Instituto Politécnico Nacional

Escuela Superior de Cómputo

TRABAJO TERMINAL

Recolector y clasificador de noticias

2018-B013

PRESENTAN:

CARLOS ANDRES HERNANDEZ GOMEZ
LUÍS DANIEL MEZA MARTÍNEZ

DIRECTORES:

M. en C. JOEL OMAR JUÁREZ GAMBINO
Dra. CONSUELO VARINIA GARCÍA MENDOZA

Ciudad de México,14 de mayo de 2019

Índice general

Capítulo 1

Introducción

El artículo periodístico o noticia, es la información de un hecho de interés ocurrido en un periodo de tiempo determinado. Constituye el elemento primordial en la información de la prensa y del género básico del periodismo (Internaútica, 2018). Conocer los acontecimientos del mundo independientemente del tema, día o lugar en el cual se han suscitado, tiene una gran importancia en la sociedad, se comparten por distintos medios de comunicación, tales como la televisión, redes sociales, diarios, blogs y la radio. Nos permiten conocer la situación económica del país, logros de la ciencia, desastres naturales, la situación en cuestión de inseguridad entre otros hechos. En el ámbito de las inversiones, crean expectativas y eso a su vez puede modificar los planes de inversión en cualquier sector, siendo así de suma importancia compartirlas de una forma eficaz (Manning et al., 2010).

El uso de páginas web como medio de comunicación está en incremento, permitiendo consultar noticias de distintos sitios como los periódicos electrónicos; su información al igual que un diario tradicional se encuentra dividida en secciones para facilitar la consulta, sin embargo, la clasificación suele variar en cada portal, incluso teniendo el mismo contenido. Un problema mayor se encuentra en los sitios independientes, los cuales no cuentan con una segmentación particular, haciendo difícil realizar una búsqueda eficaz.

1.1. Problemática

Los métodos tradicionales para la recopilación de información de los recolectores web (Crawler), están basados en las etiquetas o marcadores que los sitos añaden a su código fuente, por ejemplo, algunos artículos periodísticos son etiquetados a la sección que pertenecen (política, deporte, cultura, etc). Sin embargo, existen muchas fuentes de información que no etiquetan sus publicaciones, incluso si la tarea es realizada, dicha segmentación no indica claramente el tipo de contenido; Al consultar los portales mas visitados en México (en el giro del periodismo) se encuentra definida la sección deportes con varios sinónimos como **Universal deportes** (diario El Universal), **La afición** (Milenio), **Adrenalina** (Excelsior), etc. Como este ejemplo se encuentran más. Las noticias son segmentadas de forma tan diversa que ha complicado su búsqueda en la Internet.

Para definir las etiquetas o marcadores con los cuales se clasifica la información de los sitios web, se requiere un proceso manual de análisis de la información. Este proceso implica tiempo y esfuerzo por parte de las personas que realizan el trabajo. Por lo anterior se plantea la necesidad de crear métodos para automatizar esta tarea.

1.2. Justificación

Hoy en día existen distintas maneras de informarse acerca de los acontecimientos más recientes, por ejemplo, la televisión, blogs, redes sociales, foros, diarios, etc. Esto ha provocado que la información se encuentre dispersa y se deba acceder a múltiples recursos para ser recopilada, implicando una inversión de tiempo y esfuerzo. Para facilitar esta tarea, existen herramientas que hacen la búsqueda de noticias de interés para el usuario en forma automática. Sin embargo, dichas herramientas requieren que los sitios a consultar tengan etiquetas definidas y homogéneas.

Según el díario El Economista (Economista, 2019) el sitio web Animal Político[1] ocupa el lugar número cuatro en el ranking de medios nativos digitales, clasifica sus noticias de una manera poco habitual para los lectores como la sección **El sabueso**, **El plumaje**, **Hablemos de . . .** , entre otras, lo que hace complicado obtener los artículos con los métodos tradicionales de recopilación que, se basan sólo en las etiquetas que identifican cada sección y no el contenido de las noticias.

1.3. Solución Propuesta

Se propone crear una aplicación web que recolecte y clasifique noticias de acuerdo a su contenido y periodo de publicación. Finalmente, las noticias que satis-

[1] www.animalpolitico.com

fagan ambos filtros (Tipo de contenido y fecha de publicación) serán mostradas al usuario.

1.4. Objetivo

Crear un recolector de noticias, el cual permita recopilar información de diferentes fuentes como diarios, sitios de noticias, foros y mediante el análisis automático de su contenido muestre aquellas noticias que satisfagan los filtros establecidos por el usuario.

1.5. Objetivos Específicos

- Desarrollar un recolector de noticias, el cual permita obtener información de diferentes fuentes como diarios, sitios de noticias, blogs y foros
- Analizar de forma automática el contenido de las noticias para satisfacer los filtros establecidos por el usuario
- Mostrar las noticias que cumplieron con los filtros establecidos, así como su enlace (URL) para redirigirlos a la página de la noticia
- Afinar el clasificador de noticias realizado en el trabajo terminal 2017-A02 para utilizarlo en el contexto de esta propuesta (filtro de sección)

Capítulo 2

Estado del arte

2.1. Introducción

El uso de la información digital ha superado la producción de libros y publicaciones impresas, este fenómeno ha influenciado la producción de bibliotecas digitales, publicaciones electrónicas; Se ha incrementado el uso de las redes sociales, correos electrónicos, creando un gran repositorio de información útil, el cual puede ser analizado(Aggarwal, 2018).

Debido a la necesidad de procesar grandes volúmenes de datos recolectados de Internet, se han desarrollado diversas investigaciones entorno a esta tarea. A continuación se muestran distintos artículos nacionales e internacionales relacionados al campo de investigación (Clasificación de noticias), de igual forma se muestran herramientas web que desempeñan un trabajo similar al propuesto (Sitio web de noticias). Cabe destacar que el área de interés cuenta con un amplio desarrollo, no obstante solo se mencionan los trabajos más relevantes para este documento.

2.2. Trabajos nacionales

Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático

Este trabajo terminal de la escuela superio de cómputo (García et al., 2018) clasifica mediante técnicas de aprendizaje automático, noticias de diarios de circulación nacional en las diferentes secciones en que en estos se dividen. Se recolectaron 4,027 artículos de tres diarios de circulación nacional: **El universal**, **La jornada** y **Excélsior**. 3,624 noticias fueron utilizadas para la etapa de entrenamiento y 407 para hacer las pruebas. Los algoritmos implementados fueron los siguientes:

Selección de características

- Frecuencia
- Binarización
- TF-IDF

Clasificación

- Árboles de decisión
- Máquinas de soporte vectorial
- Naive Bayes Multinomial
- Regresión logística

El trabajo utiliza pre-procesamiento de información con la técnica tokenización y lematización (Ver Capítulo 3). El mejor resultado en las prubas se dio en la combinación del algoritmo **TF-IDF** para extraer las características y **Máquinas de soporte vectorial** para la clasificación de artículos, se obtuvo un 79.81 % de exactitud, *i.e* 8 de cada 10 noticias son clasificadas correctamente.

Clasificación automática de textos de desastres naturales en México

En este trabajo se propone clasificar noticias en el ámbito **desastres naturales** (Téllez-Valero et al., 2019), utilizando estrategias de reducción de dimensionalidad conocidas como, umbral en la frecuencia y ganancia en la información, los métodos de clasificación utilizados fueron el clasificador simple de Bayes y vecinos más cercanos.

Se utilizaron 375 noticias del periódico Reforma como conjunto de entrenamiento, se clasificaron en artículos relevantes e irrelevantes, de los cuales 11.5 % de

noticias eran relevantes y el 88.5 % restante eran irrelevantes. Una vez obtenido el conjunto de noticias se procedió con un pre-procesamiento, el cual reduce el tamaño de los documentos, eliminando la parte de los textos que no brindan información útil, posteriormente se realizó un indexado: Los documentos son representados por vectores de palabras en un espacio de dimensión n, para realizar una reducción de dimencionalidad. Finalmente se utilizaron técnicas de clasificación (Algoritmo simple de Bayes) con el cual se obtuvo un resultado de 97 % de efectividad en la clasificación de noticias.

News article classification of mexican newspapers

En este trabajo se propone clasificar noticias utilizando métodos supervisados de aprendizaje automático (Mejor conocido como *Machine learning* en lenguaje ingles) para su clasificación(García-Mendoza and Gambino Juárez, 2018). Para realizar esta tarea se recolectaron 4,027 artículos junto con su sección correspondiente de tres periódicos mexicanos duranta un periodo de 6 meses. Diferentes características fueron extraídas y un conjunto de métodos de aprendizaje fueron probados. Los resultados obtenidos muestran una exactitud de 80 % en la clasificación de los artículos en su correspondiente sección de los tres periódicos seleccionados.

Usando aprendizaje automático para extraer información de noticias de desastres naturales

Este trabajo describe un sistema basado en métodos de Aprendizaje automático que mejora la adquisición de datos de desastres naturales(Téllez Valero et al., 2009). Este sistema automáticamente llena una base de datos de desastres naturales con la información extraída de noticias de periódicos en línea. En particular, se extrae información acerca de cinco tipos de desastres naturales: huracanes, temblores, incendios forestales, inundaciones y sequías. Los algoritmos implementados para la extracción de información son los siguientes:

- Naive bayes
- Maquinas de soporte vectorial
- C4.5

Los resultados experimentales en una colección de noticias en Español muestran la eficacia del sistema propuesto tanto para detectar documentos relevantes sobre desastres naturales (alcanzando una medida-F de 98 %), así como para extraer hechos relevantes para ser insertados en una base de datos dada (alcanzando una medida-F de 76 %).

2.3. Trabajos internacionales

Clasificador de noticias usando autoencoders

En este trabajo se propone la clasificación de noticias utilizando *Deep Learning* (Farias et al., 2018), las noticias se clasificaron en las siguientes categorias:

- Deportes
- Política
- Espectáculos
- Economía
- Policía

El alcance que tiene es:

- Local (Valpara iso)
- Nacional (Chile)
- Internacional (Resto del mundo)

El clasificador se constuyó utilizando una base de datos con 542 noticias etiquetadas con los criterios anteiores, las características se obtuvieron utilizando Autoencoders (AE) para entrenar una Red Neuronal Artificial (ANN). Los resultados obtenidos con 156 noticias fue una tasa de éxito del 92.3 % para la clasificación de la categoría y un 87.2 % para el clasificador de alcance. La tasa general de éxito, categoría y alcance fue de 83.75 %.

Document classification for newspaper articles

El trabajo clasifica artículos de la universidad *Massachusetts Institute of Technology* (Ramdass and Seshasai, 2009) en las categorías: *Arts*, *Features*, *News*, *Opinion*, *Sports*, *World*. Los algoritmos implementados para la extracción de características y clasificación son los siguientes:

Selección de características

- *Multi-Variate Bernoulli Featureset*
- *Multinomial Featureset*
- *Normalized Multinomial Featureset*

Clasificación

- *Naive Bayes Classification*

- *Maximum Entropy Classification*

Para la etapa de entrenamiento se ocupó un total de 480 artículos por sección, y para realizar las pruebas 120 noticias. Los resultados relevantes se muestran en la siguiente tabla:

Selección de características	Clasificador	Exactitud
Multi-Variate Bernoulli Featureset	Naive bayes	0.7667
Multi-Variate Bernoulli Featureset	Maximun entropy	0.7233

El mejor resultado se obtiene utilizando *Multi-Variate Bernoulli Featureset* como algoritmo de extracción de características y *Naive Bayes Classification* como algoritmo clasificador ya que, obtiene un 77 % de exactitud.

Categorization of web news documents using word2vec and deep learning

El trabajo desarrolla un algoritmo utilizando *Word2Vec* como entrada a una red neuronal profunda (Yuan et al., 2014), para clasificar noticias en 6 categorías:

- *Entertainment*
- *Sports*
- *The economy*
- *It, science*
- *Domestic*
- *Overseas*

En el análisis se ha utilizado textos redactados en lenguaje japones obtenidos de sitios web como Yahoo. Se ha utilizado 600 noticias para el entrenamiento de la red y 200 para realizar pruebas. La **Tabla 1** muestra el resultado obtenido, además se ha probado el algoritmo *Naive bayes* para tener un punto de comparación con base al tiempo y exactitud de clasificación.

Tabla 1				
Algoritmo	Tiempo 1[s]	Tiempo 2[s]	Tiempo 3[s]	Exactitud[%]
Propuesto	1353.34	1390.63	1367.92	78
Naive bayes	4.97	5.10	4.97	68

Los resultados arrojados muestran que el método propuesto obtiene una mayor exactitud de clasificación con un 78 %, 10 puntos porcentuales superior a *Naive bayes* con 68 %, sin embargo este último ocupa solo 5 segundos para completar el proceso superando por mucho la propuesta del artículo el cual ocupa

1350 segundos.**?**

Category classification and topic discovery of japanese and english news articles

Este trabajo presenta algoritmos para la clasificación de noticias en categorías (Como política, deportes, tecnología) y temas (Sección de deportes: tenis, fútbol, golf), además se especializa en descubrir y clasificar temas emergentes en Internet (B. Bracewell et al., 2009). Se ocupa un método pare extraer palabras claves en cualquier idioma propuesto por Bracewell **?** que tenga herramientas de análisis morfológico. Se definieron 8 secciones posibles a las que puede ser clasificado el artículo proporcionado, las cuales son:

- *Business*
- *Politics*
- *Crime and Misfortune*
- *Health*
- *Sports*
- *Entertainment*
- *Technology* y
- *Science and Nature*

Para el desarrollo del método se ocuparon 1,000 artículos descargados de sitios como Yahoo, se recolectaron noticias redactadas en lenguaje inglés y japones. 800 se ocuparon en el entrenamiento y 200 para realizar pruebas. Los algoritmos probados fueron : **Naive bayes**, **Árboles de decisión**, **Máxima entropia** y el propuesto por este trabajo. La tabla 2 muestra los resultados obtenidos contemplando la precisión, exhaustividad y la media-F :

Algoritmo	Exhaustividad	Precisión	Media-F
Naive bayes	54.3 %	69.3 %	55.2 %
Árboles de decisión	60.2 %	60.3 %	57.2 %
Máxima entropia	15.3 %	14.8 %	12.1 %
Propuesta	63.4 %	68.6 %	65.9 %

Los algoritmos presentados en este documento se basaron en una extracción de palabras clave que es capaz de manejar múltiples idiomas y no requiere una colección de documentos o estadísticas del corpus.

Automatic news articles classification in indonesian language by using naive bayes classifier method

El artículo clasifica noticias ocupando el algoritmo clásico *Naive Bayes* (Darliani Asyárie and Wahyu Pribadi, 2009). El método propuesto consiste en 3 tareas importantes: Pre-procesamiento el cual consiste en la siguiente serie de pasos:

- *Case folding*: Proceso para convertir todas letras en minúsculas
- *Parsing*: Es el proceso de convertir oraciones en palabras
- *Stopwords elimination*: Es el proceso de eliminar palabras que se repiten con mucha frecuencia y no es información útil (Una definición mas amplia se da en el capítulo 3)
- *Stemming*: Es un proceso de corte o eliminación de afijos en una palabra. Las variantes de los afijos son prefijos, sufijos, in-fijos y con-fijos (la combinación de prefijos y sufijos)

La segunda tarea es la etapa de entrenamiento del algoritmo y por último la clasificación de artículos. Cabe destacar que el método **Frecuencia de término** (Frecuencia de aparición de una palabra en un documento dado) es utilizado en la etapa de aprendizaje. Las secciones definidas en el trabajo son: *Economy*, *Sport*, *Tecnology*, *Healt* y *Metropolitan*. Para el proceso de aprendizaje se ocuparon 50 noticias por tópico, las cuales fueron recolectadas de los sitios web *Kompas*[1], *Republika*[2] y *Suara pembaruan*[3].Las pruebas fueron realizadas con 12 noticias por sección. Además para tener una métrica en la eficiencia del método se calculó la precisión, exhaustividad y la media-F. Los resultados se muestran en la siguiente tabla (En la columna **Documentos** se coloca el número de noticias usadas para la prueba):

Documentos	Exhaustividad	Precisión	Media-F
12	90 %	90 %	90 %
24	93.4 %	91.3 %	92.33 %
36	93.75 %	90.9 %	92.30 %
48	93.1 %	93.1 %	93.1 %
60	94.11 %	90.5 %	92.26 %

Los resultados muestran que el método de **Naive bayes** es un clasificador con una exactitud alta en todos las categorías dadas.

News article text classification in indonesian language

Este documento busca el mejor algoritmo de clasificación en lenguaje Indu, comparando la eficiencia de algoritmos de selección de características (Palabras clave) y de clasificación de noticias (Wongso et al., 2017). Las secciones definidas por el artículo son las siguientes, *Economy*, *Health*, *Sports*, *Politic* y *Tecnology*;

[1]Sitio web Indu de noticias: https://www.kompas.com
[2]Sitio ya no disponible: http://www.republika.com
[3]Sitio web Indu: https://sp.beritasatu.com

El trabajo realiza pre-procesamiento de datos con los métodos *lemmatization* y *Stopwords* para reducir el ruido en la información. Para la obtención de noticias se hace uzo de la técnica *crawling*[4] en el sito *ccnnindonesia*[5]. Se obtuvieron 1,000 artículos para cada sección. 800 se usaron para la etapa de entrenamiento y 200 para realizar pruebas. Se muestra la lista de los algoritmos implementados:

- Selección de características
  - *Singular Value Decomposition*(SVD)
  - *Term frequency-inverse document frequency*(TF-IDF)
- Clasificación
  - *Support vector machine*(SVM)
  - *Naive bayes classifier*(NBC)
  - *Gaussean naive bayes*(GNB)
  - *Multinominal naive bayes*(MNB)
  - *Multivariate naive bayes*(MNB)
  - *Bernulli naive bayes*(BNB)

Los resultados obtenidos se muestran en la siguiente tabla:

Combinación usada	Precisión	Exhaustividad	Tiempo (Segundos)
TFIDF + GNB	-	-	-
TFIDF + BNB	0.9822558	0.9820000	0.7015419
TFIDF + MNB	**0.9841519**	**0.9840000**	**0.7020838**
TFIDF + SVM	0.9794023	0.9790000	74.9765017
TFIDF + SVD + GNB	0.3591179	0.3460000	11.4571054
TFIDF + SVD + BNB	0.3925421	0.3050000	11.8058102
TFIDF + SVD + MNB	-	-	-
TFIDF + SVD + SVM	0.4360882	0.3910000	1.3520746

La tabla muestra que el mejor resultado es en combinación de TF-IDF y *Multinominal naive bayes*(MNB) con la precisión y exhaustividad mas alta el cual está alrededor de 98.4 % con un tiempo de 0.702 segundos, seguido de TF-IDF y *Bernulli naive bayes*(BNB) con 98.2 % en precisión y exhaustividad con un tiempo de .701 segundos.

[4]Extracción de información en la web
[5]Sitio web de noticias: wwww.ccnnindonesia.com

2.4. Herramientas disponibles

Entre las herramientas de trabajo que son de utilidad para el procesamiento de lenguaje natural y aprendizaje automático se encuentran:

Cloud Natural Language

Google Cloud Natural Language (Google, 2019) revela la estructura y el significado del texto con modelos potentes de aprendizaje automático previamente entrenados en una API de REST fácil de usar y con modelos personalizados se puede utilizar para extraer información sobre personas, lugares, eventos y muchos otros datos, que se mencionan en documentos de texto, artículos periodísticos o entradas de blog. También se puede utilizar para comprender las opiniones sobre los productos expresadas en los medios sociales o analizar la intención en las conversaciones de los clientes que se den en un centro de atención telefónica o una aplicación de mensajería.

Googlebot

Es el crawler diseñado por Google para indexar el contenido nuevo o actualizado de Internet. Googlebot (Google, 2018) no sólo tiene la capacidad de rastrear e indexar los sitios web de Internet, sino que además puede extraer información de ficheros como pueden ser PDF, XLS, DOC, etc. Una vez el contenido está indexado, el servidor lo clasifica y establece un orden de relevancia para las distintas búsquedas que pueda efectuar un usuario, es decir, lo posiciona.

Watson natural language classifier

Watson NLC (IBM, 2017) aplica técnicas de computación cognitiva para analizar un texto y proporcionar la clase que mejor encaja entre un conjunto de clases predefinidas a partir de un texto corto. Al ser un clasificador, esta compuesto de ciertos pasos, en primera instancia se necesitan de clases las cuales son etiquetas que identificarán el texto analizado y será la salida proporcionada por el clasificador; posteriormente se debe tomar en cuenta que se necesita de una colección de textos, los cuales proporcionarán apoyo para que el clasificador logre identificar las clases ingresadas posteriormente teniendo todos estos datos se logra entrenar al clasificador, el cual proporcionará una salida dependiendo a los datos que fueron utilizados.

Capítulo 3

Marco teórico

La teoría constituye la base de sustento para el desarrollo de la ciencia. En este capítulo se expondrán de manera detallada y ordenada el conjunto de conocimientos que permitirán comprender y analizar el tema propuesto. Cabe señalar que este marco permitirá interpretar los resultados y, finalmente, formular las conclusiones del trabajo terminal.

La figura 3.1 muestra los campos abarcados por la investigación. A continuación cada área sera desarrollada con los conceptos de interés para la solución propuesta.

Figura 3.1: Campo de estudio

3.1. Inteligencia Artificial

Esta referencia no es buena. Se suelen usar referencias a sitios web cuando no hay otra fuente de donde obtener la información. En este caso hay muchisimos libros de IA. Favor de obtener la información de un libro serio de IA.

La Inteligencia Artificial (AI) es un campo de investigación y desar
por objetivo resolver problemas complejos para los cuales no se
ciones algorítmicas exactas computables en la práctica, ya sea p
dimensiones, su complejidad estructural o los niveles intrínsecos
bre de los datos que manejan (Posgrado, 2019).
Hoy en día la Inteligencia Artificial juega un papel muy importa
rrollo diversos campos de investigación, así como en la industria, finanzas, edu-
cación, transporte y más.
La tecnología ha avanzado día con día y eso implica que la Inteligencia Artificial también avanza, uno de los objetivos primordiales de la Inteligencia Artificial es construir modelos computacionales capaces de resolver las actividades que realiza el ser humano de una manera más eficiente y precisa.

3.2. Machine Learning

Machine Learning es una rama de la Inteligencia Artificial la cual permite desarrollar técnicas aprendizaje automático por parte de las computadoras, los cuales tienen la capacidad de resolver problemas, predecir continuamente los cambios que se puedan suscitar, gracias a los modelos de aprendizaje utilizados en ellos.

Machine Learning utiliza una variedad de algoritmos que aprenden iterativamente de datos para mejorar, describir datos y predecir resusltados. A medida en la cual los algoritmos de entrenamiento obtienen datos es posible obtener modelos más precisos basados en esos datos. Un modelo de Machine Learning es una salida generada cuando se tiene entrenado un algoritmo de aprendizaje automático. Después de entrenar el modelo con una entrada de datos, obtendremos una salida. Por ejemplo, un algoritmo predictivo creará un modelo predictivo. Luego, cuando se le proporciona datos al m tendrá una predicción basada en los datos que entrenaro and Massaron, 2016).

Busquen una referencia de un libro o artículo

Algunas de las áreas en las cuales Machine Learning se ha visto involucrada es (SAS, 2019):

- Servicios financieros: Las empresas pueden detectar insights en la información de sus clientes y sus operaciones, permitiendo de manera automatizada la recomendacion de productos financieros al usuario indicado y en el momento preciso.

- Salud: En combinación con sensores o dispositivos en prendas de vestir (weareble devices) un sistema puede monitorear y valorar el estado de salud de una persona en tiempo real, y si detecta una irregularidad, tomar una acción en forma automática.

- Ventas y mercadotecnia: A partir de compras previas por el usuario se pueden realizar realizar recomendaciones con alto potencial de éxito. Este conocimiento del cliente también ayuda a implementar campañas de marketing con gran nivel de precisión.

- Gobierno: El gobierno puede analizar su gigantesco cúmulo de datos, y detectar ágilmente áreas o funciones cuya mejora debe ser prioritaria –de esta forma, los recursos públicos se invierten en los ámbitos correctos, aquellos que realmente mejoran la vida de la ciudadanía y evitan procesos lentos y burocráticos.

- Transporte: Las compañías analizan sus operaciones de negocio y rápidamente detectan rutas más eficaces, lo que incrementa la rentabilidad de sus procesos (tiempos de entrega más cortos, con menor consumo de combustible, menor riesgo de desperfectos y menor desgaste de las unidades).

3.2.1. Aprendizaje supervisado

Los algoritmos de aprendizaje supervisado dependen de datos previamente etiquetados, es decir necesita de un entrenamiento para que el algoritmo pueda comprender los datos y con ello determinar que etiqueta debe asignarse a los nuevos datos en función del patron y asociando los patrones a los nuevos datos sin etiquetar. Después de ello, la maquina recibe un nuevo conjunto de datos para que el algoritmo de aprendizaje supervisado analice los datos y produzca un resultado correcto de los datos etiquetados (CleverData, 2019).

Regresión líneal

La regresión es una técnica estadística utilizada para estudiar la relación entre dos variables, permite hallar el valor esperado de una variable aleatoria a cuando b toma un valor específico. La aplicación de este método implica un supuesto de linealidad cuando la demanda presenta un comportamiento creciente o decreciente, por tal razón, se hace indispensable que previo a la selección de este método exista un análisis de regresión que determine la intensidad de las relaciones entre las variables que componen el modelo.
El pronóstico de regresión lineal simple es un modelo óptimo para patrones de demanda con tendencia (creciente o decreciente), es decir, patrones que presenten una relación de linealidad entre la demanda y el tiempo (Cleverfit, 2016).

La figura **3.2** muestra un ejemplo de los resultados obtenidos utilizando regresión líneal.

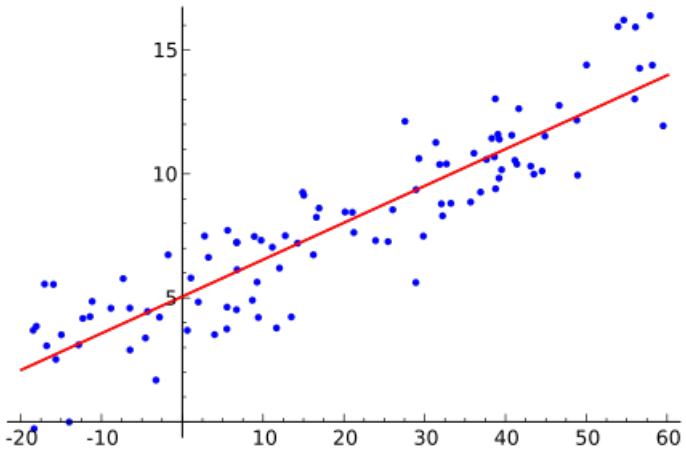


Figura 3.2: Regresión lineal con una variable dependiente y una variable independiente.

Regresión logística

La regresión logística es una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente no métrica (donde la variable es binaria o también conocida como dicotómica, es decir, solo va a dar como resultado dos alternativas posibles) y un conjunto de variables intependientes métricas o no métricas (Velasco, 2002). Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan, como una función de variables explicativas, utilizando una función logística.
La regresión logística es usada extensamente en las ciencias médicas y sociales. Otros nombres para regresión logística usados en varias áreas de aplicación incluyen modelo logístico, modelo logit, y clasificador de máxima entropía.

Naïve bayes

Naïve Bayes es un conjunto de algoritmos de aprendizaje supervisado que se basan en la aplicación del teorema de Bayes con "Naïve"(Ingenuo) la cual es la supuesta de independencia condicional entre cada par de características dado el valor de la variable de clase.
La clasificación Naive Bayes son aproximaciones probabilísticas, las cuales hacen especulaciones sobre como deben de ser generados los datos. Generalmente utilizan aprendizaje supervisado sobre el conjunto de entrenamiento para poder esimar los parámetros del modelo generativo, en tanto el conjunto de datos de entrada nuevos se realiza el teorema de Bayes, seleccionando la probable categoría que se ha generado McCallum and Nigam (1998).

Todas las características extraídas que utilizan este clasificador son independientes entre sí. La ventaja de usar este clasificador es que funciona bien tanto con datos numéricos como con datos textuales y, además, es más fácil de implementar. La desventaja de este clasificador es que su rendimiento empeora cuando las características extraídas se correlacionan entre sí.

Maquina de soporte vectorial

Las maquinas de soporte vectorial son un conjunto de algoritmos de aprendizaje los cuales se basan en el uso de un espacio de funciones lineales, el cual se encuentra con mas dimensiones inducido por un kernel, en el que las hipotesis son las entradas para el algoritmo.
El algoritmo induce separadores lineales ya sea en el espacio original de los ejemplos de entrada, si los datos no son separabales se busca un hiperplano en el que si lo sean, se hace de forna implicita con las funciones kernel.

Estos métodos están propiamente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases a 2 espacios lo más amplios posibles mediante un hiperplano de separación definido como el vector entre los 2 puntos, de las 2 clases, más cercanos al que se llama vector soporte. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase (Castro and Luis, 2014).

Random forest

Random forest es una combinación de árboles predictiroes, de modo que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de bagging que construye una larga colección de árboles no correlacionados y posteriormente los promedia (Breiman, 2001).

3.2.2. Aprendizaje no supervisado

Los algoritmos de aprendizaje no supervisado no dependen de datos previamente etiquetados, por lo cual los algoritmos tienen la tarea de agrupar la información no clasificada según sus similitudes, patrones y diferencias sin ningún entrenamiento previó de datos, por lo cual aprenden gracias a la cantidad de datos que le son ingresadas con características propias de un objeto y con ello pueda determinar los resultados basados en los datos de entrada (Zambrano, 2018).

3.3. Procesamiento de lenguaje natural

El procesamiento de lenguaje natural es una disciplina de la Inteligencia Artificial que se ocupa de la formulación e investigación de mecanismos computacionales para la comunicación entre personas y maquinas mediante el uso de Lenguajes Naturales.
El procesamiento del lenguaje natural incluye diferentes técnicas para interpretar el lenguaje humano, que van desde los métodos estadísticos y del aprendizaje basado en máquina hasta los enfoques basados en reglas y algorítmicos. Necesitamos una amplia variedad de métodos porque los datos basados en texto y en voz varían ampliamente, al igual que las aplicaciones prácticas.
Las tareas básicas de NLP incluyen la simbolización y el análisis sintáctico , lematización/derivación, etiquetado de la parte del habla, detección del lenguaje e identificación de relaciones semánticas. Si alguna vez creó diagramas de enunciados en la primaria, ya ha realizado estas tareas de forma manual antes. En términos generales, las tareas NLP dividen el lenguaje en piezas elementales más cortas, intentan entender las relaciones entre las piezas y exploran cómo funcionan las piezas juntas para crear significado (SAS, 2019).
Estas tareas implícitas se utilizan a menudo en recursos NLP de más alto nivel, como:

- Categorización de contenido. Un resumen del documento basado en la lingüística, incluyendo búsqueda e indización, alertas de contenido y detección de duplicación.
- Descubrimiento y modelado de temas. Capture con precisión el significado y temas en colecciones de texto, y aplique analítica avanzada a texto, como optimización y pronósticos.
- Extracción contextual. Extraiga automáticamente información estructurada de fuentes basadas en texto.
- Análisis de sentimiento. Identificación del estado de ánimo u opiniones subjetivas en grandes cantidades de texto, incluyendo minería de sentimiento y opiniones promedio.
- Conversión de habla a texto y de texto a habla. Transformación de comandos de voz en texto escrito y viceversa.
- Sumarización de documentos. Generación automática de sinopsis de grandes cuerpos de texto.
- Traducción basada en máquina. Traducción automática de texto o habla de un idioma a otro.

3.3.1. Tokenización

Es el proceso que descompone los textos de una colección en sus unidades mínimas, las palabras o términos propiamente dichos. A tales elementos se les

denomina tokens que conforman una lista de ítems que se utiliza para su análisis estadístico, lingüístico, de almacenamiento y posteriormente de recuperación de información. Los tokens a su vez pueden ser identificados mediante una codificación ASCII o en su defecto hexadecimal, con el objeto de facilitar la identificación uno a uno cada caracter que compone la palabra. De hecho, este proceso permite la identificación de cadenas de caracteres de forma unívoca, de cara a posteriores tratamientos de depuración, eliminación de signos de puntiación o la reducción morfológica (Ochando, 2013).
Ejemplo (3.1): Hoy es un gran día para salir.

ID	1	2	3	4	5	6	7
Token	Hoy	es	un	gran	día	para	salir

Cuadro 3.1: Ejemplo de tokenización

3.3.2. Lematización

Es el proceso lingüstico que, dada una palabra flexionada se encuentra su lema. Una palabra flexionada es cuando esta en el plural, en femenino conjugada, diminutivo o en superlativo. El lema es la palabra que esta en singular para sustantivo, singular masculino para adjetivo e infinitivo para un verbo (Gómez Díaz, 2005). Ejemplo:

- amigos, amiga, amiguitos->Amigo
- soy, son, es->Ser

Cabe mencionar que existen diversos grados de lematización

- Mórfólogica: Es la anterior mente explicada
- Sintáctica: Toma encuenta el contexto donde se encuentra la palabra

Una opción para lematizar es Freeling (Padró and Stanilovsky, 2012), este es un lematizador hecho por la universidad de catalunia.

3.4. Representación del texto

Los métodos de Machine Learning requieren que la información de la cual aprenderán esté representada en un formato que facilite su procesamiento. Generalmente esta representación es mediante vectores de valores numéricos. Cuando se requiere utilizar estos métodos con información en forma de texto, dicha información debe ser transformada para generar una representación más adecuada.

3.5. Corpus

Se le llama corpus a la recopilación de un conjunto de textos, de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos.

3.6. Crawler

Un crawler (Olston and Najork, 2010) es una herramienta la cual analiza sitios web, permitiendo recolectar las páginas web para así posteriormente extraer la información que contengan. Un crawler también conocido como como robot o spider, es un sistema para la descarga masiva de páginas web. Son uno de los componentes principales de los motores de búsqueda web, los sistemas que reúnen un conjunto de páginas web, las indexan y permiten a los usuarios realizar consultas contra el índice y encontrar las páginas web que coincidan con las consultas.

3.7. Sitios web

Un sitio web (Techopedia, 2018) es un conjunto de páginas web interconectadas y de acceso público que comparten un solo nombre de dominio. Los sitios web pueden ser creados y mantenidos por un individuo, grupo, empresa u organización para cumplir una variedad de propósitos. Todos estos sitios constituyen la World Wide Web.

3.7.1. Página web

Una página web es un documento electrónico el cual forma parte de la WWW (*World Wide Web*) generalmente construido en el lenguaje HTML (*Hyper Text Markup Language*). Este documento puede contener enlaces que nos direcciona a otra página web. Para visualizar una página web es necesario de un browser o un navegador (Emprendedor, 2019). Dentro de las páginas web podemos encontrar un sinfin de sitios los cuales pueden ser de nuestro interés.

3.7.2. Blog

Un blog es una página web en la cual el usuario no necesita conocimientos específicos del medio electrónico ni del formato digital para poder aportar contenidos de forma inmediata, ágil y constante desde cualquier punto de conexión a Internet (Bruguera, 2019). En un blog el usuario puede compartir cualquier tipo de información que sea de su agrado, teniendo una mayor libertad de expresión lo cual permite que otras personas compartan y comenten su manera de expresarse.

3.7.3. Foro

Un foro es una herramienta de comunicación asíncrona. Los foros permiten la comunicación de los participantes desde cualquier lugar en el que esté disponible una conexión a Internet sin que éstos tengan que estar dentro del sistema al mismo tiempo, de ahí su naturaleza asíncrona. Brindando una mayor interacción entre distintos participantes y permitiendo conocer la opinión sobre un tema de distintas personas.

Capítulo 4

Análisis y diseño

En este capítulo se describe el análisis y el diseño del sistema web para el trabajo terminal propuesto, mostrando la arquitectura y los modulos con los cual cuenta. Hasta este punto se presentan los requisitos que deberá cumplir el sistema así como los casos de uso, diagramas de secuencia y diagramas de flujo. Cabe destacar que se definen las paginas web utilizadas con base en los sitios web más consultados, la técnica utilizada para la recolección de noticias y la manerá en la cual se ordenan y visualizan los artículos obtenidos.

4.1. Actores y roles

Usuario: Cualquier persona que ingrese al sistema y esté interesada en consultar noticias de alguna

4.2. Secciones de noticias

- **Política:**
- **Deportes:**

- **Ciencia y tecnología:**
- **Economía:**
- **Cultura:**

4.3. Sitios web definidos para la recolección de noticias

4.3.1. Diarios

- **El Universal**: https://www.eluniversal.com.mx/
- **Azteca Noticias**: https://www.aztecanoticias.com.mx/
- **Aristegui Noticias**: https://aristeguinoticias.com/
- **Excelsior**: https://www.excelsior.com.mx/
- **La Jornada**: https://www.jornada.com.mx/ultimas
- **Milenio**: https://www.milenio.com/

4.3.2. Blogs

- **Yahoo**: https://es-us.noticias.yahoo.com/
- **Sopitas**: https://www.sopitas.com/
- **SDP Noticias**: https://www.sdpnoticias.com/
- **Uno TV**: https://www.unotv.com/inicio/

4.4. Requisitos funcionales

RF1 Recolectar noticias

- **Descripción:** El sistema debe recolectar noticias de forma automática de los sitios web definidos.

RF2 Clasificar noticias

- **Descripción:** El sistema debe clasificar las noticias recolectadas de acuerdo a su contenido, en las secciones previamente definidas.

RF3 Filtrar noticias

- **Descripción:** El sistema debe filtrar las noticias recolectadas de acuerdo a la fecha de publicación, el periodo permitido para el filtrado de noticias es: de la fecha actual de ingreso al sistema hasta tres días antes.

RF4 Mostrar resultados

- **Descripción:** El sistema debe mostrar las noticias que cumplan con los filtros de búsqueda establecidos por el usuario (Sección y fecha de publicación).

4.5. Requisitos no funcionales

RNF1 Tiempo de clasificación

- **Descripción:** El tiempo de clasificación de las noticias recolectadas no debe tardar mas de cinco segundos.

RNF2 Número de palabras

- **Descripción:** Las noticias recolectadas deben tener un mínimo de 180 palabras en ellas.

RNF3 Número de noticias mostradas

- **Descripción:** El sistema debe mostrar al menos 15 noticias clasificadas, por los filtros seleccionados por el usuario.

4.6. Reglas de negocio

En esta sección se describen las reglas de negocio implementadas en el trabajo propuesto.

RN1 Número de palabras

- **Tipo:** Dominio.
- **Descripción:** La notica debe tener al menos 180 palabras
- **Referenciado por**: CU1 Recolectar noticias

RN2 Lenguaje de noticias

- **Tipo:** Dominio.
- **Descripción:** Las noticias deben estar redactadas en lenguaje español.
- **Referenciado por: CU2 Clasificar noticias**

RN3 Diccionario de sitios

- **Tipo:** Dominio.
- **Descripción:** Solo se puede recolectar información de los siguientes sitios.
  - **El Universal**: https://www.eluniversal.com.mx/
  - **Azteca Noticias**: https://www.aztecanoticias.com.mx/
  - **Aristegui Noticias**: https://aristeguinoticias.com/
  - **Excelsior**: https://www.excelsior.com.mx/
  - **La Jornada**: https://www.jornada.com.mx/ultimas
  - **Milenio**: https://www.milenio.com/
  - **Yahoo**: https://es-us.noticias.yahoo.com/
  - **Sopitas**: https://www.sopitas.com/
  - **SDP Noticias**: https://www.sdpnoticias.com/
  - **Uno TV**: https://www.unotv.com/inicio/
- **Referenciado por**: CU1 Recolectar noticias

RN4 Umbral de grado de pertenencia

- **Tipo:** Flujo.
- **Descripción:** Solo se puede mostrar una noticia si su grado de pertenencia a una sección es mayor o igual al umbral establecido.
- **Referenciado por: CU2 Clasificar noticias**

RN5 Orden de publicación

- **Tipo:** Estructura.
- **Descripción:** Las noticias se muestrán de forma descendente de acuerdo al grado de pertenencia a la sección.
- **Referenciado por**: CU1 Recolectar noticias,CU4 Mostrar resultados

RN6 Número de noticias recolectadas

- **Tipo:** Dominio.
- **Descripción:** De cada sitio establecido se recolectan todas las noticias que se encuentren en el periodo establecido (3 días), de cada noticia se extrae **Título**, **URL al artículo**, **Fecha de publicación** y de contar con ello el **Resumen**.
- **Referenciado por**: CU1 Recolectar noticias

4.7. Casos de uso

4.7.1. Diagrama de casos de uso

La figura 4.1 muestra el diagrama de casos de uso de la aplicación. Los casos de uso marcados en color gris son descritos en el documento, sin embargo los mostrados en color rojo serán desarrollados posteriormente.

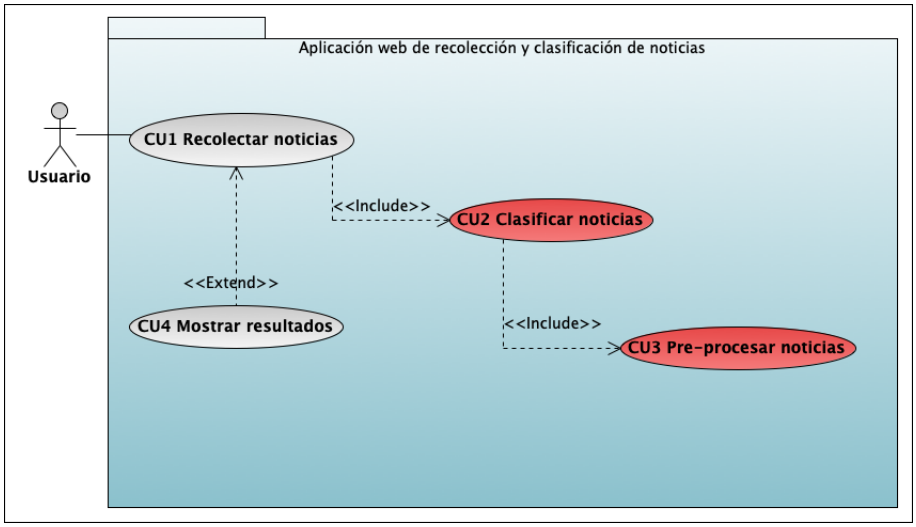


Figura 4.1: Diagrama de casos de uso

4.7.2. CU1 Recolectar noticias

Resumen

Brinda al usuario un punto de acceso para elegir una sección; Las clasificaciones son, **Ciencia y técnología**, **Política**, **Deportes**, **Economía** y **Cultura**, posteriormente se recolectan noticias de la web, tomando como punto de partida los sitios establecidos previamente. Se crea un proceso de recolección independiente por cada sitio web, para simular un ambiente de extracción en paralelo; De cada sitio se recoletan las noticas publicadas; De cada artículo se obtiene **Fecha de publicación**, **Título**, **Contenido**, **URL de la noticia**, y de contar con ello el **Resumen**. Cabe destacar que las ligas contenidas en los sitios visitados son extraidas para su posterior análisis.

Descripción

Caso de uso:	CU3 Recolectar noticias
Actor:	Usuario
Propósito:	Brindar una herramienta de recolección de noticias de Internet(Crawler).
Entradas:	URL de las paginas por consultar
Salidas:	• MSG7 Petición vacía • MSG9 Fallo en la recolección
Precondición:	EL **Diccionario de URL'S** debe contener los vínculos de los sitios a consultar
Postcondiciones:	• El usuario tendrá la facultad de visualizar las noticias clasificadas • El usuario podrá cambiar el periodo de busqueda
Reglas de negocio:	• RN1 Número de palabras • RN3 Diccionario de sitios • RN5 Orden de publicación • RN6 Número de noticias recolectadas
Errores:	• **Uno:** Cuando no se ha recuperado ninguna **dirección web** se muestra el mensaje MSG1 Catálago vacio, fin del caso de uso • **Dos:** Cuando no se ha encontrado noticias en el día seleccionado se muestra el mensaje MSG2 Petición vacía, fin del caso de uso

Caso de uso:	CU3 Recolectar noticias
Errores:	• **Tres:** Cuando no se puede extraer información de los sitios brindados, se muestra el mensaje MSG3 Fallo en la recolección, fin del caso de uso
Autor:	Carlos Andres Hernandez, Luis Daniel Meza

Trayectoria principal

1. Selecciona una opción de la pantalla UI1 Inicio; **Política**, **Economía**, **Deportes**, **Ciencia y tecnología** o **Cultura**.
2. Obtiene las **Direcciones web**.
3. Verifica que al menos se recupere una **Dirección web**. [Error Uno]
4. Muestra la pantalla Pantalla UI2 Espera de proceso. [Trayectoria A]
5. Verifica que no se haya recolectado noticias previamente. [Trayectoria B]
6. Por cada URL recuperada se extraen las noticias con base en la regla de negocio RN6 Número de noticias recolectadas . [Error Tres]
7. Incluye el caso de uso CU2 Clasificar noticias.
8. Obtiene la fecha actual.
9. Obtiene de cada noticia clasificada en el paso **??** de la trayectoria principal el **Título**, **URL al artículo**, **Fecha de difusión** y de contar con ello el **Resumen**.
10. Ordena las noticias clasificadas deacuerdo a la regla de negocio RN5 Orden de publicación .
11. Muestra la pantalla Pantalla UI3 Proceso concluido.
12. - - - - *Fin del caso de uso.*

Esto se debe redactar como una trayectoria alternativa. Corregir estos casos en TODO el documento

Esto ya se hizo en el paso 6

Trayectoria alternativa A:

Condición: *El usuario ha presionado el botón cancelar*

A-1. Presiona el botón **Cancelar** de la pantalla Pantalla UI2 Espera de proceso.

A-2. Muestra la pantalla UI1 Inicio.

A-3. - - - - *Fin de la trayectoria.*

Trayectoria alternativa B:

Condición: *Ya se han recolectado noticias*

B-1.  Continua en el paso 11 de la trayectoria principal.

B-2. - - - - *Fin de la trayectoria.*

Puntos de extensión

Causa de la extensión: El usario desea consultar las noticias clasificadas.
Región de la trayectorioa: Proviene del paso 11 de la trayectoria principal.
Extiende a : CU4 Mostrar resultados

4.7.3. CU4 Mostrar resultados

Resumen

Permite al actor vizualizar las noticias correspondiente a la sección elegida, ya sea **Política**, **Deportes**, **Ciencia y técnología**, **Economía** o **Cultura**. La consulta se realiza en un periodo establecido; El sitio muestra 15 noticias ordenadas de forma descendente deacuerdo a la fecha de publicación, cada artículo contiene el **Título** , la **Fecha de publicación**, **URL** el cual direcciona a la página fuente que ha proporcionado la noticias y de contar con ello un **Resumen de la información**.

Descripción

Caso de uso:	CU2 Buscar noticias
Actor:	Usuario.
Propósito:	Brindar una herramienta que permita consultar las noticias clasificadas
Entradas:	Ninguna.
Salidas:	Noticias clasificadas; De cada una se muestra: • **Título** • **URL al artículo** • **Fecha de publicación** • **Resumen**
Precondición:	La clasificación de las noticias debe estar completa
Postcondiciones:	Ninguna.
Reglas de negocio:	RN5 Orden de publicación
Errores:	Ninguno.
Autor:	Carlos Andres Hernandez Gomez

Trayectoria principal

1. Presiona el botón **Aceptar** de la pantalla Pantalla UI3 Proceso concluido. [Trayectoria A]
2. Muestra 15 noticias de las ordenadas previamente, deacuerdo a la regla de negocio RN5 Orden de publicación las cuales cumplan con el filtro de sección y fecha, como se visualiza en la pantalla UI2 Sección política
3. Consulta la información.[Trayectoria B]
4. - - - - *Fin del caso de uso.*

Trayectoria alternativa A:
Condición: *El usuario ha presionado el botón cancelar*

A-1. Presiona el botón **Cancelar** de la pantalla Pantalla UI3 Proceso concluido.

A-2. Muestra la pantalla UI1 Inicio.

A-3. - - - - *Fin de la trayectoria.*

Trayectoria alternativa B:
Condición: *El usuario ha cambiado el periodo establecido*

B-1. Presiona un botón del menú **Cambio de periodo** de la pantalla UI2 Sección política.

B-2. Continua en el paso 2 de la trayectoria principal con el periodo seleccionado.

B-3. - - - - *Fin de la trayectoria.*

4.8. Mensajes

MSG1 Catálago vacio

- **Tipo:** Error.
- **Objetivo:** Dar a conocer que no se tiene las lígas a los sitios web.
- **Redacción:** El catálogo **Direcciones web** se encuentra vacio.
- **Referenciado por:** CU1 Recolectar noticias

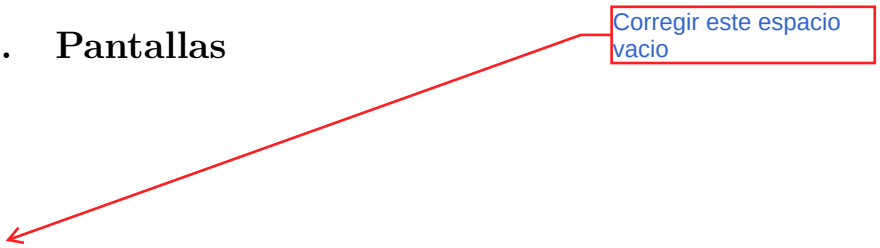
MSG2 Petición vacía

- **Tipo:** Error.
- **Objetivo:** Informar al usuario que no se ha encontrado resultados en el día seleccionado.
- **Redacción:** El tiempo de búsqueda máximo se ha alcansado y no se ha encontrado infomración en el periodo ingresado.
- **Referenciado por:** CU1 Recolectar noticias

MSG3 Fallo en la recolección

- **Tipo:** Error.
- **Objetivo:** Informar al usuario que las ligas registradas en el diccionario de **URL's** no permiten extraer información.
- **Redacción:** No se puede extraer noticias de los sitios registrados en este portal web.
- **Referenciado por:** CU1 Recolectar noticias

4.9. Pantallas



4.9.1. UI1 Sitio

Objetivo

Permite al usuario seleccionar la sección a consultar ya sea **Política**, **Deportes**, **Ciencia y técnología**, **Economía** o **Cultura** y de ser necesario, permite cambiar el periodo de consulta en un rango de 3 días.

Descripción

La pantalla muestra un menú con las secciones definidas para el sistema, en ella se puede navegar para acceder a las consultas de las noticias clasificadas y de ser necesario, ingresar al sitio web del artículo.

Referenciado por:

CU1 Recolectar noticias,CU4 Mostrar resultados

Figura 4.2: Pantalla UI1 Inico

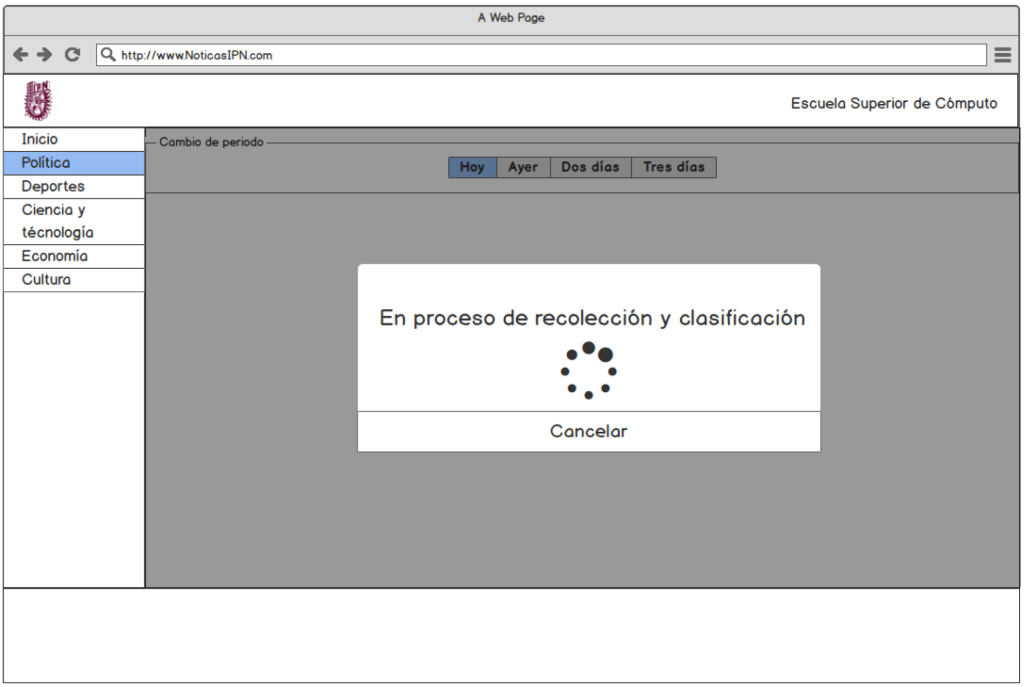


Figura 4.3: Pantalla UI2 Espera de proceso

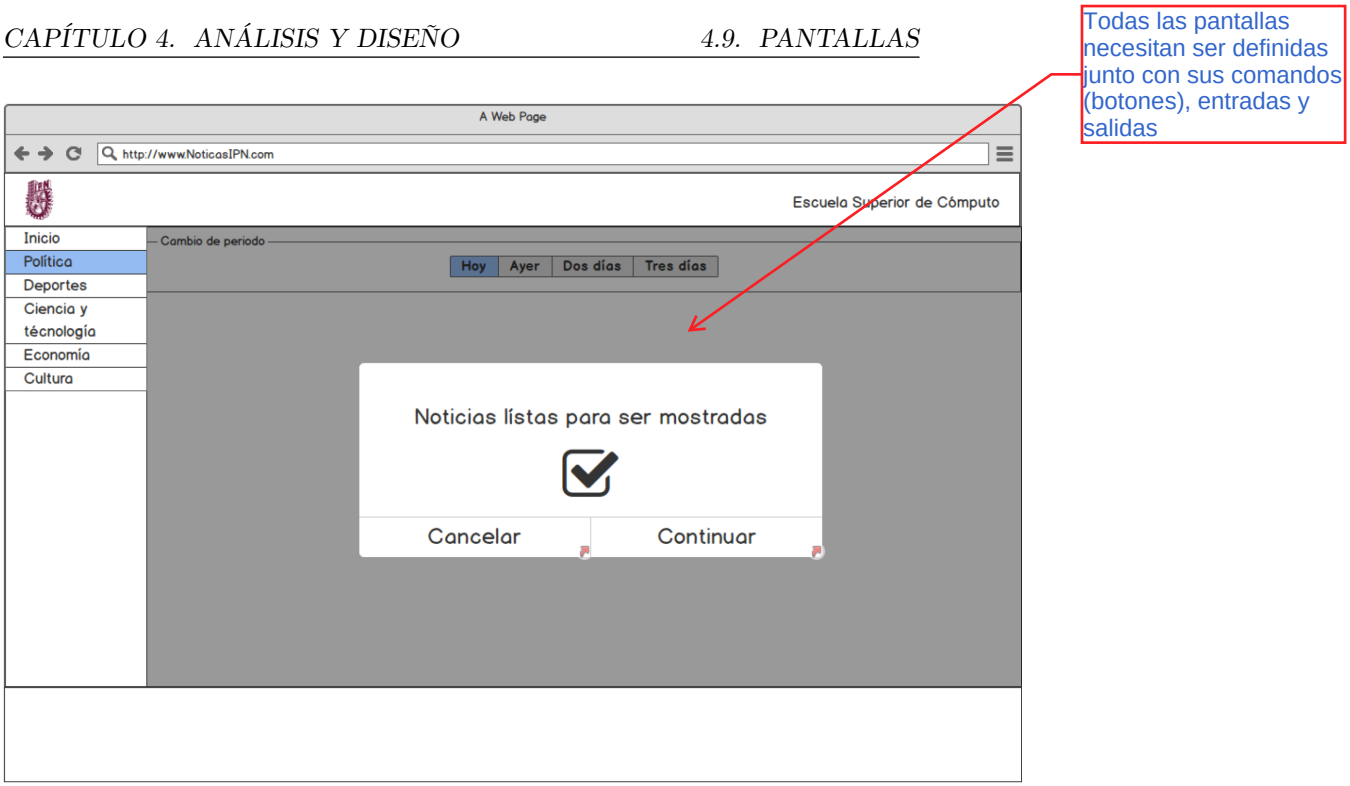


Figura 4.4: Pantalla UI3 Proceso concluido

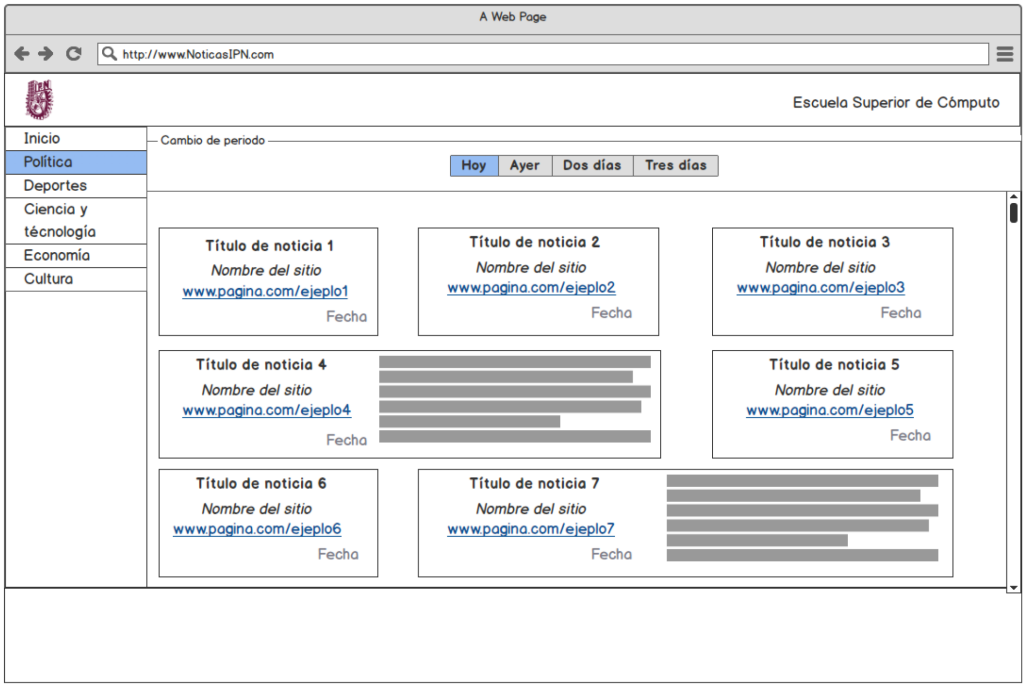


Figura 4.5: Pantalla UI4 Sección polítca

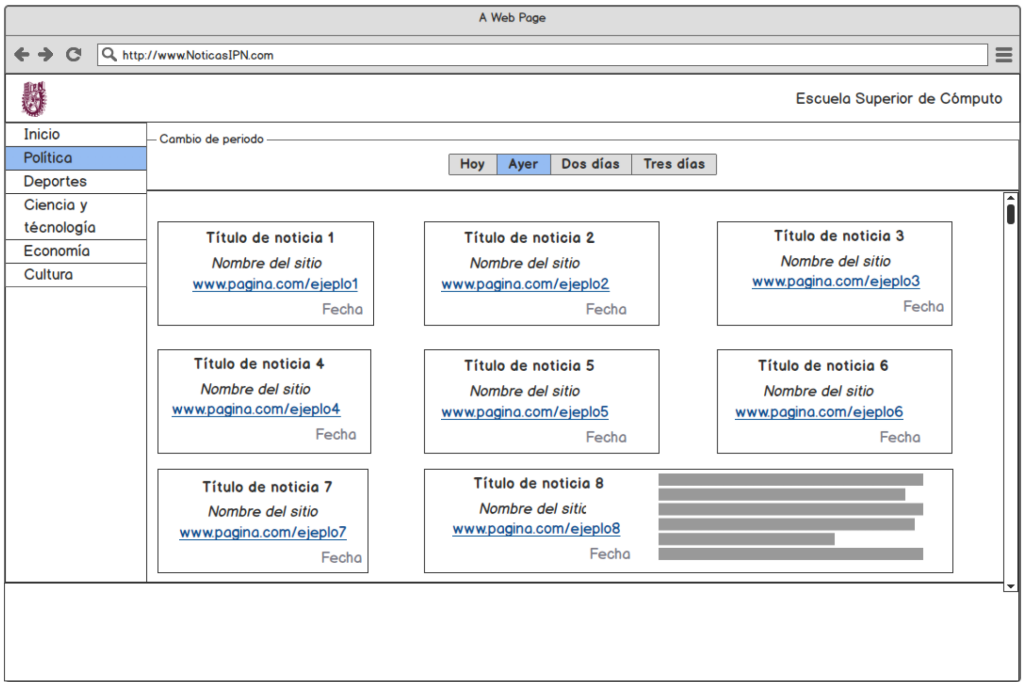
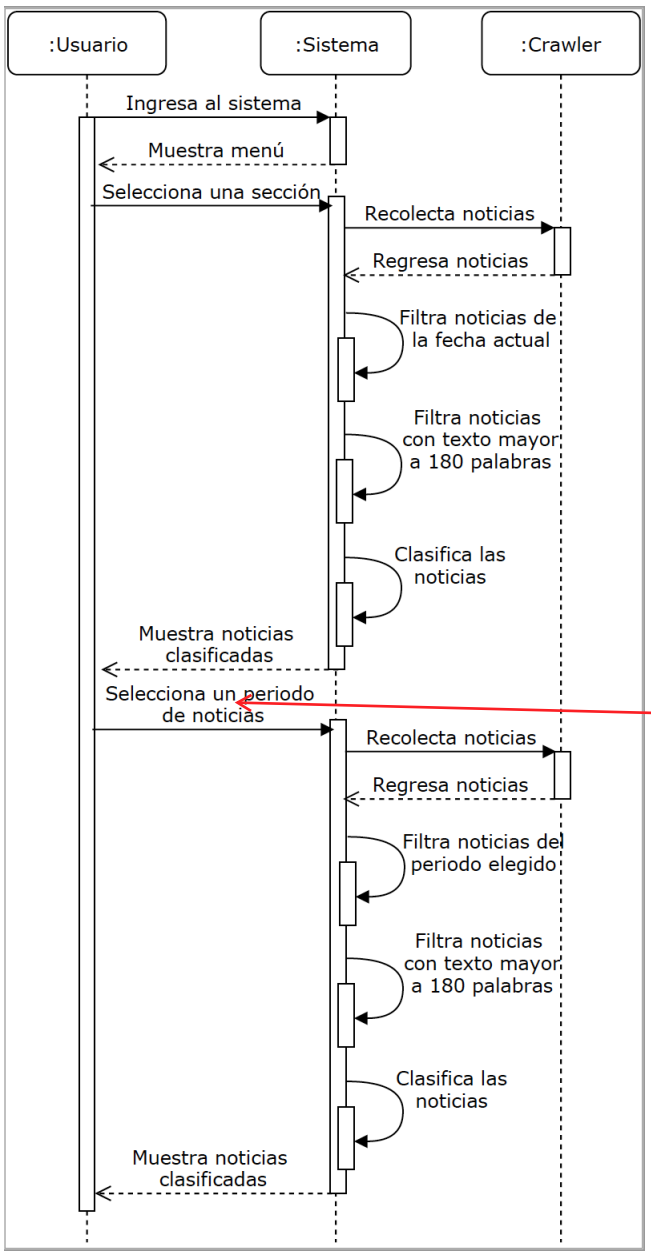


Figura 4.6: Pantalla UI5 Cambio de perido

Figura 4.7: Pantalla UI6 Página de sitio web

4.10. Diagrama de secuencia

La figura 4.8 muestra el diagrama de secuencia de la aplicación.



Esto es incorrecto. El crawler sólo recolecta las noticias al inicio del proceso y posteriormente el filtrado por fecha se hace sin la intervención del crawler. Tampoco me queda claro porque modelar al crawler como una entidad separada del sistema y al clasificador no lo modelan así.

Figura 4.8: Diagrama de casos de uso

Capítulo 5

Avances

En esta sección se mostrarán los avances que hasta el día de hoy se tienen.

5.1. Herramientas utilizadas

Para el desarrollo de los avances para el presente trabajo se utilizó el lenguaje de programación Python, se utilizó Anaconda debido a que es una distribucion de Python que se utiliza para el análisis de datos, es decir con Anaconda tenemos un ambiente de trabajo para el análisis de datos. Mientras que para el web-crawling se utilizó Scrapy, el cual es un framework que nos permite el web scraping para la extracción de información de los sitios web.

5.2. Estudio previo

Una vez que se eligieron las herramientas que nos iban a proporcionar la información de las páginas web, se procedió con el e[illegible]os cuales se extraerían noticias para el entrenamiento.
En primera instancia se eligiero 9 sitios web (5.

- 3 Sitios de foros de noticias: Aristegui No[illegible]Sopitas.

No explican por qué se eligieron estos sitios. Esa explicación es muy importante y se pueden basar en las estadísticas de los sitios más consultados que hace tiempo les había proporcionado

- 3 Sitios de diarios: El Universal, La Jornada y Excelsior.

Faltan las lineas que encuadran a la tabla

visivos: TV Azteca, Televisa y Once Noticias.

Sección	El Universal	La Jornada	Excelsior	Aristegui Noticias	SDP Noticias	Sopitas	Azteca Noticias	Televisa	Once Noticias
Nacional	Nacional	-	Nacional	México	Nacional	Noticias	-	Nacional	Nacional
Internacional	Mundo	Mundo	Global	Destacado/Mundo	Internacional	-	Internacional	Internacional	Internacional
Ciudad	Metrópoli	Capital	Comunidad	-	CDMX	-	-	CDMX	CDMX
Estados	Estados	Estados	Estados	Sociedad/México	Locales	-	Estados	Estados	Nacional
Economía	Cartera	Economía	Dinero	Economía	Economía	-	Finanzas	Economía	Economía
Deportes	Deportes	Deportes	Adrenalina	Deportes	Deportes	Deportes	-	Deportes	Deportes
Espectáculos	Espectáculos	Espectáculos	Función	-	En el show	Entretenimiento	-	Espectáculos	Deportes
Cultura	Cultura	Cultura	-	-	-	-	-	Arte y cultura	Cultura
Política	-	Política	Política	Poderes	-	-	Política	Política	-

Una vez que se analizarón las secciones con las que contaba cada sitio se procedió a homologar las secciones en las cuales la mayoría de los sitios coincidian, por lo cual se quedarón definidas 5 secciones para clasificación de las noticias extraídas.

- Política
- Deportes
- Ciencia y tecnología
- Economía
- Cultura

Estos dos párrafos están muy mal redactados, no se entiende lo que dicen y tiene errores en el uso del lenguaje. Corregir esto

Posteriormente se procedio con la recolección de noticias. La biblioteca utilizada es Scrapy la cual nos servirá para el Scrapeo.
Una de las cosas que se debía investigar sobre cada sitio web fue su estructura XML, al saber la estructura del contenido de una página nos permite realizar el Scrapeo de manera correcta y eso depende de cada sitio, ya que no todos los sitios cuentan con una página la cual tenga las mismas secciones o el mismo orden.

5.3. Recolección de noticias

5.3.1. Prerrequisitos

- Tener instalado en nuestro computador alguna versión de Linux.
- Tener instalado python en a partir de su versión 2.7
- Tener en cuenta que se necesitará utilizar entornos virtuales

Abriremos una terminal nueva con la ruta donde se desee guardar el proyecto, figura 5.1

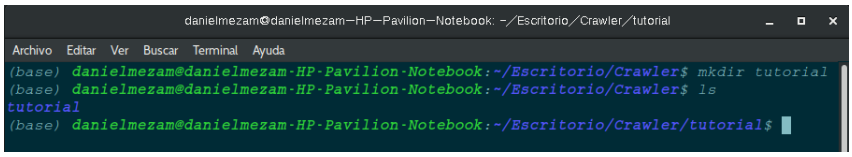


Figura 5.1: Carpeta donde se creará el proyecto.

Una vez ubicados en la carpeta donde se creará el proyecto se procede con la creación de un entorno virtual y ahí se activará el entorno virtual creado figura **5.2** .

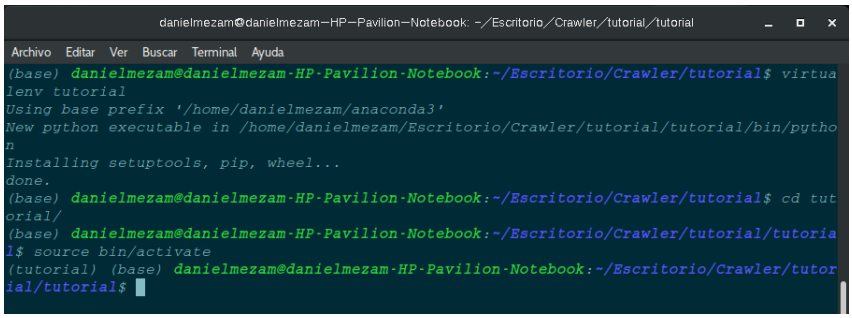


Figura 5.2: Creación del entorno virtual.

Una vez en nuestro entorno virtual instalaremos Scrapy y procederemos con la creación del proyecto con la siguiente línea ***scrapy startproject tutorial*** figura **5.3** .

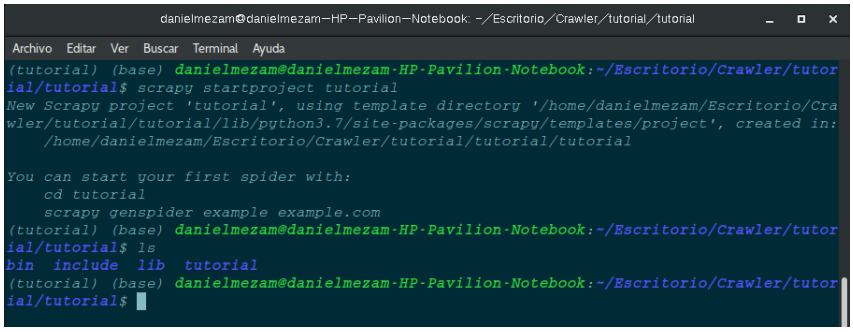


Figura 5.3: Creación del proyecto Scrapy.

Veremos que se ha creado una carpeta con el nombre del proyecto **"tutorial"** la cual contiene los siguientes archivos figura **5.4** .

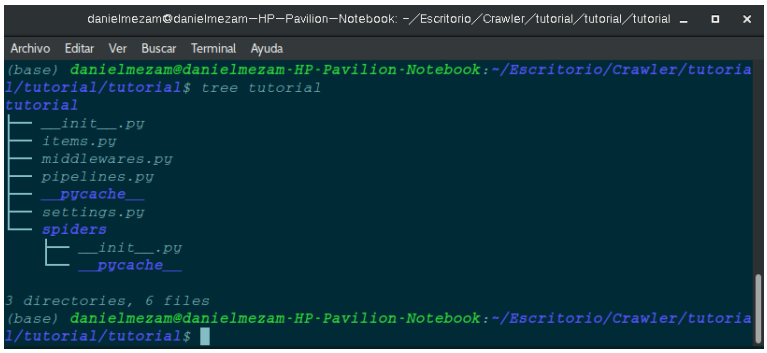


Figura 5.4: Carpetas y archivos creados.

Crearemos un archivo dentro de la carpeta Spiders llamado ***spiders.py***
Definimos las partes de la página que deseamos recolectar de cada noticias y lo guardamos en el archivo **items.py** 5.5. Para nuestro recolector es claro que necesitamos los siguientes aspectos de cada noticia, lo cual nos permitirá obtener mejores resultados

- URL: Se necesita la URL para redireccionar al usuario a la página de la noticia seleccionada.
- Sección: Necesitamos saber la sección para nuestro entrenador.
- Título: El título de la noticia permitirá al usuario saber sobre la noticia.
- Autor: Permitirá mostrarle al usuario el autor de la noticia.
- Fecha: Se necesitará la fecha para su clasificación por la fecha de publicación de la noticia.
- Descripción: Se mostrara al usuario una pequeña descripción de la noticia si es que la tiene.
- Noticia: La noticia nos ayudará a la clasificación de la misma.

```python
# -*- coding: utf-8 -*-

# Define here the models for your scraped items
#
# See documentation in:
# https://doc.scrapy.org/en/latest/topics/items.html
import scrapy

class tutorialItem(scrapy.Item):
    url = scrapy.Field()
    seccion = scrapy.Field()
    titulo = scrapy.Field()
    autor = scrapy.Field()
    fecha = scrapy.Field()
    descripcion = scrapy.Field()
    noticia = scrapy.Field()
    pass
```

Figura 5.5: Secciones de la página que deseamos recolectar.

En el archivo spiders.py se definen las reglas que debe seguir el recolector **5.6**

```python
#Importamos las librerias correspondientes
import scrapy
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor
#Importamos los datos que necesitamos saber de la página web
from scrapy.exceptions import CloseSpider
from tutorial.items import tutorialItem

class MilenioSpider(CrawlSpider):
    name = 'tutorial'
    item_count = 0
    #Se permitirá obtener información unicamente en este dominio
    allowed_domain = ['https://www.milenio.com/']
    #Aquí será el inicio del Scrapeo
    start_urls = ['https://www.milenio.com/']

    rules = {
        # Será la xpath para poder acceder a cada noticia
        Rule(LinkExtractor(allow=(), restrict_xpaths=('//div[@class="title"]/a')),
                           callback= 'parse_item', follow= False)
    }

    def parse_item(self, response):
        mi_item = tutorialItem()

        mi_item['url'] = response.url
        mi_item['seccion'] = response.xpath('normalize-space(//*[@id="main-header"]/div[3]/div/ul/li[1]/a[2]/text())').extract()
        mi_item['titulo'] = response.xpath('//h1[@class="title"]/text()').extract()
        mi_item['autor'] = response.xpath('//span[@class="author"]/text()').extract()
        mi_item['fecha'] = response.xpath('//time/text()').extract()
        mi_item['descripcion'] = response.xpath('//span[@class="summary"]/text()').extract()
        mi_item['noticia'] = response.xpath('//div[@id="content-body"]/p//text()').extract()
        self.item_count += 1
        if self.item_count >50:
            ra mi_item Spider('item_exceeded')
        yield mi_item
```

Figura 5.6: En la primera parte vemos las librerias, posteriormente se definen las reglas las cuales son diferentes para cada sitio web.

Se debe considerar que se debe analizar la página para poder obtener los datos de la misma, y se debe ser muy específico, de lo contrario no se podrá recolectar de manera correcta la información.

Una vez definidas las reglas para la extracción se procede con la exportación del documento, en el cual se guardará la información recolectada.

Posteriormente se procede a modificar el archivo **pipelines.py** el cual nos permitirá hacer la importación de la información obtenida a un archivo en formato CSV **5.7**

```python
import scrapy
from scrapy import signals
from scrapy.exporters import CsvItemExporter
from scrapy.pipelines.images import ImagesPipeline
from scrapy.exceptions import DropItem
from scrapy import Request
import csv

class tutorialPipeline(object):
    def __init__(self):
        self.files = {}

    @classmethod
    def from_crawler(cls, crawler):
        pipeline = cls()
        crawler.signals.connect(pipeline.spider_opened, signals.spider_opened)
        crawler.signals.connect(pipeline.spider_closed, signals.spider_closed)
        return pipeline

    def spider_opened(self, spider):
        file = open('%s_items.csv' % spider.name, 'w+b')
        self.files[spider] = file
        self.exporter = CsvItemExporter(file)
        self.exporter.fields_to_export = ['url','seccion','titulo', 'autor', 'fecha', 'descripcion', 'noticia']
        self.exporter.start_exporting()

    def spider_closed(self, spider):
        self.exporter.finish_exporting()
        file = self.files.pop(spider)
        file.close()

    def process_item(self, item, spider):
        self.exporter.export_item(item)
        return item
```

Figura 5.7: Código anexado para la exportación de la informacióne extraída.

Por último se debe ejecutar el siguiente comando para extraer la información **scrapy crawl tutorial -t csv** **5.8**

```
{'autor': ['Yuan Yang y Nian Liu'],
 'descripcion': ['"Optimiza" fuerza laboral. La firma echó a 900 empleados, '
                 'que representa 60 por ciento de su esfuerzo de investigación '
                 'y desarrollo en el país asiático.'],
 'fecha': ['13.05.2019 03:57:10'],
 'noticia': ['Oracle',
             ', ',
             'la compañía estadunidense de ',
             'software',
             ', ',
             'despidió a 900 empleados de su equipo de ',
             'China',
             ', que representa 60 por ciento de su esfuerzo de investigación y '
             'desarrollo allí, de acuerdo con empleados, algunos de ellos '
             'culparon de los recortes a las tensiones entre\xa0',
             'EU',
             ' y ',
             'China',
             '.',
             'El 7 de mayo se le informó al personal de ',
             'Oracle ',
             'en ',
             'China ',
             'que la compañía estaba "optimizando" su fuerza laboral y que se '
```

Figura 5.8: Muestra de la recolección por parte del crawler.

5.3.2. Resultados

Una vez que termino de ejecutarse el crawler se generó un archivo CSV, ahí se almaceno la información recolectada, posteriormente toda la información que se arecolectada será utilizada para el algoritmo de entrenamiento textbf5.9.

url	seccion	titulo	autor	fecha	descripcion
https://www.milenio.com/estados/ve-mapa-incendios-mexico-satelite-n	Estados	Así se ven los incendios en México desde satélites de la NAS	Milenio Digital	13.05.2019 17:37:16	La NASA co
https://www.milenio.com/deportes/mas-aficion/cuerpo-pedia-adrenalina	La Afición	Mi cuerpo siempre pedía la adrenalina del cuadrilátero: Silver K	Luis Salcedo Cassio	12.05.2019 03:49:29	Hijo del Dr.
https://www.milenio.com/politica/yeidckol-polevnsky-inviten-desayunar	Política	"Eso de que me inviten a desayunar... huevos me da mucha fl	Humberto Ríos Navarrete	06.05.2019 05:50:47	La dirigente
https://www.milenio.com/politica/cante-baile-hice-deporte-quise-abogad	Política	Canté, bailé e hice deporte, pero siempre quise ser abogada: C	Carolina Rivera	08.05.2019 03:36:01	La consejera
https://www.milenio.com/cultura/concibo-poesia-mala-prensa-javier-sic	Cultura	No me concibo sin poesía... aunque tenga mala prensa: Javier	José Antonio Belmont	09.05.2019 03:51:00	Nombrado P
https://www.milenio.com/politica/marko-cortes-marko-cortes-noviero-ar	Política	"Fui noviero, amiguero y pecador estándar... lo normal"	Daniel Venegas	13.05.2019 04:06:42	Al dirigente
https://www.milenio.com/politica/salvador-guerrero-chico-escapaba-cas	Política	"De chico me escapaba de mi casa por la ventana para andar	Pedro Domínguez	07.05.2019 05:13:24	El presidente
https://www.milenio.com/ciencia-y-salud/mas-ciencia-salud/mujer-ciencia-bailo-vals-rocan		Soy una mujer de ciencia, aunque igual bailo un vals que un ro	Jorge Almazán R.	10.05.2019 05:14:18	Aficionada d
https://www.milenio.com/internacional/jimmy-carter-recupera-cirugia-ca	Mundo	Jimmy Carter se recupera de cirugía de cadera tras caerse en	EFE	13.05.2019 17:59:09	El ex preside
https://www.milenio.com/politica/dichos-vicente-fernandez-refleja-homo	Política	Dichos de Vicente Fernández son reflejo de su ignorancia: Cor	Notimex	13.05.2019 20:06:24	La titular del
https://www.milenio.com/negocios/finanzas-personales/como-salir-del-	Negocios	¿Cómo salir del Buró de Crédito?	Karen Guzmán	13.05.2019 20:23:09	Si te has atra
https://www.milenio.com/opinion/diego-fernandez-de-cevallos/sin-rodeo	Opinión Nacional	Dos Bocas y la S. de I.I.		13.05.2019/01:50	
https://www.milenio.com/opinion/juan-pablo-becerra-acosta/doble-fondo	Opinión Nacional	Solo le ruego al Dios de AMLO...		13.05.2019/02:33	
https://www.milenio.com/espectaculos/game-of-thrones-cantante-actua	Hey	'Grey Worm' de 'Game of Thrones' también canta, ¡escucha su	Milenio Digital	13.05.2019 19:27:34	Si pensabas
https://www.milenio.com/deportes/extra-cancha/gol-hija-mohamed-sala	La Afición	El gol de la hija de Mohamed Salah que prendió a la afición de	La Afición	13.05.2019 18:55:27	Makkah Moh
https://www.milenio.com/content/emprender-en-mexico	InBrand				
https://www.milenio.com/videos/espectaculos/surtido-rico-invasion-am	Hey				
http://coleccionmilenioarte.milenio.com/					
https://www.milenio.com/espectaculos/game-of-thrones-8x06-pasar-ca	Hey	Daenerys, al Trono de Hierro en el avance del capítulo final de	Milenio Digital	12.05.2019 21:48:34	HBO lanzó e
https://www.mediotiempo.com/					
https://www.milenio.com/espectaculos/musica/daddy-yankee-sera-prod	Hey	Daddy Yankee será productor de televisión	Agencia AP	13.05.2019 17:44:56	El reguetone
https://www.milenio.com/espectaculos/matan-tiros-rapero-aab-hellanba	Hey	Matan a tiros a rapero AAB Hellanbandz en Miami Beach	AFP	13.05.2019 12:07:33	El músico, d
https://www.milenio.com/espectaculos/cine/x-men-day-celebran-legado	Hey	Celebran el legado de los héroes mutantes con el X-Men Day	Notimex	13.05.2019 11:53:13.M	Fans de X-M
https://www.milenio.com/espectaculos/cine/avengers-endgme-robert-d	Hey	Así fue el emotivo reencuentro entre Iron Man y Spider-Man	Milenio Digital	13.05.2019 11:17:55.M	El actor que
https://www.milenio.com/content/infraestructura	InBrand	Infraestructura			
https://www.milenio.com/espectaculos/musica/the-kooks-anuncia-cono	Hey	The Kooks regresa a México para dar dos conciertos	Milenio Digital	13.05.2019 14:42:22	La banda bri
https://www.milenio.com/content/el-espiritu-del-bosque	InBrand	El Espíritu del bosque			
https://www.milenio.com/espectaculos/cine/james-bond-25-suspende-p	Hey	James Bond 25 suspende su producción	Milenio Digital	13.05.2019 20:37:27	El actor se l

Figura 5.9: Resultados obtenidos dela extracción del sitio web almacenado en un archivo CSV.

5.3.3. Consideraciones

- Se debe tener conocimiendo de XML para poder realizar las reglas que nos permitirán extraer información de la página web que solicitemos extraer

información y no se extraíga espacios en blanco.

- Cuando la noticia consta de un video, no se obtiene ninguna información adicional de la noticia.
- Se debe tomar en cuenta que las secciones no están homogeneizadas, es decir a pesar de que de la misma página existan varias secciones
- La distribución de la información varia dependiendo a la sección y sitio web.
- Se acoto el periodo de busqueda de noticias ya que algunos sitios web muestran las noticias más recientes, lo cual no nos permite realizar el trabajo del Crawler como se había planteado en un principio.

5.4. Trabajo a futuro

- Una vez que se obtuvo la extracción de las noticias, se pretende seguir así para que todas las noticias formen parte de nuestro corpus del algoritmo de entrenamiento.
- Corregir las reglas para la extracción de la información de los sitios web, y así evitar extraer información con código HTML.

¿Esto es todo lo que les queda de trabajo futuro? ¿y el clasificador no está pendiente?

Bibliografía

Aggarwal, C. C. (2018). *Machine Learning For Text*. Springer.

B. Bracewell, D., Yan, J., Ren, F., and Kuroiwa, S. (2009). Category classification and topic discovery of japanese and english news articles. *Electr. Notes Theor. Comput. Sci.*, 225:51–65.

Breiman, L. (2001). *Machine learning*. Kluwer Academic Publishers.

Bruguera, E. (2019). Qué es un blog. http://openaccess.uoc.edu/webapps/o2/bitstream/10609/17821/5/XX0893006_01331-3.pdf.

Castro, A. and Luis, J. (2014). *Máquinas de Vectores Soporte*.

CleverData (2019). Conceptos básicos de machine learning. https://cleverdata.io/conceptos-basicos-machine-learning/.

Cleverfit (2016). Linear regression. http://www.curvefit.com/linear_regression.htm.

Darliani Asyárie, A. and Wahyu Pribadi, A. (2009). Automatic news articles classification in indonesian language by using naive bayes classifier method. pages 658–662.

Economista, E. (2019). Ranking de medios nativos digitales. https://www.eleconomista.com.mx/Ranking-de-Medios-Nativos-Digitales.

Emprendedor, G. (2019). Qué es una página web. http://www.madrid.org/cs/StaticFiles/Emprendedores/GuiaEmprendedor/tema7/F49_7.9_WEB.pdf.

Farias, G., Vergara, S., Fabregas, E., Hermosilla, G., Dormido-Canto, S., and Dormido, S. (2018). Clasificador de noticias usando autoencoders. pages 1–5.

García-Mendoza, C.-V. and Gambino Juárez, O. (2018). News article classification of mexican newspapers. In Mata-Rivera, M. F. and Zagal-Flores, R., editors, *Telematics and Computing*, pages 101–109, Cham. Springer International Publishing.

García, J., Ramirez, L., and Sanches, M. (2018). Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático. Trabajo Termianl de ESCOM con número 2017-A04 (CDMX).

Google (2018). Googlebot. https://www.humanlevel.com/diccionario-marketing-online/googlebot.

Google (2019). Google cloud. https://cloud.google.com/natural-language/?hl=Es-419.

Gómez Díaz, R. (2005). La lematización en español: una aplicación para la recuperación de información. *Gijón: Trea*.

IBM (2017). Reconocimiento del lenguaje. https://www.ibm.com/blogs/think/es-es/2017/05/16/watson-nlc-en-hogwarts/.

Internaútica, I. (2018). Importancia de las noticias. https://innovainternetmx.com/2014/12/importancia-de-las-noticias/.

Manning, C., Raghavan, P., and Schütze, H. (2010). Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103.

McCallum, A. and Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. 752:41–48.

Mueller, J. P. and Massaron, L. (2016). *Machine learning for dummies*. John Wiley & Sons.

Ochando, M. B. (2013). Técnicas avanzadas de recuperación de información. 1:4.

Olston, C. and Najork, M. (2010). Web crawling. http://infolab.stanford.edu/~olston/publications/crawling_survey.pdf.

Padró, L. and Stanilovsky, E. (2012). *FreeLing 3.0: Towards Wider Multilinguality*. *In Proceedings of the Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Posgrado, U. (2019). Inteligencia artificial. http://www.mcc.unam.mx/IA.php.

Ramdass, D. and Seshasai, S. (2009). Document classification for newspaper articles. pages 1–11.

SAS (2019). Aprendizaje automático. https://www.sas.com/es_mx/insights/analytics/machine-learning.html.

SAS (2019). Procesamiento del lenguaje natural. https://www.sas.com/es_mx/insights/analytics/what-is-natural-language-processing-nlp.html.

Techopedia (2018). Website. https://www.techopedia.com/definition/5411/website.

Téllez-Valero, A., Montes, M., Fuentes, O., and Villaseñor-Pineda, L. (2019). Clasificación automática de textos de desastres naturales en méxico. Master's thesis.

Téllez Valero, A., Montes, M., and Villaseñor-Pineda, L. (2009). Usando aprendizaje automático para extraer información de noticias de desastres naturales. *Computación y Sistemas*, 13:33–44.

Velasco, M. S. (2002). La regresión logística . una aplicación a la demanda de estudios universitarios. 1:10.

Wongso, R., Ariandy Luwinda, F., Christian Trisnajaya, B., Rusli, O., and , R. (2017). News article text classification in indonesian language. *Procedia Computer Science*, 116:137–143.

Yuan, Y., He, L., Peng, L., and Huang, Z. (2014). A new study based on word2vec and cluster for document categorization. *Journal of Computational Information Systems*, 10:9301–9308.

Zambrano, J. (2018). Machine learning. https://medium.com/@juanzambrano/aprendizaje-supervisado-o-no-supervisado-39ccf1fd6e7b.