CENTRO DE CIENCIAS BÁSICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
AUTÓMATAS II
7° "A"

FASE 2_3. PREPROCESAMIENTO DE TEXTO Y LEMATIZACIÓN

Profesor: Francisco Javier Luna Rosas

Alumnos:
Almeida Ortega Andrea Melissa
Espinoza Sánchez Joel Alejandro
Flores Fernández Óscar Alonso
Gómez Garza Dariana
González Arenas Fernando Francisco
Orocio García Hiram Efraín

Fecha de Entrega: Aguascalientes, Ags.**, 11** de octubre de 2021

Fase 2_3. Preprocesamiento de Texto y Lematización

Antecedentes

1.-Análisis de sentimientos

En muchas ocasiones, cuando hablamos de reputación online, aparece el concepto de “análisis de sentimiento” pero ¿sabemos realmente qué significa? El análisis de sentimiento se refiere a los diferentes métodos de lingüística computacional que ayudan a identificar y extraer información subjetiva del contenido existente en el mundo digital (redes sociales, foros, webs, etc.). Gracias al análisis del sentimiento, podemos ser capaces de extraer un valor tangible y directo, como puede ser determinar si un texto extraído de la red Internet contiene connotaciones positivas o negativas.

El análisis de sentimientos, también conocido como minería de opinión, se trata de una tarea de clasificación masiva de documentos de manera automática, que se centra en catalogar los documentos en función de la connotación positiva o negativa del lenguaje ocupado en el mismo.

Con las redes sociales, los usuarios tienen hoy en día todo tipo de facilidades para mostrar sus opiniones sobre cualquier tema que deseen. Tener constancia sobre las opiniones referentes a una marca o producto y medir su impacto es actualmente de vital importancia para todas las empresas, ya que es tu imagen lo que está en juego.

A toda la información que se recopila de esta forma se le denomina minería de opinión y gracias a ella, las empresas tienen una inmediata disponibilidad de la información deseada. Además, la minería de opinión no solo permite responder “qué opinan los internautas sobre su propia marca o producto” sino que facilita, mediante los medios adecuados, obtener ventajas competitivas en diferentes ámbitos.

Las organizaciones utilizan este método para obtener información que les permita comprender la forma en la que los clientes reaccionan respecto a un producto o servicio específico.

Las métricas tradicionales, como el número de vistas, clics, me gusta, compartir, comentarios, etc. se centran en la cantidad. El análisis de sentimiento va más allá de los números y se centra en la calidad de las interacciones entre el público y la organización.

Justo antes de las elecciones de cargos públicos, los partidos políticos, los medios de comunicación, los consultores y los estudiantes realizan varias encuestas y sondeos de opinión. Las personas comparten sus inquietudes, necesidades y expectativas respondiendo a encuestas previas. Los partidos pueden predecir sus posibilidades de ganar una elección haciendo análisis de opinión y minería de opinión de encuestas.

Tenemos un ejemplo de tweets que expresan el sentimiento público y lo que más les importa:

“¡Me encanta lo que hace el candidato del partido azul! ¡Vota por el partido azul!”

“El tema clave en #EleccionesPresidenciales será #ViviendaParaLosPobres y el #AumentoDeTrabajo”

Basado en el análisis del monitoreo de redes sociales y las respuestas de las encuestas, las partes pueden formular sus estrategias futuras. Los líderes pueden conocer la voz del cliente sin filtros y actuar en consecuencia.

Muchas agencias de marketing digital y relaciones públicas utilizan la herramienta de análisis de sentimiento en plataformas como Twitter para medir el reconocimiento de marca.

Puedes extraer todos los datos con el hashtag de tu marca y analizar las palabras utilizadas para expresar emociones y experiencias.

Ejemplo de filtros de datos de análisis de opinión:

Ejemplos de sentimientos positivos: Bueno, excelente, recomendar

Ejemplos de sentimiento neutral: No puedo decir, no sé, tal vez

Ejemplos de sentimientos negativos: Decepcionado, necesita mejorar, no me gustó, no lo recomiendo

Gracias al análisis de sentimiento o minería de opinión podemos recopilar información suficiente para conocer qué piensa o qué opinan los usuarios (o target) en la red Internet.

¿Qué es el análisis de sentimiento? ¿qué es minería de opinión?

En las redes sociales y en la red en general se encuentran multitud de textos, en los cuales deben aplicarse subjetividad y no únicamente clasificarlos según su naturaleza o procedencia. Existen dos formas de enfrentarse al análisis de sentimientos: aplicando un enfoque semántico o aplicando un aprendizaje automático.

Análisis del sentimiento inteligente

Cómo funciona el análisis de sentimiento

Mediante el análisis del sentimiento, queremos lograr entender cuál es la intención exacta de una frase. Saber si se refiere a una marca, a un producto en concreto o a cualquier otro aspecto.

Posteriormente queremos conocer que valoración tiene dicha frase, y para ello se le aplica la denominada polaridad, a través de la cual se clasifica el mensaje en función de la intención que tenga el autor al realizarlo, pudiendo ser este positivo, neutro o negativo. Esto permite controlar el sentimiento de los usuarios respecto a una marca o producto, con lo que obtendremos los puntos fuertes y débiles sobre ello fácilmente.

Para aplicar esta polaridad y posteriormente poder obtener datos concluyentes y predecir comportamientos futuros.

Existen básicamente dos formas de procesar la información obtenida tal como mencionábamos en el punto anterior:

El análisis manual suele darse en casos en los que las palabras claves sobre las que se quiere obtener información pueden representar diferentes significados en diferentes ámbitos, por lo que habrá que estar atento e ir clasificando cada texto en su lugar correspondiente. Un buen ejemplo sería una marca o el nombre de una empresa que se llama igual que una ciudad, de este modo se recopilarían multitud de datos que no tienen nada que ver con lo que de verdad se pretende obtener.

El análisis de sentimiento automático. Este comienza con el establecimiento de una serie de palabras clave para que cualquier texto que contenga esa palabra o combinación de ellas, quede automáticamente encuadrado en una categoría de una forma previamente definida o descartado directamente. Por ejemplo, mensajes que contengan “No me gusta”, “odio” o “no recomiendo” se clasificarán automáticamente cómo datos negativos. Mientras que, aquellos mensajes que incluyan un “excelente”, “genial” o “perfecto”, quedarán clasificados cómo positivos.

Qué limitaciones posee el análisis de sentimiento automatizado

Exactamente no hay ningún método de combinar correctamente las diferentes palabras a utilizar para que el análisis de sentimiento sea 100% fiable.

Los sistemas que se limitan a la configuración y extracción de contenido con palabras clave son incapaces de generar resultados satisfactorios de análisis de sentimiento en su totalidad. Esto viene dado por la complejidad del idioma humano. Por ejemplo, ¿cómo le inculcas a un robot la capacidad de definir si una frase es realizada con sarcasmo o no?

Anteriormente hemos mencionado el término “perfecto” cómo un adjetivo positivo, pero, dependiendo del contexto, este podría cambiar todo el significado de la frase. De esta manera, podría surgir un mensaje que dijera lo siguiente: “Perfecto mensaje a favor del machismo, os habéis lucido”. Este mensaje debería ir entonces clasificado como negativo.

Por este motivo, muchos algoritmos cometen errores, encontrándose con la imposibilidad de fijar una longitud exacta del comentario o la intención real que lleva una determinada palabra. Es decir, no son capaces de inferir de una valoración

exacta de las diferentes relaciones semánticas, y se puede afirmar que actualmente es imposible conseguir un 100% de éxito en este campo.

Sin embargo, los sistemas de análisis del sentimiento más avanzados son capaces de luchar con estos posibles errores y ofrecer resultados más ajustados.

Cómo son las plataformas para análisis del sentimiento

Es aquí donde entra en juego el aprendizaje automático (machine learning). Este término hace referencia a la creación de sistemas a través de la Inteligencia Artificial, donde lo que realmente aprende es un algoritmo, el cual supervisa los datos con la intención anteriormente mencionada: poder predecir comportamientos futuros.

Machine learning y análisis de sentimiento, juntos en el camino

CLIC TO TWEET

Esa cantidad ingente de datos son imposibles de analizar por una persona para sacar conclusiones y menos todavía para hacer predicciones. Los algoritmos, correctamente utilizados, en cambio, sí pueden detectar patrones de comportamiento.

Existen herramientas de monitorización de las redes sociales como NetOpinion que hacen de esta tarea sea sumamente fácil y rápida, por su capacidad de monitorizar en tiempo real y su gestión y procedimientos en la supervisión de los datos.

Normalmente, la estructura utilizada para la organización adecuada de los datos son los árboles binarios, a través de los cuales se pueden establecer los tres patrones de comportamiento ya comentados (positivo, neutro y negativo). Con esta estructura se van observando comportamientos, y cuando ya se han recopilado una cantidad de datos importante, el algoritmo ofrecerá un tanto por ciento de posibilidad de predecir un comportamiento u otro.

La cantidad de datos que se generan actualmente en las empresas está creciendo a un ritmo impresionante, y obtener información útil y valiosa de ellos supone una

ventaja competitiva muy importante respecto a los competidores. Pero, ¿cómo es realmente el proceso?

Se realizan los siguientes pasos:

- Filtración de datos. En primer lugar, se utilizan las palabras claves para descartar contenido no deseado, y posteriormente se establecen palabras para obtener categorías según su polaridad o su procedencia.
- Extracción del contenido. Una vez que pasen el filtro, se elimina el contenido no deseado y se comenzará a trabajar con el contenido de calidad.
- Análisis de contenido. Este proceso lo puede realizar el algoritmo o una persona física en sí. Aquí el contenido útil y de calidad quedará encuadrado en la categoría que le corresponda.
- Limpieza del contenido. Quizás se haya colado contenido erróneamente, y este es el momento de enviarlo a su categoría correcta o descartarlo directamente.
- Revisión. Se gestionarán en este apartado todos los posibles aspectos a mejorar. Tal vez encontremos una nueva palabra a incluir para descartar contenido, o nos demos cuenta de que una palabra considerada positiva se utiliza a modo negativo en determinados momentos.

Para qué sirve el análisis de sentimiento

Gracias a este proceso se consigue obtener datos de calidad,

Se evita tener multitud de datos que carecen de valor para la toma de decisiones

Hacer también, tomar decisiones en tiempo real, como, por ejemplo: para apaciguar una crisis de reputación online.

Gracias al análisis de sentimiento, se consigue desarrollar mejores estrategias empresariales.

Facilita la gestión de la reputación online y ayuda a saber qué acciones llevar a cabo en el plan estratégico de marketing online.

2.- Procesos de lematización

La lematización es un proceso lingüístico que consiste en, dada una forma flexionada (es decir, en plural, en femenino, conjugada, etc), hallar el lema correspondiente. El lema es la forma que por convenio se acepta como representante de todas las formas flexionadas de una misma palabra. Es decir, el lema de una palabra es la palabra que nos encontraríamos como entrada en un diccionario tradicional: singular para sustantivos, masculino singular para adjetivos, infinitivo para verbos. Por ejemplo, decir es el lema de dije, pero también de diré o dijéramos; guapo es el lema de guapas; mesa es el lema de mesas.

Otra manera de enfrentar la normalización es separar las palabras analizadas en un núcleo conceptual (lexema) y agregados morfológicos (morfemas). En este caso, la lematización consiste en encontrar el lexema de las palabras analizadas. Para la mayoría de las lenguas europeas, esto se traduce por encontrar una combinación raiz+sufijo en que la raíz corresponde al lexema buscado y el sufijo a un morfema. El proceso de eliminar sufijos morfológicos se conoce como stemming en la literatura técnica, sin embargo, puede considerarse como una variante de lematización.

La importancia de la lematización radica en el hecho que, para acceso por contenido a bases de datos textuales, permite superar las limitaciones de una búsqueda simple de strings, haciendo que relaciones ocultas por la variabilidad morfológica de las palabras queden manifiestas. La lematización mejora por lo tanto el recubrimiento (recall) aunque pueda ser a expensas de la precisión cuando diferentes conjugaciones morfológicas de una misma raiz están asociadas a conceptos distintos.

Lematizar implica estandarizar, desambiguar, segmentar y, en caso de usar programas de lematización automática, también etiquetar.

La lematización está muy relacionada con el etiquetado automático de textos (POS tagging), que consiste en atribuir a cada palabra su categoría gramatical, ya que la categoría puede determinarse por las flexiones o derivaciones (ej: en castellano -ar

indica un infinitivo, -ado un participio pasado masculino singular, etc.). Muchos esquemas de procesamiento de textos, aplicados a lenguas flexivas europeas, plantean un etiquetado automático previo a la lematización, de manera que al lematizar se cuente con la información de la categoría gramatical de las palabras. Sin embargo, la atribución de etiquetas correctas depende en general de una lematización implícita basada en un análisis de sufijos y prefijos, lo que permite una primera predicción que se corrige, en una segunda etapa, en función del contexto inmediato de la palabra analizada (Brill). Esta manera de proceder presenta algunos problemas: (i) requiere de un corpus manualmente etiquetado de gran dimensión para derivar reglas de etiquetado automático adecuadas, (ii) no aprovecha la existencia de paradigmas de conjugación o derivación, (iii) sólo considera raíces libres.

La lematización puede realizarse automáticamente mediante programas de análisis morfológico. Hay diversos grados de lematización posible: podemos hacer una lematización puramente morfológica, o bien hacer una lematización sintáctica que tenga en cuenta el contexto en el que aparece la palabra. Por ejemplo, en un análisis morfológico la palabra ama tendría dos lemas: el sustantivo ama y el verbo amar. Sin embargo, en un contexto sintáctico (es decir, en una oración), podemos desambiguarlo y optar por un único lema. Así, en El ama de llaves abrió la puerta, ama es sustantivo, mientras que en María ama a Pedro, ama es del verbo amar. Para poder hacer este tipo de lematización es necesario, por lo tanto, hacer un análisis sintáctico.

La lematización es una tarea propia de la Lingüística Computacional, y es útil en la tecnología aplicada a buscadores, traductores automáticos, extracción de información y demás herramientas vinculadas al Procesamiento del Lenguaje Natural.

3.- Procesamiento del lenguaje natural

El "Procesamiento del Lenguaje Natural" es una disciplina con una larga trayectoria. Nace en la década de 1960, como una subárea de la Inteligencia Artificial y la Lingüística, con el objeto de estudiar los problemas derivados de la generación y comprensión automática del lenguaje natural. La Traducción automática, por ejemplo, ya había nacido a finales de la década de los cuarenta, antes de que se acuñara la propia expresión «Inteligencia Artificial».

En sus orígenes, sus métodos tuvieron gran aceptación y éxito, no obstante, cuando sus aplicaciones fueron llevadas a la práctica, en entornos no controlados y con vocabularios genéricos, empezaron a surgir multitud de dificultades. Entre ellas, pueden mencionarse por ejemplo los problemas de polisemia y sinonimia.

En los últimos años, las aportaciones que se han hecho desde este dominio han mejorado sustancialmente, permitiendo el procesamiento de ingentes cantidades de información en formato texto con un grado de eficacia aceptable.

El Procesamiento del Lenguaje Natural es el campo de conocimiento de la Inteligencia Artificial que se ocupa de la investigar la manera de comunicar las máquinas con las personas mediante el uso de lenguas naturales, como el español, el inglés o el chino.

Virtualmente, cualquier lengua humana puede ser tratada por los ordenadores. Lógicamente, limitaciones de interés económico o práctico hace que solo las lenguas más habladas o utilizadas en el mundo digital tengan aplicaciones en uso.

Pensemos en cuántas lenguas hablan Siri (20) o Google Assistant (8). El inglés, español, alemán, francés, portugués, chino, árabe y japonés (no necesariamente en este orden) son las que cuentan con más aplicaciones que las entienden. Google Translate es la que más lenguas trata, superando el centenar… pero hay entre 5000 y 7000 lenguas en el mundo.

Actualmente existen diferentes campos en los que se puede utilizar NLP. Entre los que se pueden mencionar:

- Traducción: es la búsqueda de la forma más adecuada de expresar una frase o palabra en un idioma diferente. Quizás el mejor ejemplo sea el traductor de Google, que con el pasar de los años ha mejorado de manera gradual. Pero en un principio, su desempeño era realmente deficiente ya que utilizaba lo que se llama 'Phrase-Based Machine Translation' (PBMT). El PBMT buscaba frases similares entre los diferentes idiomas, lo que causaba que no siempre se encontraran frases con el mismo significado entre los idiomas; incluso al haber palabras inexistentes en ciertos idiomas se veía imposibilitado de traducir correctamente. Google actualmente utiliza el Google Neural Machine Translation (GNMT) que a su vez emplea Machine Learning con NLP para buscar diferentes patrones entre los idiomas.
- Reconocimiento del habla: es la habilidad de una máquina para interpretar frases o palabras de un lenguaje. Esta aplicación de NLP se puede encontrar en teléfonos móviles y hasta en casas inteligentes. Simplemente con decir "Llamar a John", el dispositivo reconoce lo que esta frase significa y automáticamente empieza a llamar a John. Aquí se puede mencionar el caso de Alexa o Siri que son bastante populares.
- Análisis de sentimientos: es la determinación del tono emocional implícito en una frase. Por ejemplo, si un texto anuncia una caída de las acciones en la bolsa, el programa va a predecir que posiblemente es un texto negativo. Por otra parte, si un texto se refiere a una fiesta y todos están invitados a asistir, se puede decir que su significado es positivo.
- Sistemas de Preguntas y Respuestas: consiste en responder de forma automática a preguntas por medio de un programa. Esta aplicación se puede encontrar fácilmente en chats de redes sociales, llamadas o herramientas como Siri o IBM Watson.
- Generación automática de resúmenes: dado un texto, el programa va a obtener las ideas principales del mismo y producir un resumen coherente. Por ejemplo, si una persona tiene que leer Don Quijote, un programa puede extraer las ideas principales del libro.

- Chatbots: son programas que entablan conversaciones con humanos. Por ejemplo: cuando una persona necesita comprar algo en una tienda virtual y tiene ciertas preguntas sobre un producto, es muy probable que esta reciba respuestas automáticas generadas por una máquina.
- Inteligencia de mercado: con base a lo que una persona ha buscado en internet, automáticamente se empieza a buscar anuncios relacionados. Un buen ejemplo sería cuando se busca un producto específico y automáticamente en redes sociales nos aparecen anuncios relacionados con dicho producto.
- Clasificación automática de textos: se define como la asignación de una etiqueta a un texto según su contenido y semántica. Por ejemplo, cuando se recibe un correo se puede decir si es spam o no basado en su contenido.
- Revisión automática de gramática: el computador reconoce los diferentes errores gramaticales o de ortografía de un texto según el contexto.

Ventajas y Desventajas

Exploramos el procesamiento del lenguaje natural, dándole un vistazo a sus usos así como a sus ventajas/desventajas.

Entre las ventajas que se pueden mencionar en la utilización de NLP son:

- Es más económico la utilización de un programa que una persona. Ya que la persona puede durar el doble o el triple en tareas como las que se mencionó anteriormente.
- Tiempo de respuesta es rápido. Normalmente cuando se emplea NLP el tiempo de respuesta, ya sea en chatbots o llamadas es bastante rápido. Usualmente, los call centers tienen personal limitado lo que restringe el número de llamadas que pueden ser atendidas. Con la utilización de NLP se puede atender más llamadas por lo que el tiempo de espera se reduce.
- Fácil de implementar. En el pasado muchas veces para utilizar NLP se tenía que realizar una ardua investigación con respecto al idioma e implementar muchas tareas de manera manual. En muchos casos, para traducir, se tenía

que crear una especie de diccionario que tuviera un conjunto de palabras que literalmente significan lo mismo en otro idioma. Por lo tanto, se tomaba mucho tiempo de desarrollo. Pero actualmente es muy fácil encontrar modelos de machine learning pre-entrenados que facilitan a los desarrolladores utilizar NLP en diferentes aplicaciones.

Entre las desventajas se pueden encontrar que:

- El entrenamiento puede tomar tiempo: en el caso de que se quiera desarrollar un modelo con un conjunto de datos nuevos, sin utilizar un modelo pre-entrenado, dependiendo de la cantidad de datos puede tomar semanas para obtener un buen desempeño.
- Su confiabilidad no es del 100%: una de las desventajas que siempre se encuentran en machine learning es que nunca se puede asegurar un 100% de fiabilidad. Por lo tanto, pueden existir errores en sus predicciones o resultados.

Las lenguas humanas pueden expresarse por escrito (texto), oralmente (voz) y también mediante signos. Naturalmente, el PLN está más avanzado en el tratamiento de textos, donde hay muchos más datos y son más fáciles de conseguir en formato electrónico.

Los audios, aunque estén en formato digital, hay que procesarlos para transcribirlos en letras o caracteres y, a partir de ahí, entender la pregunta. El proceso de respuesta es el inverso: primero se elabora la oración y luego se “sintetiza la voz”.

Por cierto, la voz artificial cada vez suena más humana, con inflexiones tonales y prosódicas que imitan la producción humana.

Modelos para procesamiento del lenguaje natural

Tratar computacionalmente una lengua implica un proceso de modelización matemática. Los ordenadores solo entienden de bytes y dígitos y los informáticos codifican los programas empleando lenguajes de programación como C, Python o Java.

Los lingüistas computacionales se encargan de la tarea de “preparar” el modelo lingüístico para que los ingenieros informáticos lo implementen en un código eficiente y funcional. Básicamente, existen dos aproximaciones generales al problema de la modelización lingüística:

Modelos Lógicos: gramáticas

Los lingüistas escriben reglas de reconocimiento de patrones estructurales, empleando un formalismo gramatical concreto. Estas reglas, en combinación con la información almacenada en diccionarios computacionales, definen los patrones que hay que reconocer para resolver la tarea (buscar información, traducir, etc.).

PLN

Estos modelos lógicos pretenden reflejar la estructura lógica del lenguaje y surgen a partir de las teorías de N. Chomsky en los años 50.

Modelos probabilísticos del lenguaje natural: basados en datos

La aproximación es a la inversa: los lingüistas recogen colecciones de ejemplos y datos (corpus) y a partir de ellos se calculan las frecuencias de diferentes unidades lingüísticas (letras, palabras, oraciones) y su probabilidad de aparecer en un contexto determinado. Calculando esta probabilidad, se puede predecir cuál será la siguiente unidad en un contexto dado, sin necesidad de recurrir a reglas gramaticales explícitas.

Es el paradigma de “aprendizaje automático” que se ha impuesto en las últimas décadas en Inteligencia Artificial: los algoritmos infieren las posibles respuestas a partir de los datos observados anteriormente en el corpus.

Componentes del procesamiento del lenguaje natural

A continuación, vemos algunos de los componentes del procesamiento del lenguaje natural. No todos los análisis que se describen se aplican en cualquier tarea de PLN, sino que depende del objetivo de la aplicación.

Análisis morfológico o léxico. Consiste en el análisis interno de las palabras que forman oraciones para extraer lemas, rasgos flexivos, unidades léxica compuestas. Es esencial para la información básica: categoría sintáctica y significado léxico.

Análisis sintáctico. Consiste en el análisis de la estructura de las oraciones de acuerdo con el modelo gramatical empleado (lógico o estadístico).

Análisis semántico. Proporciona la interpretación de las oraciones, una vez eliminadas las ambigüedades morfosintácticas.

Análisis pragmático. Incorpora el análisis del contexto de uso a la interpretación final. Aquí se incluye el tratamiento del lenguaje figurado (metáfora e ironía) como el conocimiento del mundo específico necesario para entender un texto especializado.

Un análisis morfológico, sintáctico, semántico o pragmático se aplicará dependiendo del objetivo de la aplicación. Por ejemplo, un conversor de texto a voz no necesita el análisis semántico o pragmático. Pero un sistema conversacional requiere información muy detallada del contexto y del dominio temático.

Lematización de los Tweets

En contraste con la derivación, la lematización es mucho más poderosa, va más allá de la reducción de palabras y considera el vocabulario completo de un idioma para aplicar un análisis morfológico a las palabras, con el objetivo de eliminar solo las terminaciones flexivas y devolver la forma base o de diccionario de una palabra, que se conoce como lema.

En sí, de la palabra o frase original sacamos la función de la palabra raíz (lema), para agilizar el proceso se recomienda primero hacer una extracción de palabras centrales y reunirlas, después hacer una conversión del tiempo y pasar todo a presente, y para terminar trasladamos de plural a singular todo.

Para realizar la lematización utilizamos el enfoque “WordNet(con etiqueta POS)”, Wordnet es una base de datos léxica disponible públicamente en más de 200 idiomas que proporciona relaciones semánticas entre sus palabras.

Está presente en la biblioteca nltk en Python, Wordnet vincula las palabras en relaciones semánticas por ejemplo sinónimos los cuales agrupa en forma de synsets (un grupo de elementos de datos que son semánticamente equivalentes).

Para utilizarlo se debe de descargar el paquete de nltk, posteriormente hay que descargar Wordnet desde nltk

```
import nltk
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
nltk.download('punkt')
from nltk.corpus import wordnet
```

Continuamos con el programa inicializando el lematizador, importando pandas y empezando a leer nuestra base de datos de tweets “Team10 Scrapping (Primera Limpieza).csv”.

```
lemmatizer = WordNetLemmatizer()
import pandas as pd
datos=pd.read_csv('Team10 Scrapping (Primera Limpieza).csv')
```

Para evitar que maneje todo como sustantivos utilizamos la etiquetas POS(Part of Speech), la cual sirve para definir el type de una palabra para ver si es un adjetivo, verbo, sustantivo o adverbio.

```
def pos_tagger(nltk_tag):
    if nltk_tag.startswith('J'):
        return wordnet.ADJ
    elif nltk_tag.startswith('V'):
        return wordnet.VERB
    elif nltk_tag.startswith('N'):
        return wordnet.NOUN
    elif nltk_tag.startswith('R'):
        return wordnet.ADV
    else:
        return None
```

Ahora seleccionamos y recorremos la columna que contiene los tweets, para extraerlos y guardarlos en “sentence” para posteriormente aplicar lo que es la lematización.

```
archivo="Lemantizacion.csv"
csv=open(archivo,"w")
for a in range(len(datos.iloc[:,7])):

    sentence = datos.iloc[a,7]

    pos_tagged = nltk.pos_tag(nltk.word_tokenize(sentence))

    print(pos_tagged)
```

En esta parte es donde se hace la clasificación y se exporta en el nuevo archivo de Excel.

```
35      wordnet_tagged = list(map(lambda x:(x[0], pos_tagger(x[1])),
            pos_tagged))
36      print(wordnet_tagged)
37
38      lemmatized_sentence = []
39
40      for word, tag in wordnet_tagged:
41          if tag is None:
42
43              lemmatized_sentence.append(word)
44          else:
45
46              lemmatized_sentence.append(lemmatizer.lemmatize(word, tag))
47      lemmatized_sentence = " ".join(lemmatized_sentence)
48      lemmatized_sentence = lemmatized_sentence + "\n"
49      try:
50          csv.write(lemmatized_sentence)
51      except ValueError:
52          print("")
```

Ejemplo de ejecución:

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\zente\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     C:\Users\zente\AppData\Roaming\nltk_data...
[nltk_data]   Package averaged_perceptron_tagger is already up-to-
[nltk_data]       date!
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\zente\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Traceback (most recent call last):
```

Primero hace las descargas.

```
('was', 'VBD'), ('Death', 'NNP'), ('And', 'CC'), ('the', 'DT'), ('word',
'NN'), ('was', 'VBD'), ('nigger', 'JJR'), ('And', 'CC'), ('the', 'DT'),
('word', 'NN'), ('was', 'VBD'), ('death', 'NN'), ('to', 'TO'), ('all',
'DT'), ('niggers', 'NNS'), ('And', 'CC'), ('the…', 'NN')]
[('RT', 'n'), ('@', 'n'), ('jatella', 'n'), (':', None), ('“', 'n'), ('In',
None), ('the', None), ('beginning', 'n'), ('was', 'v'), ('the', None),
('word', 'n'), ('And', None), ('the', None), ('word', 'n'), ('was', 'v'),
('Death', 'n'), ('And', None), ('the', None), ('word', 'n'), ('was', 'v'),
('nigger', 'a'), ('And', None), ('the', None), ('word', 'n'), ('was', 'v'),
('death', 'n'), ('to', None), ('all', None), ('niggers', 'n'), ('And',
None), ('the…', 'n')]
[('@', 'RB'), ('3privz', 'CD'), ('WTF', 'NNP'), ('IM', 'NNP'), ('JUST',
'NNP'), ('A', 'NNP'), ('NIGGER', 'NNP'), ('THO', 'NNP')]
[('@', 'r'), ('3privz', None), ('WTF', 'n'), ('IM', 'n'), ('JUST', 'n'),
('A', 'n'), ('NIGGER', 'n'), ('THO', 'n')]
```

Después descompone todo por palabras y las analiza.

@ Mattimatikus @ Urbanistbhd @ DracheOhneLP Ich nenne andere nigger und werde nicht gesperrt
I be go to sue Delta chinese chink for not honor NSA who say I can fly . Life or death ! God famed Democrat leave Obama nigger human trafficking ! Law suit ! Europe alread
Gon na take everything from Faye dad not to call Teddy a nigger
@ Big0047 Wakikaa kafanaa naaa ma nigger
@ matthew47138586 @ Oscaranking2 @ _littlehuman_ This nigger about to be a father http : //t.co/ydDVoNfwgA
@ akintonmi especially s more than spelling .
@ CashWunner nigger mode well
If he can ’ t help market us then fuck him market himself Typical nigger

Y los resultados son exportados a un nuevo archivo de Excel llamado “Lemantización.csv”

Fuentes de consulta

- Lima, A. (2021). Python: enfoques de lematización con ejemplos – Acervo Lima. Retrieved 12 October 2021, from https://es.acervolima.com/2021/02/09/python-enfoques-de-lematizacion-con-ejemplos/
- (2021). Retrieved 12 October 2021, from https://www.youtube.com/watch?v=lhC01D6CbVU&list=PLgHCrivozIb0ULMKfJVV-rFdRG2OeEgfq&index=5
- Análisis de sentimiento. Qué es y cómo realizarlo | QuestionPro. (2019). Retrieved 12 October 2021, from https://www.questionpro.com/blog/es/herramienta-de-analisis-de-sentimientos/#:~:text=El%20an%C3%A1lisis%20de%20sentimiento%20es,sus%20actitudes%2C%20emociones%20y%20opiniones.
- Análisis de sentimiento, ¿qué es, cómo funciona y para qué sirve?. (2017). Retrieved 12 October 2021, from https://itelligent.es/es/analisis-de-sentimiento/
- Lematización - Wikipedia, la enciclopedia libre. (2021). Retrieved 12 October 2021, from https://es.wikipedia.org/wiki/Lematizaci%C3%B3n#:~:text=La%20lematizaci%C3%B3n%20es%20un%20proceso,flexionadas%20de%20una%20misma%20palabra.
- (2021). Retrieved 12 October 2021, from https://users.dcc.uchile.cl/~abassi/ecos/lema.html
- Procesamiento del lenguaje natural - EcuRed. (2021). Retrieved 12 October 2021, from https://www.ecured.cu/Procesamiento_del_lenguaje_natural#:~:text=El%20%22Procesamiento%20del%20Lenguage%20Natural,comprensi%C3%B3n%20autom%C3%A1tica%20del%20lenguaje%20natural.
- Procesamiento del lenguaje natural ¿qué es? - IIC. (2017). Retrieved 12 October 2021, from https://www.iic.uam.es/inteligencia/que-es-procesamiento-del-lenguaje-natural/
- Procesamiento del Lenguaje Natural con Machine Learning. (2021). Retrieved 12 October 2021, from https://www.encora.com/es/blog/procesamiento-del-lenguaje-natural-con-machine-learning

Conclusiones

Andrea Melissa Almeida Ortega:

Para poder aplicar el análisis de cada tweet tuvimos que hacer varios pasos previos para poder garantizar la efectividad de estos, ya que si hay varias “distracciones” para la máquina esta no podrá dar resultados muy precisos, y por ellos hay que prepararle la información para que así la trabaje con mayor facilidad.

Joel Alejandro Espinoza Sánchez:

El procedimiento de análisis de sentimiento es una parte fundamental en el proyecto de la materia. Entre otras cosas, parte del núcleo del desarrollo será el procesamiento óptimo de la información a nuestra disposición y es muy importante realizar esta parte con un buen modelado e implementación, por lo que considero que esta actividad es de las más importantes dentro del proyecto

Óscar Alonso Flores Fernández:

Este análisis es sumamente fundamental para este proyecto ya qué este Procesamiento de información nos dará la interpretación necesaria de todo lo que hemos recolectado hasta ahora y así poder traducirlo a un lenguaje qué nosotros como humanos entendamos para posteriormente seguir con esta serie de procedimientos

Dariana Gómez Garza:

En la elaboración de esta práctica nos dimos cuenta, aunque la computadora pueda hacer todo lo que queramos, es imprescindible indicarle cómo hacerlo, lo cual logramos gracias a la investigación teórica para que nosotros comprendiéramos mejor que era lo que queríamos que la computadora hiciera.

Fernando Francisco González Arenas:

La organización y el trabajar con los tweets hubiera sido más complicado sin la implementación del código ya que analizar uno por uno, si bien llevaría su tiempo seria muy ineficiente a comparación de una máquina que puede analizar cientos de ellos es segundos.

Hiram Efraín Orocio García:

La lematización es muy útil para depurar grandes cantidades de información a su idea principal, para sí poder manejar conjuntos de información más compactos e directos para poder trabajar con más facilidad en nuestro proyecto.

Anexos

Código:

```
import nltk
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')
nltk.download('averaged_perceptron_tagger')
nltk.download('punkt')
from nltk.corpus import wordnet

lemmatizer = WordNetLemmatizer()
import pandas as pd
datos=pd.read_csv('Team10 Scrapping (Primera Limpieza).csv')

def pos_tagger(nltk_tag):
    if nltk_tag.startswith('J'):
        return wordnet.ADJ
    elif nltk_tag.startswith('V'):
        return wordnet.VERB
    elif nltk_tag.startswith('N'):
        return wordnet.NOUN
    elif nltk_tag.startswith('R'):
        return wordnet.ADV
    else:
```

```
        return None

archivo="Lemantizacion.csv"
csv=open(archivo,"w")
for a in range(len(datos.iloc[:,7])):

    sentence = datos.iloc[a,7]

    pos_tagged = nltk.pos_tag(nltk.word_tokenize(sentence))

    print(pos_tagged)

    wordnet_tagged = list(map(lambda x:(x[0], pos_tagger(x[1])), pos_tagged))
    print(wordnet_tagged)

    lemmatized_sentence = []

    for word, tag in wordnet_tagged:
        if tag is None:

            lemmatized_sentence.append(word)
        else:
```

```
        lemmatized_sentence.append(lemmatizer.lemmatize(word, tag))
lemmatized_sentence = " ".join(lemmatized_sentence)
lemmatized_sentence = lemmatized_sentence + "\n"
try:
    csv.write(lemmatized_sentence)
except ValueError:
    print("")
```