SRI sobre el corpus de los textos extraídos de las sociales cubanas

Resumen

Se necesita crear un entorno con muchos comentarios reales provenientes de las distintas redes sociales cubana, donde se puedan intercalar 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 hispanohablante. 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 el objetivo es lograr un bot que sea capaz de mezclarse en la sociedad digital cubana, de forma tal que pueda generar opiniones sobre diversos temas y que los miembros de este subgrupo de los hispanohablantes realizando el análisis más crítico posible no sean capaces de distinguir entre opiniones reales y textos autogenerados. El resultado final necesita de una evaluación en un entorno lo más parecido a lo antes descrito posible, 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. Dicho SRI es resultado de un largo proceso de desarrollo, investigación y selección de modelos, no solo para su definición sino también para evaluar el desempeño del mismo, pues este no solo debe ser eficiente en tiempo sino que también debe ser capaz de detectar dentro de todos los textos del corpus el conjunto de documentos que más coherentes sean con un tema dado

Modelos

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 implementar distintos modelos y seleccionar aquellos que recupere los textos que sean más relevantes y coherentes, con respecto a la consulta realizada. Los modelos que formarán 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.
Partiendo de la totalidad del corpus se realizan varias técnicas de procesamiento de texto tokenización, extracción de stop words, lematización y detección de entidades nombradas para detectar todos los términos del mismo, para posteriormente vectorizar el mismo realizando el proceso tf-idf de dichos términos . 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.

Modelo Probabilístico

Otro modelo a tener en cuenta es el Modelo Probabilístico, este presenta una filosofía distinta al vectorial, este no toma en cuenta la cantidad de ocurrencias de un término ni en los textos y en el corpus sino que su vectorización es binaria donde cada componente de los vectores representa un término el cual es 1 o 0 si el término aparece en el texto correspondiente, además intenta reconocer la relevancia de los textos mediante la teoría de las probabilidades, aunque en la practica su desarrollo se basa en valores estimados y en visiones frecuentistas de las probabilidades. En este caso particular el desarrollo se vaso en los valores estimados pi y ri a los que se les asignaron 0.5 e idf para cada termino respectivamente. De igual manera que el modelo vectorial, la selección de este modelo estuvo fuertemente influenciado por su

capacidad para brindar un ranking de los documentos y que realizando algunos precómputos la función de similaridad del mismo es prácticamente lineal sobre el tamaño del corpus. Otra ventaja de este modelo es que el mismo puede modificarse durante su explotación e ir aprendiendo con la experiencia, a medida que el sistema de respuesta a las distintas consultas, calificando las mismas los distintos pi se pueden ir ajustando para que este en futuras ocasiones pueda mejorar su desempeño.

Thesaurus

Los thesaurus es otra de las técnicas frecuentemente usadas en recuperación de información, sobre todo en textos. La idea más sencilla que aborda un thesaurus es la de tener un diccionarios de sinónimos para que al momento de listar los términos indexados en las distintas consultas poder agregar dichos sinónimos a la lista y que de esta manera se tengan en cuente los documentos que a pesar de que no contener términos originales de la consulta pueden ser relevantes para la misma. Para la construcción de esta herramienta existe tres enfoques principales, los manuales, los supervisados y los automatizados. En este casos tras estudiar el trabajo y resultados de [2] y [3] se ha seleccionado el enfoque automático. Los thesaurus sobre todo los del enfoque automático se salen un tanto de la definición de diccionario de sinónimos, pues es complicado definir algoritmicamente cuando dos palabras son sinónimos, en su lugar se busca otro tipo de relaciones de dependencias basadas en frecuencias y probabilidades.

Thesaurus Strength

Este thesaurus se basa en la idea de la correlación textual, bajo la tesis de que si dos palabras aparece la mayoría de veces juntas entonces es probables que este termino también sea importante para la consulta. En el caso particular del desarrollo realizado influenciado por el estudio de [2], el thesaurus consiste en una matriz de correlación termino a termino, donde la medida de correlación se define de la forma:

$$Strength(q_j, t_i) = \frac{n_{q_j t_i}}{n_{t_i}}$$

Donde $n_{q_j t_i}$ es la cantidad de documentos en los que el termino j de la consulta y el termino i del corpus aparecen juntos, mientras n_{t_i} es la cantidad de ocurrencias del termino i sin el termino j. Además se fija un alpha para definir a partir de que valor de correlación se considera dependencia entre estos términos. Este componente se debe posicionar antes de los modelos antes descritos para hacer la expansión de la consulta termino a termino y agregar todos los nuevos términos encontrados a la consulta de entrada a los modelos de recuperación de información. Esta es una herramienta muy interesante que puede agregar mucho valor a las distintas consultas, pero cuenta con la desventaja de que depende totalmente del corpus para ser efectiva, si el corpus es demasiado heterogeneo es muy probable que existan muchas palabras que solo aparecen en pocos textos estas no necesariamente tiene una relación semántica importante con el resto de los términos de dicho texto, además de que el parámetro alpha es una decision del usuario la cual es en principio totalmente arbitrario y basada en la intuición.

Thesaurus Bayesiano

El thesaurus bayesiano también es fruto del estudio del trabajo de [2], en este caso la definición de correlación entre términos se define mediante la inferencia de probabilidades de una red bayesiana. En este caso la red esta compuesta por dos capa, cada una con tantos nodos como términos se detectaron en el corpus y las conexiones en la red unicamente enlazan nodos de capaz distintas. La construcción de la red se basa en el Thesaurus Strength, dos nodos de distintas capas están conectados si y solo si estos se encuentra correlacionados según el Thesaurus Strength. Luego de la construcción, el proceso de inferencia consta de dos parte, un proceso de inicialización, donde cada nodo de la capa externa toma una probabilidad del inverso de la dimension de corpus salvo los términos que aparezcan en la consulta los cuales se inicializan con probabilidad 1, y otro de propagación de las probabilidades que se desarrollo siguiendo las descripciones realizadas en [2]. El resultado final es una distribución de probabilidades en la capa interna que al ordenarlas y filtrarlas por un umbral dado se obtienen todos los términos que tiene alguna relación con la consulta realizada. Esta técnica teóricamente es muy interesante, pero el procesos de construcción temporalmente es ligeramente superior al cuadrado de la cantidad de términos del corpus, aunque el corpus de las redes sociales en promedio son textos cortos, la dimension del corpus recolectado hace el computo de la red una tarea extremadamente lenta. Pero en las pruebas iniciales con aproximadamente el 20% del corpus final los resultados que arrojaba

este thesaurus eran bastante interesantes

Evaluación de los Modelos

Para evaluar este modelo se realizo una investigaron sobre el dataset para encontrar una estructura en el mismo que permita agrupar los documentos mediante su información semántica. Dicha investigación no obtuvo muy buenos resultados se puede ver dicha investigación aquí

Conclusión

Los distintos modelos presenta buenos resultados, pero no se pudo encontrar una buena manera de evaluarlos. Si el dataset es relativamente pequeño o se cuenta con el tiempo suficiente para entrenar el modelo los autores recomienda un combinación del modelo vectorial con un thesaurus bayesiano. En otro caso se recomienda el modelos probabilístico. Dichas recomendaciones esta basadas únicamente en la experiencia de los autores

Referencias

1 - Wikipedia
2 - Clustering terms in the Bayesian network retrieval model a new approach with two term-layers (Luis M. de Campos∗, Juan M. Fernández-Luna, Juan F. Huete)
3 - Query Expansion in Information Retrieval Systems using a Bayesian Network-Based Thesaurus (Luis M. de Campos, Juan M. Fernandez, Juan F. Huete)