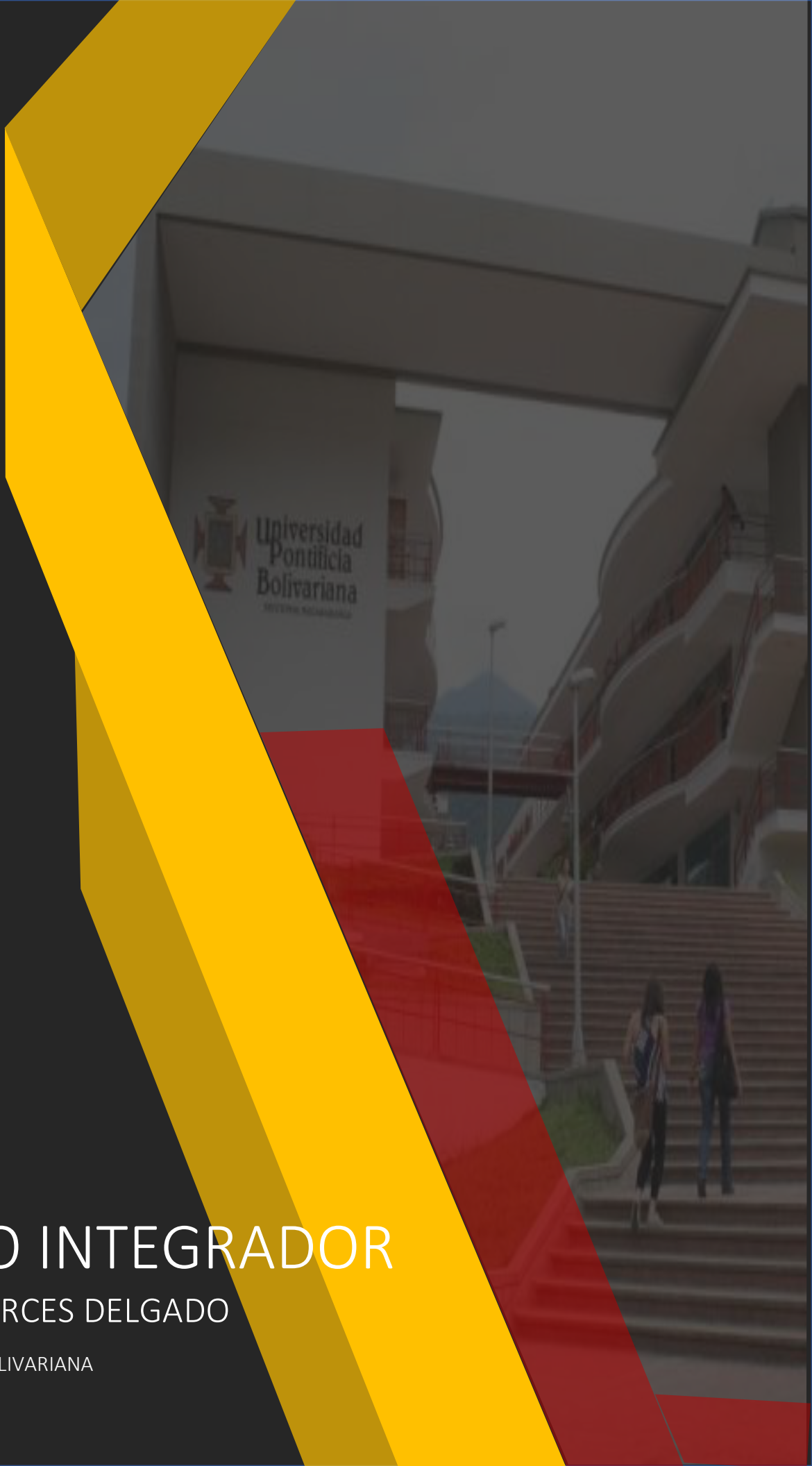

PROYECTO INTEGRADOR

Autora: JENNY GARCES DELGADO

UNIVERSIDAD PONTIFICIA BOLIVARIANA

CARACTERÍSTICAS GENERALES

NOMBRE DEL PROYECTO:

NLP DATA

MATERIAS INTEGRADAS:

OPTATIVA: SISTEMAS AVANZADOS DE BASES DE DATOS
Encargado: **JUAN SEBASTIAN GOMEZ ROSAS**

GERENCIA DE PROYECTOS TECNOLÓGICOS
Encargado: **AIZAR MEJIA JALABE**

AUTOR:

JENNY MARCELA GARCÉS DELGADO

DESCRIPCIÓN CORTA:

NLP DATA es un componente que integra las bases de datos no relacionales (mongo DB) y Python para la aplicación del procesamiento del lenguaje natural.

METODÓLOGIA USADA:

KANBAN (Tarjeta Visual)

Nota temporal: la siguiente información esta sujeta las entregas programadas en las 2 materias integradas.

TABLA DE CONTENIDO

SITUACIÓN PROBLEMÁTICA:

Con el fin de automatizar procesos, estudiantes de las materias de Estructuras de Datos y Matemáticas Discretas II de la Universidad Pontificia Bolivariana seccional Bucaramanga desarrollan un chatbot capaz de entender, interpretar y dar solución a preguntas referentes a procesos desarrollados por su universidad; Dicho proceso requiere de la conexión a una base de datos avanzada que contenga la información requerida para que dichos estudiantes puedan implementar sus algoritmos de procesamiento de lenguaje natural (NLP).

PREGUNTA PROBLEMÁTICA

¿Cómo desarrollar un componente capaz de conectarse con una base de datos no relacional, usando el lenguaje de programación Python?

MARCO CONCEPTUAL:

NLP

El procesamiento del lenguaje natural (NLP, por sus siglas en inglés) es una rama de la inteligencia artificial que ayuda a las computadoras a entender, interpretar y manipular el lenguaje humano. NLP toma elementos prestados de muchas disciplinas, incluyendo la ciencia de la computación y la lingüística computacional. [1]

BASES DE DATOS NOSQL

Las bases de datos NoSQL están diseñadas específicamente para modelos de datos específicos y tienen esquemas flexibles para crear aplicaciones modernas. Las bases de datos NoSQL son ampliamente reconocidas porque son fáciles de desarrollar, por su funcionalidad y el rendimiento a escala. [2]

PYTHON

es un lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código.2 Se trata de un lenguaje de programación multi paradigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma. [3]

MONGODB

MongoDB es una base de datos distribuida, basada en documentos y de uso general que ha sido diseñada para desarrolladores de aplicaciones modernas y para la era de la nube. MongoDB es una base de datos documental, lo que significa que almacena datos en forma de documentos tipo JSON. [4]

CHATBOT

Por lo tanto, un chatbot es un software que utiliza mensajes estructurados para emitir respuestas desde una máquina hacia un interlocutor humano. [5]

REGEX

Las expresiones regulares son patrones utilizados para encontrar una determinada combinación de caracteres dentro de una cadena de texto. [6]

KANBAN

es una de las llamadas metodología ágiles, aquellas que buscan gestionar de manera generalizada cómo se van completando las tareas.

Las principales ventajas que aporta la metodología Kanban son que, dada su representación a través de tarjetas, es una metodología muy visual y sencilla, por lo que es fácilmente incorporable al sistema y procesos de una empresa. [7]

WEB SCRAPING

Web scraping es una técnica utilizada mediante programas de software para extraer información de sitios web. Usualmente, estos programas simulan la navegación de un humano en la World Wide Web ya sea utilizando el protocolo HTTP manualmente, o incrustando un navegador en una aplicación. [8]

OBJETIVOS

GENERAL

Desarrollar un componente capaz de conectarse a una base de datos no relacional, usando el lenguaje de programación Python.

ESPECÍFICOS

Diseñar el modelo no relacional de la base de datos aplicando conceptos de colecciones, registros y documentos.

Sembrar documentos en la base de datos usando herramientas como Node.js

JUSTIFICACIÓN:

El siguiente proyecto se enfocará en la importancia que representa el manejo de los datos a la hora de aplicar ramas de la ciencia de la computación, tal como lo es la inteligencia artificial, en este caso aplicada a chatbots desarrollados con algoritmos de machine learning en el lenguaje Python. De igual forma el sembrado de datos

para obtener información y poder generar conocimiento, usando técnicas como el web scraping para la extracción de información de la web; Todo esto para fortalecer los conocimientos adquiridos en clase y el uso de herramientas para el manejo de nuevas tecnologías que actualmente son tendencia en el ámbito profesional.

METODOLOGÍA:

Para el desarrollo de este proyecto se usa la metodología de Kanban la cual consiste el uso de un tablero con cada una de las actividades a realizar.

IMAGEN 1-TABLERO KANBAN

Unos de sus principios son:
calidad garantizada: las cosas deben salir bien a la primera, se tarda más en arreglar algo que sale mal, además de consumir más recursos, que cuando sale bien a la primera. Por ello, lo más importante no es que se haga rápido, sino que se haga bien.
Reducción del desperdicio: no se necesita hacer nada extra o superficial, solamente lo necesario para que salga bien. De este modo se optimizan recursos. **Mejora continua:** aprovechando la realización de tareas, se busca mejorar los procesos, a través de un sistema de mejora continua.
Flexibilidad: se dispone de capacidad de respuesta ante tareas no previstas, de forma que exista una "cola de espera" de tareas en las que ir priorizando su realización en función de las necesidades de cada momento y de la urgencia de cada una de ellas.

CRONOGRAMA DE ACTIVIDADES:

El proyecto se divide en cuatro etapas principales, para cada actividad se plantean sus actividades principales y el nivel de prioridad de cada una de ellas.

Primer Avance		Segundo Avance		Tercer Avance		Entrega Final	
prioridad	actividad	prioridad	actividad	prioridad	actividad	prioridad	actividad
	Propuesta		Conexión a la base de datos		Pruebas de python		Consultas finales
	Analisis del problema		Pruebas de consultas		Aplicación de web scrapin		Pre-Sustentacion
	Investigacion de herramientas		Analisis de datos		Estudio de mercados		Sustentación
			Estudio tecnico		Sembrado de la base de datos		

TABLA 1-CRONOGRAMA DE ACTIVIDADES

Color	Tipo de Prioridad
	Minima
	Media
	Máxima

TABLA 2-NIVELES DE PRIORIDAD

DESARROLLO DEL PROYECTO

ESTUDIO DE MERCADOS

FICHA TÉCNICA:

En el estudio de mercados se realiza la ficha técnica para brindar a los clientes una breve descripción del producto ofrecido.

Ficha Técnica	
Nombre de Producto	NLP Data
Descripción	NLP DATA es un componente que integra las bases de datos no relacionales (mongo DB) y Python para la aplicación del procesamiento del lenguaje natural.
Requisitos	Instalación de Python superior a 2.7; Instalacion de la librería pymongo en python, en su defecto el software de Anaconda con Jupyter; En algunos casos la instalacion de mongodb(Herramienta avanzada)

Condiciones de uso	El codigo fuente no podra ser alterado, de lo contrario el componente no funcionara de acuerdo a lo establecido.
Imagen	N/A
Información adicional	Para su correcto uso consulte el manual.
Licenciamiento	GNU GPL : Libre, abierta y gratuita
Presentación	Se presenta como un archivo .PY el cual puede consumirse dependiendo de los requerimientos
Lugar de elaboración	Universidad Pontificia Bolivariana

TABLA 3-FICHA TECNICA

FICHA TÉCNICA DE LA ENTREVISTA:

En el caso de este estudio no se realizan encuestas, ya que el componente desarrollado es para un cliente especifico, el cual a partir de una entrevista se determinan los requisitos de la aplicación.
Sesión de preguntas:

1. **¿Es necesario la implementación de algún lenguaje de programación especifico para el desarrollo?**

Sí, los estudiantes requieren de Python para que la conexión con los algoritmos sea mas fácil.

2. **El gestor de base de datos que uso es mongo Atlas ¿Requieren de algún otro sistema de administración?**

El proyecto esta pensado para bases de datos relacionales, pero también se puede implementar usando mongoDB.

3. **¿Cuál es la procedencia de los datos?**

Las preguntas para el chatbot son otorgadas por los estudiantes.

4. **¿Actualmente hay fuentes para la extracción de datos?**

Los datos pueden obtenerse mediante la aplicación de web Scraping

5. **¿Las expresiones regulares están dadas por los estudiantes o el lenguaje de programación?**

Son las consideradas por Python.

notas para considerar: los estudiantes no aplicaran conceptos de mongo a menos que sea necesario; El desarrollo debe tener un enfoque desde el ámbito de la gestión de la base de datos no relacional.

Con respecto a esto se realizan los requerimientos del proyecto.

ALCANCE DEL PROYECTO:

La aplicación debe tener la capacidad de conectar el lenguaje de programación de Python con una base de datos no relacional, la cual va a contener diferentes colecciones que puedan ser consultadas por el usuario de forma ágil.

FUNCIONALIDADES DEL PRODUCTO:

Conexión con una base de datos no relacional dependiendo del tema (Bienestar Universitario, etc.).

Guardar, modificar y consultar tanto palabras claves como respuestas.

CARACTERISTICAS Y USUARIOS:

La base de datos va a estar alojada en un Cluster de mongo Atlas, para que pueda ser consultada de forma remota, el acceso estará embebido en el código según el tema que corresponda para su chatbot.

Se entregará un archivo .py el cual contiene los métodos para la consulta, inserción y modificación de los documentos de las colecciones correspondientes.

REGLAS DEL NEGOCIO:

La modificación del código puede alterar el funcionamiento de este, el código base no esta exento de cambios. Se recomienda hacer copia del archivo original y hacer los debidos cambios en un nuevo archivo.

REQUERIMIENTOS DE INTERFACES EXTERNAS:

Para ser ejecutado el archivo debe contar con el software de Python o la instalación por consola, para su correcto funcionamiento la versión de este debe ser superior a 2.7.

MODELAMIENTO DE DATOS

Antes de empezar con el modelado de base de datos debemos tener en cuenta que este proyecto esta desarrollado en lenguaje noSQL y la herramienta usada en este caso es mongoDB, el cual cuenta con algunos conceptos que pueden confundirse a la hora de revisar el modelo de datos.
De manera general estos son algunos que pueden ser de interes:

Lenguaje SQL	Lenguaje noSQL
Database	Database
Table	Collections
Rows	Documents
Index	Index

TABLA 4-CONCEPTOS DE MONGODB

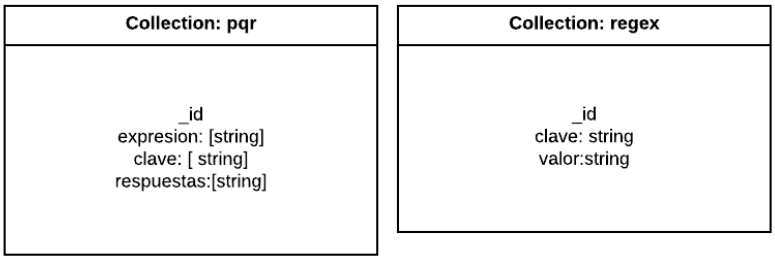

TABLA 5-MODELO DE LA BASE DE DATOS

Debido a los requerimientos se plantean dos colecciones una de preguntas y respuesta y la otra será una colección de expresiones regulares.

ESTUDIO TECNICO

En el estudio tecnico se contemplan los aspectos tecnicos operativos necesarios en el uso eficiente de los recursos disponibles para la elaboracion del proyecto y en el cual se analizan la distribucion del espacio, las localizaciones y los procesos de la empresa.

LOCALIZACIÓN:

El primer aspecto que analizamos es la localización, este estudio esmuy util para determinarel éxito o fracaso de un negocio, ya que la decisión aceca de donde ubicar el proyecto no solo considera criterios economicos, sino tambien estrategicos, tecnicos y sociales, entre otros.
Para evaluar este aspecto usamos el método de los factores ponderados, donde se presenta un analisis cuantitativo en el que se comparab entre si las diferentes alternativas para conseguir determinar una o varias localizaciones aceptables.

Nº	Factores	Peso relativo(%)	alternativas							
			De la cuesta oficinas(Piedecuesta)		EcoparqueEmpresarial Anillo Vial		Real de Minas apto		metropolitan bucaramanga	
1	Sistemas de transporte	15%	10	1,5	6	0,9	8	1,2	8	1,2
2	Impuestos	10%	5	0,5	7	0,7	7	0,7	5	0,5
3	Servicios Públicos	10%	6	0,6	5	0,5	8	0,8	6	0,6
4	Disponibilidad de mano de obra	10%	5	0,5	7	0,7	8	0,8	10	1
5	Salarios de la región	5%	8	0,4	7	0,35	7	0,35	7	0,35
6	Proximidad a clientes	5%	5	0,25	6	0,3	7	0,35	9	0,45
7	Seguridad de la zona	10%	8	0,8	10	1	7	0,7	9	0,9
8	Conectividad (TI)	20%	8	1,6	9	1,8	9	1,8	10	2
9	Ambiente laboral	15%	8	1,2	9	1,35	7	1,05	8	1,2
		100%		7,35		7,6		7,75		8,2

TABLA 6-TABLA DE FACTORES PONDERADOS

En este caso tomamos 4 diferentes ubicaciones del area metropolitana de Bucaramanga, primero se determinan los factores relevantes y se asigna un peso en porcentaje, luego se fija una escala (en este caso de 1 a 10), la puntuacion se multica por el procentaje para luego sumar y obtener la mejor localización.

DISTRIBUCIÓN Y DISEÑO DE LAS INSTALACIONES

Para que la distribución y diseño de las instalaciones de un proyecto provean condiciones de trabajo aceptables, es preciso tomar en cuenta dos especificaciones en particular: funcionalidad y estetica que proporcionen y optimicen el trabajo en cada una de las areas. El diseño de esta oficina se hace teniendo en cuenta el ambiente laboral que se desea tener.

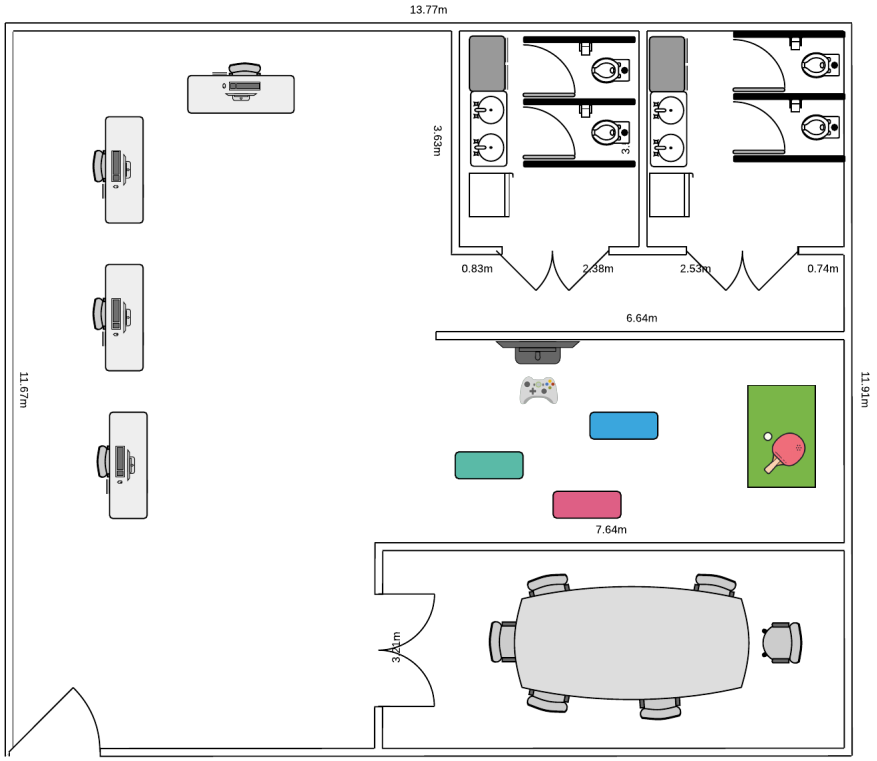

IMAGEN 2-LAYOUT

PROCESOS DEL DESARROLLO

Los diagramas de flujo es una manera de representar gráficamente ciertos procesos a través de una serie de pasos estructurados y vinculados que permiten su revisión. La representación grafica de estos procesos emplea en los diagramas de flujo una serie determinada de figuras geométricas que representan cada paso o procedimiento del desarrollo.

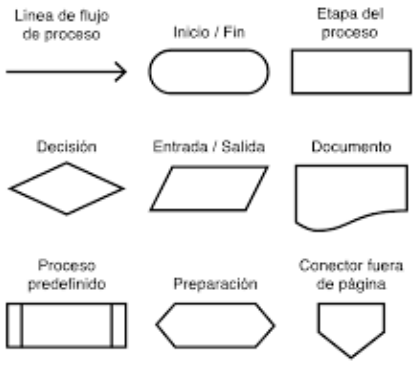

IMAGEN 3-ELEMENTOS FLUJOGRAMA

Para entender un poco estos diagramas podemos visualizar algunos de los elementos y su significado o uso.

El siguiente diagrama de flujo se realizan con el fin de dar a entender el funcionamiento general de la aplicación NLP data.

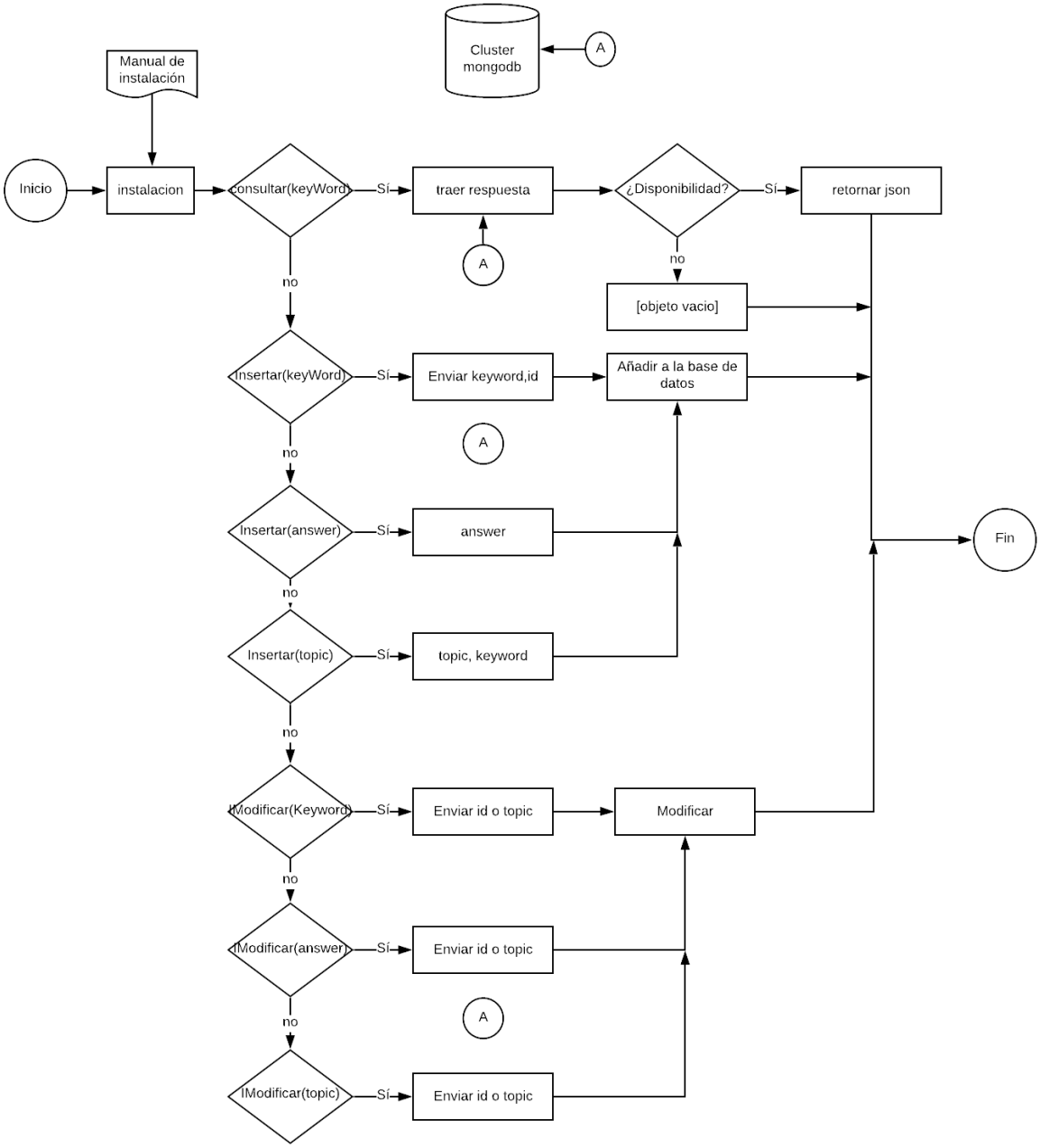

IMAGEN 4-FLUJOGRAMA NLP

Bibliografía

[1] sas.com, «https://www.sas.com,» [En línea]. Available: https://www.sas.com/es_co/insights/analytics/what-is-natural-language-processing-nlp.html.

[2] amazon.com, «https://aws.amazon.com,» [En línea]. Available: https://aws.amazon.com/es/nosql/.

[3] .wikipedia.org, «https://es.wikipedia.org,» [En línea]. Available: https://es.wikipedia.org/wiki/Python.

[4] mongodb.com, «https://www.mongodb.com,» [En línea]. Available: https://www.mongodb.com/es.

[5] J. Charlan, «www.esic.edu,» 08 08 2018. [En línea]. Available: https://www.esic.edu/rethink/2018/08/04/que-es-un-chatbot-y-para-que-sirve/.

[6] mozilla.org, «developer.mozilla.org,» [En línea]. Available: https://developer.mozilla.org/es/docs/Web/JavaScript/Guide/Regular_Expressions.

[7] getbillage.com, «www.getbillage.com,» [En línea]. Available: https://www.getbillage.com/es/blog/metodologia-kanban-ventajas-y-caracteristicas.

[8] wikipedia.org, «es.wikipedia.org,» [En línea]. Available: https://es.wikipedia.org/wiki/Web_scraping.

[9] concepto.de, «concepto.de,» [En línea]. Available: https://concepto.de/diagrama-de-flujo/.