Diseño De Una Metodología De Análisis De Texto
-Una Implementación Desde La Complejidad Y Teoría De Redes Complejas-

Edilberto Mejía Meza

Universidad De Pamplona

Facultad De Ingenierias Y Arquitectura

Programa De Ingeniería De Sistemas

Pamplona, 2019

Diseño De Una Metodología De Análisis De Texto
-Una Implementación Desde La Complejidad Y Teoría De Redes Complejas-

Edilberto Mejía Meza

Director
Yesid Alexander Madrid Carrillo

Codirector
Nelson Fernández Parada
PhD Ciencias Aplicadas

Universidad De Pamplona

Facultad De Ingenierias Y Arquitectura
Programa De Ingeniería De Sistemas
Pamplona, 2019

Dedicatoria

A mi hermana que con el sacrificio que ha hecho día a día para que mi sueño de ser profesional haga realidad, que sin ella no lo hubiese podido lograr. A mis padres y hermanos por el apoyo constantemente.

Agradecimientos

Gracias a Dios por llegar esta meta, ya que con sus bendiciones he podido culminar cada uno de los procesos durante mi carrera

También agradezco a mi familia por ser mi fuente inspiradora para el alcance de este logro Apoyo, por su acompañamiento y esfuerzos durante este largo proceso de formación.

De antemano también agradezco a mi hermana Mercy Mejia Meza que con sus esfuerzos ha sido el motor principal para mí ya que con sus sacrificios y apoyo incondicional empecé hacerme profesional y cumplir mi sueño de ser in gran Ingeniero de Sistemas.

A mi director de Trabajo de Grado Yesid Madrid, por su dedicación y orientación brindándonos las bases necesarias durante el desarrollo investigativo.

Al programa de Ingeniería de Sistemas y su cuerpo docente por guiar nuestro proceso de formación personal y profesional.

Resumen

El presente trabajo consiste en desarrollar una metodología para el análisis textual aplicando nociones de minería de texto, redes complejas y complejidad. La técnica propuesta en el trabajo tiene como objetivo brindar una herramienta que permita el análisis estadístico por medio de Big Data en una base de datos de texto, que más adelante será evaluada a través de un prototipo codificado en R con el fin de utilizar las ventajas que este lenguaje provee respecto a la estadística. La herramienta realizará la minería de texto de documentos seleccionados en los que es necesaria su limpieza y corrección, debido a errores de tipo ortográfico. Los datos obtenidos, serán procesados de manera estadística, será calculada su complejidad y se visualizarán a través de redes complejas. El análisis automatizado utilizará técnicas de aprendizaje automático (Machine Learning), inteligencia artificial y técnicas de ciencia de datos. Con ello se logrará el tratamiento ágil de textos de difícil lectura y tratamiento a través de la hibridación de técnicas en un ambiente de software libre.

Palabras claves: Redes Complejas, Complejidad, minería de texto, Auto-Organización, Redes de Co-ocurrencia, Emergencia, Preprocesamiento.

Abstract

The present work consists in developing a methodology for textual analysis applying notions of text mining, complex networks and complexity. The technique proposed in the work aims to provide a tool that allows statistical analysis through Big Data in a text database, which will later be potentially through an R-coded prototype in order to use the advantages that this language provides regarding statistics. The tool used for text mining of selected documents in which cleaning and correction is necessary, due to spelling errors. The updated data will be processed statistically, its complexity will be calculated and displayed through complex networks. Automated analysis will use machine learning techniques, artificial intelligence and data science techniques. With this, the agile treatment of difficult-to-read texts and treatment will be achieved through the hybridization of techniques in a free software environment.

Keywords: Complex Networks, Complexity, text mining, Self-Organization, Co-occurrence Networks, Emergency, Preprocessing.

Tabla de Contenido

Lista de figuras

Justificación

¿Cómo realizar análisis textual utilizando las nociones de complejidad y redes complejas?

Es conocido el grado de complejidad que tiene el ser humano al momento de analizar un gran volumen de información, en algunas ocasiones es necesario buscar soluciones que ayuden a optimizar los trabajos relacionados con este gran volumen de datos, pero que no cumplen con las expectativas. El manejo de los datos por medio del Big Data es un área que ha ido tomando fuerza en el mundo de la investigación, es por eso que trabajos adyacentes a esta área tienen con facilidad alto grado de impacto en la sociedad.

La propuesta encontrada en este trabajo, se concentra en la formulación de una metodología de análisis de texto usando nociones de minería de texto, redes complejas y complejidad. Estas técnicas, presentan una novedad ante la posibilidad de generar un análisis estadístico más completo de acuerdo a las necesidades del usuario, ya que actualmente no es conocido en Colombia un procedimiento similar en la industria que ayude a la evaluación de datos por medio del análisis estadístico. El trabajo tendrá por otra parte un caso de aplicación por medio de un prototipo codificado en R para evaluar el funcionamiento de la metodología.

Objetivos

Objetivo General

Diseñar una metodología de análisis textual utilizando las nociones de complejidad y redes complejas.

Objetivos Específicos

Desde el diseño e implementación de técnicas mejoradas de minería de texto, y la aplicación de medidas formales de emergencia, auto-organización y complejidad, sumada a la implementación de redes complejas este proyecto pretende específicamente:

1. Identificar aspectos relacionados con la obtención de información a partir de texto.
2. Estudiar las técnicas de minería de datos aplicadas al texto.
3. Estudiar las nociones de complejidad y redes complejas teniendo en cuenta sus aplicaciones en el análisis estadístico.
4. Diseñar una metodología estadística utilizando las nociones estudiadas.
5. Evaluar la metodología por medio de un prototipo codificado en el lenguaje R.

1. Marco Referencial

1.1 **Descubrimiento de conocimiento en texto (KDT, Knowledge Discovery in Text)**

Tradicionalmente, el descubrimiento de conocimiento se ha venido realizando sobre los datos almacenados en base de datos estructuradas. Sin embargo, la mayoría de la información de las que disponen los organizadores está en formato textual y, puede estar en páginas web, informes de trabajo, publicaciones, correos electrónicos, etc [1].

El procesamiento de esta información suele ser complejo debido a la gran cantidad de documentos [2], su heterogeneidad y su falta de estructura [3]. Durante los últimos años, se han realizado operaciones sobre documentas tales como catalogación, se han generado referencias, se han creado índices, se han extraído términos relevantes y se han extraído resúmenes con el fin de agilizar las búsquedas de información sobre ellas sin tener que volver a leer y estudiar el documento [1].

Según Frawley, en su artículo [4] define el descubrimiento de conocimiento como la extracción de información no trivial, implícita, previamente desconocida y potencialmente útil a partir de los datos. En nuestro trabajo, adaptaremos esta definición considerando que los datos son textuales.

Es precisamente esta consideración de datos textuales y no de datos estructuradas de bases de datos la que hace emerger una serie de problemas que necesitan un tratamiento adicional, pudiendo requerir algún método o proceso que las trate convenientemente [1]. Estos problemas son:

La falta de estructura del texto. Incluso en lenguajes semiestructurados como el HTML, el texto sigue siendo carente de una estructura homogénea procesable de forma automática sin que se produzca pérdida de información [1].

La naturaleza heterogénea y distribuida de los documentos. Como ya hemos mencionado, las fuentes externas que nutren los almacenes de texto pueden ser tan diversas como podamos imaginar: intranets, bases de datos documentales, redes sociales, censos, páginas web, informes empresariales, publicaciones, correos electrónicos, etc [1].

El multilingüismo presente no sólo en diferentes conjuntos de documentas, sino también dentro de una misma colección de textos [1].

El análisis de texto depende del contexto y del dominio de la aplicación, lo cual implica el uso de diccionarios, tesaurus u ontologías específicas de dicho contexto para poder llevar a cabo el procesamiento de los textos [1].

El proceso de Descubrimiento de Conocimiento en Textos implicará, por lo tanto, a diferentes ámbitos de conocimiento, principalmente: Recuperación de Información (para filtrar y reunir documentos adecuados), Extracción de Información (que selecciona hechos específicos sobre tipos de entidades y relaciones de interés), procesamiento del Lenguaje Natural (para realizar el preprocesamiento y etiquetado de los textos) y Minería de Datos (para descubrir asociaciones desconocidas entre hechos desconocidos) [1].

1.1.1 **Fases del KDT**

1.1.1.1 Limpieza de datos

Fase del proceso en la que se elimina el ruido y los datos inconsistentes obtenidos de las fuentes externas. Para ello se utilizan casi exclusivamente métodos estadísticos: histogramas (detección de datos anómalos)

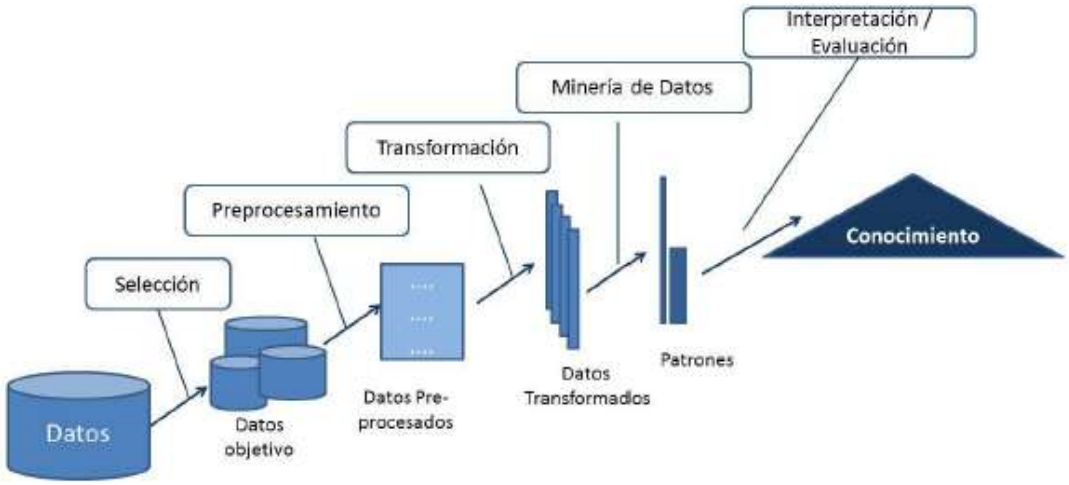


Figura 1 Fase del Proceso del KDT (Fayyad et al., 1996])

selección de datos (muestreo, ya sea verticalmente, eliminando atributos, u horizontalmente, eliminando tuplas), redefinición de atributos (agrupación o separación) [1].

Hay dos situaciones irregulares a considerar en esta fase debifo a datos anómalos o datos que faltan. Las actuaciones ante el primer tipo de datos pueden ser: ignorarlos, filtrar la columna (puede ser útil no eliminarla sino sustituirla por una columna con datos discretos donde el valor anómalo se identifique como tal), filtrar la fila o reemplazar el valor. Cuando se está ante una situación de falta de datos, una posible solución puede ser esperar a que estos datos estén disponibles [1].

1.1.1.2 Integración de datos

Proceso en el que se combinan múltiples fuentes de datos externas y heterogéneas. Uno de los principales caballos de batalla con los que se topa el analista cuando intenta realizar KDD es el de las fuentes externas, normalmente heterogéneas, que nutren los almacenes de datos y que pueden ser tan diversas como [1]:

- Bases de datos relacionales de trabajo diario dentro de la empresa.
- Datos del área geográfica, de economía.
- Bases de datos externas de otras organizaciones.
- Censos, información extraída de Internet, directorios.

1.1.1.3 Selección de Datos

No todos los datos almacenados en bases de datos serán útiles para extraer Información En esta fase se recuperan de la de datos los datos relevantes para la tarea de análisis [1].

1.1.1.4 Transformación de Datos

Los datos se transforman o se consolidan en formas apropiadas para la minería. Este proceso implica perder las relaciones de integridad y la normalización. Cuando se trabaja con Data Marts, se puede optar por un modelo en estrella o un modelo de copo de nieve, que agilizan las consultas posteriores y permiten buscar sobre información resumida [1].

1.1.1.5 Minería de Datos

Fase principal del proceso KDD Se aplican métodos inteligentes para extraer patrones de datos Habrá que elegir los algoritmos de minería adecuados en fucinón de datos y del tipo de información que se quiere descubrir [1].

1.1.1.6 Evaluación de los Patrones

Extraídos en la fase anterior Se identifican los patrones interesantes [1].

1.1.1.7 Representación del conocimiento

Visualización del conocimiento obtenido en la minería [1].

De forma general, podemos resumir las fases principales del KDT en estas tres: Preprocesamiento, Minería de Textos y Visualización. Hay que tener en cuenta, no obstante, que en la literatura está mucho más extendido el término Minería de Textos que el término Descubrimiento de Conocimiento en Textos. Esto es debido a que la mayoría de las veces se utiliza la fase Minería de Textos para identificar al proceso de descubrimiento completo. Este fenómeno nace de una metonimia semántica: la fase principal del proceso identifica al proceso genérico [1].

En la siguiente sección presentamos las fases principales del proceso del Descubrimiento de Conocimiento en Textos: la fase de Preprocesamiento, la Minería de Textos y la fase de Visualización [1].

1.1.2 Fases principales del KDT

1.1.2.1 Preprocesamiento de los Documentos: cómo Obtener una Forma Intermedia Adecuada

La reprocesamiento de datos es un paso preliminar durante el proceso de minería de datos. Se trata de cualquier tipo de procesamiento que se realiza con los datos brutos para transformarlos en datos que tengan formatos que sean más fáciles de utilizar. [22]

El propósito del preprocesamiento de datos es principalmente corregir las inconsistencias de los datos que serán la base de análisis en procesos de minería de datos. En el caso de las fuentes de datos estructuradas, el propósito no es distinto y pueden ser aplicadas diversas técnicas estadísticas y de aprendizaje computacional. [23]

Ya sabemos que el texto no presenta una estructura fácil para aplicarle las técnicas de Minería de Textos directamente, así que habrá que realizar una serie de operaciones sobre él hasta conseguir una representación adecuada a nuestras necesidades [1].

Para conseguir realizar la ansiada minería, antes hay que preparar el texto de forma que admita ser procesado por técnicas de extracción de patrones útiles. No basta sólo con homogeneizar dichos documentos (procedentes de distintas fuentes de información y que pueden estar en distintos formatos e idiomas), ya que, si nos quedamos aquí, se trataría de un procesamiento de textos avanzado, que puede llevarse a cabo mediante lingüística computacional [1]. Sin embargo, y como indica Hearst, la lingüística computacional trata sobre la comprensión del lenguaje, computa estadísticas sobre grandes colecciones de texto para

descubrir patrones útiles que se usan para informar subproblemas de procesamiento de lenguaje natural: etiquetas de part of speech, desambiguedad del sentido de las palabras y creación de diccionarios bilingües; no obstante, los patrones que se obtienen en la lingüística computacional carece del valor desde el punto de vista empresarial. Lo interesante son los patronse que aportan conocimiento no explícito en el texto, como ocurre con la Minería de Textos [1].

La fase en la que realizamos operaciones sobre el conjunto de documentos objeto de estudio, es el Preprocesamiento (algunos autores, como por ejemplo Tan la llaman Text Refining). Este paso es primordial ya que, en función de cómo representemos los datos, el resultado de la minería puede variar. Este problema queda reflejado en la Figura2, donde podemos observar que, si en la fase de Preprocesamiento se elige una técnica determinada, ésta establecerá el tipo de información obtenida en el descubrimiento [1].

Teniendo en cuenta esta determinación, habrá que elegir cuidado qué técnicas

Preprocesamiento	Representación	Descubrimiento
Categorización	Vector de tópicos	Relaciones entre tópicos
Análisis de texto completo	Sequenciamiento de palabras	patrones de lenguaje
Extracción de Información	Tabla de base de datos	Relaciones entre entidades

Figura 2 Relación entre Preprocesamiento, Representación Interna y Descubrimiento ([Montes y Gomes et al., 2002)

Figura 3 Fases por las que pasa un documento ([S. liritano and Rullo, 2003])

Utilizaremos para darle una representación interna al texto. Se pueden definir estructuras menos simples que pueden ser extraídas del texto con un coste razonable y de forma automática. Esta estructura debe conservar la riqueza del documento y debe permitir que se realicen operaciones sobre ella. [1]

1.1.2.2 Minería de Textos

La minería de texto es un área de investigación del procesamiento automático de la información. Se define como el proceso de descubrimiento de patrones interesantes y nuevos

conocimientos en una colección de textos, es decir, es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos. [19]

Para Witten, la minería de texto es el proceso de analizar escritos o conjuntos de enunciados para extraer información que resulta útil para propósitos particulares. [20]

La minería de textos es el proceso de analizar colecciones de materiales de texto con el objeto de capturar los temas, conceptos claves y comunes. Descubrir las relaciones ocultas y las tendencias existentes sin necesidad de conocer las palabras o los términos exactos que los autores han utilizado para expresar dichos conceptos. La minería de textos y la acción de recuperar información son conceptos que a veces se confunden, aunque son bastante diferentes. Una recuperación precisa de la información y su almacenamiento supone un reto importante, pero la extracción y administración de contenido de calidad, de terminología y de las relaciones contenidas en la información son procesos cruciales y determinantes. [8]

Figura 4 Actividades del Proceso Minería de Textos. *Tomado de* **(Salamanca, 2018)**.

Minería de textos basada en lingüística, por otro lado, aplica los principios de procesamiento de lenguaje natural NLP, análisis asistido por sistema de lenguajes humanos, al análisis de palabras, frases y sintaxis, o estructura, del texto. Un sistema que incorpora tecnología NLP puede extraer conceptos de forma inteligente (incluidas frases compuestas). Además, el conocimiento del lenguaje subyacente permite la clasificación de conceptos en grupos relacionados (como por ejemplo, productos, organizaciones o personas) utilizando el significado y el contexto[1].

1.1.2.2.1 Técnicas de minería de texto

A continuación, se presentan las diferentes técnicas de minería de texto para obtener conocimientos de los datos en los documentos:

1.1.2.2.1.1 Reglas de asociación

La generación de reglas de asociación es una técnica potente de minería de datos utilizada para buscar en un conjunto de datos, por reglas que revelan la naturaleza y frecuencia de las relaciones o asociaciones entre las entidades de los datos. Las asociaciones resultantes pueden ser utilizadas para filtrar la información por análisis humano y posiblemente definir un modelo de predicción basado en el comportamiento observado [7].

1.1.2.2.1.2 Métodos de agrupamiento

Es utilizado en el paso de pre-procesamiento de los datos, debido a la característica de aprender semejanzas sin supervisión entre objetos y reducir el espacio de búsqueda a un conjunto de los atributos más importantes parta la aplicación o a un conjunto finito de objetos. El método más frecuentemente utilizado para agrupar es el k-means el cual identifica un cierto número de grupos u objetos similares el cuál puede ser utilizado conjuntamente con el método de la Vecindad más próxima (K-Nearest Neighbor k-NN), esta técnica coloca un objeto de interés dentro de clases o grupos examinando sus atributos y agrupándolo con otros cuyos atributos son cerrados a él. k-NN es una técnica clásica para descubrir asociaciones y secuencias cuando los atributos de los datos son numéricos. Con atributos no numéricos o variables es difícil aplicar esta técnica por la dificultad de definir una medida que pueda ser utilizada para cuantificar la distancia entre un par de valores no numéricos [7].

1.1.2.2.1.3 Árboles de decisión

Un árbol de decisión es una estructura en forma de árbol que visualmente describe una serie de reglas (condiciones) que causan que una decisión sea tomada [7].

1.1.2.2.1.4 Algoritmos genéticos

Los algoritmos genéticos son técnicas de optimización que pueden ser utilizadas para mejorar otros algoritmos de minería de datos obteniendo como resultado el mejor modelo para una serie de datos. El modelo resultante es aplicado a los datos para descubrir patrones escondidos o para realizar predicciones [7].

1.1.2.2.1.5 Redes Neuronales Artificiales

Estos son modelos de predicción no lineales que aprenden como detectar un patrón para emparejar un perfil particular a través de un proceso de entrenamiento que envuelve aprendizaje iterativo, utilizando un conjunto de datos que describe lo que se quiere encontrar. Las redes neuronales son conocidas en la estructura del aprendizaje automático cómo "aproximaciones universales" con un gran carácter paralelo de cálculo y buenas capacidades de generalización, pero también como cajas negras debido a la dificultad para penetrar dentro de las relaciones aprendidas. Son utilizadas en el la minería de datos: para generar modelos de regresión que puedan predecir comportamientos futuros, sobre la base de pares de datos de entrada – salida de información numérica histórica continua (la red neuronal asocia salidas numéricas (outputs) con cualquier nuevo objeto de valores de atributos conocidos), y automáticamente representa un conjunto de datos por un pequeño número de prototipos representativos, preservando las propiedades topológicas del espacio original del atributo (aprendizaje sin supervisión) [7].

1.1.2.2.1.6 Técnicas estadísticas

Una variedad de técnicas puede ser utilizadas para identificar patrones, los cuales pueden ser entonces utilizados para predecir el futuro. Estas incluyen las regresiones lineales, los modelos aditivos generalizados (GAM) y las regresiones adaptativas multivariadas por splines [7].

1.1.2.2.1.7 Árboles e Inducción de reglas

La inducción de reglas es el proceso de extraer reglas (si-entonces) de datos, basadas en significados estadísticos. El aprendizaje de máquinas (ML, de sus siglas en inglés), es el centro del concepto de la minería de datos, debido a su capacidad de ganar penetración física dentro del problema, y participar directamente en la selección de datos y en los pasos de búsqueda del modelo. Para dirigir problemas de clasificación (árboles de decisión claros y borrosos), regresión (árboles de regresión), predicción temporal (árboles temporales), el campo del aprendizaje de máquinas, básicamente se centra en el diseño automático de reglas "si-entonces", similares a aquellas utilizadas por los expertos humanos. La inducción de árboles de decisión es capaz de manejar problemas de gran escala debido a su eficiencia computacional, dar resultados interpretables y en particular identificar los atributos más representativos para una tarea dada [7].

1.1.2.2.1.8 Lógica Borrosa (Fuzzy Logic)

La lógica borrosa maneja conceptos imprecisos (como pequeño, grande, joven, viejo, alto, bajo) y es más flexible que otras técnicas. Proporciona la noción de un conjunto borroso más que una clara demarcación de límites, por ejemplo, en vez de 0 o 1 hay también 0.9, 0.85, 0.93, 0.21, 0.05 etc [7].

1.1.2.2.1.9 Técnicas de visualización

Histogramas (estimando la distribución de probabilidad para ciertos atributos numéricos dados en un conjunto de objetos), gráficas de dispersión (proporcionan información sobre la relación entre dos atributos numéricos y unos discretos), gráficas tridimensionales, dendrogramas (análisis de correlación entre atributos u objetos) [7].

1.1.2.2.1.10 Conjuntos Aproximados (Rough Sets)

La teoría de conjuntos aproximados es adecuada para problemas que pueden ser formulados cómo tareas de clasificación y ha ganado un significante interés científico como estructura de minería de datos y KDD (Ohrn, 1999). La base de la teoría de los conjuntos aproximados está en la suposición de que cada objeto del universo de discurso tiene rasgos característicos, los cuales son presentados por información (conocimiento, datos) acerca del objeto. (Pawlak, 2002). Los objetos que tienen las mismas características son indiscernibles. La teoría ofrece herramientas matemáticas para descubrir patrones escondidos en los datos, identifica dependencias parciales o totales, es decir relaciones causa – efecto, en bases de datos, elimina redundancia en los datos, da aproximaciones a valores nulos o inválidos, datos perdidos, datos dinámicos etc [7].

1.1.2.3 Visualización

Una vez obtenido el conocimiento deseado (tras las fases de Preprocesamlento, Representación Intermedia y Minería de Textos) existe una fase de representación necesaria para finalizar el proceso de Descubrimiento de Conocimiento en texto que puede resultar importante para el usuario final a la hora de Interpretar los resultados de la minería: la visualización [1].

En esta fase se proporciona un entorno de exploración de datos guiado para el usuario que sea lo más amigable posible Las últimas tendencias presentan resultados mediante gráficas tridimensionales, páginas web o tag clouds [6].

El proceso que nos lleva hasta la fase de visualización, podemos resumirlo como: a un conjunto de documentos se le aplica una serie de transformaciones (como Preprocesamiento, Representación Interna y Minería de Textos) y se obtienen, como resultado, Reglas de Asociación que se representan gráficamente en 3D), en ese caso [1].

Una vez obtenidos conceptos, los términos, las o cualquiera que sea el resultado, se pueden utilizar métodos automáticos de visualización o bien pueden interpretarse los resultados directamente [1].

Si se ha elegido realizar el proceso de visualización, pueden ser útiles las técnicas utilizadas por ([Wong et al., 1999], [Wong et al., 2000]), que presentan sistemas para visualizar reglas de asociación aplicadas a Minería de textos, permitiendo así una mejor detección de tópicos, ([Dubois and Quafafou, 2002]) que sugiere algoritmos de visualización basados en gráficos SOM para mostrar estructuras dinámicas obtenidas por la minería de textos, ([Parahc and Bednar, 2003]) que muestra, vía html, los resultados obtenidos en el clusterlng, ([Mei and Zhai, 2005]) que Visualiza temas mediante un plot, et al , 20071) que afirma que las técnicas de visualización de la información deberían jugar un papel primordial en el descubrimiento y sugieren el uso de técnicas VTM (Visual Text Mining) que van desde mapas cognitivos hasta SOM, ([Federico et al., 2011]) que sugiere una solución basada en capas para la Visualización de los resultados de minería aplicados a redes sociales utilizando algoritmos de Teoría de Grafos, ([van Eck and Waltman, 2011]) que utiliza la aplicación VOSviewer, utilizada para explorar mapas de bibliométricos ([KontopouIos et al., 2013]) que Visualiza las ontologías utilizando la herramienta OntoGen [1].

([Chen, 2005]) menciona algunos problemas existentes a la hora de representar la semántica de la información, puesto que debe trabajar con datos no numéricos, no espaciales y altamente dimensionados, entre ellos es necesario mencionar: la visualización del conocimiento del dominio, inferencia visual y predicción, medidas de cualidad intrínsecas o la propia estética de la representación. Aun así, sugiere que la navegación a través de la información visual requiere [1].

1.2 Teoría de Grafos

Si queremos entender un sistema complejo, en primer lugar, hay que saber cómo sus componentes interactúan entre sí. Una red es un catálogo de componentes de un sistema a

menudo llamados nodos o vértices y las interacciones directas entre ellos, llamados enlaces o aristas. [10].

El estudio de la teoría de redes tiene sus inicios en la teoría de grafos. Los grafos fueron utilizados para resolver acertijos y problemas. Entre los problemas resueltos con grafos, se destaca el problema del puente de Königsberg, abordado por el matemático Leonard Euler, para interpretar [11].

Definición 4.1 Según Brandes (2005) Un grafo **G** consiste de un conjunto **V** de vértices y un conjunto **E** de pares de vértices llamados aristas, usualmente denotamos al conjunto de vértices y aristas por **V(G)** y **E(G)** y sus cardinales por **n** y **m** respectivamente.

Con soporte en la anterior definición, se puede estimar que una red es una estructura matemática formal que representa a los elementos y sus relaciones. Las redes, particularmente, se han convertido en una herramienta central en el estudio de sistemas complejos, donde los nodos son considerados como "agentes".

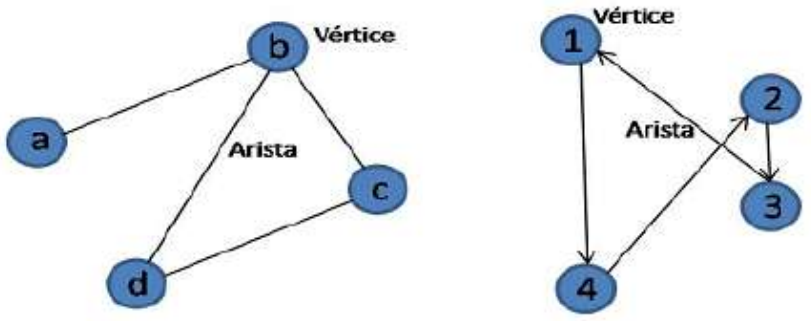


Figura 5 Grafo no Dirigido (Izquierda) y Grafo Dirigido (Derecha). Tomado de Leal (2009)

Las redes han tenido mucha importancia debido a su desempeño al momento de ser aplicadas a diferentes tipos de sistemas, entre ellos destacamos sistemas tales como, biológico, ecológicos, sociales e informáticos. [12]

1.3 Matriz de Adyacencia

Esta matriz es una propiedad correspondiente a cada red, la cual es una matriz cuadrada que establece las relaciones binarias de las conexiones de los nodos, en donde 1 implica conexión y 0 no conexión. Esta matriz tiene la característica que en redes no dirigidas es una matriz

simétrica, por lo cual actuar sobre este tipo de matriz con alguna función matemática no afecta si se realiza con sus columnas o sus filas, caso contrario si la red es dirigida.

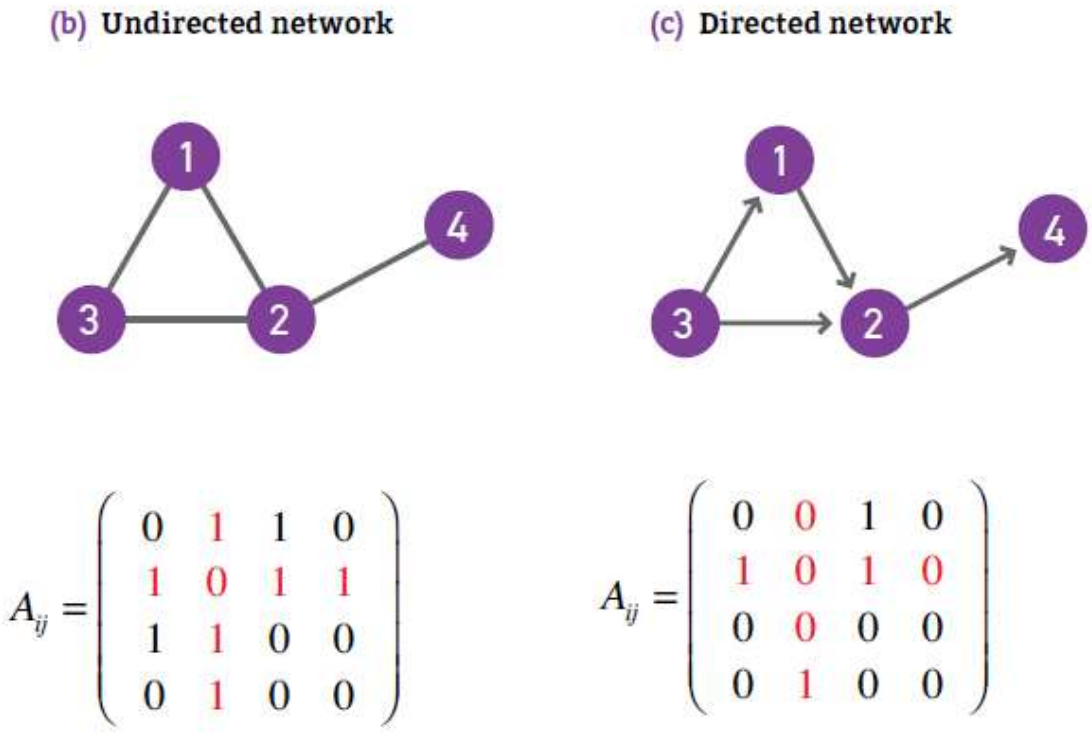


Figura 6 (a)Matriz de Adyacencia de una red no dirigida, (b) Matriz de Adyacencia de una red dirigida, [10]

1.4 Distribución de Grado

Una propiedad importante en la teoría de grafos, es el grado de un nodo. Esta característica representa la conectividad que tiene el nodo con respecto a los otros de la red. En una clase de colegio, cuantos alumnos hay en un aula de clases representan que tan conectado está el profesor en ese momento, este es un posible ejemplo de esto. Esta característica en redes dirigidas puede corresponder al grado de entrada o de salida de un nodo, en el cual dice cuántas conexiones tiene el nodo de dirección hacia él o desde el respectivamente.

El grado de distribución, proporciona la probabilidad de que un nodo seleccionado al azar en la red tiene k grados [8], siendo k el número de grados seleccionado.

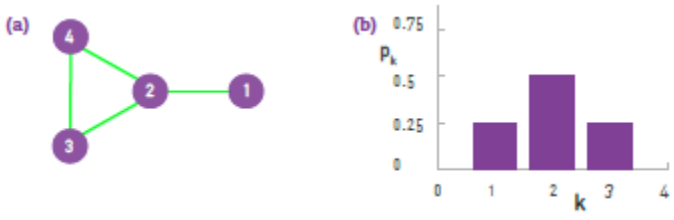


Figura 7 : Histograma Distribución de Grado(b) de la red (a),[10]

1.5 Redes de Co-ocurrencia

Las redes de co-ocurrencia son la interconexión colectiva de términos basados en su presencia emparejada dentro de una unidad de texto especificada. Las redes se generan conectando pares de términos usando un conjunto de criterios que definen la co-ocurrencia. Por ejemplo, se puede decir que los términos A y B "co-ocurren" si ambos aparecen en un artículo particular. Otro artículo puede contener términos B y C. Vincular A a B y B a C crea una red de co-ocurrencia de estos tres términos. Las reglas para definir la co-ocurrencia dentro de un corpus de texto se pueden establecer de acuerdo con los criterios deseados [32].

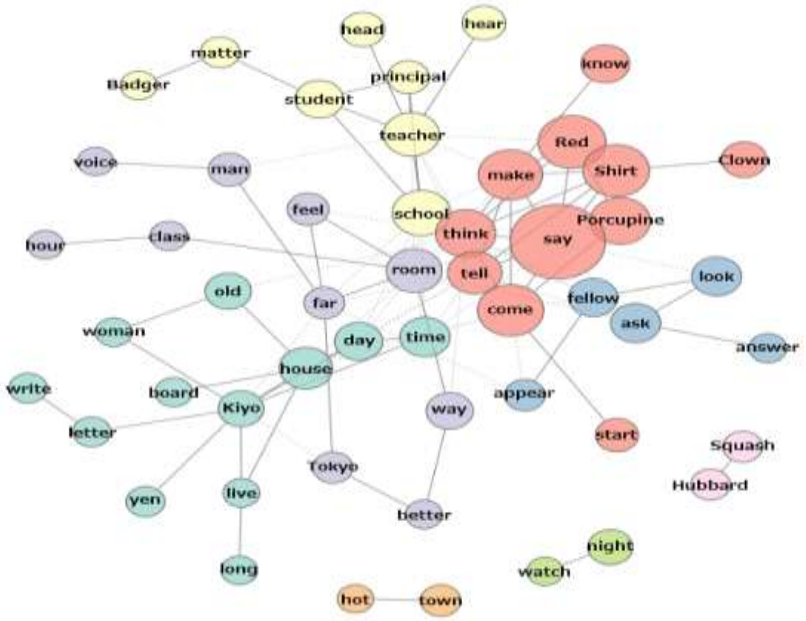


Figura 8 Red de Co-ocurrencia Creada en KH-Coder. Tomado de (Universidad Nacional del Sur, 2016)

Las redes de co-ocurrencia pueden ser creadas para cualquier lista de términos (cualquier diccionario) en relación con cualquier colección de textos (cualquier corpus de texto). Los pares co-occurrentes de términos se pueden llamar "vecinos" y éstos agrupan a menudo en "barrios" basados en sus interconexiones. Los términos individuales pueden tener varios vecinos. Los barrios pueden conectarse entre sí a través de al menos un término individual o pueden permanecer desconectados [32].

Los términos individuales son, en el contexto de la minería de textos, representados simbólicamente como cadenas de texto. En el mundo real, la entidad identificada por un término normalmente tiene varias representaciones simbólicas. Por tanto, es útil considerar los términos como representados por un símbolo primario y hasta varios símbolos sinónimos alternativos. La ocurrencia de un término individual se establece mediante la búsqueda de cada

representación simbólica conocida del término. El proceso puede ser aumentado a través de algoritmos de procesamiento de lenguaje natural (NLP) que interrogan segmentos de texto para posibles alternativas como orden de palabras, espaciado y separación de palabras. El NLP también se puede usar para identificar la estructura de oraciones y categorizar las cadenas de texto de acuerdo con la gramática [32].

La representación gráfica de las redes de co-ocurrencia permite visualizar las inferencias sobre las relaciones entre entidades en el dominio representado por el diccionario de términos aplicados al corpus de texto. Una visualización significativa requiere normalmente simplificaciones de la red. Por ejemplo, las redes pueden ser dibujadas de manera que el número de vecinos que se conectan a cada término sea limitado. Los criterios para limitar los vecinos podrían basarse en el número absoluto de co-ocurrencias o criterios más sutiles como la "probabilidad" o "frecuencias" de co-ocurrencia o la presencia de un término descriptivo intermedio. [13]

1.6 Emergencia

Este concepto es uno de los más difíciles para poder aclarar debido a que su significado ha sido propagado a nivel social de una manera muy excluyente de lo que en realidad es, para nuestra aplicación en el trabajo la emergencia es tratada como la propiedad que posee un sistema para no estar presente en sus componentes, en otras palabras, la acción que puedan tener los componentes para adecuarse al cambio [14]. Ahora bien, en general las propiedades de un sistema son emergentes, si ellas no están presentes en sus componentes. En otras palabras, propiedades globales que son producidas por las interacciones locales son emergentes. La emergencia se da a una escala dada, y no puede ser descrita sobre la base de las propiedades de una escala inferior [15].

Si describimos un fenómeno en términos de información, para tener nueva información, la información vieja debe haber sido transformada. Esta transformación puede ser dinámica, estática, activa, o estimérgica [16]. Sucesivamente, se establece la relación que la información trae al sistema, como aquel aporte necesario para empezar la adaptación al cambio que es

presentado por el ambiente. En nuestro contexto de trabajo nos basaremos en la información de Shannon I como la emergencia, utilizando la base 10.

$$E = I$$

Al estar E basada en I, tenemos que es una medida probabilística. E = 1 significa que cuando cualquier variable binaria se vuelve conocida, un bit de información emerge. Sí E = 0, entonces no emergerá nueva información, incluso variables aleatorias se vuelven "conocidas" (se conocen de manera anticipada, de antemano). [14].

1.7 Auto-Organización

Esta cualidad que poseen los sistemas es conveniente de analizar, debido a que puede resultar muy dependiente para la evaluación y respectivo análisis, la auto organización tenida en cuenta desde un punto de vista de complejidad, se expresa como la regularidad que pueden tener los componentes de un sistema, que puede derivar de un conjunto de patrones [14]. Este proceso viene desde la dinámica interna del sistema, definida por las interacciones entre los elementos. Su resultado es un grado determinado de regularidad, que puede derivar en la expresión de un conjunto de patrones [14].

La auto-organización en otras palabras más contextuales, es un proceso en el cual las características del sistema terminan en un orden global a través de las interacciones locales de sus componentes. El proceso se lleva a cabo de una manera autónoma sin ningún método de control n_i dirección, la organización resultante está completamente descentralizada o distribuida sobre todos los componentes del sistema; esta organización resulta típicamente muy robusta, capaz de sobrevivir y auto-reparar danos o perturbaciones sustanciales [31].

La auto-organización ha sido asociada, correlacionada, con el incremento de orden. Por ejemplo, con la reducción de entropía [14]. Si la emergencia implica un incremento de información, la cual es análoga a la entropía y al desorden, la auto-organización estaría anti-correlacionada con la emergencia [15]. Bajo las propuestas hechas en [14] bajo los axiomas necesarios se sigue la medida:

$$S = 1 - I = 1 - E$$

Donde intuitivamente se denota que si $S = 0$o $S = 1$ no existe ninguna información aportada por el sistema.

1.8 Complejidad

Etimológicamente, complejidad viene del latín plexus, lo cual significa entrelazado. Algo complejo es algo difícil de separar. Esto significa que los componentes del sistema son interdependientes, de manera que su futuro está parcialmente determinado por sus interacciones. [17] Dado el caso conocido de que lo complejo viene dado a través de las interacciones de los componentes de un sistema, por tanto, la complejidad puede estar asociada a los aspectos emergentes y auto-organizantes que presente el sistema. [14]

Tradicionalmente se ha pensado que la complejidad se refiere a aquella condición que limita al modelista para formular el comportamiento completo de los sistemas dinámicos SD, en un lenguaje dado. En realidad, lo complejo se debe a las interacciones en el SD que generan información nueva y relevante no presente en las condiciones iníciales ni de frontera y que influyen en el desarrollo del SD. La complejidad, por tanto, puede estar asociada a los aspectos emergentes y auto-organizantes del SD. [18]

Un ejemplo de complejidad clásico se halla en el conocido "Juego de la Vida" de Jhon Conway, en el cual, 4 reglas simples generan dinámicas ricas como estructuras estables, móviles y/u oscilatorias, que son difíciles de predecir. En base a [18] se afirma que la complejidad está dada por el balance entre la emergencia y la auto-organización existente en el sistema. [19] Por lo tanto, se sigue la ecuación propuesta:

$$C = 4 * E * S$$

1.9 Teoría de la Computación

Esta teoría comprende un conjunto de conocimientos y técnicas diseñadas con el propósito de estudiar la abstracción de sistemas reales, y poder reproducirlos en sistemas formales o matemáticos, es decir, a través de algoritmos creados con axiomas necesarios para representar el sistema real en un lenguaje conocido para la máquina. En otras palabras, la teoría de la computación tiene como propósito desarrollar modelos matemáticos formales de computo que reflejan los ordenadores del mundo real [29].

1.9.1 Teoría de la Computabilidad

Esta teoría nace de la apreciación de los problemas imposibles de resolver mencionados anteriormente. Esta teoría explora los límites de la posibilidad de solucionar problemas mediante algoritmos, y con esta llegar a la posibilidad de reducir tiempo y recursos tratando de resolver problemas imposibles. Según esta teoría se pueden clasificar los problemas en solucionables y no solucionables [29].

1.9.2 Teoría de la Complejidad Computacional

La principal cuestión planteada en esta área es "¿Que hace que algunos problemas sean computacionalmente difíciles y otros problemas fáciles?". De manera informal, un problema que se llama "fácil", si es eficiente solución. Por otro lado, un problema que se llama "duro", si no se puede resolver de manera eficiente, o si no se sabe si se puede resolver de manera eficiente [29].

1.9.3 Computación Paralela

En la última década, el cambio drástico que ha tenido la computación con respecto a los procesadores, ha desarrollado la necesidad de utilizar nuevas herramientas que permitan aprovechar al máximo lo que la maquina brinda [31].

Los grandes aumentos de la potencia de cálculo que hemos estado disfrutando desde hace décadas han estado en el corazón de muchos de los avances más espectaculares en campos tan diversos como la ciencia, la Internet, y el entretenimiento [30].

La computación paralela es una forma de computo en la que muchas instrucciones se ejecutan simultáneamente, operando sobre el principio de que problemas grandes, a menudo se pueden dividir en unos más pequeños, que luego son resueltos simultáneamente (en paralelo) [31].

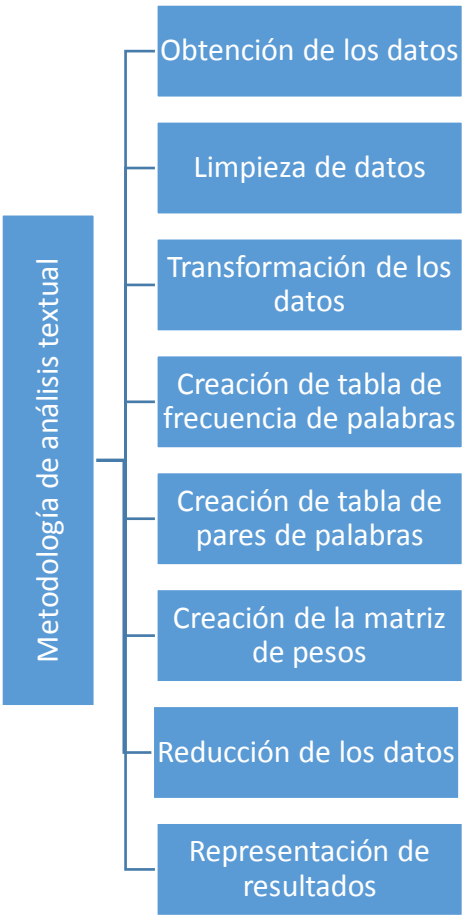
1.10 R

Es un lenguaje para el análisis estadístico y gráfico R es un ambiente de programación formado por un conjunto de herramientas muy flexibles que pueden ampliarse fácilmente mediante paquetes, librerías o definiendo nuestras propias funciones. Además, es gratuito y de código abierto, un Open Source parte del proyecto GNU, como Linux o Mozilla Firefox. [25]

Metodología

Esta metodología tiene como principal objetivo hacer un buen análisis estadísticos de los datos de tipo textuales a través de un estudio exploratorio. Utilizando pasos de la técnica de minería de texto estadística, ya que a lo largo de su implementación serán usadas técnicas de análisis estadístico y tratamiento de datos por medio de R. Esta metodología será evaluada al finalizar por medio de una herramienta a través de unas encuentas realizadas por el establecimiento que ha proporcionado los datos (por motivos de provacidad de datos las

encuentas no son mostradas dentro de este trabajo)



Mapa conceptual de las fases de la metodología

Fase1 Obtención de los datos

Para el cumplimiento de los objetivos se evidenció como requisito tener un archivo de texto de extensión txt, por lo cual se tiene como caso de aplicación una base de datos de un restaurante de comidas rápidas a domicilio, donde se podrá encontrar una cantidad de datos escritos por personas del común.

Fase2 Limpieza de datos

Es un proceso necesario para asegurar la calidad de los datos que se emplearán para *analytics*. Este paso es fundamental para minimizar el riesgo que supondría el basar la toma

de decisiones en información poco precisa, errónea o incompleta. [26] En esta limpieza se eliminarán palabras vacías(stopword), datos de tipo numéricos y caracteres, url, correos electrónicos, hashtag, espacios en blancos.

Fase3 Transformación de datos

En la transformación se adecúa la información. En este proceso es típico duplicar tablas que contienen la información correcta y la creación de nuevos campos o nuevas tablas con datos agregados y/o calculados. Por ejemplo, para agrupar información por criterios geográficos, temporales, o de estructura jerárquica o comercial que serán útiles para el análisis. [27] En esta transformación, implica perder la integridad y normalización de los datos originales que están mal escritos.

Fase4 Creación de tabla de frecuencia de palabras

En esta fase se separan los datos en forma individual y se agregan a una tabla con su correspondiente frecuencia apara poder graficar en nube e histogramas.

Fase5 Creación de tabla de frecuencia de pares de palabras

En esta fase se separan los datos en pares de datos y se agregan a una tabla con su respectiva frecuencia y así poder graficarla en redes, para poder visualizar las relaciones entre ellos.

Fase6 Creación de la matriz de pesos

Después de haber realizado el pre-procesamiento se representan los datos y su relación en una matriz donde las columnas serán las variables y las filas serán los individuos, además la relación representará el peso que hay entre ellas. Por lo tanto, esta matriz será no dirigida, adicionalmente se utilizarán varias herramientas como análisis de componentes principales (PCA), agrupación jerárquica en componentes principales(HCPC), redes, emergencia, complejidad y auto-organización.

Fase7 Reducción de datos

Encontrar las características más significativas para representar los datos, dependiendo del objetivo del proceso. En este paso se pueden utilizar métodos de transformación para reducir

el número efectivo de variables a ser consideradas o para encontrar otras representaciones de los datos. [28] En este proceso reducimos el tamaño de la matriz de pesos y dejamos aquellos datos que tiene cuyas sumas de los pesos sean las mayores.

Fase8 Representación de resultados

Después de obtener los datos necesarios, se procede a encontrar una manera de visualizar y analizar correcta y eficazmente los datos, por lo cual, se utiliza R como lenguaje propio para la interpretación de los mismos [31]. Primero se representan los datos individuales mediante nubes de palabras e histograma de frecuencia de palabras. Luego se realizará un análisis de componentes principales(PCA) y un análisis de agrupamiento jerárquico de componentes principales(HCPC) de los datos. El PCA se utilizó para resumir y visualizar la información contenida en la matriz de pesos. El HCPC se utilizó para identificar los grupos de datos contenido en la matriz de pesos. Después se visualizarán mediante redes complejas las relaciones de los datos que se encuentran en la tabla de frecuencia de pares de palabras y en la matriz de pesos, entre estas redes se puede observar la red de co-ocurrecia por pesos, dirigida y no dirigida que se encuentra en la tabla de frecuencia de pares de palabras, por otra parte, se visualiza la red de co-ocurrencia por nodos y frecuencia obtenida de la matriz de pesos, además se graficaran mediante un diagrama de barra la emergencia, auto-organización y complejidad de los datos obtenidos de la matriz de pesos.

2.5 Resultados

1. Para el análisis de las palabras se estableció un parámetro de valor de frecuencia, de esta manera se pudo observar las palabras más frecuentes en el archivo cargado.

Figura 9 Nube de palabras frecuentes

En la figura 6 se muestra una nube de palabras más co-ocurrentes de forma individual del archivo cargado, donde se le coloco como parámetro de frecuencia que sean mayores que 100, donde las palabras con mayor frecuencia resaltaran más que las de menos.

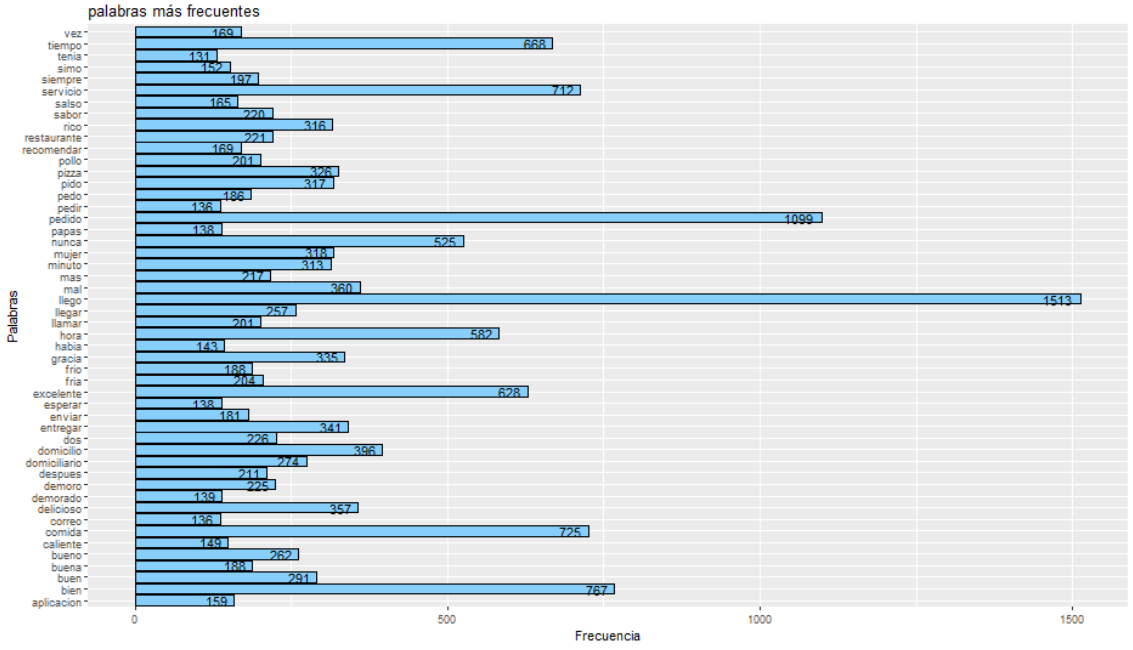


Figura 10 Histograma de frecuencia de palabras

En la figura 7 se muestra un histograma de las palabras más frecuentes, donde se le coloco como parámetro de cantidad las 50 palabras más frecuentes en el archivo cargado, lo cual cada palabra tiene su propio número de frecuencia.

2. Para el análisis de palabras relacionadas se estableció un parámetro de valor de frecuencia, de esta manera se pudo observar las palabras relacionadas más frecuentes en el archivo cargado, donde se representa en redes.

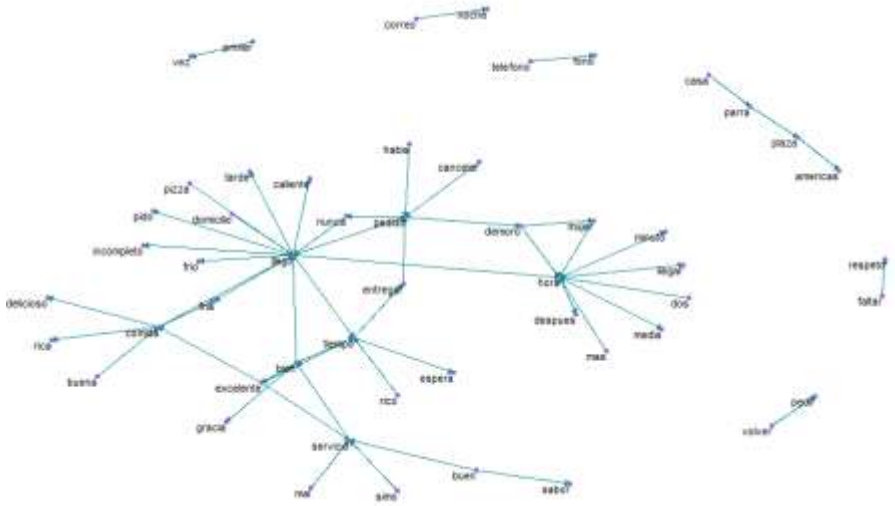


Figura 11 Grafo dirigido de palabras relacionadas

En la figura 8 se muestra un grafo dirigido de las palabras relacionadas, donde se le coloco como parámetro de frecuencias 30, lo cual graficara pares de palabras que están conectada a una frecuencia mayor o igual a 30 en el archivo cargado.

Para el caso estudiado es observado como palabras negativas y positivas son acopladas por igual. Dejando un grado de duda para el análisis final, aunque se puede encontrar una gran conexión entre las palabras de satisfacción, por lo que deja saber que el restaurante tiene más seguidores que antes.

Figura 12 Grafo no dirigido de palabras relacionadas

En la figura 9 se muestra un grafo no dirigido de las palabras relacionadas que se exportaran en un archivo html para mejor visualización de los nodos, donde se le coloco como parámetro de frecuencias 30, lo cual graficara pares de palabras que están conectada a una frecuencia mayor o igual a 30 en el archivo cargado.

3. Creación de matriz de pesos nxn

Obteniendo los datos preprocesados se comienza a crear la matriz de pesos cuadrada o nxn , donde su tamaño fue proporcionada por la cantidad de palabras descritas durante el preprocesamientos y tanto las filas como las columna deben tener los mismo identificadores, esta matriz constará de pesos cuyos pesos serán las frecuencias con las que se relacionan.

4. Para el análisis de componentes principales(PCA) se utilizó las columnas como variables y las filas como individuos, dándole formas de calcular como son la de contribución(“contrib) y coseno cuadrado(“cos2”) para saber las relaciones entre las variables.

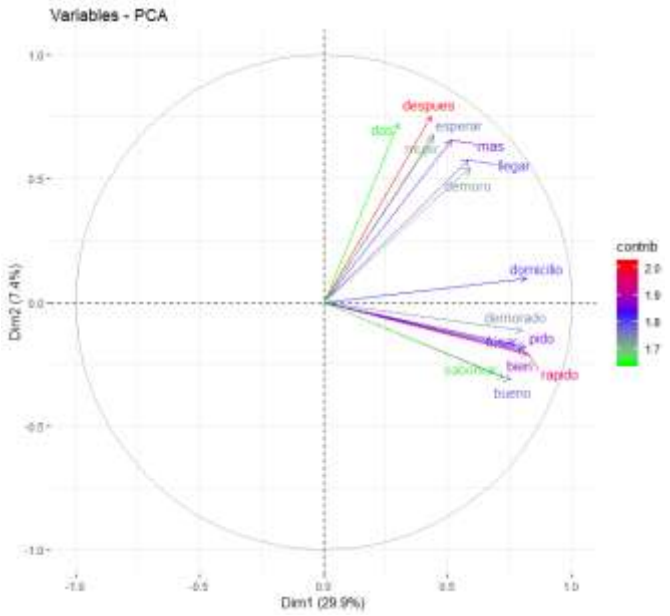


Figura 13 Análisis Pca con contribución

En la figura 10 se muestra el análisis de componentes principales para las 100 variables cuyas sumas de pesos sean las mayores de la matriz de pesos con el método de contribución(“contrib), pero en este caso le colocamos como parámetro 15 para mostrar las primeras 15 de esas 100 variables que se relacionan entre sí.

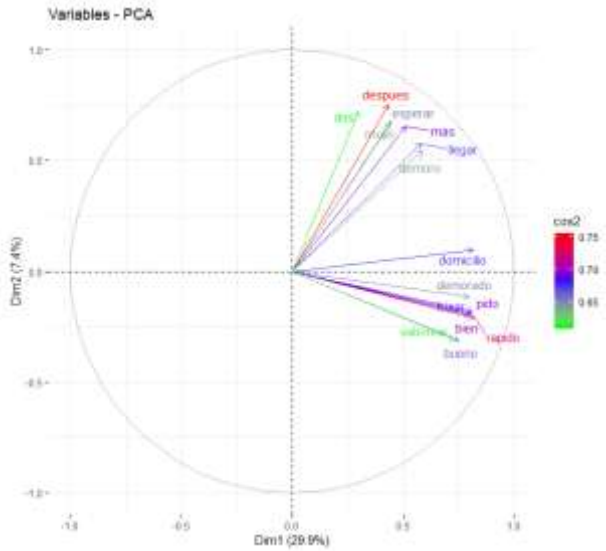


Figura 14 Análisis Pca con coseno cuadrado

En la figura 11 se muestra el análisis de componentes principales para las 100 variables cuyas sumas de pesos sean las mayores de la matriz de pesos con el método de coseno cuadrado(“cos2), pero en este caso le colocamos como parámetro 15 para mostrar las primeras 15 de esas 100 palabras que se relacionan entre sí.

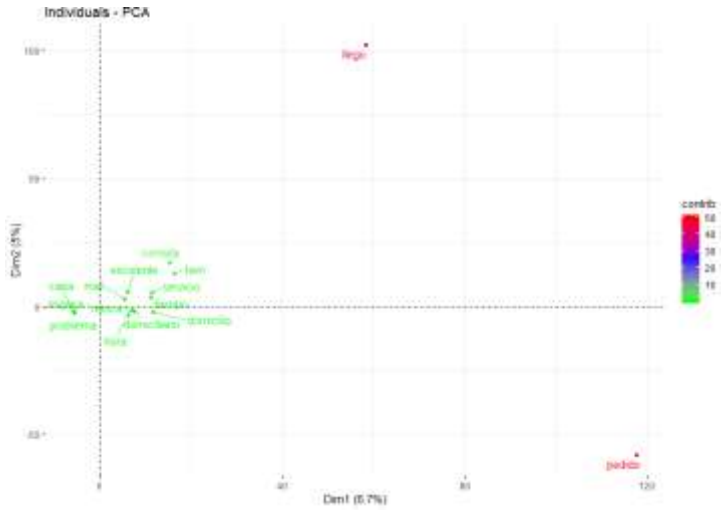


Figura 15 Análisis Pca con contribución

En la figura 12 se muestra el análisis de componentes principales para los 100 individuo cuyas sumas de pesos sean las mayores de la matriz de pesos con el método de contribución(“contrib), pero en este caso le colocamos como parámetro 15 para mostrar las primeras 15 de esas 100 palabras que se relacionan entre sí.

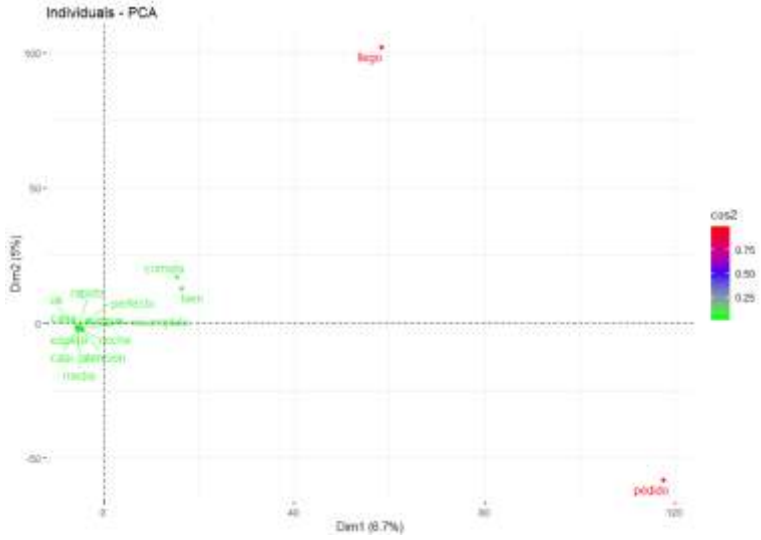


Figura 16 Análisis Pca con coseno cuadrado

En la figura 13 se muestra el análisis de componentes principales para los 100 individuos cuyas sumas de pesos sean las mayores de la matriz de pesos con el método

de coseno cuadrado(“cos2), pero en este caso le colocamos como parámetro 15 para mostrar las primeras 15 de esas 100 palabras que se relacionan entre sí.

Para el análisis de agrupamiento jerárquico de componentes principales, se utilizó las columnas como variables y las filas como individuos para clasificar grupos de datos similares dentro del archivo cargado, cuyos datos son los individuos, dándole formas de calcular con el numero clúster que se compone de un valor entero, las métricas (“euclidean”, “manhattan”) y el método (“ward”, “average”, “complete”, “single”). Todas las anteriores usadas como métricas y métodos de análisis de datos definidos por la estadística.

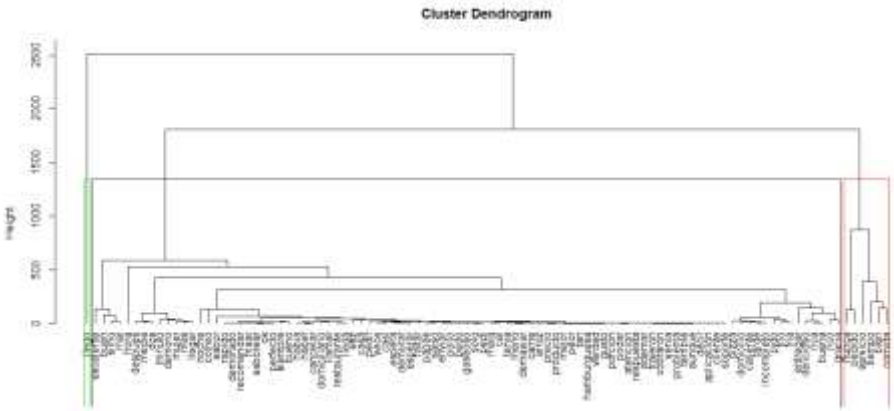


Figura 17 Árbol jerárquico

En la figura 14 se muestra un árbol jerárquico con los datos de los primeros 100 individuos que se extrajeron de la matriz de peso, dándole como parámetro el número de clúster igual a 3, la métrica euclidean y el método Ward que se basa en la variancia multidimensional como el análisis de componentes principales.

En este tipo de análisis es importante resaltar los grupos formados entre palabras, definiendo así las regiones de similitud entre los comentarios de las personas sobre el restaurante. En el ejemplo estudiado se puede demostrar como existen grupos de satisfacción, y cuales palabras definen estos grupos. Así como también los grupos formados por comentarios negativos presentados por la comunidad.

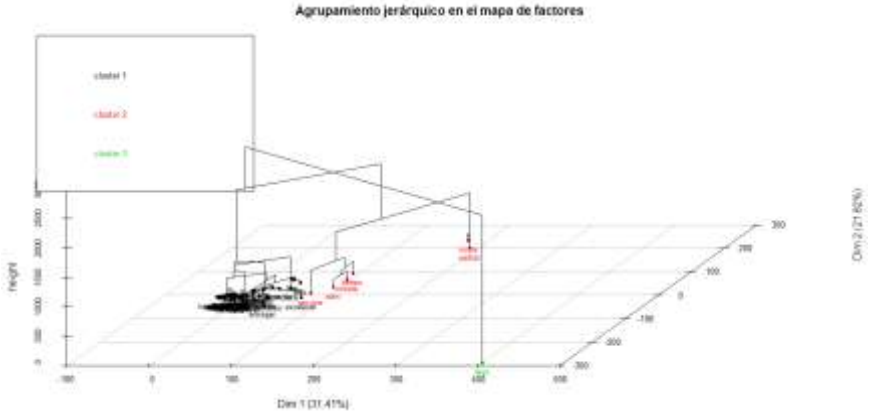


Figura 18 Mapa de factores 3d

En la figura 15 se muestra el Mapa de factores de forma tridimensional donde se clasifica estos individuos en los componentes principales, lo cual se tomaron los primeros 100 individuos que se extrajeron de la matriz de peso, dándole como parámetro el número de clúster igual a 3, la métrica euclidean y el método Ward que se basa en la variancia multidimensional como el análisis de componentes principales.

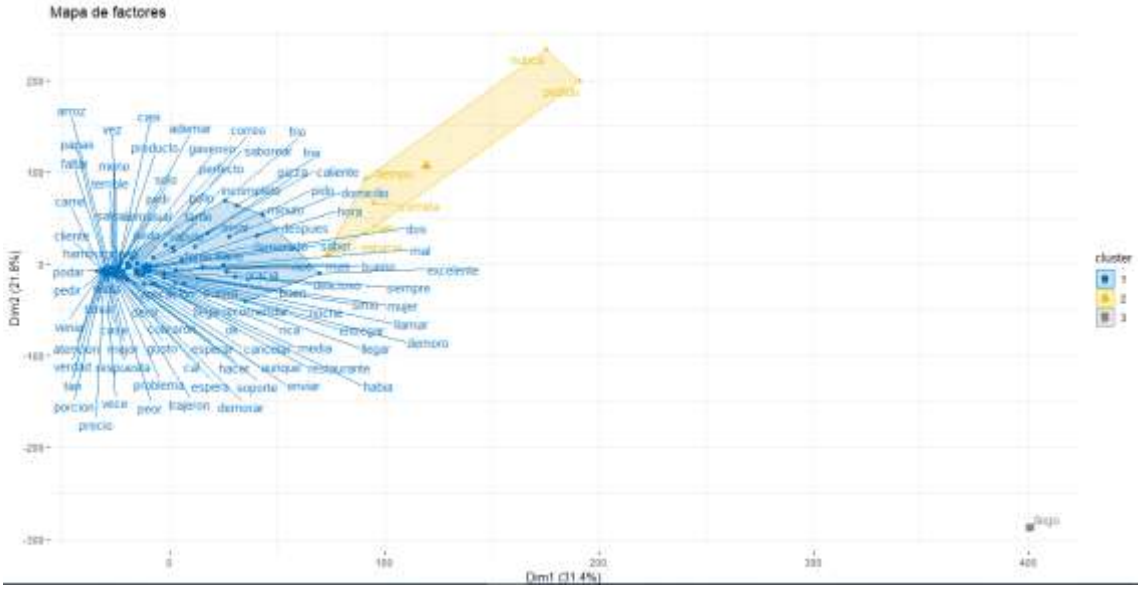


Figura 19 Mapa de factores en 2d

En la figura 16 se muestra el Mapa de factores en 2d donde se clasifica estos individuos en los componentes principales y se dibujan en grupos, también se colorean de acuerdo al grupo que pertenecen, lo cual se tomaron los primeros 100 individuos que

se extrajeron de la matriz de peso, dándole como parámetro el número de clúster igual a 3, la métrica euclidean y el método Ward que se basa en la variancia multidimensional como el análisis de componentes principales.

5. Para el análisis de redes de frecuencia se utilizó los pares de palabras relacionadas, se estableció un parámetro de valor de frecuencia, de esta manera se pudo observar los pares de palabras relacionadas más frecuentes en red, donde se dibujará los tamaños de enlace de acuerdo a la frecuencia.

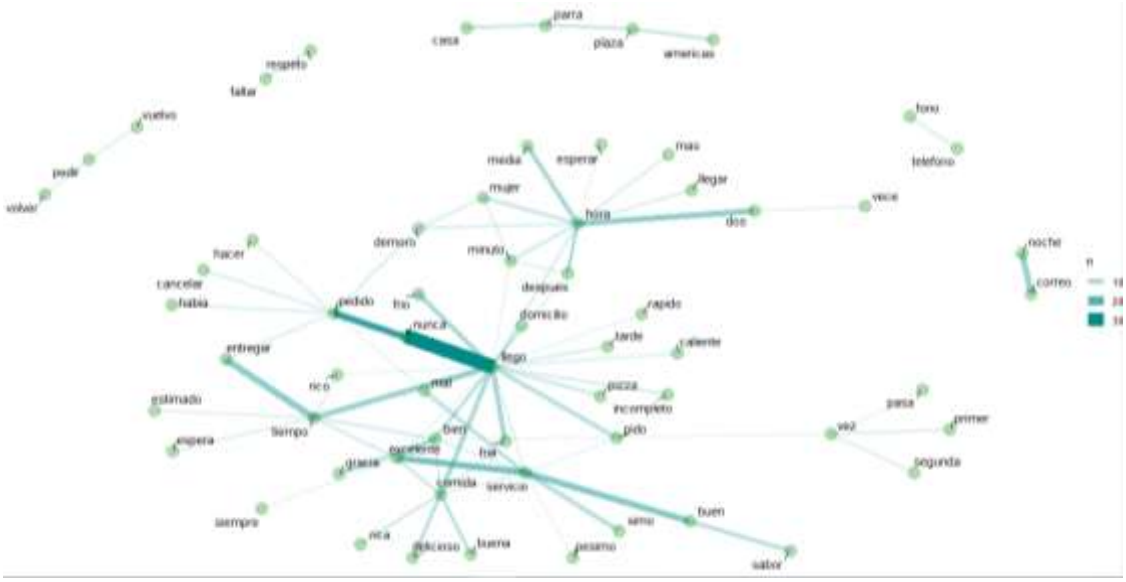


Figura 20 Red de frecuencia con tamaño de enlace

En la figura 17 se muestra una red de palabras relacionada donde se le coloco como parámetro de frecuencia mayor o igual a 20, con lo cual se dibujarán aquellos pares de palabras que sean mayor o igual a la frecuencia dada, pero se dibujarán los enlaces con los tamaños de frecuencia de cada par.

6. Para el análisis de redes de frecuencia se utilizó la matriz de pesos, se estableció un parámetro de valor de frecuencia para eliminar aquellos enlaces, de esta manera se pudo observar las relaciones de cada nodo de la red, pero se dibujarán los nodos con los tamaños y color a la frecuencia de cada uno.

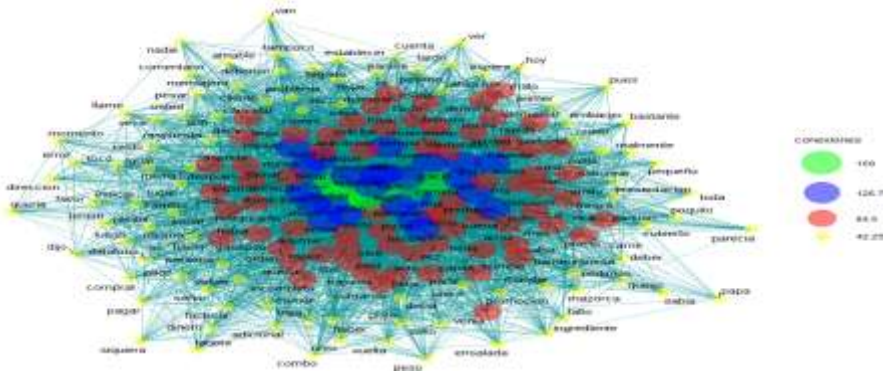


Figura 21 Red de frecuencia con tamaño de nodos

En la figura 18 se muestra una red de palabras relacionada donde se le coloco como parámetro de frecuencia menor que 50, con lo cual se eliminarán aquellos nodos con enlaces menores a la frecuencia dada, pero se dibujarán los nodos con los tamaños de frecuencia de cada uno, así con una etiqueta a la derecha con la frecuencia de cada uno.

7. Para el análisis de emergencia, complejidad y auto-organización se utilizó las 100 variables cuyas sumas de pesos sean las mayores de la matriz de pesos, se estableció como parámetros la cantidad de variables más frecuentes y la base a la cual calcular el logaritmo.

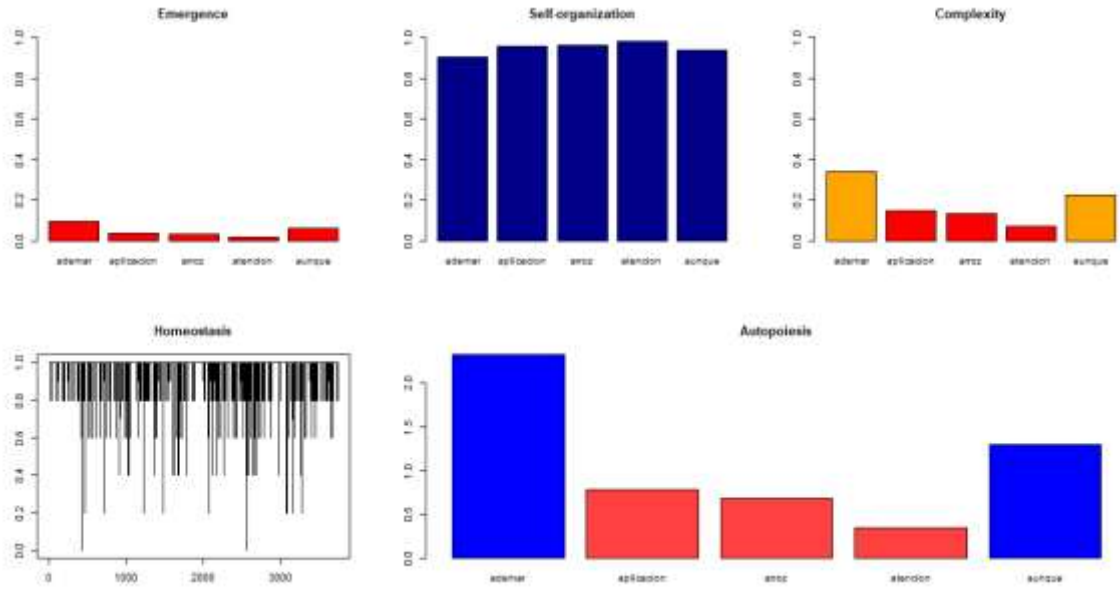


Figura 22 Emergencia, Complejidad, Auto-organización

En la figura 19 se muestra Emergencia, Complejidad, Auto-organización de las variables, donde se le coloco como parámetro la cantidad de palabras igual a 5 y la base 10.

En el ejemplo trabajado, se puede visualizar como pocas palabras presentan emergencia, dando a conocer que la comunidad está orientada en un camino mutuo, y observando las gráficas se puede establecer un gran grado de homogeneidad entre los usuarios.

El restaurante tiene buenas referencias dentro de los usuarios, con una gran polaridad dentro de lo común. Sin embargo, el uso de la herramienta logró evidenciar falta de compromiso en los domicilios, y que a su vez induce a una elevación en la emergencia fluctuante de los usuarios al no estar complacidos.

El restaurante cuenta con un gran grupo de usuarios satisfecho debido a la comida, por lo que puede estar tranquilo con respecto a las referencias que infieren el manejo de los menús hacia los clientes.

2.6 Herramienta en R para minería de texto

2.6.1 Aplicación De Minería De Texto En Shiny.

En esta aplicación se muestra primero una entrada llamada seleccione un archivo, es donde se busca y se sube un archivo de nuestro equipo para el análisis.

Luego una casilla de selección llamada leer archivo separado por símbolo (,) o (;), que sirve para buscar la información en una tabla si esta tiene cabeceras que este separado por estos símbolos (esta tabla tiene que estar en un archivo txt), enseguida le despliega las columnas de la tabla y así poder seleccionar la de agrado.

Nota: si no es una tabla puede deshabilitar la casilla y procesarlo como un texto normal.

Después una casilla de selección llamada Ejecutar preprocesamiento lo cual hay que habilitarla para hacer el preprocesamientos de los datos del archivo subido, luego se le

despliega un menú con todos los análisis que se realizaran posteriormente. Solo hay que habilitar la casilla de su análisis requerido o todos.

Para el ejemplo: Se utilizó un archivo llamado comentarios, lo cual es una base de datos (tabla que está separada por punto y coma (;)) de un restaurante de comidas rápidas a domicilios, luego se buscó la columna donde están los datos que es Comment, para posterior habilitar la casilla de ejecutar preprocesamiento y comenzamos hacer cada análisis.

Figura 23 Menú de Interfaz

2.6.2 Nube De Palabras

En esta parte se mostrará dos barras deslizantes, el grafico de la nube de palabras y un botón de descarga de la imagen en formato PNG. La primera barra que es frecuencia de palabras nube, tomará automáticamente el tamaño máximo, con la frecuencia mayor dentro de las palabras y el valor inicial será la mitad del tamaño, lo cual se podrá deslizar hasta buscar una frecuencia deseada, y la segunda que es cantidad de palabras tomara

automáticamente la cantidad de palabras que tienen la frecuencia escogida, la cual se mostrara en el gráfico.

Para el ejemplo: El tamaño máximo de frecuencia es 1513, donde pudimos deslizar hasta la frecuencia 11 que contenía un total de 52 palabras.

Figura 24 Nube de palabras SHINY

2.6.3 Histograma De Palabras.

En esta parte se mostrará una barra deslizante, el grafico de histograma de palabras y el botón de descargar imagen en formato PNG. En la barra deslizante llamada cantidad de palabras se escogió un tamaño máximo de 50, cuyo valor será fijo siempre,

en esta parte se podrá deslizar la barra para mostrar la cantidad de palabras deseadas en el gráfico con mayor frecuencia.

Para el ejemplo: el tamaño máximo de la barra será 50 como ya habíamos mencionado anteriormente, pudimos deslizar la barra hasta colocarla en 21, donde se observó las 21 palabras con mayor frecuencia

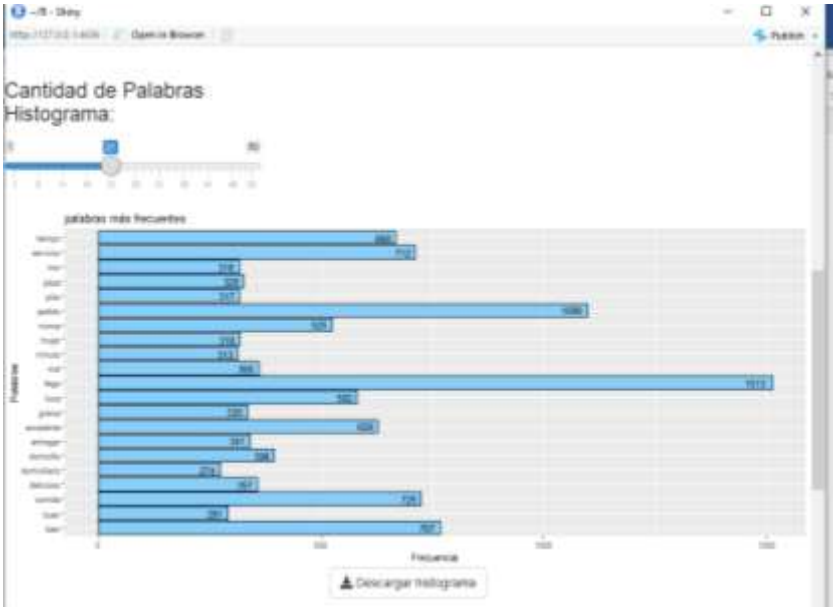


Figura 25 Histograma de Palabras Shiny

2.6.4 Red 2d Dirigida.

En esta parte se mostrará una barra deslizante, el grafico de red dirigida y el botón de descargar imagen en formato PNG. En la barra deslizante que es frecuencia de palabras Red2D, tomará automáticamente el tamaño máximo, con la frecuencia mayor de los pares de palabras más frecuentes y el valor inicial será la mitad del tamaño, en esta parte se podrá deslizar la barra para mostrar la cantidad de pares de palabras deseadas en el gráfico con mayor frecuencia.

Para el ejemplo: el tamaño máximo de la barra es de 302, pudimos deslizar hasta una frecuencia de 40, se observó la cantidad de palabras relacionadas

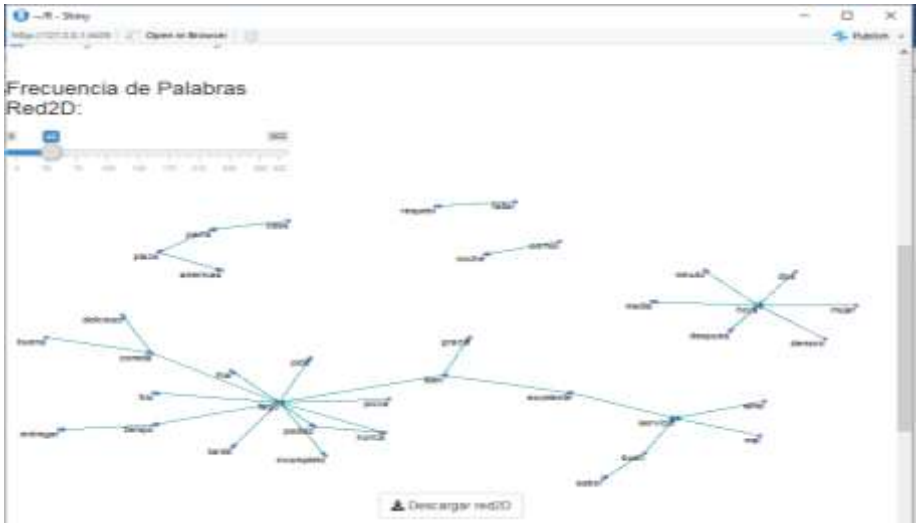


Figura 26 Red dirigida 2D Shiny

2.6.5 Red 3d No Dirigida.

En esta parte se mostrará una barra deslizante, el grafico de red no dirigida y el botón de descargar imagen en formato HTML. En la barra deslizante que es frecuencia de palabras Red3D, tomará automáticamente el tamaño máximo, con la frecuencia mayor de los pares de palabras más frecuentes y el valor inicial será la mitad del tamaño, en esta parte se podrá deslizar la barra para mostrar la cantidad de pares de palabras deseadas en el gráfico con mayor frecuencia.

Para el ejemplo: el tamaño máximo de la barra es de 302, pudimos deslizar hasta una frecuencia de 55, se observó la cantidad de palabras relacionadas.

Figura 27 Red No Dirigida Shiny

2.6.6 Análisis De Componentes Principales (PCA).

En esta parte se mostrará 2 imágenes con sus respectivas configuración y botones de descargar imagen en formato PNG. Imágenes PCA para variables y PCA para individuos.

2.6.6.1 PCA para variables.

En esta parte se muestra dos casillas de selección, una casilla de entrada y una barra deslizante, el grafico de PCA en variables y un botón de descargar imagen en formato PNG. En la primera casilla llamada seleccionar circulo los cual contiene dos métodos de calcular como (“contribución”, “cos2”), la segunda casilla llamada ver mapa de factores variables contiene formas de visualización como dibujar (“flechas y etiquetas”, “etiquetas”, “flechas”, en la casilla de entrada solo para números es para saber la cantidad de palabras a mostrar con el método escogido y la barra deslizante llamada tamaño de las etiquetas que tiene un valor de tamaño fijo de 20 que es el tamaño a darles a las etiquetas.

Por ejemplo: se escogió el método de contribución, con una cantidad de palabras a mostrar 10, mostrar flecha y etiquetas, con un tamaño de etiqueta igual a 5.

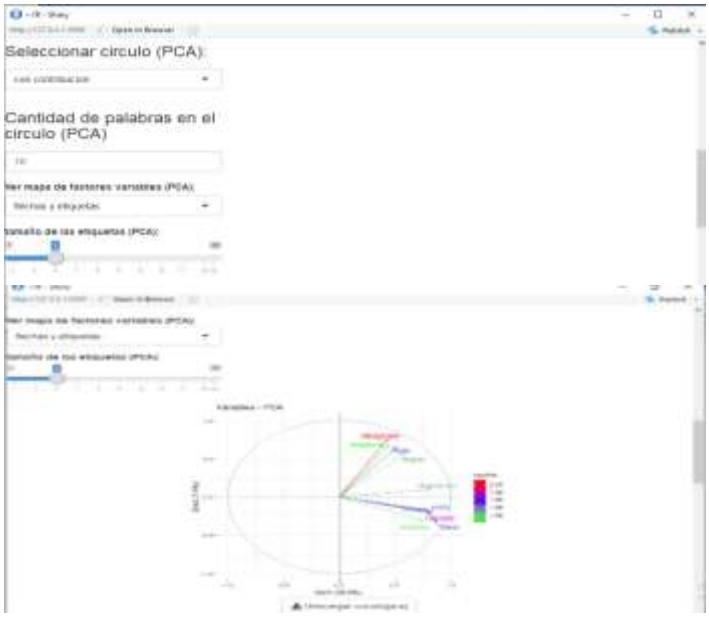


Figura 28 PCA variables Shiny

2.6.6.2 PCA para individuos.

En esta parte se muestra dos casillas de selección, una casilla de entrada y dos barras deslizantes, el grafico de PCA en individuos y un botón de descargar imagen en formato PNG. En la primera casilla llamada seleccionar plano los cual contiene dos métodos de calcular como (“contribución”, “cos2”), la segunda casilla llamada ver mapa de factores variables contiene formas de visualización como dibujar (“puntos y etiquetas”, “etiquetas”, “puntos”, en la casilla de entrada solo para números es para saber la cantidad de palabras a mostrar con el método escogido, en la primera barra deslizante llamada tamaño de los puntos que tiene un valor de tamaño fijo de 10 que es el tamaño a darles a los puntos en el gráfico y la segunda barra deslizante llamada tamaño de las etiquetas, que tiene un valor de tamaño fijo de 20 que es el tamaño a darles a las etiquetas en el gráfico .

Por ejemplo: se escogió el método de contribución, con una cantidad de palabras a mostrar 10, mostrar puntos y etiquetas, con un tamaño de etiqueta igual a 5 y con un tamaño de nodos igual a 1.

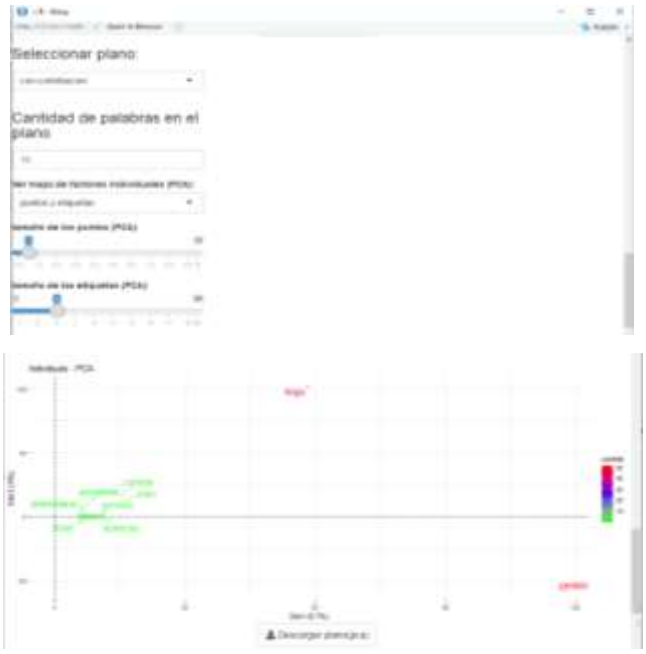


Figura 29 PCA Individuos Shiny

2.6.7 Agrupamiento Jerárquico De Componentes Principales (HCPC).

En esta parte se mostrará 3 imágenes con sus respectivas configuración y botones de descargar imagen en formato PNG. Imágenes como cluster dendrogram, mapa de factores 3D y mapa de factores 2D.

En esta parte se muestra una barra deslizante, dos barras de selección. En la barra deslizante llamada número de clúster se escogió un tamaño fijo de 10, en la primera barra seleccionable llamada seleccionar métrica se tiene varias selecciones las cuales son (“euclidean”,” manhattan”), la segunda barra es el método los cuales tienen (“ward”,”average”,”complete”,”single”).

Por ejemplo: se le dio un número de clúster de 3, métrica igual a euclidean y el método Ward.

Figura 30 Configuración HCPC Shiny

2.6.7.1 cluster dendrogram.

En esta parte se muestra dos casillas de selección, el grafico de árbol y un botón de descargar imagen en formato PNG. La primera casilla llamada rectángulos de los grupos del árbol es para graficar en el árbol la cantidad de clúster dado anteriormente, la segunda casilla llamada diagrama de barras de inercia es para graficar la inercia interna perdida.

Por ejemplo: se le dio que mostrara los números de clústeres

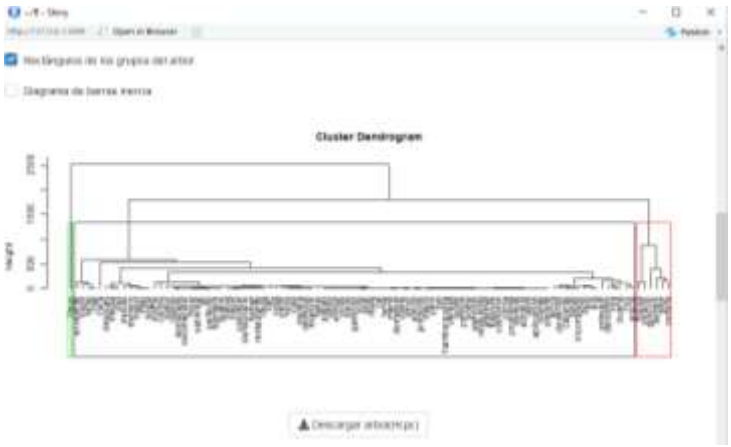


Figura 31 Árbol Shiny

2.6.7.2 Mapa de factores 3D

En esta parte se muestra dos casillas seleccionables, una barra deslizante, el grafico de mapa de factores en 3D y un botón de descargar imagen en formato PNG. En la primera casilla llamada nombres de los individuos 3D es para dibujar los nombres de los individuos en el plano 3D, la segunda casilla llamada centros de grupos 3D es para dibujar los clústeres en el plano 3D y la barra deslizante llamada ángulo de vista plano 3D es para visualizar las diferentes perspectivas del plano.

Por ejemplo: se le dio que mostrara los nombres y los centros de grupo con un ángulo de 60 grados.

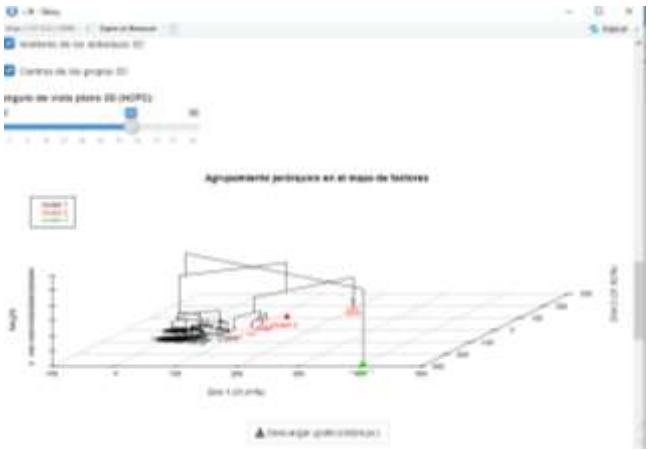


Figura 32 Mapa de Factores 3D Shiny

2.6.7.3 Mapa de factores 2D

En esta parte se muestra una casilla de selección, dos barras deslizantes, el grafico de mapa de factores en 2D y un botón de descargar imagen en formato PNG. En la casilla llamada ver mapa de factores es para seleccionar los elementos a ver en el gráfico como (“puntos y etiquetas”,” puntos”,” etiquetas”), en la primera barra deslizante llamada tamaño de los puntos se le dio un tamaño máximo fijo de 10, es para darle tamaño a los puntos graficado y en la segunda barra deslizante llamada tamaño de las etiquetas se le dio un tamaño máximo fijo de 20, es para darle tamaño a las etiquetas graficadas.

Por ejemplo: se le dio que mostrara puntos y flecha con un tamaño de punto igual a 1 y un tamaño de etiqueta igual 12.

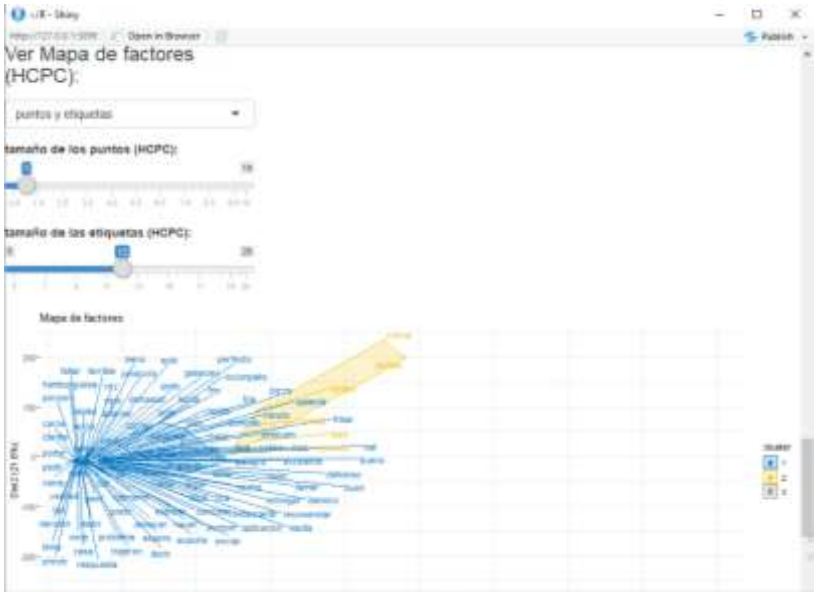


Figura 33 Mapa de Factores 2D Shiny

2.6.8 Red De Coocurrencia Por Pesos De Enlaces.

En esta parte se mostrará una barra deslizante, el grafico de red no dirigida y el botón de descargar imagen en formato PNG. En la barra deslizante que es frecuencia de palabras Red2D coocurrencia de pesos, tomará automáticamente el tamaño máximo, con la frecuencia mayor de los pares de palabras más frecuentes y el valor inicial será

la mitad del tamaño, en esta parte se podrá deslizar la barra para mostrar la cantidad de pares de palabras deseadas en el gráfico con mayor frecuencia.

Para el ejemplo: el tamaño máximo de la barra es de 302, pudimos deslizar hasta una frecuencia de 41, se observó la cantidad de palabras relacionadas.

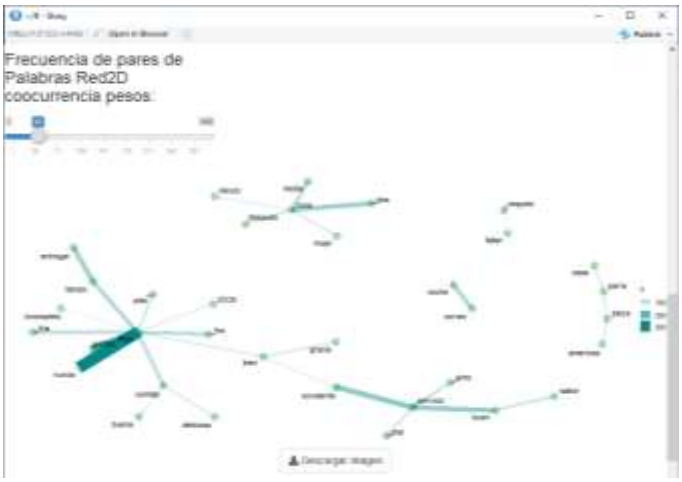


Figura 34 Red de Coocurencia Peso Shiny

2.6.9 Red De Coocurrencia Por Nodos.

En esta parte se mostrará una barra deslizante, el grafico de red no dirigida y el botón de descargar imagen en formato PNG. En la barra deslizante que es frecuencia coocurrencia por nodo, tomará automáticamente el tamaño máximo, de la suma de los pesos de la matriz de pesos por cada palabra y el valor inicial será la mitad del tamaño, en esta parte se podrá deslizar la barra para mostrar la cantidad de pares de palabras deseadas en el gráfico con mayor frecuencia.

Para el ejemplo: el tamaño máximo de la barra es de 326, pudimos deslizar hasta una frecuencia de 46, se observó la cantidad de palabras concurrentes

Figura 35 Red De Coocurencia Por Nodos Shiny

2.6.10 Red Por Grados Del Nodo

En esta parte se mostrará una barra deslizante, el grafico de red no dirigida y el botón de descargar imagen en formato PNG. En la barra deslizante que es frecuencia coocurrencia por nodo, en esta parte se podrá deslizar la barra para eliminar aquellos nodos que estén por debajo de la frecuencia dada, se muestra una guía que indica la cantidad de conexiones de los grados.

Para el ejemplo: el tamaño máximo de la barra es de 100, pudimos deslizar hasta una frecuencia de 70, se observó la cantidad de palabras concurrentes.

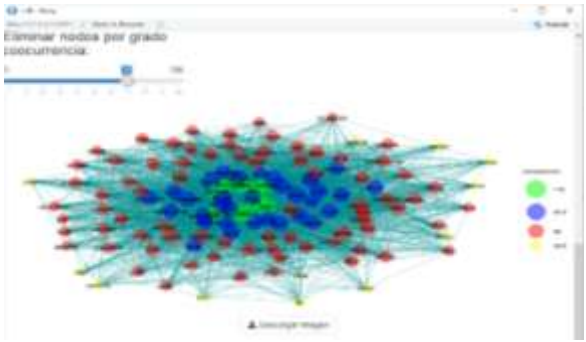


Figura 36 Red de Nodo por Grado Shiny

2.6.11 Emergencia, Auto-Organización, Complejidad.

En esta parte se muestra dos barras deslizantes, el grafico de EMERGENCIA, AUTO-ORGANIZACIÓN, COMPLEJIDAD y un botón de descargar imagen en formato PNG. En la primera barra deslizante llamada cantidad de palabras a mostrar se escogió un número máximo de palabras los cuales fuero 10 fijos a graficar, en la segunda barra deslizante que es base se escogió un número máximo de 20 fijos, que es la base del logaritmo a calcular.

Para el ejemplo: Pudimos deslizar la barra a un total de 5 palabras y una base de 10, luego se muestra las gráficas EMERGENCIA, AUTO-ORGANIZACIÓN, COMPLEJIDAD de las 5 palabras escogidas.

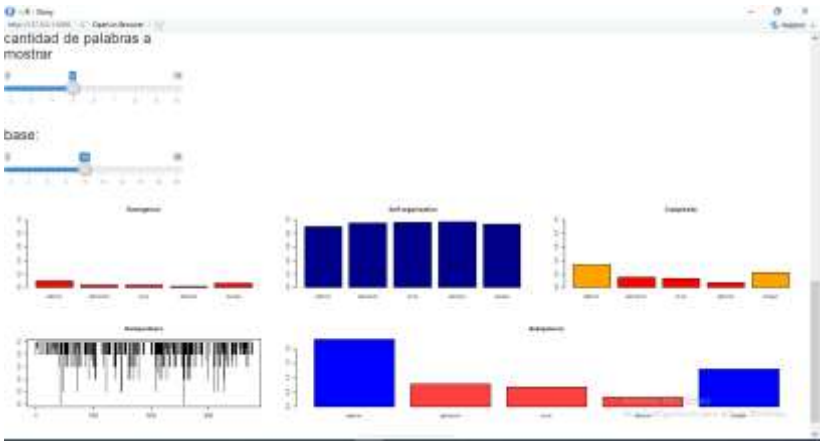


Figura 37 Emergencia, auto - organización, Complejidad

2.6.12 Diagrama de flujo del Algoritmo.

A continuación, se presenta gráficamente el funcionamiento del algoritmo ejecutado por la herramienta codificada en R, la idea es representar cada funcionalidad llevada a cabo durante el preprocesamiento y análisis de los datos montados a la herramienta.

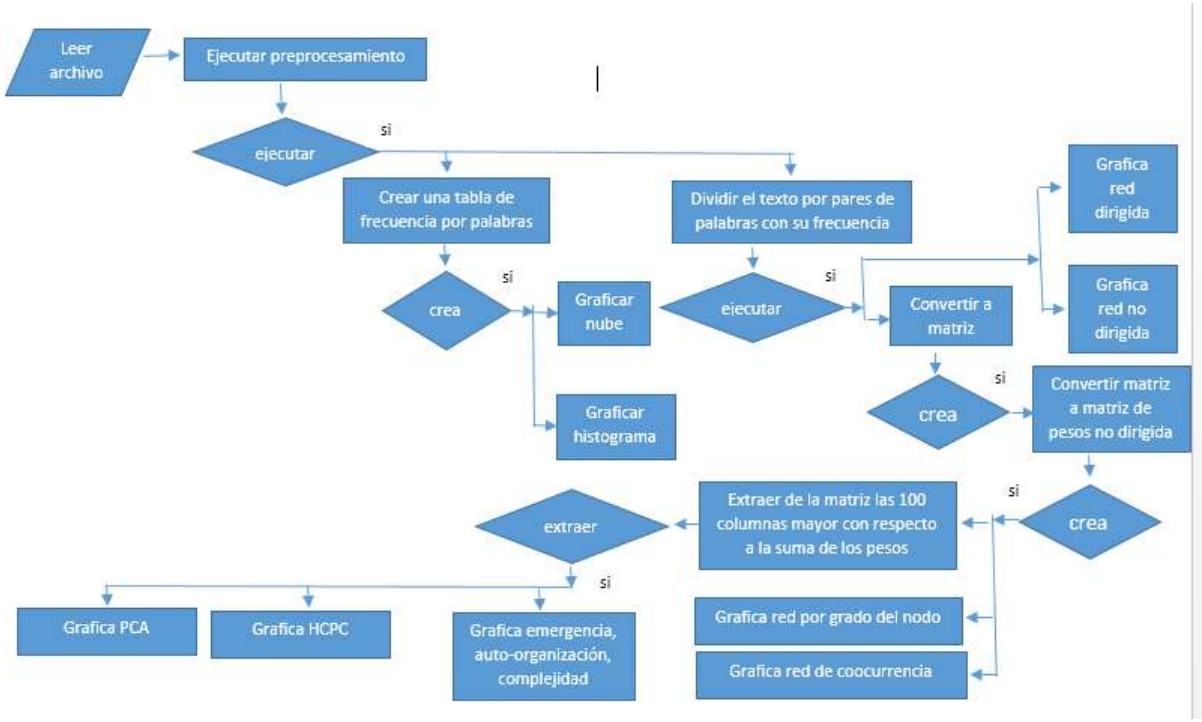


Figura 38 Diagrama de flujo

2.6.13 Costo computacional del algoritmo de preprocesamiento de datos

A continuación, se presenta gráficamente el costo computacional de las funciones que más tardan en procesar en el algoritmo de preprocesamiento de datos, donde se grafica el comportamiento de estas, lo cual se compara la cantidad de datos o palabras y es el tiempo que tarda en ejecutarse en segundos.

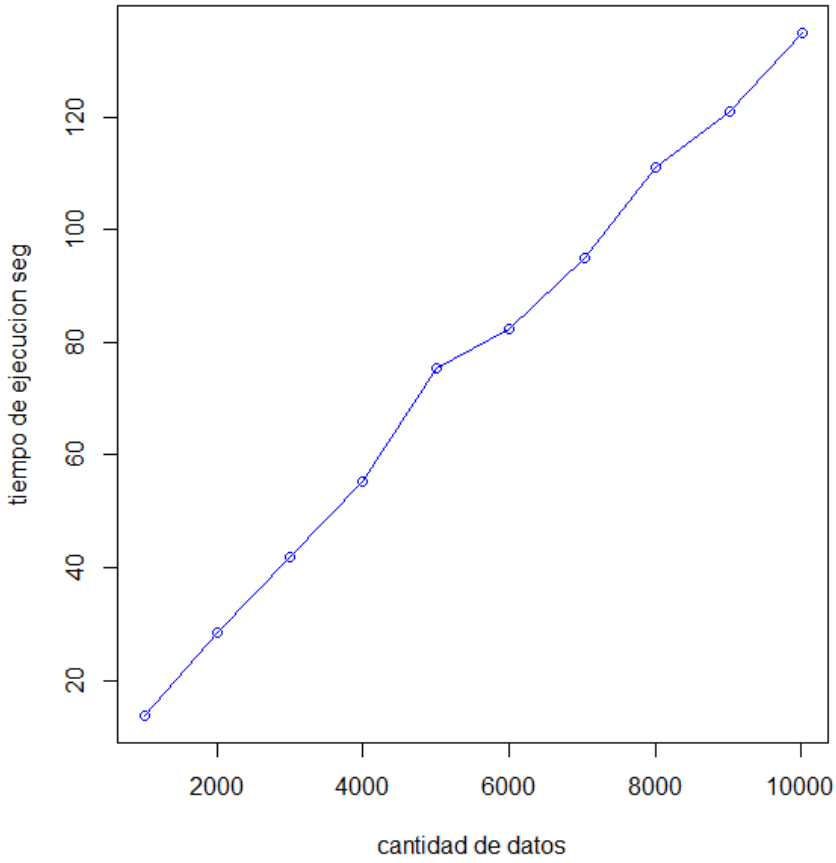


Figura 39 comportamiento de las funciones que demoran en algoritmo de preprocesamiento

En la figura 36 se observó que la cantidad de datos inicial es de 1005 palabras con un tiempo de ejecución 13.20 segundos y la cantidad de datos final, es de 10013 palabras con un tiempo de ejecución de 134.81 segundos. Y además se observó que el comportamiento de estas funciones es lineal.

2.6.14 Comparación del algoritmo de preprcesamiento

A continuación, se presenta gráficamente el costo computacional del algoritmo de preprocesamiento de datos con los datos de la base de datos del restaurante de comidas rápidas a domicilio, donde se grafica el comportamiento del algoritmo, donde se compara la cantidad de datos o palabras y es el tiempo que tarda en ejecutarse en segundos y la comparación del algoritmo de forma normal trabajando con los núcleo del procesador deshabilitado y la forma paralela trabajando con cuatro núcleos del procesador.

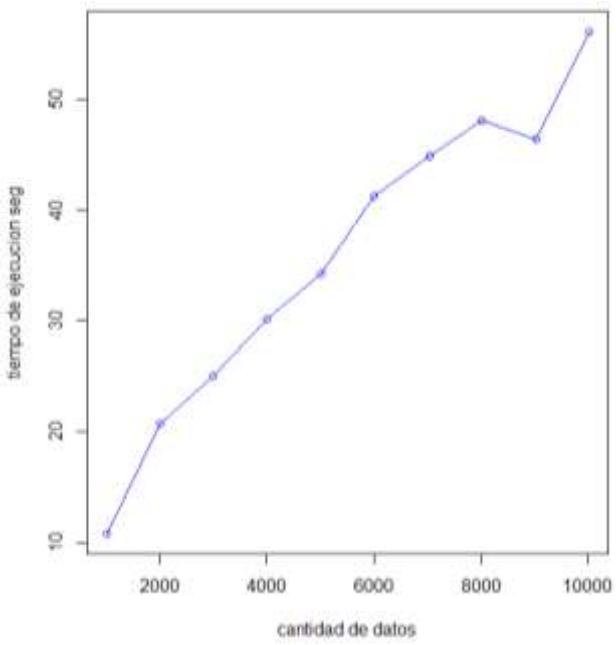


Figura 40 sin núcleo del procesador

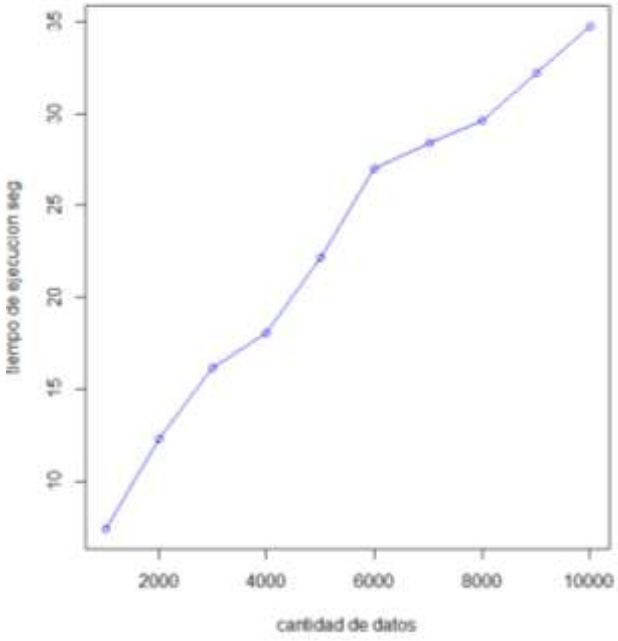


Figura 41 paralelo con 4 núcleo del procesador.

En la figura 37 se observó que la cantidad de datos inicial es de 1005 palabras con un tiempo de ejecución 10.85 segundos y la cantidad de datos final, es de 10013 palabras con un tiempo de ejecución de 56.06 segundos.

En la figura 38 se observó que la cantidad de datos inicial es de 1005 palabras con un tiempo de ejecución 7.41 segundos y la cantidad de datos final, es de 10013 palabras con un tiempo de ejecución de 34.75 segundos.

Conclusiones

A continuación, se abordarán las principales conclusiones dejadas en este trabajo por medio de la creación de la metodología:

- Se logró definir un estudio conceptual de temas relacionados con complejidad y autoorganización, además de las técnicas de minería de datos, brindando el soporte adecuado para el tratamiento de texto por medio de una metodología estadística para analizar grandes volúmenes de datos utilizando diversas técnicas de innovación en el área de ciencia de datos.

- Se obtuvo satisfactoriamente la implementación de una herramienta codificada en R que pudo verificar el funcionamiento de la metodología a través de un ejemplo real. Además, se demostró el uso de técnicas de auto-organización y complejidad dentro del área de Big data y ciencias de datos, comprobando su importancia como técnicas viables dentro de la inteligencia artificial para el análisis estructurado de datos.

- Finalizando el análisis se determinó exitosa la metodología ya que cumplía con los parámetros fijados por el modelo inicial, ya que mostraba una interpretación cualitativa y cuantitativa adecuada de los datos. Demostrando así que la metodología promueve el uso de las TICs en zonas importantes como el análisis de texto.

Referencias Bibliográficas

[1] Justicia, M.D.C. (2017). Nuevas Técnicas de Minería de Texto: Aplicaciones. (Tesis Doctoral). Universidad de Granada. España.

[2] Tan, A.-H. (1999). Text mining: Promises and Challenges. In Pacic Asia Conf On Knowledge Discovery and data Mining PAKDD´99 workshop on Knowledge Discovery from Advance Databases.

[3] Feldman, R. and Dagan, I. (1995). Knowledge discovery in textual database(kdt). In KDD, volume 95.

[4] Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1992) Knowledge discovery in tatabases: An overview. AI magazine.

[5] Hearst. M. A. (1999b). Untangling text data mining. In Proceesings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics.

[6] Kim, K., Ko, S., Elmqvist, N., and Ebert, D. S. (2011). Wordbridge: Using composite tag clouds in mode-link diagrams for visualizing content and relation in text corpora. In System Sciences (HICSS), 2011 44 th Hawaii International Conferences. IEEE.

[7] Arévalo, J. L, Perez, R, (n.d.). ESTADO DEL ARTE EN LA UTILIZACIÓN DE TECNICAS AVANZADAS PARA LA BUSQUEDA DE INFORMACIÓN NO TRIVIAL A PARTIR DE DATOS EN LOS SISTEMAS DE ABASTECIMIENTO DE AGUA POTABLE [Archivo PDF]. Recuperado de http://www.lenhs.ct.ufpb.br/html/downloads/serea/4serea/serea2002/trabalhos/A15_15.pdf

[8] IBM. (s.f.). IBM Knowledge Center. Obtenido de https://www.ibm.com/support/knowledgecenter/es/SS3RA7_18.1.1/ta_guide_ddita/textmining/shared_entities/tm_intro_tm_defined.html

[9] Salamanca, U. d. (2018). Universo Abierto. Obtenido de Blog de la biblioteca de Traducción y Documentación de la Universidad de Salamanca: https://universoabierto.org/2018/02/22/que-es-la-mineria-de-textos-como-funciona-y-por-que-es-util/

[10] ALBERT-LASZLO BARABASI. Network Science, Chapter 2 - Graph Theory.

[11] David Easley and Jon Kleinberg. Networks, crowds, and markets: Reasoning

about a highly connected world. Cambridge University Press, 2010.

[12] Mark Newman, Albert-Laszlo Barabasi, and Duncan J Watts. The structure and dynamics of networks. Princeton University Press, 2011.

[13] Universidad Nacional del Sur, (2016). Recuperado de http://ars-uns.blogspot.com/2016/11/ars-101-redes-de-co-ocurrencia.html. Obtenido de Curso de postgrado "Analisis de redes sociales ".

[14] Un enfoque Auto-organizado and Homeostático Emergente. Modelado multiagentes de sistemas dinámicos.

[15] NELSON FERN_ANDEZ, JOSE AGUILAR, and OSWALDO TERAN. Formalization of emergent properties in computing agent networks.

[16] Carlos Gershenson and Nelson Fern_andez. Complexity and information: Measuring emergence, self-organization, and homeostasis at multiple scales. Com-plexity, 18(2):29{44, 2012.

[17] Carlos Gershenson and Nelson Fernández. Complexity and information: Measuring emergence, self-organization, and homeostasis at multiple scales. Com-plexity, 18(2):29{44, 2012.

[18] Fernandez, N. F. (2015). Modelado Multi-agentes de Sistemas Dinámicos.

[19] Madrid, Y. A. (2016). Complejidad Estructural y Dinámica en Redes Libres de Escala y Pequeño Mundo. Pamplona - Colombia.

[20] Swanson, Don R., idem.

[21] Text mining in a digital library. Ian H. Witten, Katherine J. Don, Michael Dewsnip, Valentin Tablan.

[22] PowerData(2016). Calidad de datos en minería de datos a través del preprocesamiento

[23] Hernández y Rodríguez (2013). PREPROCESAMIENTO DE DATOS ESTRUCTURADOS, Universidad Distrital Francisco José de Caldas

[24] ALBERT-LASZLO BARABASI. Network Science, Chapter 2 - Graph Theory.

[25] Ferrero-López. Data Science. Máxima formación. Recuperado de https://www.maximaformacion.es/blog-dat/que-es-r-software/#comments

[26] Logicali(12 de febrero 2015). Data cleansing y sus fases: contra los problemas de calidad de datos. Recuperado de https://blog.es.logicalis.com/analytics/data-cleansing-y-sus-fases-contra-los-problemas-de-calidad-de-datos

[27] Nae_(23 DICIEMBRE, 2015). Caso práctico: el proceso de transformar los datos en información útil. Recuperado de https://nae.global/caso-practico-el-proceso-de-transformar-los-datos-en-informacion-util/

[28] Universidad de Carlos III de Madrid (s.f). METODOLOGIA DE ANALISIS DE DATOS. Recuperado de http://ocw.uc3m.es/ingenieria-informatica/analisis-de-datos/transparencias/metodologia.pdf

[29] Anil Maheshwari and Michiel Smid. Introduction to theory of computation. Free Online, 2012.

[30] Peter Pacheco. An introduction to parallel programming. Elsevier, 2011.

[31] Madrid, Y. A. (2016). Complejidad Estructural y Dinámica en Redes Libres de escala y Pequeño Mundo. Pamplona - Colombia.

[32] Ortega, O. (2019 SISTEMAS CON DINÁMICA ACOPLADA (SCDs) - UN ENFOQUE METODOLÓGICO DESDE EL ANÁLISIS FORMAL DE CONCEPTOS PARA LA INTEGRACIÓN DE TEORÍAS MULTIPLES AGENTES, REDES Y JUEGOS -. *(Tesis Pregrado). Universidad de Pamplona. Colombia.*