounen* (少年): anime for male teens, which includes the action genre with comic elements. Examples include Akira Toriyama's *Dragon Ball*, and *Naruto*, by Masashi Kishimoto. This demographic also encompasses stories with adventures about sports, adventure, and fighting imbricated in the action genre (Butler, 2019). Anime in this demographic are aimed at a more adult audience, with storylines that carry elements of obscene humor and sexuality (without being explicit). They commonly feature narratives about becoming a stronger man (Galbraith & Schodt, 2009), following the narrative line of the hero's journey or purposefully subverting this narrative (Bryce & Davis, 2010). Galbraith and Schodt (2009) mark a trend of contemporary works with the perspective of dialogue with elements of *Shoujo* demographics.
- iv. *Josei* (女性): young adult and adult females anime (Bryce & Davis, 2010), with titles such as *Be Love* and *Office You* (Bryce & Davis, 2010). The *Shoujo* and *Josei* demographics, especially the latter, encompass works with plots focused on existential, emotional, and multidimensional sexual dimensions, overlapping in many cases with heteronormative experiences (Bryce & Davis, 2010). The interpersonal dynamics of being in a relationship, as well as the pursuit of romantic love, are portrayed not only from the point of view of female characters, but are also adopted from a male point of view in *Josei* works (Bryce & Davis, 2010).

- v. *Seinen* (青年): anime for young adults and male adults (Galbraith & Schodt, 2009) (Bryce & Davis, 2010), usually over the age of 15 (Butler, 2019). This genre operates with more dramatic plots and themes aimed at a more adult audience, about the challenges of life (Bryce & Davis, 2010) or considered dark, i.e., dealing with violence, psychological disturbances, and existential themes. For example, we can cite the anime *Elfen Lied* and *Psycho-Pass*, characterized by narratives aimed at an adult audience and by graphic representations of violence, sex and/or nudity (Galbraith & Schodt, 2009).

Another subgenre is anime consisting of harem stories, where a young man surrounds himself with attractive girls that elicit explicit, or not, sexual and romantic situations, such as *Love Hina* (Bryce & Davis, 2010). Finally, there are stories that mix the *Shoujo* and *Seinen* demographics (Bryce & Davis, 2010), usually presenting elements such as cute girls' anime that involve adult themes, often dark or with some elements in this vein, with works that prioritize *Seinen* elements over the *Shoujo*, such as *Puella Magi Madoka Magica*, and others doing the contrary (Butler, 2019). Recently, more *Seinen* anime productions with *Shoujo* elements can be identified, such as *Love Live*, *Girls und Panzers* and *Youjo Senki*, which indicating a demographic scaling of the *Sentou Bishoujo* subgenre, now representing consumers groups that significantly outnumber the original *Shoujo* demographic (Saitou & Azuma, 2011).

Finally, it is noteworthy that this demographic is based on a pattern established by Japanese publishers, where manga and anime consumption spans virtually all age groups (Bryce & Davis, 2010). In contrast, the Western static range is more restricted (Katsuno & Maret, 2004). This issue will be central to the analysis of the results obtained from the logical segmentation and automatic consociation of the works, since the origin of the adopted databases will be an important variable.

Contemporary anime distribution and consumption market

After the initial success of exporting anime in the 1990s through cable television (Iwabuchi, 2002), throughout the 2000s the increase in terms of access conditions and transfer rates on the Internet allowed various fans to share and make available online content of their favorite anime works, performing the tasks of translating, subtitling, and distributing on a voluntary and informal basis (Urbano & Araujo, 2021).

In the 2010s, streaming services such as Netflix and Crunchyroll (anime-focused) increasingly promoted Japanese television content beyond Japan, especially anime (Urbano & Araujo, 2021); for example, Crunchyroll was born in 2006 as an informally subtitled anime website and became a licensed anime streaming service, with over 800 licensed anime in 2017 (Urbano & Araujo, 2021).

Currently, most streaming services make anime available, with many investing in their own original productions, such as Netflix, Disney+, and Amazon Prime, with the value of the anime market doubling between 2009 and 2019, reaching US\$ 22.1 billion, which continued to receive new multimillion investments even during the pandemic (Brzeski, 2022). In 2020, for the first time in history, the international anime consumption market, with US\$10.89 billion in revenues, surpassed the Japanese market, with US\$10.41 billion, a historic milestone in the industry (Pineda, 2021), demonstrating anime's shift from a former domestic niche market to a global mass market.

Anime's databases

Regarding the databases related to the most popular anime, the portals *MyAnimeList* (<https://myanimelist.net/>) and *Anime Data Base*, AniDB (<https://anidb.net/>) were adopted for the search and compilation. Both are built collaboratively by fans and supporters, being mostly used by Westerners, as they are written in English. These fan-maintained portals describe and catalog various information about the works, such as demographics, genre, themes, tags, images, episodes, dubbing actors in different languages, duration, age group, among others, as well as textual descriptions of anime and characters. It was decided that the most relevant information would be short terms that summarize the work, allowing a better interpretation by the algorithm in the data analysis stage.

It is worth noting that the existence of such extensive fan-created databases demonstrates the strength of engagement of the anime fan audience, which is highly determined in the consumption of various narratives, both official and fan-created in various media. It is also identified that these anime works follow a database-driven development and creation pattern (Azuma, 2009). I.e., their development is based on small narratives formed by the collection of archetypes, stories, scenarios, and characters and their possible associations, with fans and consumers, because of their need to consume narratives, cataloging this imaginary database of narratives and characters and, consequently, making it digitally tangible (Azuma, 2009). Incidentally, this need to consume narratives induces consumers to them as well, thereby feeding back the ecosystem around the otaku industry (Azuma, 2009).

MyAnimeList is an anime and manga social network and, in parallel, a social cataloging application managed by volunteers, anime fans, which allows users to create lists to organize and rate anime and manga, and facilitate finding users who share similar tastes (Orsini, 2018), calling itself the largest anime and manga database and community in the world (<https://myanimelist.net/>). The *MyAnimeList* catalog has more than 145,000 characters from 20,000 anime cataloged, from

those with the highest engagement such as *Shingeki no Kyojin*, with 3,473,107 user followers, to older and lesser-known anime. All general information can be seen on the anime's homepage, as exemplified by the anime *Tonikaku Kawaii* in figure 1.

AniDB is a detailed, non-profit, non-advertising, user-maintained database aimed at cataloging detailed anime and characters information of works from China, South Korea, and Japan, focusing on generating the most detailed database of anime on the Internet (<https://anidb.net/>), having in its catalog, as of August 4, 2022, 13,866 anime, 238,524 episodes, 123,457 characters, 105,707 songs, and 4,476 anime and character tags cataloged (<https://anidb.net/>). Its comprehensive and detailed tag system caught our attention for allowing a brief and credible description of the anime, improving the performance of the LDA algorithm later on. The tag system involves several pieces of information, as seen in the distribution of information from the *Tonikaku Kawaii* anime, exemplified in figure 1.

The distribution of anime in these databases follows the long tail theory (Anderson, 2006), which states that the distribution of products in a market in the 21st century, especially in cultural and entertainment products, follows a Pareto distribution, in which the upper part of the curve concentrates mass products and a long tail of niche products that becomes increasingly extensive with the decrease in production and distribution costs and with new technologies, which link supply to demand, applicable also to the anime market.

As a cut-off for the segmentation of anime to be used, it was decided to include all those with more than 42,500 followers in the *MyAnimeList*, on August 4th, 2022, and that were serial anime, i.e., anime with serialized episodes for TV, Internet, and streaming (ONA, online network anime).

The adoption of this reference took place after the sequential analysis of the number of followers of each anime, starting from the most popular, and its positioning dynamics on the portal. When the anime reached the range of 42,000 followers, the positioning of these works varied with each query made on the *MyAnimeList* portal, changing dynamically. It was found that, in this range, in short periods of time, an anime was positioned in another position for having added a small number of followers, surpassing others, and therefore changing the extraction link in the table (not generating compliances for using the RPA-based tool), since it lists 50 anime per page. Thus, it was necessary to establish a stopping criterion to stabilize the classification for the extraction. The observation of the changes in the links in this classification allowed us to verify that the adoption of a minimum limit of 42,500 followers stabilized this dynamic, so this number was then adopted.

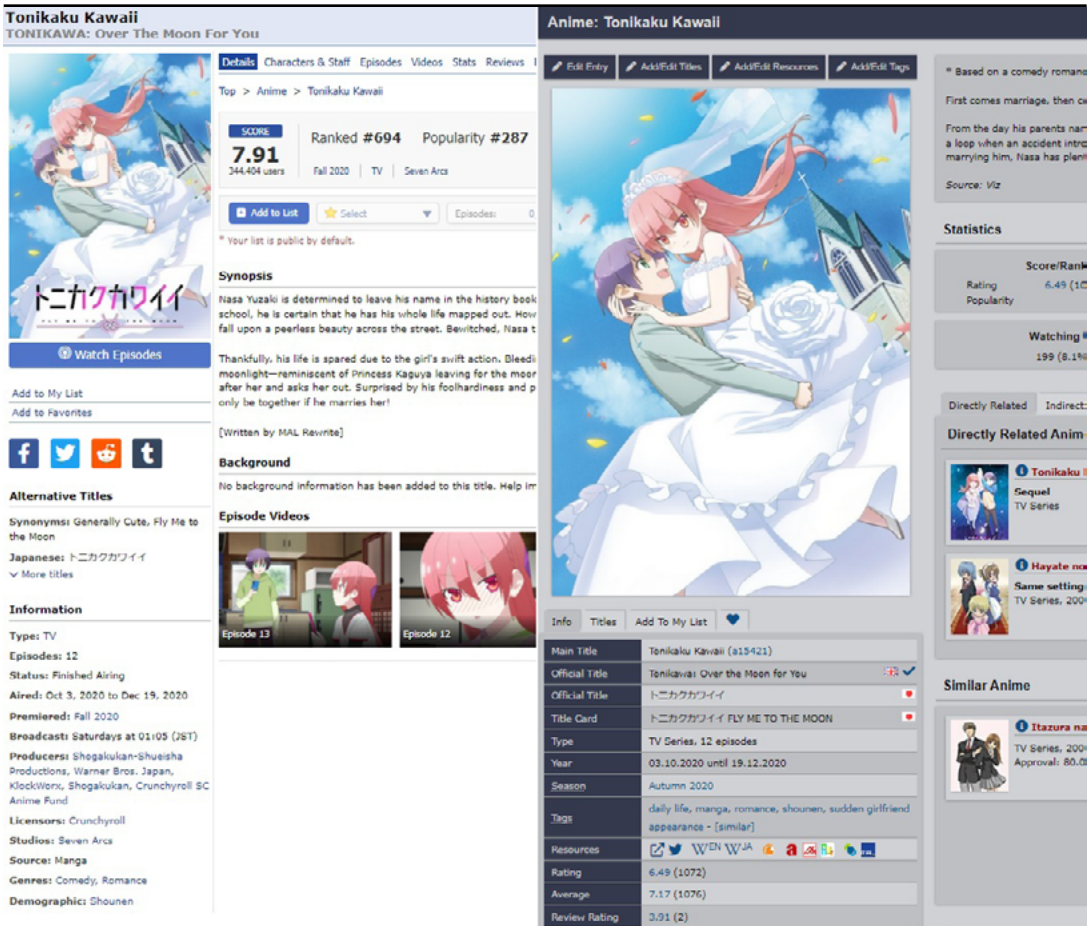


Figure 1. Tonikaku Kawaii anime page on MyAnimeList (left) and on AniDB (right)
 Source: Own elaboration based on the information on MyAnimeList (left) and AniDB (right), accessed on August 4th, 2022.

On the other hand, and corroborating the choices previously reported, when we plotted the numbers of anime followers (figure 2), the long tail phenomenon was geometrically verified, starting with the anime *Shingeki no Kyojin*, with more followers (3,473,107), which is an action fighting *Shounen* anime. When checking the position of anime with at least 42,500 followers, they were already in the niche market region (figure 2), according to Anderson (2006). For example, the last work considered in the analysis was *Over Drive*, a cycling anime with 42,593 followers, which ranked 1,887 in a universe of the MyAnimeList portal, with more than 20,000 anime series cataloged.

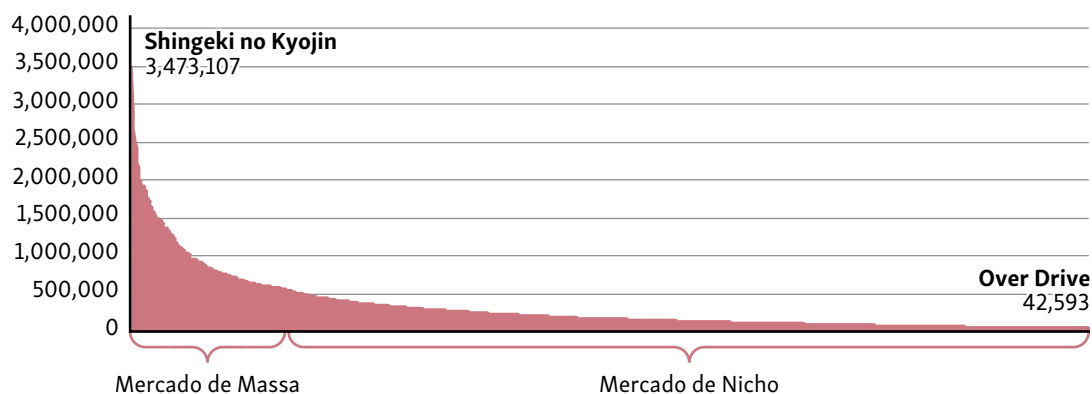


Figure 2. Distribution of the 1,887 anime series on MyAnimeList, by followers

Source: Own elaboration.

METHODOLOGY RESULTS PRESENTATION

The results will be presented through an explanation about RPA and its application, followed by development of the application of the LDA algorithm methodology and, at the end, the analysis and interpretation of LDA's topic list results and discussion.

RPA application

Once the two databases were selected, we chose the types of data to be used in the study. First, from *MyAnimeList*, we extracted: (i) genre (action, drama, romance, comedy, adventure, etc.); (ii) theme (*mecha*, military, school, daily life, fighting, fantasy, psychological, etc.), and (iii) demographics (*Kodomo*, *Shoujo*, *Shounen*, *Josei*, and *Seinen*). On the other hand and thanks to its structure, *AniDB* allowed to obtain the anime tags, which would be more specific information about the plot, such as violence, harem, school romance, black humor, girlfriend suddenly appearing, *otaku* culture, love triangle, bounty hunter, shooting, post-apocalyptic, university, among others.

The manual collection and processing data from 1,887 anime is a daunting task that would be time consuming, with a high probability of mistakes, if conducted by persons. Briefly, it would be a tedious, repetitive, low-quality job and high-volume job, so the use of automated computational proves to be a viable alternative to deal with this robust database processing, namely the RPA method (Tripathi, 2018).

Therefore, for massive data mining, we chose the RPA method, which represents the technique of construction and use of an automated digital artifact that resorts to the user's interface on the computer to perform activities that could be derived manually, indicated for the completion of simple tasks that require operation time, such as massive processing (Van der Aalst et al., 2018; Tripathi, 2018). A tool based on the RPA technique is endowed with functionalities that enable interaction

with information systems once operated by humans, allowing their replacement, in addition to enabling the aggregation of other structuring elements of cyber-physical systems, such as artificial intelligence and machine learning (Van der Aalst et al., 2018).

The extraction process was based on the logic of automatically going from the browser with the linked portal to go to each anime page in *MyAnimeList* and *AniDB*, and perform data scraping of the designated public information. By data scraping we mean the process of systematically extracting and combining relevant content from the Web using software (Glez-Peña et al., 2014). Data scraping is one of the most relevant ways of investigating social research in virtual environments, and today there are several software programs capable of automatically extracting large amounts of data available online, enabling new means to collect, analyze, and visualize data, which can be processed from other structuring elements of cyber physical systems such as Big Data Analytics (Marres & Weltevrede, 2013). However, there are challenges in this field, such as the use of biased databases, demanding attention to the data quality (Marres & Weltevrede, 2013).

In this case, the software used for the development of the process and the local RPA robot was the software UiPath Studio, considered one of the best RPA platforms available in the market, especially with Google Chrome (Tripathi, 2018; Mullakara & Asokan, 2020). Thus, two automated processes were mapped and developed, which would be executed in the following sequence: (i) extraction of the table of the most popular anime, and (ii) extraction of data from each anime.

The process (i) was conducted from the main page of *MyAnimeList*, which allowed us to make tables of anime by number of followers, indicating their name and the number of followers, while process (ii) used the names extracted from the previous process to open the *MyAnimeList* page for each anime. From there, genre, theme, and demographic data were extracted (Marres & Weltevrede, 2013). After extracting the anime page data in *MyAnimeList*, the RPA software continued the process by (ii) clicking on the anime's URL in the *AniDB* portal (which was already open) to extract the respective anime's tag line in the general work's data chart.

With the extracted data, the robot performed text processing to remove data irrelevant to this research, such as dubbers, advertisements, and HTML programming text lines. The robot ran for a total of 16 hours, with some intervals for bug fixes that interrupted the operation, such as those in which the portal site crashed, where it was necessary to insert page refresh redundancies and browser reset into the robot's operation logic. Data from all 1,887 selected anime were extracted, processed, and saved in an Excel spreadsheet format.

LDA algorithm application for clustering

Considering the selected base of 1,887 anime, manually separating these into groups and determining the characteristics that make them belong to the chosen group is not a trivial task. For this purpose, in this proposal, we used the Latent Dirichlet Allocation (LDA) topic modeling technique to cluster the anime by tags, themes, demographics, and genres. The programming for the clustering was supported by the Python language and the NLTK and Sklearn libraries.

LDA is a probabilistic model widely used to document, classify, and segment textual documents, as well as to perform recommendation tasks (Blei et al., 2003), consisting of finding groups of keywords called topics. LDA's main theory is that textual documents are made up of a mixture of these topics. Thus, finding these topics is a way of clustering textual documents. Finally, LDA assumes that textual documents are made up of a mixture of keywords, which in turn can be grouped into a topic. Therefore, according to LDA's logic, a textual document is composed of a distribution of such topics (Blei et al., 2003). The operation of LDA consists of creating a distribution of textual documents by topic, and a separate distribution of words by topic (Nijhawan et al., 2022).

The pre-processing applied to the data obtained in the previous step consisted of combining the columns for tags, theme, demographics, and gender into a single column, so that the terms were separated by commas, which was referred to as "complete".

Next, stopwords were removed, expression tokenization was performed, and analysis weights measured by applying the Term frequency-inverse document frequency (TF-IDF) via the `TfidfVectorizer` function found in the Sklearn library. Stopwords are non-noun connectives that do not provide data for this analysis. They are removed so as not to increase the dimensionality of the matrix generated after applying TF-IDF (Mhatre et al., 2017). Such an operation is provided by the NLTK library function `stopwords.english()`.

The tokenization process consists of separating sentences into words, which are called tokens. This operation is necessary for the simplification of inputs in algorithms that perform natural language processing tasks, thus increasing their efficiency (Mhatre et al., 2017). For this process, we adopted the `ngram_range` parameter (1,1), considering tokens of only one word.

The central thesis of TF-IDF in weighting token weights is that words that are repeated in many documents do not serve well to describe the clusters that are generated (Zhang et al., 2011). After applying TF-IDF, the partial result is a table where the rows are the documents, and the columns are the tokens. The values in the table are the numbers of times the word appears in the textual document multiplied by the TF-IDF weighting.

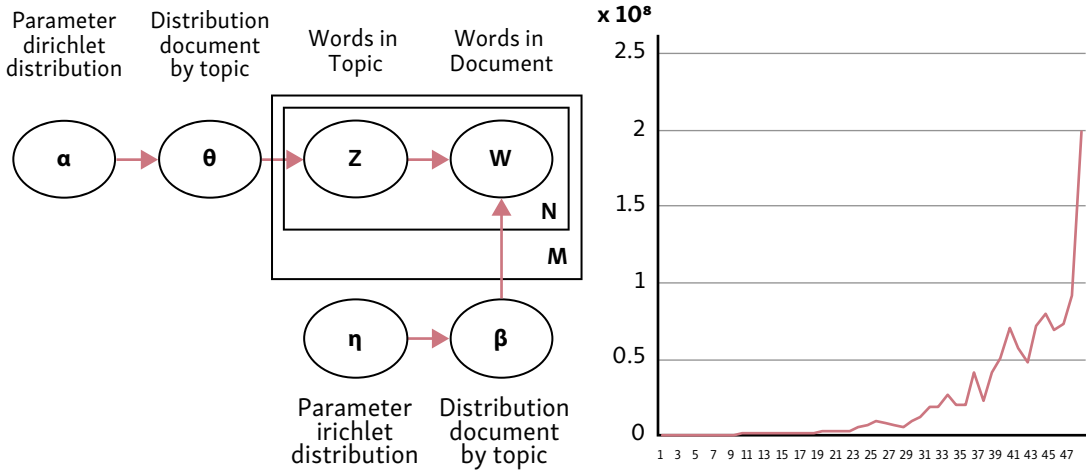


Figure 3. Visual Schematic of LDA's Operation (left); Variation of Perplexity by the number of topics k (right).

Source: Own elaboration, based on Nijhavan and colleagues (2022) (left); own elaboration, (right).

At the end of this process, what we have is a matrix where the rows are the anime and the columns are the words present in all the textual documents (after eliminating those in the stopwords list). This matrix is the input for the LDA.

Figure 3 shows schematically the principle of operation with LDA. Parameters α and η are parameters of Dirichlet distributions. Parameter α controls the distribution of documents per topic and parameter η , the distribution of words per topic (Nijhawan et al., 2022). These parameters can be modified to obtain topics with more or fewer keywords, and documents composed of more or fewer topics.

In the case study at hand, LDA was run with parameters α and η set to 0.00001. The goal was that topics could have few words and that a document should not be identified with many topics. To find the optimal value of topics, we ran the model with different numbers of topics and adopted the perplexity metric for the measurement, used to evaluate the quality of generated topics (Blei et al., 2003). The perplexity metric is used to measure how well a statistical model describes the original data (Zhao et al., 2015), as outlined in equation 1:

$$Perplexity = exp - \frac{\sum_{d=1}^M \log p(w_d)}{\sum_{d=1}^M N_d} \quad (1)$$

where M is the set of textual documents in the database, w is the words t contained in the documents and used (also called the corpus), and N is the textual document (Zhao et al., 2015).

Figure 3 shows, for this case study, the perplexity levels (ranging from zero to 2×10^8) of the LDA runs with k topics, ranging from 2 to 50. The lower the perplexity, the better the model quality (Blei et al., 2003). However, the model with the lowest perplexity is not always the one that generates topics best understood by humans (Zhao et al., 2015). Therefore, qualitative knowledge about the subject of study is determining the best modeling.

Figure 3 shows that, for the computational routine implemented in this case study, the range of number of topics varies between two and 25. Since anime are divided into five demographics, the strategy was to apply the proposed methodology with a larger number. Initially, the formation of eight and 10 topics was tested, and a number of 12 demographics was finally adopted.

It should be noticed that since the database is relatively small (1,887 anime), the use of a high number of topics tends to present a dispersion of these anime in topics with few examples and representativeness. In addition to this consideration, excessively repeated terms occur in more than one topic, with a high possibility of not responding to a demography whose differences between elements are easy to perceive, since it was intended to delineate topics that had a unique association of terms.

As already mentioned, in this stage, the qualitative analysis is important for the refinement of the final proposal coming from the application of digital artifacts. It is a refinement that necessarily passes through the sieve of the modeler. With few topics, they would be cluttered with terms that would have few connectivities between them. With many topics, there would be excessive repetition and, instead of clarifying, it would be a source of conformity failure for future analysis.

For example, by using 10 topics, the application of the proposed methodology suggested a demographic of idol singers only. The issue is that these anime target audiences with a high degree of distinction. When simulating with the number 12 (figure 4), the idol topic was removed and the anime targeting male audiences were allocated to the *Bishounen* demographic, while those targeting female audiences were placed in the cute girl anime topic. This increase in topics caused the LDA to consider cute girl or cute man as more representative than being an idol. Finally, in the adoption of 12 topics, two were with lower number of works in relation to the others (almost half of the immediately superior) and one with average adherence among the other works (this issue will be scrutinized in the next item of the text, thus requiring qualitative refinement by the authors of this proposal, being then the indication that we had reached a higher level of demographic intelligibility of our results (figure 4).

Another important issue in the process of configuring the topic items (figure 4) was the need to dismember terms with more than one word. Although examples words like Sci Fi, Slice of Live, and Fiction Science have semantic meaning when

written with this word composition, they were separated in the tokenization step, because, by adopting tokens with more than one word, incomprehensible terms were generated due to the random order of tag placement in the textual document.

Certain terms described in figure 4 are specific to the otaku reality and Japanese subculture (Azuma, 2009), and were indicated with a number in parenthesis and their explanation can be found in annex 1 for further understanding.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	fantasy	mystery	game	harem	life	fantasy
2	isekai (1)	horror	visual	romance	school	action
3	adventure	psychological	erotic	vampire	comedy	mythology
4	action	fantasy	novel	school	slice	shoujo
5	magic	action	romance	comedy	daily	proxy
6	novel	supernatural	seinen	time	manga	battles
7	world	violence	organized	reverse	romance	demon
8	comedy	seinen	crime	travel	iyashikei (4)	comedy
9	parallel	detective	harem	culture	koma (5)	contemporary
10	game	angst	school	otaku	seinen	supernatural
11	reincarnation	novel	mafia	fantasy	cgdct (6)	manga
12	rpg	suspense	maid	novel	workplace	girl
13	anthropomorphism	drama	tragedy	supernatural	high	magical
14	shounen	tragedy	drama	bishounen (3)	short	adventure
15	juujin (2)	gore	life	contemporary	episodes	mahou
	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	sports	ecchi (10)	fi	historical	power	school
2	mecha (7)	harem	sci	military	super	music
3	team	breasts	fiction	samurai	action	romance
4	action	comedy	science	retribution	fantasy	life
5	piloted	large	action	violent	martial	polygon
6	robot	school	mecha (7)	accidental	arts	love
7	military	fantasy	alien	infringement	adventure	comedy
8	school	romance	new	action	shounen	arts
9	shounen	pantsu (11)	space	warfare	manga	drama
10	new	nudity	military	feudal	swordplay	clubs
11	fiction	parody	apocalyptic	swordplay	supernatural	high
12	science	manga	post	medical	violence	manga (8)
13	manga (8)	action	law	meiji (12)	comedy	coming
14	manhua (9)	novel	order	period	tragedy	age
15	fi	seinen	human	bakumatsu (13)	contemporary	idol (14)

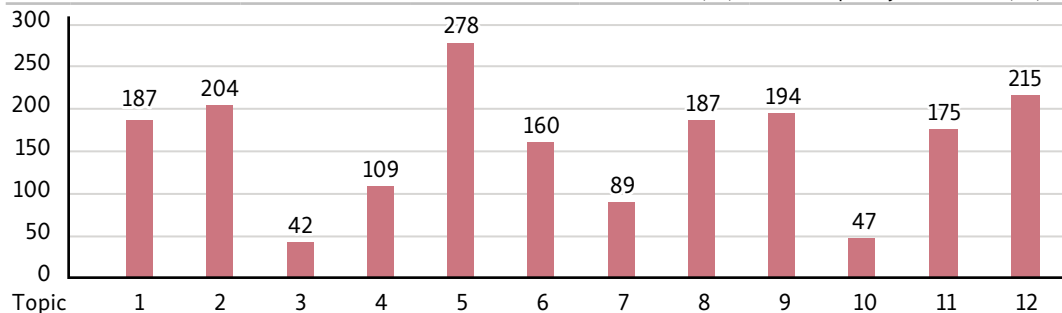


Figure 4. Distribution of words per topic and number of anime segmented per topic

Source: Own elaboration.

Overview and qualitative analysis of demographics via LDA clustering

With the number of topics defined and the automatic positioning of anime in each of these demographics, the next step in the methodological assertion was the purposeful consolidation of a new demographic for anime classification, supported by an automatic LDA clustered classification.

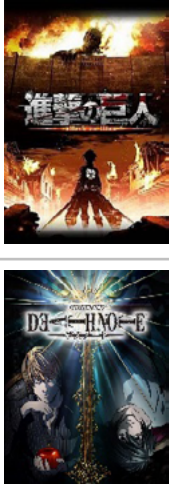
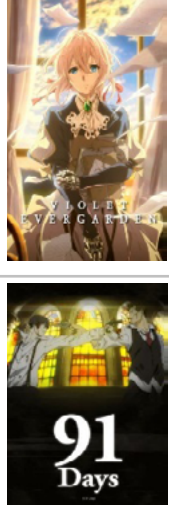
Initial qualitative analysis of the results reveals that the computational procedure configured 10 topics with a high degree of consistency. The exceptions were topics 3 and 4, with different thresholds. Thus, without any human intervention, 10 out of 12 demographic themes were demarcated with their boundary conditions stable and well established. As for the numerical distribution of cataloged anime by demographics, looking at figure 4, topics 3 and 10 presented the smallest numbers. Topic 3 was automatically configured as a demographic that would house anime not contemplated in other topics. In the case of topic 10, this occurs because of its tags, which are very specific to the historical theme domain, not appearing in other topics.

Concerning the topics with some degree of nonconformity, the analysis allows us to state that they were specific issues, not invalidating the proposal presented in this work.

Topic 3 was diffusely structured anime based on games and novels, aimed at adults audiences, which had not been contemplated in other demographics. This demography deals with anime with themes of dramas, crime action stories, and narratives inspired by electronic games, having common terms present in other topics, such as school, novel, life, and drama.

Topic 4 was characterized by two themes (reverse harem and time travel), which are not always related and dialogued in anime, thus generating a demographic with relative diffusivity. This topic resulted in the grouping of *Bishounen* romance anime, which often involves vampires, fantastic creatures, reverse harem (with boys surrounding a girl leading character), and school themes. However, the analysis revealed that there was an association between these romantic anime, aimed primarily at a mostly female audience, with time travel anime, as a significant portion of the latter type of work are recorded on websites with tags linked to romance narratives such as harem, game, and romance.

After these observations, as part of the proposed methodology, we conducted a qualitative interpretation and analysis of the terms and anime in each topic, to make a comprehensive textual description of each of the new LDA output themes, resulting in a new anime's segmentation indicated in figure 5. This reveals that qualitative human's interpretation is important to test whether the LDA output has resonance with reality and the respective cultural sector.

T	Name	Description	Example
1	Isekai magic fantasy anime	Magical fantasy adventure anime, with almost the entirety of Isekai, with anthropomorphic characters and RPG and game elements, in which fantastic elements such as elves, magic, dragons, kings, and princesses are the character's reality in the world, or in his new world.	 <ul style="list-style-type: none"> •Konosuba •Gate •Fairy Tail •Black Clover •Overlord •Fullmetal Alchemist
2	Psychological	Animes with a strong content of mystery and horror, where psychological terror and gore are often present, with fantastic and supernatural elements as the driving force of the story, with drama and tragedies, in which the suspense and anguish of what will happen later attract the audience attention, with action.	 <ul style="list-style-type: none"> •Shingeki no Kyojin •Death Note •Sword Art Online •Monster •Fate/Zero •Berserk
3	Dramatic romances and crime stories based on novels and games	Less coherent topic, it includes dramatic novels and action dramas, and adult comedies involving organized crime, with less or no element of fantasy. All based on novels and adult games. Created based on the terms Novel and Game.	 <ul style="list-style-type: none"> •Violet Evergarden •91 Days •Grisaia no Rakuen •Gokudolls •Romeo x Juliet •Loveless

4

Time travel, otaku, and bishounen reverse harem stories

Time travel dramas and novels, reverse harems (men who surround a female protagonist), and drama novels involving bishounen and/or vampires, and comedies about otaku themes. However, the theme of harem and romance is central in almost all works.

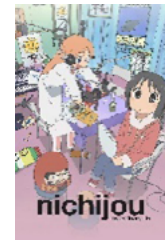


- Steins; Gate
- Kaichou wa Maid-sama!
- Tokyo Revengers
- Vampire Knight
- Himouto! Umaru-chan

5

Slice of Life

Anime about everyday life, about living an ordinary life, usually about work, school and family, in which even some are not slice of life, present strong aspects of this subgenre. Also include all iyashikei anime. All have comedy, and drama is rare.

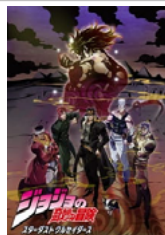


- Nichijou
- Yuru Camp
- Shokugeki no Souma
- Citrus
- Spy x Family
- Blend S

6

Magic Fight Animes and Sentou Bishoujo

Basically action and fighting anime, including all action shoujo (Sentou Bishoujo), and some shounen and seinen, with strong fantasy appeal in all works, involving the use of spells, powers, spells, and magical abilities, usually set in the current modern world.



- Sailor Moon S
- JoJo no Kimyou na Bouken
- Yu Gi Oh!
- Ao no Exorcist
- Soul Eater
- Sakura Cardcaptor

7	Sports anime and teamwork	Teamwork anime, mostly sports, where the protagonist must work with the group to win a sports tournament, as well as military-themed anime, usually of giant robots, where a group or platoon works together, presenting a lot of action.		<ul style="list-style-type: none"> •Haikyuu!! •Full Metal Panic! •Yuri!!! on Ice •Megalo Box •Re:Creators •Captain Tsubasa
8	Erotic anime	Anime with explicit or quasi-explicit erotic content, categorized as ecchi and/or harem. They also include works with drama and story development, but with lots of nudity and ecchi content. Comedy and school themes appear frequently.		<ul style="list-style-type: none"> •Monster Musume •No Game No Life •High School DxD •Ishuzoku Reviewers •Eromanga-sensei •Sankarea
9	Dramatic science fiction anime	Complex, philosophical, and questioning science fiction adventures, where, science and fantasy mix creating new worlds where characters have a high degree of individual development. Very recurring drama.		<ul style="list-style-type: none"> •Neon Genesis Evangelion •Cowboy Bebop •Plastic Memories •Made in Abyss •Darling in the FranXX •Psycho-Pass






<p>10</p> <p>Historical and scientific-themed anime</p>	<p>Anime with a historical or scientific theme based on reality, whether comedy, action, drama, or adventure, which usually try to teach something about the described subject, ranging from European history to Japan in the 90s. Scientific animes have a great diversity of concepts and topics such as health.</p>	 	<ul style="list-style-type: none"> •Vinland Saga •Hataraku Saibou •Dororo •Golden Kamuy •Kingdom •Last Exile
<p>11</p> <p>Shounen and seinen fighting anime</p>	<p>Shounen and seinen fighting anime based on super powers, in which fantasy is not mandatory. In these anime it is common to have the narrative of the hero's journey and characters getting stronger and stronger to fight stronger enemies, with the fights being the focus of the anime.</p>	 	<ul style="list-style-type: none"> •One Punch Man •Dragon Ball Z •One Piece •Akame ga Kill! •Naruto •Hunter x Hunter
<p>12</p> <p>Cute girl anime of romances and Seinen</p>	<p>Cute girl anime and idol anime in friendly romances with almost no ecchi or erotic aspect, with romantic and character development being as the focus, as well as cute girl seinen anime in the most diverse challenges, comedies, and dramas, with also by little or no erotic aspect.</p>	 	<ul style="list-style-type: none"> •Girls and Panzers •Love Live! Sunshine!! •Toradora! •K-On! •Horimiya •Tonikaku Kawaii

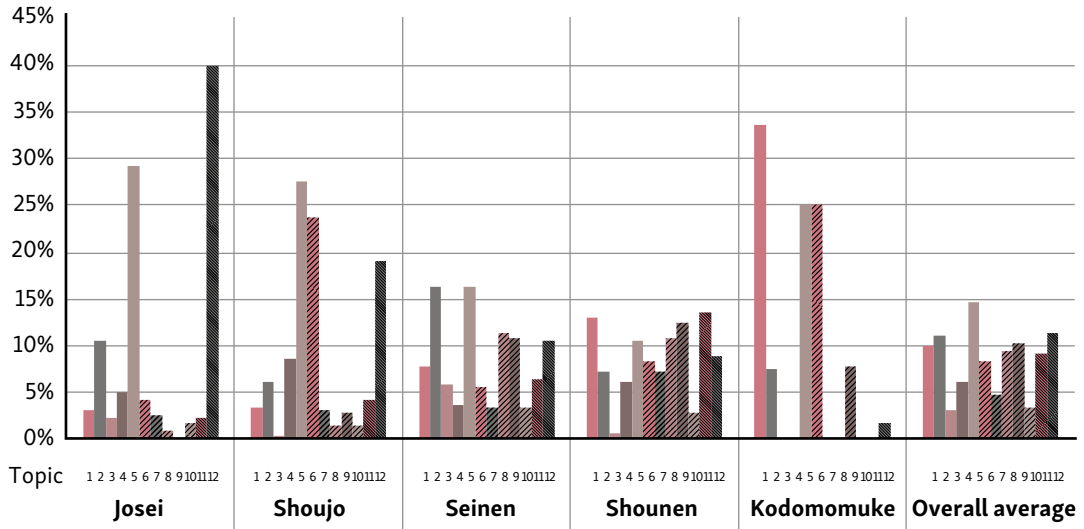
Figure 5. Analysis of the 12 topics indicated by the LDA

Source: Own elaboration.

Clustering was able to find more recurrent and common topics, grouping them together. Nevertheless, when checking figure 4, certain tags are found in more than one topic. However, the process involves the use of the probabilistic proportion of each tag's references in the configured topics to perform the automatic targeting of anime in a certain demographic. For example, in figure 4, we see that the terms related to daily slice or slice of life were among the highest average tag holders in topic 5. When examining the anime assigned to this topic, the analysis reveals a conformity with works aimed at this demographic group, consisting of anime of the daily slice and *Iyashikei* types. However, it should be noted that, although this topic turns revolves around the daily slice tone, it has tags closely associated with this content and appearing in other topics, such as school, which also is present in topics 3, 4, 7, and 8.

As an example of the occurrence of the probabilistic proportion among the topics, the analysis of the results for the anime *Knight's & Magic* can be presented. The work has tags distributed across topics in the following proportion: 32% in topic 1 (*isekai* magic fantasy anime), 13% in topic 5 (slice of life) and 53% in topic 9 (dramatic science fiction anime). Thus, this anime was classified as belonging to topic 9. The plot of this anime is about a character fan of *mechas* (giant robots) who travels to another universe to fight humanoid robots. The tags describing this anime are fantasy, *isekai*, magic, *mecha*, military, novel, piloted robot, power suit, science fiction, *shota*, and *shounen*. It can be observed that, although it's an anime strongly identified with topics 1 and 9, it also breaks into topic 5. I.e., according to figure 5, *Knight's & Magic* anime has traits that identify it as *isekai* magic fantasy anime (topic 1), slice of life anime (topic 5), and dramatic science fiction anime (topic 9). This example demonstrates that the proposed methodology has endowed the segmentation operations with plastic capabilities, ceasing to be a deterministic routine. The result of the probabilistic LDA model by order of relevance for each topic helps in further qualitative analyses, allowing the observation of relevance ratios, of common characteristics, and to make more accurate critiques of the segmentations, helping in the segmentation decision-making and in the realization of qualitative improvements. It's also noteworthy that the LDA model provides results with the quality of the adopted databases, thus allowing a methodology that enables non-deterministic results, essential for the study of cultural products, which have a high degree of fluidity in their elaboration, to meet increasingly expressive portions of the consumer public.

Finally, the analysis of the anime in relation to their demographics was conducted, comparing their current classification in relation to the methodological proposal of this work (figure 6).



Note: the y-axis is the average percentage of participation of the topics in the textual topics.

Figure 6. Distribution of words per topic

Source: Own elaboration.

Considering the used base of 1,889 animes, the current demographic distribution displays the following results: *Kodomomuke* (1.0%), *Josei* (3.3%), *Shoujo* (10.7%), *Seinen* (41.6%), and *Shounen* (43.3%). It is observed that the demographics with the lowest number of records are more concentrated in fewer topics of the demographics proposed in this paper. In the case of the *Kodomomuke* topic, we see most of the works in topics 1, 5 and 6. The anime of the *Josei* demographics are mostly allocated in topics 5 and 12.

Considering the demographics currently focused on the young adult female audience, by observing figure 6 it was possible to visualize the similarity between the topics’ distribution between the anime of the *Josei* and *Shoujo* demographics. Two relevant differences were noted and reflect the specific characteristics of these demographics when compared, showing that the automated methodology organized the anime in a new, broader, and more refined proposal, without losing the constitutive references of the demographics used so far. In the first case, it was verified that, in topic 6, there is a much higher participation of *Shoujo* anime (aimed at female teenagers) in relation to *Josei* (aimed at young and adult females), since in this demographic anime of fanciful fights featuring cute teenagers were addressed. In the second case, there is a more substantial share in topic 12 for *Josei* anime compared to *Shoujo*, as this demographic was targeted at drama and romantic comedy anime with cute, non-erotized characters and plots more valued by adult audiences.

In the demographics aimed at the young adult male audience, a similar perception is verified, at closer levels, in the comparison of distribution between

Shounen (aimed at teenage males) and *Seinen* (aimed at young adult males) anime. In this case, the same has also occurred with regard to the anime aimed at the female young adult audience. I.e., the new demographic proposal did not deviate from the current one. The most relevant cases were found in the topics. In topic 1, aimed at *Isekai* and fantasy anime, there is double the participation of *Shounen* anime compared to *Seinen*. The topic 2 is a demographic intended for action and psychological horror anime, thus targeting an adult audience, there is almost three times the participation of *Seinen* anime compared to *Shounen* anime. On the topic 3, the allocation of *Shounen* anime is irrelevant and the *Seinen* anime represent the demographic with the highest composition percentage, being dedicated to dramatic novels works and crime stories, based on novels and games. In topic 5, with anime reserved for stories of everyday life, *Seinen* anime have a share of over 50% compared to *Shounen* works.

Finally, it should be noted that *Kodomomuke* anime were basically concentrated in topics 2 (*Isekai* and fantasy anime), 5 (everyday life) and 6 (fantasy fights involving cute teenagers). However, there is an allocation of (few) anime from this demographic in topic 2 (action and psychological horror). In this case, what could be analyzed as an inconsistent result, actually reveals the importance of a broader and more refined database, because any addition of tags in these anime and targeting this topic end up directing the algorithm to this targeting operation.

FINAL CONSIDERATIONS

This paper showed the proposal and feasibility of a methodology for classifying cultural works from the extraction and automatic clustering of data based on RPA and LDA added to human qualitative analysis.

As a case study, this proposal was adopted for the automatic classification of anime, establishing the assertion of a more refined demographic compared to the one now adopted. Considering the current panorama of anime production, the current demographics, based on five topics, and extremely grounded in a sexual segmentation, is no longer sufficient to contemplate the expansion of new themes. From the incorporation of new contents and/or the combination of elements of the current demographics, the application of the proposal outlined in this work allowed the composition of a new demographic chart with a higher level of detail and in an automated way, producing a robust and consistent result.

The methodology presented in this work has the potential for an extended adoption in classification operations of any type of cultural product, just by having public and structured databases to perform data scraping operations, which supports the use of LDA to establish a probabilistic scoring system that supports automated decision-making.

REFERENCES

- Anderson, C. (2006). *The Long Tail: Why the Future of Business is Selling Less of More*. Hachette Books.
- Azuma, H. (2009). *Otaku: Japan's Database Animals*. University of Minnesota Press.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022. <https://dl.acm.org/doi/10.5555/944919.944937>
- Bryce, M., & Davis, J. (2010). *Manga: An Anthology of Global and Cultural Perspectives*. A&C Black.
- Brzeski, P. (2022, May 16). Cannes: How Japanese Anime Became the World's Most Bankable Genre. *The Hollywood Reporter*. <https://www.hollywoodreporter.com/business/business-news/cannes-japanese-anime-worlds-most-bankable-genre-1235146810/>
- Butler, C. (2019). *Shoujo Versus Seinen? Address and Reception in Puella Magi Madoka Magica* (2011). *Children's Literature in Education*, 50, 400–416. <https://doi.org/10.1007/s10583-018-9355-9>
- Condry, I. (2013). *The Soul of Anime: Collaborative Creativity and Japan's Media Success Story*. Duke University Press.
- Galbraith, P. W. & Schodt, F. L. (2009). *The Otaku Encyclopedia: An Insider's Guide to the Subculture of Cool Japan*. Kodansha International.
- Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2014). Web scraping technologies in an API world. *Briefings in Bioinformatics*, 15(5), 788–797. <https://doi.org/10.1093/bib/bbt026>
- Hui, Y. (2016). *On the Existence of Digital Objects*. University of Minnesota Press.
- Iwabuchi, K. (2002). *Recentering Globalization: Popular Culture and Japanese Transnationalism*. Duke University Press.
- Jenkins, H., Ford, S., & Green, J. (2013). *Spreadable Media*. New York University Press.
- Katsuno, H. & Maret, J. (2004). Localizing the Pokémon TV Series for the American Market. In J. Tobin (Ed.), *Pikachu's Global Adventure: The Rise and Fall of Pokémon* (pp. 80–107). Duke University Press.
- Marres, N. & Weltevrede, E. (2013). SCRAPING THE SOCIAL?: Issues in live social research. *Journal of Cultural Economy*, 6(3), 313–335. <https://doi.org/10.1080/17530350.2013.772070>
- Mitcham, C. (1994). *Thinking through Technology: The Path between Engineering and Philosophy*. The University of Chicago Press.
- Mhatre, M., Phondekar, D., Kadam, P., Chawathe, A., & Ghag, K. (2017). Dimensionality reduction for sentiment analysis using pre-processing techniques. In *2017 International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 16–21). IEEE. <https://doi.org/10.1109/ICCMC.2017.8282676>
- Mullakara, N., & Asokan, A. K. (2020). *Robotic process automation projects: build real-world RPA solutions using UiPath and automation anywhere*. Packt Publishing.
- Nijhawan, T., Attigeri, G., & Ananthakrishna, T. (2022). Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*, 9, 33. <https://doi.org/10.1186/s40537-022-00575-6>

- Orsini, L. (2018, May 30) MyAnimeList passes Third Day of Unexpected Downtime. *Forbes*. <https://www.forbes.com/sites/laurenorsini/2018/05/30/myanimelist-passes-third-day-of-unexpected-downtime/?sh=36c56d987e9e>
- Pineda, R. A. (2021, November 7). AJA: anime industry contracted 3.5% in 2020: overseas market overtakes domestic market for 1st time. *AnimeNewsNetwork*. <https://www.animenewsnetwork.com/news/2021-11-07/aja-anime-industry-contracted-3.5-percent-in-2020/179142>
- Saitou, T. & Azuma, H. (2011). *Beautiful Fighting Girl*. University of Minnesota Press.
- Scolari, C. A. (2018). *Las leyes de la interfaz: diseño, ecología, evolución, tecnología* (The laws of the interface: design, ecology, evolution, technology). Editorial Gedisa.
- Tripathi, A. M. (2018). *Learning Robotic Process Automation: Create Software robots and automate business processes with the leading RPA tool-UiPath*. Packt Publishing.
- Urbano, K. & Araujo, M. (2021). O fluxo midiático dos animês e seus modelos de distribuição e consumo no Brasil (Anime's media flow and the distribution and consumption's models in Brazil). *Ação Midiática - Estudos em Comunicação, Sociedade e Cultura*, (21), 81-101. <https://revistas.ufpr.br/acaomidiatica/article/view/71589>
- Van der Aalst, W. M., Bichler, M., & Heinzl, A. (2018). Robotic Process Automation. *Business & Information Systems Engineering*, 60, 269-272. <https://doi.org/10.1007/s12599-018-0542-4>
- Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, 38(3), 2758-2765. <https://doi.org/10.1016/j.eswa.2010.08.066>
- Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, 16, S8. <https://doi.org/10.1186/1471-2105-16-S13-S8>

ANNEX 1

Word	Meaning
(1) Isekai (異世界)	It means "different world". It is a genre with its own entity in Japanese fiction with usually <i>otaku's</i> protagonists, generally fans of Japanese pop culture works such as manga, anime, novels and video games etc., who are transported from their reality to other world in another time or parallel reality.
(2) Juujin (獣人)	People with mild animalistic features, who are not anthropomorphized animals. An example of a <i>juujin</i> would be a human with cat ears. (https://anidb.net/).
(3) Bishounen (美少年)	Means pretty boys, a subgenre of manga focusing on handsome male characters (often with androgynous traits), usually attributed in homoerotic works. (Galbraith & Schodt, 2009).
(4) Iyashikei (癒し系)	A subgenre of slice of life meaning "healing style". They are animes about characters living their normal and everyday lives in peaceful environments to de-stress and have a healing effect on the audience, bringing feelings of peace and spiritual satisfaction, as an antidote to stressful modern life. As an example, we can mention the anime <i>Yuru Camp</i> and <i>Yokohama Kaidashi Kikou</i> .
(5) Koma	Full name being <i>Yonkoma Manga</i> (四コマ漫画). They are short mangas, composed only by four sequential frames.
(6) Cgdct	Acronym for cute girls doing cute things (https://anidb.net/).
(7) Mecha	Name for giant robots or sub-genre of giant robot anime and manga.
(8) Manga (漫画)	Japanese comics.
(9) Manhua	Chinese and taiwanese comics.
(10) Ecchi (エッチ)	It means obscene, as well as euphemism for erotic and sex. This subgenre features erotic material (Galbraith & Schodt, 2009).
(11) Pantso (パンツ)	Women's and men's underwear.
(12) Meiji	Meiji Jidai (明治時代), Japanese Era between 1868 and 1912, marked by the modernization and industrialization of Japan, centralization of power in the figure of the emperor and abandonment of the isolationist policy of the Edo Era.
(13) Bakumatsu (幕末)	Term to define the end of the Edo Era, between 1850 and 1868, in which the first reforms in Japanese society began and marked by political divisions.
(14) Idol (アイドル)	High-tech singers and singers whose music and dance performances appeal directly to the audience's adoration and fantasies (Galbraith & Schodt, 2009).

Explanation chart about japanese and otaku terminology

Source: Own elaboration.

SOBRE LOS AUTORES

JÚLIO CÉSAR VALENTE FERREIRA, Profesor del Programa de Posgrado en Cultura y Territorialidades (PPCULT) de la Universidad Federal Fluminense (UFF) y de la Licenciatura en Ingeniería Mecánica (COEMEC) del Centro Federal de Educación Tecnológica Celso Suckow da Fonseca (CEFET/RJ). Doctor en Memoria Social por la Universidad Federal del Estado de Río de Janeiro (UNIRIO). Líder del Grupo de Investigación Producción y Economía de Comunión y miembro del Centro de Estudios Culturales Orientales. Coordinador Científico del Encuentro de Ingeniería en Entretenimiento (3E/UNIRIO). <http://lattes.cnpq.br/3396805392800659>

 <https://orcid.org/0000-0001-9732-7939>

THIAGO RIBEIRO FURTADO, Estudiante de maestría en Ciencias de la computación del Centro Federal de Educación Tecnológica Celso Suckow da Fonseca (CEFET/RJ). Miembro del Centro de Estudios Culturales Orientales. <http://lattes.cnpq.br/7729024086558924>

 <https://orcid.org/0000-0002-2558-0836>

RAFAEL DIRQUES DAVID REGIS, Graduado en Ingeniería Industrial por la Universidad Federal del Estado de Rio de Janeiro (UNIRIO). Desarrollador de Software RPA en Smarthis Ltda. Profesor auxiliar de japonés en Shirai Idiomas. Miembro del Centro de Estudios Culturales Orientales. Redes sociales: Twitter - @rafaeldr1, Instagram: @rafaeldirques <http://lattes.cnpq.br/5744561521271932>

 <https://orcid.org/0000-0001-8298-2973>

GABRIELA RODRIGUES DINIZ, Estudiante de Derecho en la Universidad Estácio de Sá (UNESA). Miembro del Centro de Estudios Culturales Orientales. <http://lattes.cnpq.br/4636974603291745>

 <https://orcid.org/0000-0001-6484-7356>

PAULA GONÇALVES, Licenciada en Estudios de Medios de la Universidad Federal Fluminense (UFF). Asistente de comunicación en la Fundación Heinrich Böll Brasil. Miembro del Centro de Estudios Culturales Orientales. R <http://lattes.cnpq.br/7593233800279718>

 <https://orcid.org/0000-0002-7154-7762>

VITOR PEDRO DA SILVA CASTELO TAVARES, Estudiante de Geografía en la Universidad Federal Rural de Río de Janeiro (UFRRJ). Miembro del Centro de Estudios Culturales Orientales.

 <https://orcid.org/0000-0003-1146-3273>