Automation of Topic Generation in Government Information Requests in Mexico

Hermelando Cruz-Pérez and Alejandro Molina-Villegas, and Edwin Aldana-Bobadilla

Abstract In Mexico, legislation guarantees public access to information, empowering citizens to request data from the government. This research delves into the National Transparency Platform's extensive archive, which includes over 2 million requests for information, with the goal of discerning the primary interests of citizens in government actions from 2003 to 2020. Through the analysis of 2,518,875 requests, Genetic Algorithms were employed to fine-tune three crucial hyperparameters of the Latent Dirichlet Allocation (LDA) model: alpha, beta, and the number of topics. This optimization aimed at enhancing the model's accuracy in topic identification, measured by the coherence metric of the topics identified. Additionally, Generative Pre-trained Transformer (GPT) technology facilitated the automatic generation of titles and descriptions for these topics. The investigation revealed 4,131 topics of public interest throughout the Mexican Republic, with significant emphasis on environmental management, public policies, the response to the COVID-19 health crisis, labor issues, and education in 2020. These findings underscore the critical role of proactive transparency and the provision of open data in advancing the analysis of vast quantities of government data. This study paves the way for future data-driven decision-making and policy development research. It highlights the profound influence of sophisticated data analysis in promoting government transparency and stimulating citizen engagement. Using genetic algorithms to refine

Conahcyt-Centro de Investigación en Ciencias de Información Geoespacial
Mexico City 14240, Mexico, e-mail: hcruz@centrogeo.edu.mx , ORCID: https://orcid.org/0000-0003-3740-7632

Alejandro Molina-Villegas
Conahcyt-Centro de Investigación en Ciencias de Información Geoespacial
Mexico City 14240, Mexico,e-mail: amolina@centrogeo.edu.mx , ORCID: https://orcid.org/0000-0001-9398-8844

Edwin Aldana-Bobadilla
Conahcyt-Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional—Unidad Tamaulipas, Ciudad Victoria, Tamaulipas 87130, Mexico e-mail: edwyn.aldana@cinvestav.mx , ORCID: https://orcid.org/0000-0001-8315-1813

the LDA model and large language model technology for content generation, this study innovates analyzing public information requests, contributing significantly to improving governmental transparency.

Key words: Genetic Algorithm, Latent Dirichlet Allocation, Governmental Transparency, Open Data, Data Processing.

1 Introduction

In an era where government data swells at an unprecedented rate, the skill to meticulously parse and derive insights from these extensive datasets is a cornerstone for advancing public understanding and underpinning data-driven policy-making. In the governmental sphere, this reality is especially notable, as public administrations generate immense amounts of data in their daily interactions with citizens. A clear example is the numerous requests directed to the Mexican government through the National Transparency Platform (NTP).

This study focuses on the application of advanced data processing techniques. Specifically, a genetic algorithm has been developed to automate the selection of key hyperparameters, including Alpha, Beta, and the number of topics, which are crucial for the efficient training of the LDA (Latent Dirichlet Allocation) model. This model is fundamental in the thematic analysis of large datasets, enabling the extraction of meaningful topics from unstructured text data. Further, we have applied optimization techniques to the vocabulary to enhance the quality of topic analysis. This includes Text Mining techniques to ascertain topic similarity across different geographic areas. Additionally, in line with the principles of Zipf's Law, our approach to vocabulary optimization has significantly improved the precision of our results, confirming the law's relevance in the context of natural language processing and topic modeling.

With this study, we aim to establish a solid framework for future research and practical applications in the processing and analysis of large data sets, thus marking a milestone in the management of public information and its efficient interpretation.

1.1 Background of Access to Information

The history of access to public information has its roots in Sweden, where, in 1766, the Freedom of the Press Act and the right to access public records were enacted. Anders Chydenius, a Swedish-Finnish priest who also played roles as a member of parliament and economist, drove this pioneering legislation. This event marked a significant milestone in the administrative centralization of Sweden, laying the groundwork for future developments in this field [26].

However, the modern concept of access to information underwent a fundamental change after the conclusion of World War II. In 1948, the Universal Declaration of Human Rights was proclaimed, which revisited the notion of freedom of expression. This declaration established that no citizen should be disturbed because of their opinions and fundamentally recognized the rights of citizens to seek, disseminate, and receive information. These historical events laid the foundations for the the current concept of access to information, which is intrinsically linked to the right of individuals to seek, disseminate, and receive information by any means available. Today, information is considered an indispensable resource for the exercise of other fundamental rights[14].

In Mexico, freedom of expression is a fundamental right recognized in the Constitution of 1917. Article 6 states that the expression of ideas is protected, except when it affects morality the rights of others, incites crime, or disturbs public order. This balanced approach seeks to promote freedom of expression while protecting other important values[10].

However, it was in December 1977, during López Portillo's political reforms, that the phrase "The right to information is guaranteed by the State" was added to Article Six of the Constitution, marking a milestone in the recognition of this fundamental right in Mexico [14]. Despite this significant inclusion, there persisted a notable ambiguity concerning the practical implications of this enshrined right. Citizens could access governmental information, but the Supreme Court of Justice of the Nation considered that this reform did not fully guarantee access to information generated by the government. This process reflects the evolution of the right to information in Mexico and the ongoing need to protect it effectively.

The landscape of transparency in Mexico underwent a pivotal transformation on April 30, 2002, when Congress unanimously enacted the Federal Law of Transparency and Access to Governmental Information (LFTAIG). This landmark legislation introduced institutions and procedures enabling individuals to request federal governmental information from designated entities, marking a significant milestone in the country's commitment to open governance. Additionally, a one- month deadline was established for these entities to provide the requested information. The Federal Institute for Access to Information (IFAI) was designated as the entity responsible for ensuring the right to access information [15],[8].

The most significant national milestone was the approval in 2015 of the General Law of Transparency and Access to Public Information (LGTAIP). Its main objective is to promote transparency in institutions and strengthen participation and accountability [24]. Furthermore, this law changed the name of IFAI to INAI to enhance its role as a national-level guarantor institution.

1.2 Role of the National Institute for Transparency, Access to Information, and Personal Data Protection (INAI)

The National Institute of Transparency, Access to Information and Protection of Personal Data (INAI) in Mexico plays a key role as an independent body in promoting government transparency and safeguarding personal information. Its main function is to ensure public access to information and to oversee the proper handling of personal data, fostering accountability of authorities, and ensuring the privacy of citizens.

According to the General Law of Transparency and Access to Public Information [11], the (INAI) performs the following functions:

- Resolve citizen complaints who submit requests for review against obligated subjects to safeguard the right to access information.
- Resolve disputes from citizens who submit resources of non-compliance against decisions made by local guarantor organizations; when these contravene the principle of maximum publicity of information.
- Establish sanctions and take appropriate measures in cases where the rights of access to public information and the protection of personal data are violated.
- Ensure that personal data in the possession of government entities or companies are properly safeguarded and secure, and that they are not shared without the knowledge, explicit consent, and approval of their owners.

The purpose of the INAI in Mexico is to ensure that all Mexican citizens can exercise their right to access information while being protected from any undue interference by other governmental actors. The INAI conforms to policies proposed by international organizations that seek to promote democratic governance through inclusive political institutions[9].

The General Law of Transparency and Access to Governmental Public Information (LGTAIPG), published on May 5 strengthens the autonomy of the INAI by granting it can review resolutions of local guarantors bodies in cases of controversy. Additionally, it consolidates the position of the INAI as the highest authority on the matter and prevents authorities from challenging its decisions before administrative or judicial bodies. The law also expands the powers of the guarantor bodies by allowing the INAI to resolve review requests filed by individuals against state guarantor bodies, as well as to take coercive measures and sanctions when necessary [17], [4].

The INAI is committed to promoting and contributing to the creation of a state in Mexico that is more transparent and accessible. The aim is to restore the public's trust in the authorities and to promote active participation of society in all processes related to the formulation of public policies. By working together to address public challenges, the likelihood of democracy being effective and generating positive results is increased[25].

1.3 The importance of the National Transparency Platform (PNT)

The National Transparency Platform (PNT) is the digital strategy established by the Mexican Government to make transparent the operations of more than 8,000 Obligated Subjects (SO) across the country. The PNT is a fundamental part of the information dissemination work coordinated by the National Institute for Transparency, Access to Information, and Protection of Personal Data (INAI), which is responsible for liaising with and linking the 32 institutes and/or commissions that, in turn, work in coordination with the SOs of each state.

The PNT plays an essential role in the Mexican Government's efforts to promote transparency in the management of more than 8,000 Obligated Subjects (SO) in all states of the Mexican Republic. This digital platform has been designed to ensure that relevant information is available to the public, promoting greater accountability at all levels of government.

The INAI plays a central role in the implementation and coordination of the PNT. This body is responsible for establishing links with the 32 state institutes and/or commissions, who, in turn, collaborate closely with the SOs in each state. Together, they work synchronously to ensure that public information is accessible, verifiable, and understandable for all citizens. The PNT not only seeks to ensure transparency in the management of public resources but also promotes access to relevant information in areas such as government procurement, public finances, government policies, and other topics of public interest. This allows citizens to be better informed and actively participate in decision-making, thus strengthening democracy and citizen control over government actions.

It is crucial to highlight that the National Transparency Platform (PNT) consists of four key systems, each playing specific roles in establishing a comprehensive and highly interoperable digital space. Firstly, the Information Access Request System (SISAI) connects users with over eight thousand public institutions, allowing them to request public and personal information. In addition to facilitating this interaction, SISAI also manages the registration of requests and ensures compliance with deadlines established by law. In cases of potential delays or omissions, this system presents complaint resources to expedite the process [1].

Complementing the Information Access Request System, the PNT includes three other essential systems: the Management System of Means of Contestation, the System of Transparency Obligations Portals, and the Communication System between Guarantor Bodies and Obligated Subjects (SO). These systems collaborate to establish an efficient and highly integrated cyberspace that simplifies the participation of all actors involved in access to information and governmental transparency.

2 State of the Art

Some significant works stand out in information request analysis using advanced techniques. [2] used supervised Latent Dirichlet Allocation (LDA) techniques to predict the Mexican government's response capability to federal information requests between 2003 and 2015. Similarly, [3] applies sLDA to investigate the Mexican government's response (or lack thereof) to public information requests during the same period, identifying the topics most linked to governmental responses.

Another author [6] analyzes a large dataset consisting of hundreds of thousands of information requests, comparing similar cases in terms of dependencies, themes, and timing and considering the complexity and sensitivity of the requests. This study reveals a higher success rate for requests from regions with strong support for ruling parties, especially on issues of public relevance. This suggests an attempt to mitigate political risks rather than favor supporters. Similarly, [5] examines over a million information requests, using unsupervised methods to categorize them based on their thematic diversity. This analysis demonstrates the variability of requests, from public transparency to more private and micropolitical interests, and highlights public concern for environmental impact and violence.

Our methodology is distinguished by proposing a new automated approach to identify topics with the optimal values of alpha and beta parameters based on their coherence. In addition, we conduct differentiated analyses by state and year.

Regarding the automation of hyperparameters using another type of information, We can mention [20], which focuses on optimizing LDA configurations to surpass the results of the conventional LDA model. The work introduces and applies the SA-LDA algorithm, which leverages Simulated Annealing (SA) to determine the optimal values of LDA parameters. Analogously, [28] explores the improvement of LDA model parameters through the development of a parallel differential evolution algorithm that incorporates two cost functions, LDADE and Word2Vec, emphasizing the ongoing commitment to refining techniques for thematic modeling.

3 Theoretical Framework: Automation of Topic Generation

The Automation of Topic Generation using Genetic Algorithms for Hyperparameters in Latent Dirichlet Allocation (LDA) Modeling represents a fascinating and advanced approach in the field of natural language processing and text mining. This methodology combines the robustness of LDA Topic Analysis with the efficiency of genetic algorithms, offering a synergistic approach that enhances the capability to uncover latent topics within large text corpora. LDA is a widely used topic modeling technique for uncovering hidden structures or topics within large text data sets. However, the performance of this model depends greatly on the optimization of its hyperparameters, mainly alpha, beta, and the number of topics (topic).

In this context, genetic algorithms, which are search and optimization methods inspired by natural selection and genetics present themselves as a powerful solution. These algorithms iteratively adjust and fine-tune the values of the LDA hyperparameters, seeking the optimal the combination that maximizes the coherence and relevance of the generated topics. The process involves generating a population of LDA models with different hyperparameters configurations, evaluating their performance, and applying genetic operations such as selection, crossover, and mutation to evolve the hyperparameter configurations over several generations.

This automatic approach not only improves the quality of the generated topics but also frees users from the the tedious and often subjective task of manually adjusting hyperparameters. In this section, in addition to presenting a solid theoretical framework, we will delve into the key concepts that support this approach, providing a more complete understanding of its functioning and applicability.

3.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is an unsupervised topic analysis algorithm used to identify underlying themes in a set of documents. The algorithm functions by assuming that each document is composed of a mixture of latent topics, and that each topic is comprised of a set of words [7]. The idea behind LDA is to find a representation of documents in terms of a probability distribution over a a finite set of latent topics, where each topic is represented by a probability distribution over the vocabulary. To generate a new document, LDA first selects a mix of topics according to the topic distribution of the document. Then, for each word in the document, a topic is selected according to the chosen mix, and a word from that topic is chosen according to its probability distribution in the vocabulary.

In [7], Blei et al. introduced the concept of Latent Dirichlet Allocation by employing the Dirichlet distribution, as delineated in equation (1), as the prior distribution of the two generative models: topics in documents and words in topics. In equation (1), the Dirichlet distribution's parameters are leveraged to control the mixture of topics within documents. Here, Γ denotes the generalization of the factorial function to the real number domain, $\mathbb{R}^n$, effectively extending $n!$ to non-integer dimensions. The variable T represents the total number of topics within the document corpus, while α serves as the hyperparameter influencing the distribution's shape, effectively dictating the topic mixture's concentration and diversity.

The Dirichlet distribution, defined by the equation (1) below, plays a pivotal role in the generative process of LDA, specifying the prior probabilities of topic distributions within documents. The term p_j, within the product, signifies the probability assigned to each topic j, with the exponent $\alpha - 1$ modulating the influence of each topic based on the hyperparameter α. The objective is to find the probability values of the distribution so that the original documents can be generated.

$$Dir(\alpha) = \frac{\Gamma(\alpha T)}{\Gamma(\alpha^T)} \prod_{j=1}^{T} p_j^{\alpha-1} \tag{1}$$

In equation for the Dirichlet distribution within LDA:

- $Dir(\alpha)$: Dirichlet distribution parameterized by α.
- α: Controls uniformity across topics; higher values mean more evenly distributed topics.
- T: Number of topics.
- Γ: Gamma function, for normalization.
- p_j: Probability of topic j.
- $\prod$: Product of topic probabilities, influenced by α.

3.2 Topic Evaluation

In topic modeling, coherence assessment initiates with identifying ”pivotal words” for a fixed corpus. The ”pivotal words” are selected based on frequency and thematic pertinence. The interrelations among these words are scrutinized using metrics like semantic similarity, thereby establishing a foundation for thematic cohesion. Subsequently, coherence is quantified by evaluating aspects such as the semantic proximity of words and their distribution across the corpus. This analytical procedure was integrated using statistical methodologies, including the arithmetic mean or the median, to formulate a comprehensive coherence score for the theme [23].

As delineated in [18], the role of coherence in topic modeling is instrumental for ascertaining the interpretability of themes by human analysts. By characterizing themes through the most likely words, the coherence score gauges the degree of similarity among these words, serving as a crucial indicator of thematic coherence. One of the most commonly used metrics is the UMass coherence score Equation 2. This metric calculates the frequency with which two words appear together in the corpus.

$$C_{UMass} = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)} \tag{2}$$

The components of the UMass coherence equation are delineated below, elucidating their roles in assessing topic coherence within the realm of topic modeling:

- C_{UMass}: The UMass coherence final score evaluates semantic coherence among topic words.
- N: The count of unique words in the evaluated topic.
- w_i, w_j: The topic’s indexed words for pairwise comparison.
- $D(w_i, w_j)$: Document frequency of both w_i and w_j, indicating co-occurrence.
- ϵ: A negligible positive value to prevent undefined logarithms.
- $D(w_j)$: Document frequency of w_j, the occurrence count of w_j.
- log: Logarithmic function, normalizing the coherence score.
- $\sum$: Sum over word pairs.

Crucially, C_{UMass} quantifies thematic cohesion by summing over word pairs to evaluate their co-occurrence $(D(w_i, w_j))$ and individual frequencies $(D(w_j))$, adjusted by ϵ to ensure mathematical stability. This calculation, underpinned by the logarithmic ratio, offers insights into word semantic connections. The aggregate coherence score, derived from averaging these pairwise scores, reflects the topical harmony anchored in the corpus frequencies that originally informed the topic models. This approach ensures a nuanced, context-sensitive analysis of topics [27].

3.3 Zipf's Law in Vocabulary Refinement

The frequency distribution of words in human languages follows a fundamental and enthralling phenomenon within linguistic studies, epitomized by Zipf's Law presented in Formula (3). This distribution follows a systematic and universally acknowledged pattern: specific words occur with high frequency – such as ”a”, ”the”, and ”you” – dominating functional words, whereas terms related to document semantics surface sporadically [21]. This pattern exhibits remarkable consistency across diverse natural languages and contexts, significantly influencing fields like natural language processing. We used such properties to automatize the distinction between functional words and useful terms for topic modeling.

Zipf's Law articulates this distribution as:

$$f(r) \propto \frac{1}{r^{\alpha}} \tag{3}$$

where:

- $f(r)$ represents the frequency of the term with rank r.
- α is an exponent, which in natural languages typically is close to 1.
- $\propto$ denotes proportionality, indicating that the frequency of a word is inversely proportional to its rank in the frequency list, raised to the power of α.

3.4 Genetic Algorithm

Genetic algorithms are search procedures that are based on the principles of natural selection and genetics. These algorithms combine the survival of the fittest chain structures with certain innovative aspects of human search. In each generation, a new set of artificial ”creatures” (chains) is generated using fragments from the fittest of the previous generation. Occasionally, new parts are incorporated for a better fit [12].

In the 1960s, John Holland pioneered genetic algorithms (GA) and collaborated with his students and colleagues at the University of Michigan during the 1960s and 1970s. Unlike approaches such as evolution strategies and evolutionary programming, Holland's original goal was not to design specific algorithms to solve particular problems. His primary focus was conducting a formal investigation of the adaptation process, as it manifests in nature, and developing methods to incorporate natural adaptation mechanisms into computer systems. Holland's approach and contributions are extensively discussed in Melanie Mitchell's work, ”An introduction to genetic algorithms” where it is well explained that genetic algorithms guaranteed function optimization despite their random nature [16].

Genetic algorithms stand out compared to their conventional counterparts in the search for robustness, in four fundamental aspects. First, they employ a coded representation of the parameter set instead of individual parameters. Second, instead of focusing on a single point, they explore a population of points. Additionally, they rely exclusively on the information of the objective function (Payoff) without using derivatives or other additional knowledge. Lastly, they guide their search process through probabilistic transition rules instead of deterministic rules [12].

According to [31], a genetic algorithm designed to address a specific problem consists of several essential components First, a genetic representation is required to define how to encode potential solutions to the problem. Subsequently, a procedure is established for generating an initial the population of solutions, serving as a starting point for the algorithm. An evaluation function acting as a virtual environment plays a crucial role in rating the solutions in terms of their suitability, guiding the search. In addition, genetic operators are used that alter the composition of future generations, facilitating the evolution and adaptation of solutions throughout the process. Lastly, the values of various parameters of the genetic algorithm, such as the size of the population and the probabilities of applying the genetic operators are adjusted to optimize its performance in the specific context of the problem to be solved. In this sense, most genetic algorithms (GAs) share a series of fundamental components, such as populations of chromosomes, selection based on the fitness of individuals, crossover processes that generate new generations of offspring, and the introduction of random mutations in these offspring [16].

4 Optimizing Topic Generation in LDA: A Genetic Algorithm Approach for Automated Hyperparameter Tuning

This section details the methodological approach used to automate the hyperparameter tuning, which produced many topics found in the National Transparency Platform.

To clarify the methodological process used, Figure 1 schematizes the main phases and the sequence followed by the proposal. This process ranges from collecting information requests, data preprocessing, and vocabulary optimization through Zipf's law to automating the LDA model's hyperparameters. The latter allows for identifying the main results: the prevalent topics in the citizens' information requests. Each stage is explained in detail in the rest of the section.

4.1 Information Gathering

The dataset for this study was meticulously compiled from the open data section of the National Transparency Platform, which focuses on federal entities obligated to adhere to transparency requirements. This research method deliberately omitted data from non-federal entities, including state and local governments, legislative and judicial branches, political parties, unions, and similar bodies, to maintain a federal scope. The data were accessible in various formats, such as SQL, XML, CSV, and JSON,

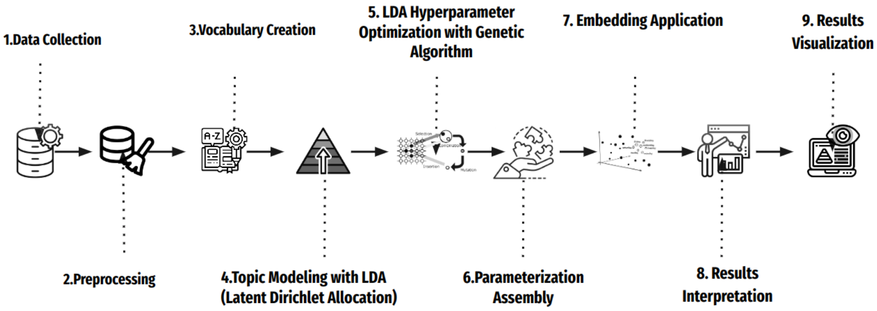
1.Data Collection
2.Preprocessing
3.Vocabulary Creation
4.Topic Modeling with LDA (Latent Dirichlet Allocation)
5. LDA Hyperparameter Optimization with Genetic Algorithm
6.Parameterization Assembly
7. Embedding Application
8. Results Interpretation
9. Results Visualization

Fig. 1 Methodological process.

facilitating comprehensive analysis.The time frame for data collection spanned from 2003 to 2020, during which data retrieval initially utilized Infomex green [29], the precursor to the National Transparency Platform, until 2015. Subsequently, data collection shifted to the National Transparency Platform [30] for the remainder of the period the Figure 2 shows the two platforms. Throughout this extensive timeframe, approximately 2,518,875 public information requests were submitted to various federal departments, underscoring the vast scope of data analyzed.

PNT
PLATAFORMA NACIONAL DE TRANSPARENCIA
inai
lunes 18 de septiembre de 2023
Inicio
Datos Abiertos
Consulta pública de solicitudes
Gráfica de tipo de respuestas a solicitudes
Gráfica de respuestas por Dependencia o Entidad
Guía de uso
Costos de Reproducción
Preguntas frecuentes sobre el acceso a la información pública
Calendario de días inhábiles de la Plataforma Nacional de Transparencia
Padrón de Sujetos Obligados
Portal de Obligaciones de Transparencia
Descarga Adobe Acrobat Reader

En este apartado puedes descargar la Base de Datos de solicitudes de información realizadas a las dependencias de la Administración Pública Federal a través de la Plataforma Nacional de Transparencia Gobierno Federal. Se incluyen las respuestas emitidas. Las solicitudes pueden ser descargadas por periodos o por sector. Las dependencias de la Administración Pública Federal están agrupadas en **sectores**.

Los archivos de gran tamaño toman tiempo para su descarga, por favor tómelo en cuenta.

Información actualizada al 31 de diciembre del 2020.

Resoluciones de Recursos de Revisión

Todas las solicitudes de información y sus respuestas

Solicitudes y respuestas por periodo

Solicitudes de información 2003 — 1568 Descargas
Solicitudes de información 2004 — 674 Descargas
Solicitudes de información 2005 — 871 Descargas
Solicitudes de información 2006 — 753 Descargas
Solicitudes de información 2007 — 746 Descargas
Solicitudes de información 2008 — 786 Descargas
Solicitudes de información 2009 — 745 Descargas
Solicitudes de información 2010 — 818 Descargas
Solicitudes de información 2019 — 1215 Descargas
Solicitudes de información 2020 — 1186 Descargas

Fig. 2 Infomex Platform and National Transparency Platform.

The database contains full details on citizen inquiries sent to various governmental entities and their responses. It includes around 20 attributes; for more detailed information, see Table 1. This table provides an example of a request related to COVID-19. In this instance, the request involves seeking information on the availability of free tests, existing treatments and their costs, as well as the procedures for receiving medical care in Mérida, Yucatán, Mexico. This instance highlights the diverse and complex nature of unstructured information and the importance of understanding specific needs and queries of citizens. In this research, we analyze and offer a detailed description of the requests submitted by interested parties. These requests, expressed through texts provided by the applicants, represent unstructured information.

INAI Attribute	**Description**	**Example**
Folio	Unique 13-digit request identifier.	310572322000377
Fecha de solicitud	Date and time when the request was filed.	05/07/2022, 14:23
Dependencia	Name of the government department the request is sent to.	Servicios De Salud De Yucatán
Status	Current status of the request, e.g., in process.	Terminada
Medio de entrada	How the request was submitted: electronic or manual.	Electrónica
Tipo de solicitud	Nature of the request: public info, personal data, data correction.	Información pública
Descripción	Applicant's explanation of requested information.	Detalles de pruebas y tratamientos de COVID-19 en Mérida, Yucatán.
Otros datos	Additional details to aid information location.	N/A
Archivo adjunto de solicitud	URL for supplementary files provided by the applicant.	N/A
Medio de entrega	Preferred method for receiving the requested information.	Electrónico a través del sistema PNT
Fecha límite de respuesta	Deadline for the department to respond.	20/07/2022
Respuesta	Type of response provided by the government department.	Entrega de información vía PNT
Archivo de respuesta	Digital file provided by the department as part of the response.	N/A
Fecha de respuesta	Date the department responded.	14/07/2022
País	Country of the applicant's location.	México
Estado	State of the applicant's location within the country.	Yucatán
Municipio	Municipality of the applicant's location.	Mérida
Código Postal	Postal code of the applicant's location.	97098
Sector	Government sector of the addressed department.	Descentralizados

Table 1 Attributes of the INAI requests database

4.2 INAI Information Request Preprocessing

The preparation of descriptions is a pivotal step in data management, involving the rigorous cleaning, transformation, and organization of the data to set the stage for further analysis and modeling. This process starts with creating JSON files, carefully categorized by federal entity and year. These files are crafted based on our database's ”state” column classification, excluding descriptions that fail to specify a particular federal entity, thereby focusing solely on those that explicitly identify one.

Subsequent steps include data cleaning removing special characters, numbers, and punctuation that could detract from the text's clarity. Following this, tokenization is performed, a crucial procedure in text processing that is especially vital for data analysis and natural language processing (NLP). For instance, after cleaning preprocessing and tokenization, the sentence ”I request information about COVID vaccines.” would dissect into tokens: [”I”, ”request”, ”information”, ”about”, ”COVID”, ”vaccines”]. This division into smaller units simplifies the later stages of analysis and manipulation of the text, enabling more effective and efficient processing.

Moreover, this stage involves identifying and removing stop words that contribute limited informative value. A custom word filter is employed for this task, designed to exclude general terms, prepositions, polite expressions, legal terminology, and typical phrases found in descriptions. This filter is critical for refining the content, allowing the analysis to concentrate on the text's most pertinent elements, ensuring the analysis remains focused and significant.

4.3 Vocabulary Optimization.

During this phase, we constructed a specialized vocabulary from terms in texts. We evaluated how often each term appears in our dataset and prioritized the most common ones, previous filtering of functional words. Zipf's law enabled us to streamline our vocabulary. According to Zipf's law, the frequency f of a word is inversely proportional to its rank r in the frequency table, We express this relationship with Equation (4):

$$f(r) = \frac{a}{r^s}, \tag{4}$$

where a is a constant and s is an exponent usually close to 1. By eliminating infrequent terms and categorizing less common terms as ”unknown,” we efficiently condensed the vocabulary size, boosting the model's performance without sacrificing accuracy or coverage.

To refine our vocabulary further, we applied a knee-point identification method to the frequency distribution. This technique involves steps such as importing data from a CSV file, calculating word frequencies, and applying a logarithmic transformation to these frequencies, as described in Equation (5):

$$\log_f = \log(frequency), \tag{5}$$

and organizing them in descending order by rank. The transformed frequencies ($\log_f$) facilitate an easier examination of the distribution.

Following this, we illustrated the relationship between the words' ranks and their logarithmic frequencies in a graph (refer to Figure 3). This visualization is crucial for observing the frequency distribution and locating the "knee point" — the juncture that signifies the ideal vocabulary size. The KneeLocator tool, which uses an algorithm to detect significant changes in the curve's slope, aids in identifying this point, marking where the frequency of word usage sharply declines.

Identifying the knee point allows us to define an optimal vocabulary size that includes the most impactful words up to the point of significant frequency reduction. This distinction aids in differentiating between common and rare terms, ensuring an effective analysis that retains the relevancy and integrity of the study on public information requests.

Subsequently, we filter the words to include only those at or below the knee point, exporting this refined selection to a text file. Additionally, we create a text representation of this streamlined vocabulary, formatting each word in quotes and separated by commas, readying it for integration into natural language processing (NLP) models.

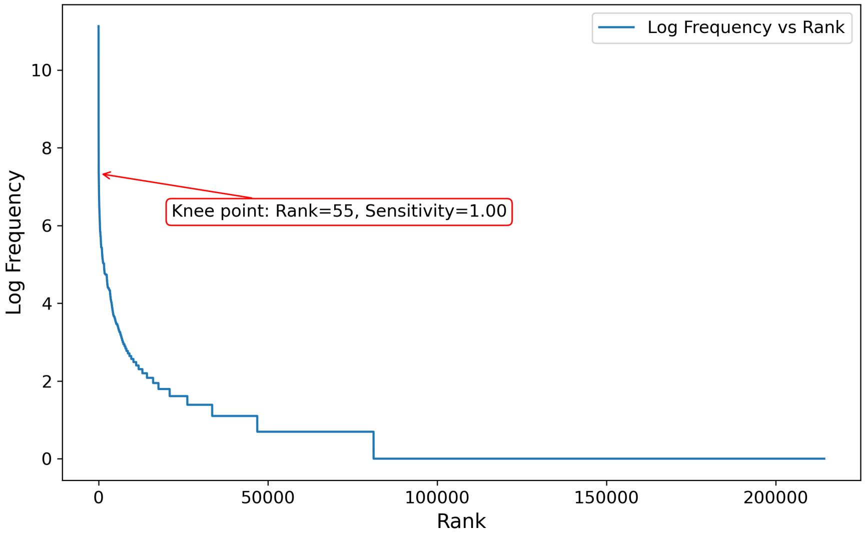

Fig. 3 Example of vocabulary optimization for requests from the state of Jalisco during 2020.

4.4 Topic modeling with LDA

In this crucial phase of the process, we implemented the Latent Dirichlet Allocation (LDA) model to generate a series of key files essential for the project's advancement. These key files include 'Corpus.mm', representing a structured collection of textual data; 'Dictionary.dict', a file mapping each unique word to a numeric identifier; 'Discarded Words.txt', containing a list of terms filtered out during the data cleaning phase; 'titles.txt', storing the titles or headings associated with each information request; and 'vocabulary.txt', a compilation of all relevant words identified in the corpus.

The Equation (6) is employed for topic modeling based on LDA and other topic mixture models. Statistical modeling calculates how documents in a collection are generated in terms of a mix of several topics, where a "topic" is a distribution over a fixed vocabulary, and each document is considered a mix of various topics. The purpose of the formula is to calculate the probability of observing a specific word.

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j) \cdot P(z_i = j) \tag{6}$$

- $P(w_i)$: Represents the marginal probability of encountering the word w_i.
- $\sum$: Signifies the summation over all categories j, from 1 to T, where T is the total number of topics.
- $P(w_i|z_i = j)$: Indicates the probability of the word w_i given topic j.
- $P(z_i = j)$: Reflects the a priori probability of selecting topic j.

For example, topic modeling in text analysis of public information requests reveals latent topics such as Waste Management, Biodiversity Conservation, and Renewable Energies, enhancing our understanding of key environmental concerns expressed in these requests.

- $P(w_i|z_i = j)$: This probability demonstrates how related a specific word w_i, like "recycling", is to a latent topic $z_i = j$, in this case, Waste Management. A high value indicates a significant association between the word and the topic.

- $P(z_i = j)$: This probability assesses how represented a specific latent topic z_i is within a request. Thus, if a request addresses Renewable Energies themes, the probability of this topic being represented (z_i = Renewable Energies') will be high.

4.5 Deployment and Implementation of a Genetic Algorithm.

We leveraged a specific version of the genetic algorithm for the efficient optimization of α and β parameters and the precise selection of the optimal number of topics, with the process evaluated on the strength of the topic coherence score. Figure 4 provides a flowchart outlining this optimization. The genetic algorithm harnesses a specific configuration to navigate the search space effectively: a population size of

20 for diversity, 50 generations to ensure ample evolution, selection of 4 parents for a balanced genetic mix, retention of 2 elite solutions to preserve superior genes, and a mutation rate of 0.05 to foster new traits. Furthermore, it explores a gene space that includes α and β parameters within a range of 0.01 to 1.0, and the number of topics between 2 and 50, ensuring the algorithm meticulously tunes the model parameters and identifies the most coherent topics for analysis. The effectiveness of the algorithm is supported by statistical tests, details of which are provided in the referenced work [13]. This streamlined approach guarantees that the genetic algorithm not only fine-tunes the model with precision but achieves this with optimal efficiency.

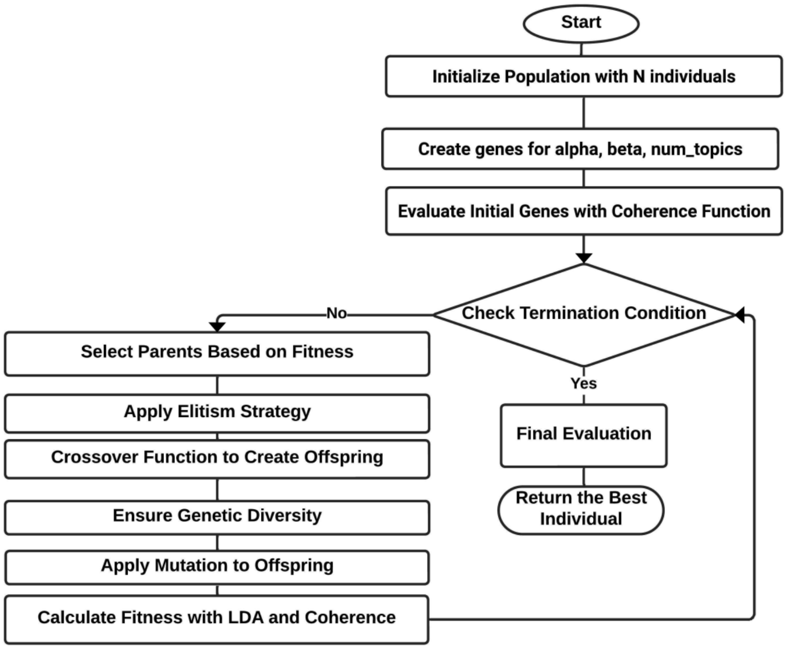
Start
Initialize Population with N individuals
Create genes for alpha, beta, num_topics
Evaluate Initial Genes with Coherence Function
Check Termination Condition
No
Yes
Select Parents Based on Fitness
Apply Elitism Strategy
Crossover Function to Create Offspring
Ensure Genetic Diversity
Apply Mutation to Offspring
Calculate Fitness with LDA and Coherence
Final Evaluation
Return the Best Individual

Fig. 4 Genetic algorithm flowchart.

4.5.1 Initializing the population.

The ‘initialize_population‘ function creates an initial population precisely, as detailed in Table 2. It utilizes two critical inputs: ‘pop_size‘, which specifies the total number of individuals to be generated, each representing a potential solution within the designated parameter space, and ‘parameter_ranges‘, offering a framework to ensure solutions comply with predefined constraints.

Starting with an empty ‘population,‘ the function iterates until ‘pop_size‘ is reached, each iteration creating an ‘individual‘ by randomly assigning values within 'parameter_ranges' for each parameter using the 'random.uniform' function. This method guarantees that the values fall within the allowed limits. Special focus is placed on the 'num_topics' parameter, which is converted to an integer to match the optimization problem's requirements. This step ensures the solution's integrity, confirming that each individual's parameters meet the model's standards. This approach details the population's generation process and emphasizes adhering to the model's parameters to produce viable solutions.

Identifier	**Definition**
`pop_size`	The population size to be initialized.
`parameter_ranges`	A dictionary with parameter names as keys and tuples as values, where each tuple contains the minimum and maximum range for the corresponding parameter.
`population`	A list that stores the population, where a list of parameter values represents each individual.
`individual`	A list representing a single individual in the population, generated within the specified parameter ranges. The last parameter is cast to an integer, assuming it's the number of topics ('num_topics').

Table 2 Identifier definitions for the population initialization function.

4.5.2 Coherence fitness function

The relevance and coherence of topics derived from topic modeling algorithms are crucial for their application in real-world scenarios. The evaluate_coherence function addresses this by offering a measurable indicator of an LDA model's coherence, a concept emphasized in research by [27], [19]. This metric is essential for assessing the utility of topic models.

The evaluate_coherence function aims to verify the coherence of topics produced by an LDA topic model, ensuring they are both interpretable and cohesive. This verification confirms the topics' relevance. Key parameters that influence the model's structure and its coherence evaluation include `alpha`, `beta (eta)`, and `num_topics`, as detailed in Table 3.

The LDA model is created using Gensim's LdaModel module. This module focuses on the algorithm's iteration count through the dataset during training, a vital factor for model quality. The CoherenceModel class from the Gensim python module is then used to set up a coherence_model_lda, which aids in assessing topic coherence.

Moreover, a fitness function is introduced to calculate and return the fitness values for model populations, evaluating each based on its ability to generate coherent and meaningful topics. This process, which begins with creating a fitness list to store each model's coherence scores, underscores the model's success in producing relevant and interpretable topics.

Identifier	Definition
alpha	Hyperparameter that controls the mixture of topics within documents. A lower value leads to documents containing fewer topics.
beta (eta)	Hyperparameter that controls the distribution of words across topics. A lower value means fewer words represent a topic.
num_topics	The number of topics to extract from the corpus.
corpus	A collection of documents represented as a list of lists of tuples (document, word frequency).
id2word	A dictionary mapping word IDs to words.
texts	The set of original documents, preprocessed and ready for analysis.
lda_model	A Latent Dirichlet Allocation (LDA) model built on the corpus.
coherence_model_lda	A model for calculating the coherence of the LDA model, assessing how well the inferred topics correspond to the texts.

Table 3 Identifiers and their definitions in the evaluate coherence function

4.5.3 Parent selection

Parent selection is a critical component in genetic algorithms, significantly influencing the evolutionary process by favoring individuals with higher fitness to breed the next generation. This approach promotes the evolution of the population towards better solutions by assuming that fitter individuals will likely transfer their superior traits to their offspring. The implementation leverages numpy for efficient array management, where individuals are ranked by fitness using ‘np.argsort()‘ on the `fitness` array, and the top `num_parents` are chosen for reproduction. This selection ensures the perpetuation of the most capable traits, aligning with the principles outlined in the Table 4.

Identifier	Definition
`population`	The complete set of individuals in the current generation, from which parents are selected.
`fitness`	An array or list of fitness values, each corresponding to an individual in the population, used to assess their suitability.
`num_parents`	The specified number of parents to be selected from the population for the next generation.
`parents`	The subset of the population selected as parents for the next generation, chosen based on their high fitness values.

Table 4 Identifier definitions for the parent selection function.

4.5.4 crossover

Genetic crossover, pivotal in genetic algorithms, combines traits from two parents to produce offspring, facilitating solution space exploration and potential fitness improvements in future generations. It employs a midpoint crossover technique within the offspring’s parameter vector, delineated in Table 5, to divide genetic contributions from both parents.

Offspring generation involves iterating a loop corresponding to the desired number of offspring, with each cycle creating a new individual. Parental selection for crossover uses indices from the table, ensuring equitable parent contribution and cyclic selection. Offspring results from merging genes before and after the crossover point from respective parents via np.concatenate, enabling inheritance of beneficial traits and aiming for ongoing population fitness enhancement.

Identifier	**Definition**
`parents`	An array containing the genetic information of the current generation's parents from which offspring will be derived.
`offspring_size`	A tuple specifying the desired number of offspring (`offspring_size[0]`) and the dimensionality of each offspring (`offspring_size[1]`), guiding the size and structure of the generated offspring array.
`offspring`	The resulting array of offspring generated by the crossover operation, each inheriting genetic traits from two parents.
`crossover_point`	A calculated point (half the offspring's second dimension size) that determines the split in genetic information between two parents for each offspring.
`parent1_idx`, `parent2_idx`	Index variables used to select pairs of parents for genetic material combination, ensuring a diverse genetic mix in the offspring generation.

Table 5 Identifier definitions for the crossover function in a genetic algorithm.

4.5.5 Mutation

To enhance the genetic diversity within a population and mitigate premature convergence to local optima, mutation is a pivotal mechanism in genetic algorithms. It systematically introduces genetic variations by adjusting individual parameters, thereby facilitating the exploration of new solution spaces. The mutation process is governed by a mutation rate, dictating each individual's mutation probability to ensure a dynamic yet controlled exploration.

The mutation operation commences with evaluating each individual against the mutation rate, determined by generating a random number between 0 and 1. Should this number fall below the mutation rate threshold, the individual is earmarked for mutation. A parameter for mutation is then randomly selected from the set of mutable parameters, as delineated in the table referred to by ‘6‘. For parameters requiring specific conditions—such as ’num_topics’, which necessitates an integer value—a new value is randomly assigned within the predefined range. This procedure ensures the introduction of genetic variability and adheres to the constraints of valid parameter values, thereby enriching the genetic algorithm's potential for uncovering novel solutions.

Identifier	Definition
`offspring`	The array of offspring to be mutated, representing individuals of the current generation.
`parameter_ranges`	A dictionary specifying the allowable range for each parameter that can be mutated, used to ensure mutations result in valid values.
`mutation_rate`	The probability of any given individual undergoing mutation, controlling the frequency of mutations within the population.
`param_to_mutate`	The parameter selected for mutation, chosen randomly from the keys of the `parameter_ranges` dictionary.
`index`	The position of the mutating parameter within an individual's parameter list, used when applying mutations to parameters other than 'num_topics'.

Table 6 Identifier definitions for the mutation function in a genetic algorithm.

4.5.6 Genetic diversity

Genetic diversity is essential for averting premature convergence to local optima and enhancing the algorithm's capacity to navigate the solution space efficiently. This mechanism assesses each individual's contribution to the population's genetic variance by examining their fitness against a stipulated threshold. When an individual's fitness falls below this threshold, it signifies a potential shortfall in augmenting the population's diversity. Consequently, such individuals undergo a regeneration process, receiving a new array of parameters generated within predefined limits, as detailed in Table 7.

This rejuvenation process entails assigning each parameter a random value from its allowable range, thus ensuring the revamped individual adheres to the established parameter boundaries. By substituting individuals with fitness levels beneath the set threshold with newly generated ones, this approach systematically promotes genetic heterogeneity within the population. It guarantees that the population retains a baseline level of quality while reintroducing variation, facilitating ongoing exploration and mitigating the risks of evolutionary stagnation.

Identifier	Definition
`population`	The array of individuals constituting the current generation, subject to diversity control based on fitness evaluation.
`fitness`	An array of fitness scores corresponding to each individual in the population, utilized for assessing the need for diversity introduction.
`threshold`	A predetermined fitness score threshold; individuals with fitness below this value are considered for genetic diversity enhancement.
`parameter_ranges`	A dictionary outlining the permissible range for each genetic parameter, ensuring modifications remain within viable bounds.

Table 7 Identifier definitions for the function aimed at controlling population diversity based on fitness in a genetic algorithm.

4.5.7 Elitism

Elitism in genetic algorithms preserves high-quality solutions across generations by protecting them from being lost during random selection, crossover, or mutation. This method starts by sorting individuals based on their fitness values using ‘np.argsort(fitness)‘, from lowest to highest, to identify the most fit individuals or elites. The top individuals, indicated by the ‘num_elites‘ parameter, are then selected by slicing the sorted array. These elites are extracted from the population and maintained for future generations, as detailed in Table 8. This approach ensures the continuous preservation of superior solutions, reducing the risk of their loss due to genetic operations.

Identifier	Definition
population	The collection of all individuals in the current generation.
fitness	A list of fitness scores for each individual in the population.
num_elites	The number of top performers to select.
elites_idx	The indices of the selected elites based on fitness.
elites	The selected top-performing individuals.

Table 8 Identifiers and their definitions in the elitism function

4.6 Topic Identification by State

The generation of specific topics for each state is underway. The key to this generation lies in the values obtained automatically through the implementation of the genetic algorithm. By adjusting the hyperparameters of alpha, beta, and topic, the corresponding topics for the 32 states of the Mexican Republic is obtained for each year of the analyzed period. The optimization of hyperparameters is paramount, exerting a direct influence on both the granularity and applicability of the derived topics. Ensuring that these topics are not only statistically robust but also of substantive practical relevance is critical. This rigorously tailored approach facilitates an in-depth understanding of regional nuances, empowering policymakers and researchers to discern and interpret trends and transitions in public interests and issues effectively.

4.7 Interpretation of Results

In the field of text analysis through Latent Dirichlet Allocation (LDA), we face the significant challenge of assigning accurate titles and formulating appropriate descriptions for identified topics. This challenge arises from the crucial need to interpret and synthesize the thematic complexities and patterns extracted through

LDA analysis in a coherent and concise manner. To address this issue, the study referenced in [22] employs the Generative Pre-training Transformer (GPT) model to enhance the interpretation of identified topics. This method proposes a three-phase strategy: initially, topics are identified using LDA; then, a domain expert assigns specific themes to these topics; and, finally, GPT is used to generate clear and comprehensible descriptions.

To optimize the task of interpreting and describing topics, our methodology emphasizes the detailed analysis of keywords and their associated probabilities, aiming for an in-depth understanding of each topic. The automatic generation of titles and descriptions is carried out using GPT models, complemented by a human validation phase to ensure the accuracy and relevance of the results. In this process, word clouds emerge as an indispensable tool, providing an intuitive visual representation that highlights the most significant words of each topic. This approach not only facilitates the creation of more accurate titles and descriptions but also ensures a deeper connection with the analyzed content, thereby improving the clarity, structure, and coherence of the text. Below, we present an example illustrated in Figure 5, where the word cloud reveals a variety of terms associated with environmental themes, informational content, and legal or regulatory action. Words like “resources”, “marine”, “protected”, “environment”, and “natural” suggest a focus on conserving or regulating natural or marine areas. For a more detailed interpretation, keywords and their interpretation are used, which are processed through the GPT-3.5-turbo model, as shown in Table 9.

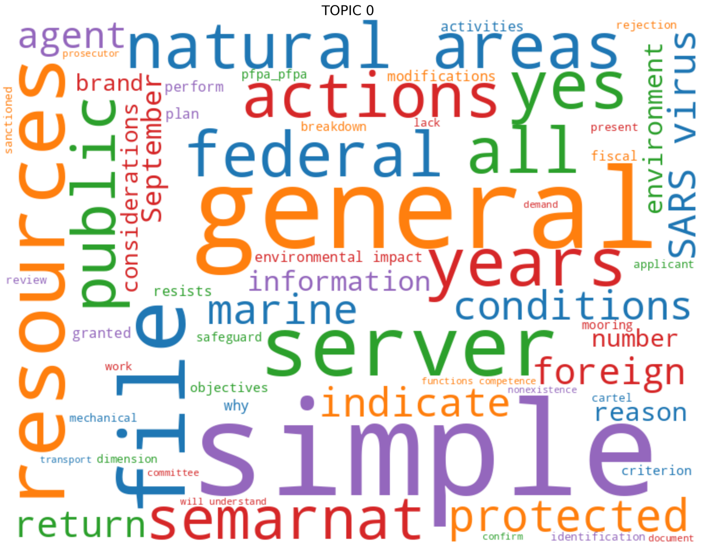

Fig. 5 Example topic: environmental resources in protected areas

Title	Keywords and Probability	Description
Management of Water Resources and Rights	simple (0.090), general (0.034), resources (0.033), yes (0.020), years (0.020), natural areas (0.018), all (0.018), actions (0.017), protected (0.015), federal (0.014), marine (0.013), indicate (0.012), SARS virus (0.011), information (0.010), September (0.010), number (0.008), foreign (0.007), return (0.007), agent (0.007), plan (0.007), environment (0.007), perform (0.007), activities (0.007), etc.	This topic suggests a strategy focused on the conservation of natural resources, highlighting the simplicity and scope of marine and terrestrial protection meaasures. It focuses on the planning and review of federal policies, possibly influenced by environmental and regulatory factors, in a context marked by digital information management and ecosystem health surveillance.

Table 9 Topic example: title and description generated by GPT-3.5-turbo.

4.8 Application of Embedding to Identify Topic Similarity

At this stage, embeddings for the topic descriptions are generated, which are vector representations of the themes' interpretation using the word embedding model known as ”sentence- embeddings-BETO”. The process involves calculating the cosine similarity between each pair of topic descriptions, setting a similarity threshold greater than 0.8. Subsequently, topics are clustered based on their similarity using the K–means algorithm, which allows for organizing and segmenting the themes along with their respective status. Finally, based on the clustered titles, a name that encompasses the topics is assigned to each group. Figure 6 shows an example of topic similarity by status.

4.9 Visualization of Results

Ultimately, our goal is to present the findings of our analysis in an intuitive and captivating way. To achieve this, we will harness various visual tools, each carefully chosen for its unique ability to simplify and effectively convey complex information. Word clouds will spotlight the dominant themes by visually emphasizing key terms based on their frequency, offering a snapshot of the dataset's core subjects. Geographic maps will contextualize spatial data, illustrating trends and distributions that reveal regional insights. In contrast, heat maps will visualize data intensity across different dimensions, pinpointing high activity or concentration areas. Additionally, similarity diagrams will map out the relationships and patterns among data entities, uncovering clusters and connections that elucidate the underlying structure of the dataset. This multifaceted visual approach aims to transform intricate datasets into accessible insights, bridging the gap between detailed analysis and practical under-

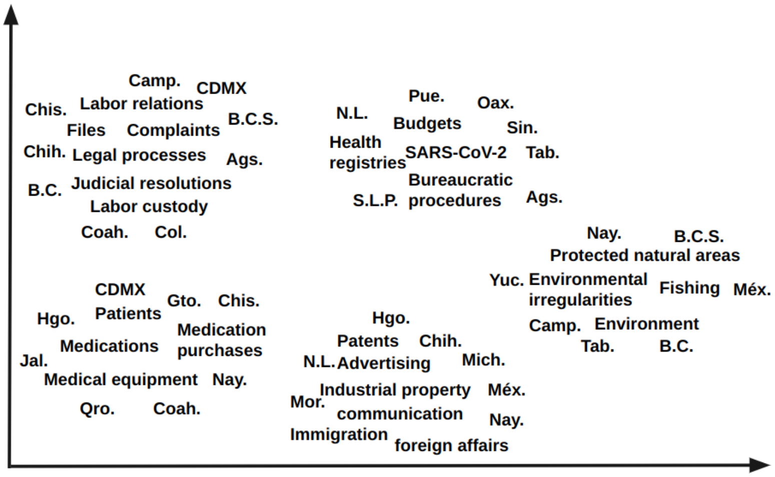

Fig. 6 Example of topic similarity.

standing for a diverse audience, thereby ensuring that our findings' significance is recognized and deeply comprehended.

5 Results

The Figure 7 illustrates the fitness evolution within a genetic algorithm over 50 generations, applied to a dataset of 2,580 instances from the state of Puebla. An upward trend is noticeable in the early stages, suggesting that the algorithm was capable of quickly identifying promising candidates and significantly improving the quality of solutions compared to the initial generations. Following this initial increase, the fitness curve stabilizes, which is typical in genetic algorithms, as the population tends to converge towards a local or global optimum solution.

The best individual that emerged from this process exhibited a parameter configuration with an alpha of 0.18 and a beta of 0.81, handling 8 different topics in its model. This set of parameters reached a coherence value of 0.73, which is a robust indicator that the solution found is significant and well-structured, according to the metrics of the LDA model used to measure the coherence among topics. The coherence value is particularly notable, as it indicates strong relevance and distinction between the topics generated by the LDA model. This means that the topics are mutually exclusive and each captures a distinct set of information from the dataset.

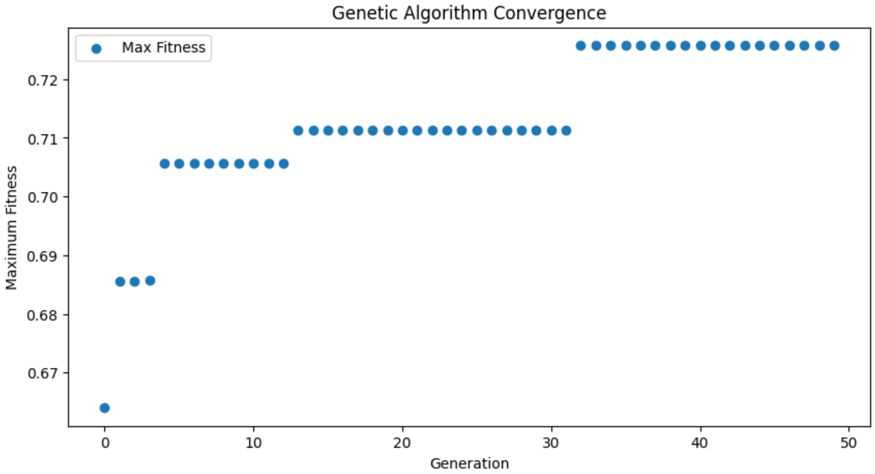

Fig. 7 Evolution of performance across generations.

To obtain results and validate the effectiveness of the the proposed methodology, data from the year 2020 were analyzed, which included a total of 97,360 public information requests. In Table 2, the values of the hyperparameters alpha, beta, and topic are presented, along with their respective coherence values for each federal entity. The detailed analysis of this table reveals an interesting trend: although in some cases it is observed that a higher number of requests corresponds to an increase in the number of topics, this relationship does not remain uniform across all instances. This finding suggests variability in the distribution of topics that is not directly correlated with the volume of requests in all entities.

Table 10 Hyperparameter Results by Entity in 2020.

Entity	Num. of Requests	Alpha	Beta	N Topics	Coherence score
Ags.	533	0.28	0.49	5	0.62
B.C.	2,540	0.68	0.22	9	0.84
B.C.S.	483	0.39	0.43	4	0.70
Camp.	205	0.9	0.39	4	0.74
Chis.	370	0.71	0.66	5	0.61
Chih.	1,354	0.85	0.87	4	0.52
CDMX	56,091	0.26	0.79	56	0.82
Coah.	651	0.56	0.45	10	0.54
Col.	234	0.67	0.47	6	0.59
Dgo.	346	0.85	0.6	3	0.60
Mex.	9,740	0.28	0.75	9	0.66
Gto.	916	0.9	0.17	9	0.50
Gro.	351	0.86	0.54	7	0.65
Hgo.	859	0.77	0.46	4	0.51
Jal.	4,109	0.92	0.6	5	0.46
Mich.	386	0.2	0.19	3	0.64
Mor.	904	0.99	0.14	5	0.52
Nay.	633	0.3	0.61	3	0.51
N.L.	1,111	0.91	0.95	7	0.45
Oax.	554	0.84	0.35	6	0.49
Pue.	2,580	0.18	0.81	8	0.73
Qro.	1,503	0.76	0.84	9	0.54
Q. Roo	847	0.18	0.76	4	0.60
S.L.P.	410	0.53	0.93	3	0.64
Sin.	1,590	0.51	0.6	7	0.55
Son.	1,139	0.91	0.27	6	0.63
Tab.	2,150	0.23	0.48	3	0.55
Tamps.	1,003	0.72	0.84	7	0.58
Tlax.	480	0.05	0.81	4	0.68
Ver.	1,767	0.85	0.1	7	0.48
Yuc.	1,339	0.72	0.94	8	0.58
Zac.	182	0.29	0.07	5	0.51

Source: Own elaboration with data from INAI (2023).

In Figure 8, a detailed analysis of the predominant topics in the state of Baja California throughout the year 2020 is presented. Based on the analysis of keywords and their respective probabilities, it is identified that Topic 0 focuses on the management of resources in protected natural areas and environmental issues. This topic encompasses crucial aspects such as biodiversity preservation and marine ecosystem protection.

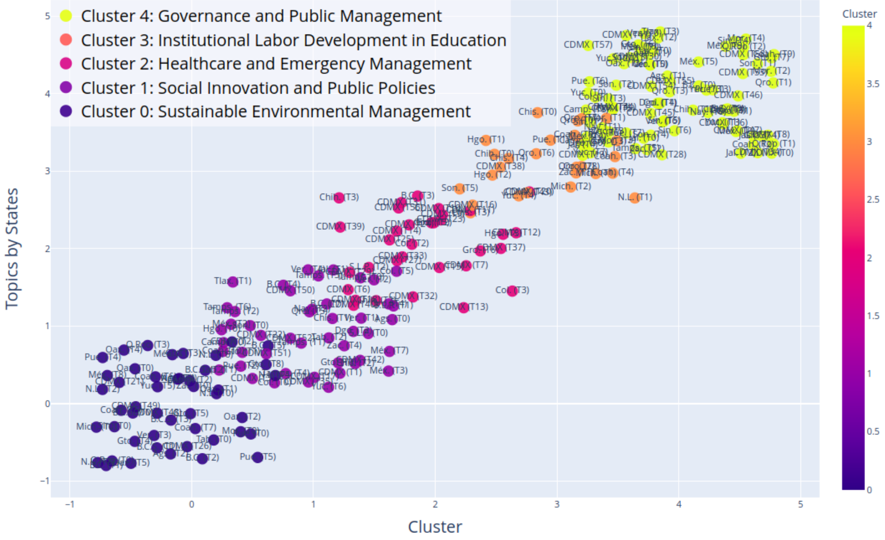

Fig. 8 Topics from the state of Baja California during 2020.

The various topics identified in the analysis weave together a comprehensive narrative on the intricate dynamics of social, environmental and economic activities. Topic 1 unveils irregularities in social and environmental resources, spotlighting key aspects such as results, projects, programs, and expenditures, setting the stage for a deep dive into governance and resource management. Transitioning smoothly into Topic 2, the focus shifts to the fishing industry, encapsulated through reports and fishing records, indicating a specialized examination of this sector's operational facets.

As the narrative progresses, Topic 3 brings to light discussions from social cabinet sessions and presidential decisions, with keywords like discussions and agreements underlining the political and decision-making processes that impact both social and environmental policies. This the political and administrative lens is complemented by Topic 4, which narrows down on fishing activities during October and September, especially in the coastal areas of Campeche and Veracruz and the challenges posed by the SAR virus, reflecting on the sector's resilience and adaptability to health crises.

Further deepening the analysis, Topic 5 explores the SAR virus's impact on the fishing sector and the health of workers in August, revealing the virus's broader implications on occupational health and sector stability. Topic 6 then takes a different turn, focusing on the development project on Salsipuedes Island, where aspects such as transportation plans, budgets and execution details highlight the logistical and financial planning essential for environmental and infrastructural projects.

In addressing the accountability and labor aspects, Topic 7 delves into project responsibility and labor demands, emphasizing the need to identify responsible parties and manage labor and costs efficiently, showcasing the human resource and financial aspects critical to the project success. The narrative culminates with Topic 8, which zooms in on the details about employees, including names, salaries, and expenses, providing a granular view of the workforce component in these endeavors. Together, these topics paint a multifaceted picture of the challenges and considerations involved in managing social, environmental, and economic activities within a complex and interconnected framework.

By applying embedding techniques to the topic descriptions and their subsequent grouping using the K-Means algorithm based on similarities, as illustrated in Figure 9,we have achieved significant results, especially in thematic categorization by geographical location.

The first cluster, related to Environmental Management and Sustainable Development encompasses a wide variety of topics focusing on environmental management, conservation of natural resources, sustainable development, and regulations in various sectors, such as energy, agriculture, tourism, and water management. Topics range from the authorization and management of activities in official bodies to the environmental impact assessment, development of national projects, water resource management, species conservation, and environmental compensation strategies in the gas industry. It also includes aspects related to sustainability in fishing, regional development in Oaxaca, handling of clandestine operations, and environmental protection in urban and rural projects and developments.

The second cluster, Social Innovation and Public Policies in Mexico, addresses a broad spectrum of topics related to implementing and evaluating social programs, social security management, labor and migration dynamics, and other key aspects of public policy and administration in Mexico.

The third cluster, Healthcare Management and Emergency Response, focuses on a series of topics primarily linked to public health management, sanitary and commercial regulations, the impact of the COVID- 19 pandemic on various sectors, and the distribution and oversight of medications. It also covers aspects related to public transport, import and export regulations, and protective measures and management during health emergencies. The topics reflect a concern for ensuring the population's well-being, adapting policies and procedures to emerging challenges, and maintaining effectiveness in the management of essential services in times of crisis.

The fourth cluster, Institutional and Labor Development, focuses on a wide range of topics related to institutional and labor management, electoral processes, labor regulations, professional and educational development, and the impact of community and educational activities on various societal sectors. Topics span from public administration and general management in specific regions like Michoacán and Coahuila to the professional integration of students and interns through to specific challenges in higher secondary education and the management of educational projects.

The fifth cluster centers on topics related to public information management, copyright, document management in public entities, sanitary regulations, transparency, auditing, and control in various public sector areas, including education, health, and judicial. It also addresses resource and contract management, anti-money laundering efforts, import and export regulations, labor rights, and the management of files and official documentation, among others.

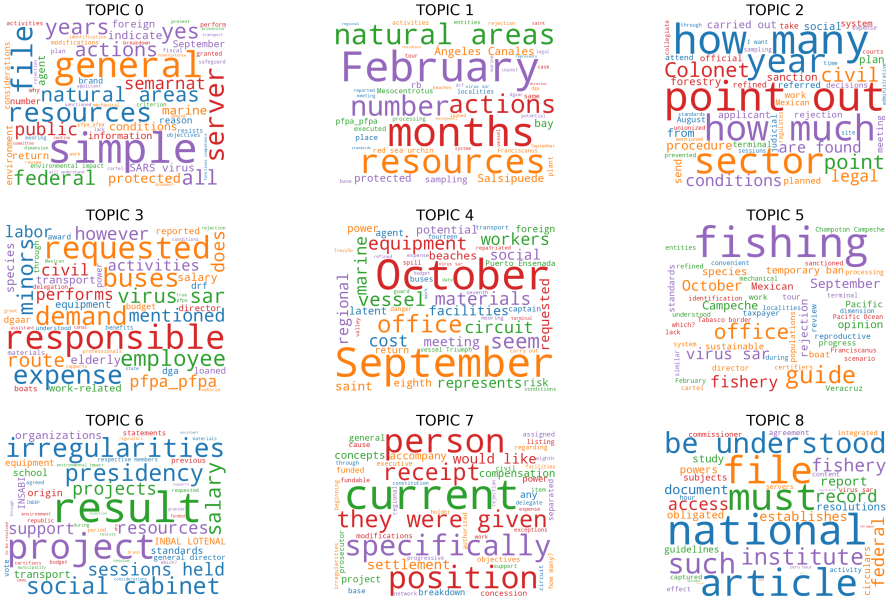

Fig. 9 Similar topics by state during 2020.

6 Discussion

The genetic algorithm has shown remarkable effectiveness in optimizing the LDA model's Alpha, Beta, and Topic parameters, significantly boosting its ability to analyze data. In parallel, embedding techniques have revealed thematic connections, offering great potential for developing proactive transparency strategies and improving open data access. This innovative approach, particularly applied to data from the National Transparency Platform, has unveiled hidden patterns and trends, marking a substantial step forward in enhancing transparency and the accessibility of public information in Mexico.

This method has identified 4 131 topics across various federal entities from 2003 to 2020 by optimizing key LDA hyperparameters. These findings are crucial for meeting citizen needs and promoting transparency and open data availability, showcasing remarkable progress in processing and analyzing large volumes of government data. Furthermore, these advancements empower the government to analyze and utilize data more effectively and bolster public trust by making governmental processes more transparent and accessible. This approach serves as a benchmark for refining data management practices. It catalyzes increased public engagement in governance, strengthening the foundations of democracy and accountability in Mexico.

7 Conclusion and Future Research Directions

In conclusion, this study has successfully demonstrated the feasibility of automating the selection of hyperparameters (alpha, beta, and the number of topics) in the Latent Dirichlet Allocation (LDA) model through the implementation of a genetic algorithm, using topic coherence as the fitness criterion. This methodology has efficiently identified 4,131 relevant topics annually and by state throughout the analyzed period. Moreover, the integration of GPT models for generating titles and descriptions of topics has facilitated their validation by transparency experts, achieving significant time savings. Although the analysis encountered limitations in states with few requests or little thematic variability, this work lays the groundwork for applying it to any unstructured data set requiring topic identification.

The methods and insights from this study could be applied across a wide range of research fields, such as social media analysis, market research, and biomedical literature review, where understanding large volumes of text data is crucial. The potential real-world applications of our work are vast, including enhancing search engine algorithms, improving recommendation systems, and aiding in curating content for personalized newsfeeds, significantly impacting how information is organized and retrieved.

One limitation of our study is the reliance on available digital data, which may not fully represent the spectrum of topics of interest, particularly in areas with lower digital footprints.

Future research could focus on adapting the genetic algorithm to address the issue of redundant topics in LDA models by incorporating mechanisms that penalize or eliminate overlapping topic clusters. This adjustment could significantly improve the distinctiveness and relevance of the identified topics. Additionally, efforts could be directed toward fine-tuning the algorithm to minimize the inclusion of topics characterized by words with low probability, thereby enhancing the overall coherence and interpretability of the topics. Exploring these modifications would refine the precision of topic identification and contribute to a deeper understanding of topic distribution and separation in large datasets.

Drawing from these findings, future comparisons of this methodology with other topic modeling techniques, such as BERTopic, represent a promising direction to expand these findings. Additional research could also explore the application of these methodologies to different datasets or within varied geopolitical landscapes, which could unveil distinct patterns of public interest and information solicitation on a global scale.

The topics of public interest identified in this study, including environmental management, public policy initiatives, and responses to health emergencies, underscore a significant opportunity for shaping public policy formulation. Future endeavors could investigate avenues through which these insights can be systematically woven into the fabric of policy-making processes, ensuring that governmental actions are in tight congruence with the populace's concerns, thereby fostering policies that are both responsive and reflective of the citizens' needs.

Data Availability Statement

All data and code pertaining to this study are published by INAI. To access the datasets, please visit https://www.plataformadetransparencia.org.mx/. For the associated code, visit https://github.com/hermecp/AutoTopicGen-INAI (last accessed on May 29, 2024).

References

1. Aguilar Miranda, A., Ramírez González, K.: La plataforma nacional de transparencia en méxico y la gestión municipal. In: XXIV Congreso Internacional del CLAD sobre la reforma del Estado y de la Administración Pública. Buenos Aires, Argentina (2019)
2. Bagozzi, B.E., Berliner, D., Almquist, Z.W.: Predicting government (non) responsiveness to freedom of information requests with supervised latent dirichlet allocation (2016). URL https://api.semanticscholar.org/CorpusID:43093099
3. Bagozzi, B.E., Berliner, D., Almquist, Z.W.: When does open government shut? predicting government responses to citizen information requests. Regulation and Governance (2019). DOI: 10.1111/rego.12282
4. Bautista-Farías, J.: La nueva ley general de transparencia: alcances y retos. DFH - Revista Análisis Plural (2015)
5. Berliner, D., Bagozzi, B.E., Palmer-Rubin, B.: What information do citizens want? evidence from one million information requests in mexico. World Development (2018). DOI: 10.1016/j.worlddev.2018.04.016
6. Berliner, D., Bagozzi, B.E., Palmer-Rubin, B., Erlich, A.: The political logic of government disclosure: Evidence from information requests in mexico. The Journal of Politics **83**, 229 – 245 (2020). DOI: 10.1086/709148
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
8. Caballero, J.A., Carbonell, M., Fix-Fierro, H., López Ayllón, S., Roldán Xopa, J., Salazar Ugarte, P.: El futuro del instituto federal de acceso a la información pública y protección de datos personales. consideraciones sobre su autonomía constitucional. México, UNAM (2012)
9. Cárdenas Sánchez, J., Gildo de la Cruz, M.G.: Eficacia institucional de los organismos independientes pro-rendición de cuentas: el caso del instituto nacional de transparencia, ac-

ceso a la información y protección de datos personales (inai). Estudios En Derecho a La Información **1**(17), 35–59 (2023). DOI: 10.22201/iij.25940082e.2024.17.18781. URL https://doi.org/10.22201/iij.25940082e.2024.17.18781

10. De Diputados, C.: Constitución política de los estados unidos mexicanos. México: Cámara de diputados (2012)
11. DOF: Ley general de transparencia y acceso a la información pública. Recuperada de http://www.diputados.gob.mx/LeyesBiblio/pdf/LGTAIP.pdf (2015)
12. Goldberg, D.E.: Optimization, and machine learning. Genetic algorithms in Search (1989)
13. Kuri-Morales, A.F., Aldana-Bobadilla, E., López-Peña, I.: The best genetic algorithm ii: A comparative study of structurally different genetic algorithms. In: Mexican International Conference on Artificial Intelligence, pp. 16–29. Springer (2013)
14. López Ayllón, S.: El derecho a la información como derecho fundamental. Derecho a la información y derechos humanos (2000)
15. López-Ayllón, S.: La creación de la ley de acceso a la información en méxico: una perspectiva desde el ejecutivo federal. Hugo A. Concha Cantú, Sergio López Ayllón y Lucy Tacher Epelstein,(Coords.), Transparentar al Estado: La experiencia Mexicana de Acceso a la Información, Instituto de Investigaciones Jurídicas, UNAM (2005)
16. Mitchell, M.: An introduction to genetic algorithms. MIT press (1998)
17. Nash Rojas, C., Chacón Fregoso, G., Rodríguez Atero, M.: Estudio comparado sobre el impacto que tienen las instituciones que resguardan el acceso a la información pública en chile y méxico sobre los derechos humanos en la ciudadanía. Available at https://repositorio.uchile.cl/handle/2250/142492 (2016)
18. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108 (2010)
19. Omar, M., On, B.W., Lee, I., Choi, G.S.: Lda topics: Representation and evaluation. Journal of Information Science **41**(5), 662–675 (2015). DOI: 10.1177/0165551515558783
20. Pathik, N., Shukla, P.: Simulated Annealing Based Algorithm for Tuning LDA Hyper Parameters, pp. 515–521 (2020). DOI: 10.1007/978-981-15-4032-5_47
21. Piantadosi, S.T.: Zipf's word frequency law in natural language: A critical review and future directions. Psychon. Bull. Rev. **21**(5), 1112–1130 (2014). DOI: 10.3758/s13423-014-0585-6
22. Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P., Kaymak, U.: Towards interpreting topic models with chatgpt. In: The 20th World Congress of the International Fuzzy Systems Association (2023)
23. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining, pp. 399–408 (2015). DOI: 10.1145/2684822.2685324
24. Ruelas, A.: La transparencia en méxico: un trabajo colectivo. Consultada en http://library. fes. de/pdf-files/bueros/mexiko/12451. pdf (2016)
25. Salas Suárez, J.: El papel de los órganos garantes del acceso a la información pública en el contexto del Estado abierto. No. 44751 in Libros de la CEPAL. Naciones Unidas Comisión Económica para América Latina y el Caribe (CEPAL) (2017). URL https://ideas.repec.org/b/ecr/col015/44751.html
26. Sandoval Ballesteros, I.E.: Leyes de acceso a la información en el mundo. Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (2008)
27. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pp. 952–961 (2012)
28. Tekin, Y.: Optimization of lda parameters. 2020 28th Signal Processing and Communications Applications Conference (SIU) pp. 1–4 (2020). DOI: 10.1109/SIU49456.2020.9302034
29. Instituto Nacional de Transparencia, A.a.l.I.y.P.d.D.P.I.: Plataforma infomex (2007). URL https://www.infomex.org.mx/gobiernofederal/homeOpenData.action. Plataforma utilizada para solicitar información pública en México.

30. Instituto Nacional de Transparencia, A.a.l.I.y.P.d.D.P.I.: Plataforma nacional de transparencia (2016). URL https://www.plataformadetransparencia.org.mx/web/guest/datos_abiertos. Portal para el acceso a información pública y datos abiertos del Gobierno de México.
31. Zbigniew, M.: Genetic algorithms+ data structures= evolution programs. Comput Stat pp. 372–373 (1996)