Marco de Referencia

Antecedentes Teóricos

En este primer apartado se debiera esperar la presentación de la o las “teorías”, desde las cuales se va a abordar el estudio. En realidad, este estudio solo pretende analizar dentro de modelos o categorías previamente motivadas al enfoque de la analítica y procesamiento de datos. La ciencia de los datos fue definida como un “cuarto paradigma de la ciencia” enfocada tanto empírica, teórica, computacional y ahora basada en datos, por Jim Gray en 2007, ganador del premio Turing, afirmando así que "todo lo relacionado con la ciencia está cambiando debido al impacto de la tecnología de la información el diluvio de la información y datos". Con este propósito se trabajará sobre un conjunto de noticias bajo diferentes medios de comunicación o diversas fuentes contenidos en archivos de tipo ‘csv’ en el que se van a preparar los datos, se analizarán, se aplicarán técnicas de analítica de texto NLP (Natural Language Processing) para entender cómo funcionan los datos que se disponen y cómo se pueden utilizar para hacer modelos de predicción.

Se realizará este estudio utilizando un notebook de Databricks con Spark. Databricks es una empresa fundada por los creadores originales de Apache Spark. Esta empresa proporciona una plataforma de análisis que acelera la innovación al unificar ciencia de datos, ingeniería y negocios. Se creó mediante un proyecto AMPLab en la Universidad de California, Berkeley, que participó en la fabricación de Apache Spark. Se entiende por Spark que es un marco de cómputo distribuido de big data de código abierto basado en la velocidad, su facilidad de uso y en el análisis sofisticado de los datos.

El tema de análisis de datos comprende varios retos a tratar para hacerlo de la manera más eficiente posible, haciendo referencia a Aleksandra Sirovatko, CEO y fundadora de Data Science UA, comprende que los retos más importantes a tener en cuenta al ejercer la analítica de datos “son la calidad de los datos, la multifuncionalidad esperada y la maldición que constituye la complejidad del algoritmo”. Con este fin debemos utilizar las herramientas adecuadas para que los datos y la información no sean

corrompidos, como estipula Sirovatko, la calidad de los datos es lo más importante para evitar errores que puedan afectar el funcionamiento de las tareas o programas que utilicen dichos datos.

Existen diversas técnicas de analítica de texto NLP de los diferentes modelos y aplicaciones de SparkML-NLP, pero se tomó en consideración utilizar la técnica de “análisis de sentimientos” y la técnica enfocada a la “regresión logística”. En cuanto al procedimiento sobre el análisis de sentimiento, hace referencia a diversos métodos de lingüística computacional que permiten identificar y extraer información subjetiva del contenido ya existente en el mundo digital, dicho de otra forma, permite utilizar y analizar el comportamiento de los datos en diferentes campos que uno mismo define. Dicho lo anterior, podemos ser capaces de extraer un valor directo y específico para determinar si las noticias utilizadas contienen una connotación positiva o negativa al respecto. Sin embargo, la regresión logística trata los datos de manera un poco diferente, se utilizará esta metodología para poder predecir en cómo serán los futuros artículos, en sí darán un resultado con una variable categórica, de la misma manera brindando una respuesta determinando si la noticia tiene una connotación positiva o negativa, mediante los datos extraídos anteriormente y tomando en cuenta el análisis de sentimientos ya experimentado con dicha información.

Antecedentes Conceptuales

Definición de conceptos

- AMPLab: laboratorio de la Universidad de California, Berkeley, enfocado en análisis de big data. sus siglas hacen referencia a Algorithms, Machines and People Lab.

- Análisis de sentimiento: proceso de determinar el tono emocional que hay detrás de una serie de palabras, y se utiliza para intentar entender las actitudes, opiniones y emociones expresadas en una mención online.

- Big Data: conjuntos de datos de mayor tamaño y más complejos, procedentes de nuevas fuentes de datos. Estos conjuntos de datos son tan voluminosos que el software de procesamiento de datos convencional no puede administrarlos. Sin embargo, estos volúmenes

masivos de datos pueden utilizarse para abordar problemas empresariales que antes no hubiera sido posible solucionar.

- Databricks: empresa que proporciona una plataforma de análisis para los datos que acelera la innovación al unificar ciencia de datos, ingeniería y negocios.

- Data Science: campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer cierto conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sean estructurados o no.

- Datos: representación simbólica de un atributo o variable, ya sea cuantitativa o cualitativa describiendo hechos o sucesos.

- Natural Language Processing: rama de la inteligencia artificial que ayuda a las computadoras a entender, interpretar y manipular el lenguaje humano para cerrar la brecha entre la comunicación humana y el entendimiento de las computadoras.

- Notebook: un entorno pensado para satisfacer necesidades concretas y ajustarse al flujo de trabajo de la ciencia de datos y la simulación numérica. En esta, los usuarios pueden realizar cálculos, ejecutar códigos y visualizar los datos y resultados.

- Pandas: librería de python enfocada al análisis de datos, que proporciona unas estructuras de datos flexibles que permiten trabajar con los datos de forma muy eficiente.

- Premio Turing: premio de las Ciencias de la Computación que es otorgado anualmente por la Asociación para la Maquinaria Computacional (ACM) a quienes hayan contribuido de manera trascendental al campo de las ciencias computacionales.

- PySpark: framework open source para la computación en paralelo utilizando clusters. Se utiliza especialmente para acelerar la computación iterativa de grandes cantidades de datos o de modelos muy complejos.

- Regresión logística: tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica en función de las variables independientes o predictoras.

Referencias Bibliográficas

1. Zakir, J., Seymour, T., & Berg, K. (2015). *Issues in Information Systems* [Ebook] (16th ed.). Retrieved from http://www.iacis.org/iis/2015/2_iis_2015_81-90.pdf
2. Russom, P. (2011). *Big Data Analytics* [Ebook]. TDWI. Retrieved from https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf
3. Análisis de Datos. (2016). Retrieved 18 November 2019, from https://www.questionpro.com/es/analisis-de-datos.html
4. Lemmatization Approaches with Examples in Python. Retrieved 19 November 2019, from https://www.machinelearningplus.com/nlp/lemmatization-examples-python/
5. Fonseca, E. (2019). State-of-the-art Multilingual Lemmatization. Retrieved 18 November 2019, from https://towardsdatascience.com/state-of-the-art-multilingual-lemmatization-f303e8ff1a8
6. Interview with Aleksandra Sirovatko. Retrieved 20 November 2019, from https://365datascience.com/interview-aleksandra-sirovatko/
7. Data science. Retrieved 18 November 2019, from https://en.wikipedia.org/wiki/Data_science
8. Glosario de términos sobre Inteligencia Artificial, Big Data & Data Science. (2018). Retrieved 20 November 2019, from https://itelligent.es/es/tag/analisis-de-sentimiento/
9. Sánchez, J. Qué es y cómo interpretar una regresión logística. Retrieved 19 November 2019, from https://conceptosclaros.com/que-es-regresion-logistica/