SRI sobre el corpus de comentarios realizados en redes sociales cubanas

Intergrantes

- Daniel Orlando Ortiz Pacheco
- Aldo Javier Verdecia Delgado

Resumen

Para crear un entorno con muchos comentarios reales extraídos de las distintas redes sociales cubana, donde se puede mezclar textos autogenerados por una inteligencia artificial, y que un grupo de persona de dicha nacionalidad intente reconocer a la IA entre todos estos textos. Para automatizar la recuperación de los comentarios reales que estén relacionado con algún tema en especifico, se realiza una investigación para seleccionar el mejor Modelo de Recuperación de Información para listar los comentarios más coherentes con la consulta planteada de inicio

Introducción

Según [1] *el español es hablado principalmente en España e Hispanoamérica, como también entre las comunidades de hispanohablantes residentes en otros países, destacando Estados Unidos con más de 40 millones de hablantes de español.* Dicho idioma con el paso del tiempo y la intensa mezcla cultural, en cada una de las localizaciones mencionadas, a experimentado diversas alteraciones que han terminando caracterizando los distintos "lenguajes" dentro de la comunidad hispano hablante. Debido a las diferencias existentes entre los distintos grupos que representan al que es un de los idiomas más hablados en el mundo, el estudio de los detalles y características de cada uno de estos subideomas es muy interesante para muchas investigaciones en los campos como la *Recuperación de Información* y el *Procesamiento de Lenguaje Natural*, sobre todo cuando el resultado que se busca tiene el objetivo de impactar a las personas pertenecientes a un grupo específico de la comunidad hispana En este caso particular motivado por el objetivo de lograr un **bot** que sea capaz de mezclarse en la sociedad digital cubana, de forma tal que puede generar opiniones sobre diversos temas y que los miembros de este subgrupo de los hispano hablantes realizando el análisis más crítico posible no sean capaces de distinguir entre opiniones reales y textos autogenerados. El resultado final necesita de una presentación en sociedad, una sociedad formada por cubanos, que sean consientes de la existencia del **bot** y que estén en disposición de detectar al mismo entre un conjunto de opiniones sobre un tema dado. Aunque sería ideal colocar al generador de opinión en las distintas redes sociales a una exposición máxima de personas y con un conjunto mucho más amplio de ejemplos reales, en estos ambientes los usuarios no se encuentran con la alerta y disposición necesaria para evaluar el desempeño de la inteligencia artificial. Un marco más acorde a las necesidades de la investigación

sería un espacio interactivo en el que el usuario se sienta retado a detectar a la entidad automática. Este escenario plantea un nuevo reto sobre la mesa, pues el mismo no cuenta con ejemplos reales de forma natural con los que mezclar los comentarios automatizados (como si lo tiene las redes sociales), en este marco el sistema no solo debe aportar los textos generados sino que también debe contar con un corpus de opiniones reales de la comunidad y recuperar los más acordes para cada tema planteado.

Para dar respuesta a la necesidad complementaria de la investigación antes descrita, se desarrolló un Sistema de Recuperación de Información(SRI) usando toda la información obtenida del procesos de minería de datos de la investigación. Optimizando la eficacia de este SRI en el sentido más semántico posible, se obtienen opiniones reales escritas por cubanos y relacionadas con el tema propuesto. Gracias al mismo se puede obtener un conjunto de ejemplos con los que mezclar los textos generados tan grande como se quiera, dentro del las dimensiones de corpus y teniendo en cuenta que a medida que crece la cantidad de comentarios recuperados aumenta la posibilidad de obtener comentarios incoherentes respecto al tema.

Corpus

El corpus de los comentarios realizados por cubanos en las distintas redes sociales, presenta distintas características a tener en cuenta para realizar recuperaciones de información sobre el mismo. La característica más determinante para el desarrollo del SRI es que todos los documentos del corpus tiene un contexto, todo comentario es la respuesta al algo que con anterioridad se publicó, con lo cual en varias ocasiones el contenido semántico del documento no se encuentra necesariamente en lo que se dice. Esto se debe analizar con detalle en al definición del modelo a implementar, pues los modelos clásicos basados únicamente en la ocurrencia y la frecuencia de las palabras de la consulta e interior de los distintos documentos no son del todo ideal, a menos que todo comentario se almacene y analice unido con si contexto. Luego el corpus de estudio también se encuentra caracterizado por ser propenso al empleo de diminutivos y siglas, además contener errores ortográficos, detalles pueden tener un gran impacto en el SRI si el mismo cuenta con mecanismos para detectar dichas equivalencias o si el corpus es previamente procesado (lo cual podría restarle expresividad a los resultados recuperados). Otras características un poco menos relevantes son; los documentos del corpus en promedio son textos cortos relativamente, debido a la variedad de las fuentes (las distintas redes sociales y usuarios) presenta una gran variedad de estilos, no se conoce el número de temas y dominios específicos a los que este se refiere

Modelo

Dadas las características de corpus y la descripción del problema, la resolución del mismo pasa por una amplia fase de investigación en la que se deben imple-

mentar distintos modelos y seleccionar aquellos que recupere los textos que sean más relevantes y coherentes, con respecto a la consulta realizada. Los modelos que formaran parte de esta fase experimental deben tener algunas características mínimas para ser considerados soluciones potenciales. Aquellos que sean seleccionados deben poder incorporar el concepto de ranking entre sus características, pues la inmensidad del corpus unido a que el tamaño de la lista de resultados solicitados no será muy grande entonces el sistema debe encontrar un orden para ofrecer los mejores documentos ante cualquier consulta. Ademas debe contar una función de similaridad fácil y rápida de computar, ya que en principio, para resolver cual es el resultado correcto para una consulta dada dicha función se debe computar tantas veces como documentos contenga el corpus, una función muy compleja puede provocar que el sistema sean ineficiente respecto a la experiencia del usuario. Como el SRI seleccionado se despegará en un ambiente interactivo, entonces los modelos que tengan capacidades de retroalimentación también deben ser tenidos en cuenta, y pueden ser una gran elección final aun no siendo los que mejores resultados tenga.

Modelo Vectorial

Uno de los modelos que encajan en la descripción anterior y por tanto, en principio es una solución al problema planteado es el **Modelo Vectorial**. Es un modelo basado en el álgebra vectorial, cuenta con el concepto de ranking desde su propia definición y preprocesando el corpus para obtener su representación de **indices invertidos**, además de almacenar otros cálculos que pertenecen a la función de similaridad y no depende de la consulta, el procesos de clasificación de todos los documentos del corpus es aproximadamente lineal con respecto al tamaño del corpus, teniendo en cuenta que la cantidad de términos de la consulta es mucho menor que dicha dimensión. Teniendo en cuenta que el objetivo principal de sistema es lograr recuperar comentarios que sean coherentes con la consulta que se realizó, con lo cual no solo es importante el texto del comentario sino que también puede ser determinante el contexto en el que este se realizo, en el preprocesamiento del corpus, no solo se tuvo en cuenta el comentario en si, sino que se unieron publicación y comentario como si de un solo documento se tratará. De esta manera aunque el un comentario y una consulta no contenga términos en común aun tiene posibilidad de aparecer en los primeras posiciones, si su "contexto" se encuentra próximo al tema de la consulta.

La lista de términos se conformó a partir de la estructura antes descrita (publicación + comentario), a partir de la cual se realizaron varias técnicas de procesamiento de texto tokenización, extracción de *stop words*, lematización y detección de entidades nombradas, dicho procesos se le realiza a todo el corpus y a cada consulta por igual. Este modelo además gracias a su sencillez es fácilmente integrable con otra técnicas (expansión de consultas, clustering del corpus, ...) que puedan ayudar a una mejor clasificación de los distintos textos según su semántica. De unirse dichas técnicas, la experiencia que el sistema puede acumular por la interacción de los usuarios puede ser de gran ayuda para que los

resultado a largo plazo pueden ser extremadamente buenos.

Conclusiones

A simple vista se puede notar que los resultados de las primeras implementaciones no son los mejores, en dichos enfoques se obtiene comentarios que no son del todo coherente con el contexto planteado, haciendo análisis muy profundos se puede pensar que en ciertas interpretaciones de la consulta los comentarios podrían tener relación pero no es el efecto deseado.

Referencias

1. Wikipedia