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.

Proyecto 3

Clasificación de Tweets

**HERRAMIENTAS:**

**NLTK, Natural Language Toolkit**

El NLTK (Natural Language Toolkit) es una biblioteca de Procesamiento de Lenguaje Natural que utiliza el lenguaje de programación Python. NLTK es software libre, lo que permite a estudiantes y al personal académico realizar estudios con la herramienta sin necesidad de realizar una inversión económica. Esta herramienta es también de código 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 académico, cuya mayor parte de integrantes se encuentra familiarizado con este lenguaje de programación.

En el caso del proyecto se usa nltk para descargar una lista de stop words en español del corpus, esta técnica se usa en la limpieza del texto.

**Listas de Categorías**

Estas listas de categorías son creadas con el fin de alimentar al clasificador y al modelo. Se tomaron estas categorías basados en una inspección de ciertos tweets donde se lograba distinguir a que hacen referencia y también en relación y en torno a cultura.

**Lista de Hashtags**

Esta lista de hashtags es creada a partir de la observación de ciertos tweets y su frecuencia de uso de menciones en estos. Esto, al igual que las listas de categorías nos ayuda a alimentar al clasificador y al modelo.

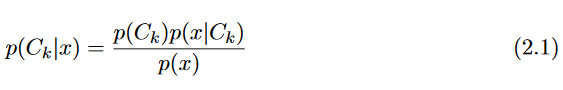
**CLASIFICADORES:**

La mayoría de autores o paginas recomiendan:

* Naive Bayes Classifier

Los clasificadores de Bayes ingenuos (Naive Bayes classifiers) son clasificadores utilizados ampliamente en la actualidad para problemas de clasificación de textos, tanto en ámbitos académicos como industriales. Los caracteriza una gran simplicidad, que conlleva la ventaja de que son de fácil comprensión y por tanto de fácil 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 probabilístico, el Teorema de Bayes.



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

Aplicados al problema de la clasificación, 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 características de este elemento como un suceso a priori. Los clasificadores de Bayes ingenuos reciben su nombre debido a que toman en cuenta la contribución de cada característica hacia la posibilidad de una clasificación de manera independiente, en contraposición a correlacionar varias características para realizar una estimación más 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.

**TRABAJO FUTURO O DE MEJORA:**

Mejora del sistema de clasificación: lo que ya se tiene planteado más una conexión a un servicio como Twitter, básicamente lo que se hace dentro del tweet es un 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.

**BIBLIOGRAFIA:**

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