María Alejandra Pérez

mperezo@eafit.edu.co

Enunciado

Realizar una tarea de clasificación para los tweets adjuntos en seleccionados y no seleccionados usando únicamente el campo de “texto”. Para esto es necesario, como mínimo, programar su propia red neuronal y usar otros dos clasificadores o combinaciones de los mismos. Se espera que todos los clasificadores tengan un mejor resultado que el hacer una clasificación aleatoria.

Proyecto 3

Clasificación de Tweets

**TRABAJO FUTURO O DE MEJORA:**

* Mejora del sistema de clasificación: lo que ya se tiene más una conexión a un servicio como TWITTER, básicamente lo que se hace dentro del tweet proceso de limpieza para obtener palabras claves, como hashtags para retroalimentar la base de datos del clasificador, cada hashtag que se agrega va a mejorar el clasificador de esa categoría.

**HERRAMIENTAS:**

**NLTK, Natural Language Toolkit**

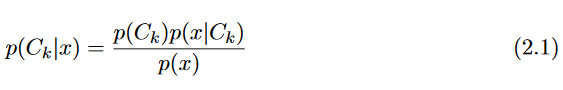
El NLTK (Natural Language Toolkit) es una biblioteca de Procesamiento de Lenguaje Natural que utiliza el lenguaje de programacion Python. NLTK es software libre, lo que permite a estudiantes y al personal academico realizar estudios con la herramienta sin necesidad de realizar una inversion economica. Esta herramienta es tambien de codigo abierto, lo que lo hace ideal para expandir sus funcionalidades en caso de necesitarlo. El hecho de estar implementada como una biblioteca Python reduce la curva de aprendizaje, y la acerca al mundo academico, cuya mayor parte de integrantes se encuentra familiarizado con este lenguaje de programación.

**CLASIFICADORES:**

* Naive Bayes

Los clasificadores de Bayes ingenuos (Naive Bayes classifiers) son clasificadores utilizados ampliamente en la actualidad para problemas de clasificacion de textos, tanto en ambitos academicos como industriales. Los caracteriza una gran simplicidad, que conlleva la ventaja de que son de facil comprension y por tanto de facil uso. A pesar de esta simplicidad, los clasificadores de Bayes ingenuos son bastante eficaces y tienen una gran cantidad de aplicaciones prácticas.

Los clasificadores de Bayes ingenuos se basan en un modelo probabilistico, el Teorema de Bayes.



Según este teorema, la probabilidad de un suceso Ck, dada una situacion x viene determinada por el número de veces que se ha producido el suceso Ck en esa situacion x, dividido por el número de veces total que se ha encontrado esta situacion x (Eq. 2.1).

Aplicados al problema de la clasificaci ́on, nos permite inferir la posibilidad que tiene un elemento de pertenecer a cada una de las clases consideradas, teniendo en cuenta cada una de las caracter ́ısticas de este elemento como un suceso a priori. Los clasificadores de Bayes ingenuos reciben su nombre debido a que toman en cuenta la contribuci ́on de cada caracter ́ıstica hacia la posibilidad de una clasificaci ́on de manera independiente, en contraposici ́on a correlacionar varias caracter ́ısticas para realizar una estimaci ́on m ́as precisa a la hora de clasificar un elemento.

* Enfoques analíticos

Los enfoques anal ́ıticos son aqu ́ellos que intentan realizar un an ́alisis morfosint ́actico sobre el texto de entrada, obteniendo informaci ́on relativa a las construcciones sint ́acticas de las oraciones y la morfolog ́ıa de las palabras utilizadas. Esta clase de enfoques se basa en el an ́alisis morfosint ́actico cl ́asico, que lleva realiz ́andose durante siglos de forma manual, como herramienta docente para dar a conocer las bases de un lenguaje natural.

Sin embargo, se trata de enfoques de mayor complejidad en comparaci ́on con los enfoques estad ́ısticos, debido a que requieren la implementaci ́on de conocimientos sobre la gram ́atica, sintaxis y sem ́antica del lenguaje objetivo. Por ejemplo, se deben de conocer todas las conjugaciones posibles de todos los verbos del lenguaje objetivo, y atender al contexto de las oraciones para utilizar la acepci ́on correcta de palabras ambig ̈uas.

Por otra parte, las herramientas de PLN deben contener corpus morfosint ́acticos y sem ́anticos para cada uno de los idiomas, adem ́as de requerir para cada uno de ellos diferente para paliar las idiosincracias de cada uno de los lenguajes.

**BIBLIOGRAFIA:**

* <https://riull.ull.es/xmlui/bitstream/handle/915/3082/Analisis%20automatico%20de%20textos%20en%20espanol%20utilizando%20NLTK.pdf?sequence=1>