Datos Generales

Proyecto	Sistema para análisis de sentimientos en reseñas de películas basada en ANEW		
Estado	ACTIVO		
Semillero	UNIAUTONOMA		
Área del Proyecto	Ingenierías	**Subárea del Proyecto**	Ingeniería de Sistemas
Tipo de Proyecto	Proyecto de Investigación	**Subtipo de Proyecto**	Investigación en Curso
Grado	Pregrado	**Programa Académico**	Ingeniería de Sistemas
Email	semilleros@uac.edu.co	**Teléfono**	3015085005

Información específica

Introducción

El presente proyecto se refiere al manejo de la información referente en la web 2.0, que hoy en día encontramos en blogs, redes sociales, wikis, entre otros. Sabemos que existe una gran cantidad de información que se encuentra en internet, pero ¿Qué tanta información está almacenada en la web? Es una pregunta que hoy en día sería casi imposible de responder, lo cierto es que anualmente se crean cerca de 2,5 trillones de bytes de datos esto sería alrededor del equivalente de aproximadamente 3 veces las partículas de arena que encontramos en una playa. Muchas personas se preguntan ¿Qué se debe hacer con toda esa información? Está tiene un costo invaluable para las empresas que la utilizan como un método para conocer su imagen corporativa desde el punto de vista del usuario, es decir realizar un análisis de datos con respecto sus productos y/o servicios en cuanto a los comentarios de los consumidores y obtener una calificación de estos. Uno de las ramas en donde más abundan opiniones de los usuarios es en el dominio de las películas de cine. Este proyecto tiene como enfoque principal la construcción de un sistema de análisis de sentimientos automático en español sobre las opiniones dadas por los espectadores de películas. El sistema trabaja bajo el enfoque léxico utilizando la adaptación al español de las normas afectivas para las palabras en inglés.

Planteamiento

Hoy en día el almacenamiento de datos en la web ha crecido enormemente. Según (Cañabate, 2014) se estima que en 2020 habrá en el mundo 26.000 millones de dispositivos conectados que generarán el 40% de la totalidad de los datos creados, aumentando así la información que debe ser procesada. La gran mayoría de estos datos provienen de redes sociales, blogs, foros o correos electrónicos. Esta información, subjetiva en la mayoría de casos, resulta muy interesante por su potencialidad de aplicación y la gran cantidad de texto no estructurado en la Web no analizado. Esta gran cantidad de datos, la subjetividad y su heterogeneidad hacen que su procesamiento sea complicado, difícil y tedioso si se hace de manera manual, pero que, son de gran utilidad para las organizaciones ya que pueden encontrar comentarios acerca de sus empresas o productos. Sin embargo, hay escasas herramientas computacionales, sobre todo en español, que sean capaces de procesar todos estos datos automáticamente de la Web y brindarle a las empresas datos confiables con el fin de tomar decisiones acertadas. Cabe resaltar que en el idioma español específicamente en Colombia existen hoy en día muy pocas herramientas que sean capaces de analizar y clasificar los sentimientos debido a que la mayoría de la información sobre esta temática está enfocada a desarrollarse para el idioma de inglés. Teniendo en cuenta lo anterior, existe un área de investigación dedicada al análisis de las opiniones contenidas en un documento. Esta área es conocida como análisis de sentimiento (AS) o minería de opinión (MO) (Pang & Lee, 2008). El trabajo presentado en este documento está motivado principalmente por la necesidad de construir un sistema para clasificación de sentimiento en reseñas de películas basado en léxico en el español. La formulación del problema sería: ¿Cómo se puede analizar las opiniones de los espectadores de películas y determinar el sentimiento asociado a ellas de forma automática en un texto en idioma español?

Objetivo General

Construir un sistema para análisis de sentimientos en reseñas de películas basadas en ANEW.

Objetivos Específicos

Identificar las etapas de un sistema para análisis de sentimientos. Asociar el análisis de sentimientos con las palabras afectivas en inglés (ANEW). Diseñar un sistema para análisis de sentimientos basados en ANEW. Implementar una herramienta computacional que valide el diseño planteado. Evaluar el sistema en el dominio de reseñas de películas en español.

Referente

En la realización del proyecto se ha encontrado una gran cantidad de datos en el mundo digital. Este proyecto busca trabajar con los datos de diferentes portales de la Web 2.0 conformada por foros, redes sociales, blogs, correos, entre otros. El término de Web 2.0 nace en 1999 cuando fue propuesto por Darcy DiNucci quien lo describió como Web 2.0 describe la World Wide Web con énfasis en el contenido generado por los usuarios, la usabilidad y la interoperabilidad y fue popularizado por Tim OReilly y Dale Dougherty en la conferencia the OReilly Media Web 2.0. Hoy 10 años más tarde una gran cantidad de páginas web hacen parte de la denominada Web 2.0. La información en estos sitios web puede no encontrarse organizada, así como estar incompleta o ser ambigua. Es por esto que para la elaboración del proyecto es necesario apoyarse en herramientas del área de inteligencia artificial que buscan la creación de sistemas automáticos los cuales sean capaces de realizar tareas las cuales, hasta el momento, habían estado reservadas en su desempeño exclusivamente a los seres humanos (Luis Amador Hidalgo, 1996); más específicamente en la disciplina de procesamiento de lenguaje natural (PLN) la cual trata de programar una computadora para que ésta pueda interpretar y entender todo tipo de texto escrito por los seres humanos. Dentro del área de PLN se encuentra el análisis de sentimiento (AS). El AS busca analizar las opiniones, sentimientos, valoraciones, actitudes y emociones de las personas hacia entidades como productos, servicios, organizaciones, individuos, problemas, sucesos, temas y sus atributos (Liu, 2012). Según Jeonghee Yi un analizador de sentimientos es extraer sentimientos sobre un tema específico utilizando técnicas. Para la construcción de una herramienta que permita realizar AS se debe tener en cuenta varios aspectos. Primero se debe extraer el texto que contiene la opinión y luego determinar que sentimiento tiene asociado (clasificación de sentimiento). Para la extracción se elige un conjunto de datos de sitios web donde abundan opiniones y comentarios en línea. Para la clasificación, normalmente positiva o negativa, se debe escoger el enfoque de clasificación que puede estar en: técnicas basadas en aprendizaje de máquinas (ML), basadas en léxico (LEX) o una combinación de las dos anteriores (hibridas) (Medhat, Hassan, and Korashy, 2014). Para el primer enfoque se realiza una subdivisión en aprendizaje supervisado y aprendizaje no supervisado. Por el lado del enfoque basado en léxico se subdivide basado en corpus y basado en diccionarios. Las diferencias fundamentales radican que el primero utiliza algoritmos o estrategias para aprender a partir de textos o corpus determinados y el segundo modelo utiliza diccionarios, léxicos y corpus de palabras, frases o su combinación. Por último el trabajo aquí presentado utiliza ANEW como recurso basado en LEX. ANEW provee un conjunto de 1.034 palabras etiquetadas emocionalmente y está clasificado en tres dimensiones: valencia, excitación y dominio. La valencia va desde un valor de 1 (malo) hasta 10 (bueno), siendo 5 el valor considerado como neutro

Metodología

El tipo de investigación es descriptiva puesto que se realizará un análisis estadístico de datos obtenidos, y se identificarán las principales propiedades para el procesamiento de críticas sobre películas, con respecto a la recolección de datos, se hará una revisión documental apoyada por una ficha bibliográfica. La ficha es un instrumento que permite recopilar la información obtenida en libros, revistas, periódicos, documentos personales y públicos y de cualquier testimonio de carácter histórico. Se construye con base en un trabajo de carácter intelectual del investigador en donde manifiesta su capacidad de análisis y de crítica. Es el resultado de la lectura reflexiva y minuciosa para obtener los aspectos relevantes que son útiles tanto para la formación de un marco teórico, sustentación de las hipótesis y por otra parte es de suma utilidad para el trabajo final de redacción de un informe.(Robledo, 2010). Después de la revisión documental se creará la ficha en donde se describirá la literatura idónea en el campo de estudio. Luego se hará un análisis de la información registrada que será insumo para el diseño e implementación del sistema.

Resultados Esperados

Los resultados parciales de la investigación son: • Se identificaron las etapas de un sistema para análisis de sentimientos. • Se caracterizó cómo el ANEW puede integrarse a un sistema de análisis de sentimientos. • Se realizó el diseño del sistema para análisis de sentimientos basados en ANEW. Los resultados esperados son: • Desarrollo de software con un enfoque total hacia el área de cine haciendo que reconozca la mayor cantidad de terminología utilizada en éste ámbito. • Validación del software que analice críticas de películas con una alta precisión.

Conclusiones

• Se ha propuesto un sistema de análisis de sentimientos que busca analizar las opiniones acerca de películas de cine y determinar el sentimiento asociado a éstas. Los resultados logrados hasta el momento son el diseño general de sistema y la identificación de herramientas, técnicas y procedimientos para el posterior desarrollo. El diseño del sistema está basado en el recurso léxico en español conocido como ANEW.

Bibliografía

? J. Redondo, I. Fraga, I. Padrón and M. Comesaña, "The Spanish adaptation of ANEW," Behavior Research Methods, vol. 3, p. 39, 2007. ? Cañabate, E. P. (2014). Big data: ¿solución o problema? Obtenido de http://arantxa.ii.uam.es/~epulido/bigdata.pdf. ? Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. . Foundations and trends in information retrieval, 1-135. ? Liu, B. (2012). Sentiment Analysis and Opinion Mining. . Synthesis Lectures on Human Language Technologies., 1-167. ? Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 1093-1113.

Integrantes

Documento	Tipo	Nombre	Email
1140891957	PONENTE	MIGUEL CELIS	semilleros@uac.edu.co
1234088558	PONENTE	ANDRES ARROYO	semilleros@uac.edu.co

Instituciones

NIT	Institución
8901025729	UNIVERSIDAD AUTÓNOMA DEL CARIBE