GLOBAL PROJECT

ALUMNO:

TETIANA IANIUK

PROGRAMA:

POSGRADO EN INTELIGENCIA ARTIFICIAL Y MACHINE LEARNING

NOMBRE DEL PROYECTO:

“EXTRACCIÓN DE NECESIDADES Y METAS DE CONSUMIDORES A PARTIR DE RESPUESTAS A ENCUESTAS MEDIANTE TÉCNICAS COMBINADAS DE PLN”

Contenido

RESUMEN

El resumen describe en máximo 300 palabras cuál es el problema abordado por el Global Project, la solución planteada por el alumno, por qué esta solución es una buena solución al problema (justificación) y los resultados obtenidos en el desarrollo del Global Project.

La meta de este trabajo resultó ser no trivial, considerando los requerimientos, limitaciones de los recursos tanto financieros y computacionales, como los humanos. Sin embargo, gracias a este tipo de limitaciones, las soluciones que se trabajan y popularizan mediante el trabajo presente y los que habrán parecidos (tesis, artículos, investigaciones, etc.), hacen evidente la demanda de unos algoritmos más asequibles, para que se democraticen las tecnologías NLP y se amplíen posibilidades de pequeñas empresas y pequeños grupos de investigación.

La meta de este trabajo consiste en satisfacer una necesidad práctica del negocio de analizar las fuentes textuales escritas en un lenguaje coloquial en idioma español, con faltas gramaticales y ortográficas para resumir de manera automática las ideas contenidas en estas, como también captar el estilo original del auditorio, extraer ideas inusuales.

Actualmente, hay varias soluciones disponibles en el mercado con las que un emprendedor podría analizar un conjunto de respuestas o comentarios de sus clientes potenciales. Sin embargo, estas herramientas a veces no pueden analizar un volumen mediano-grande de datos de una vez, requieren que el usuario trabaje su prompt para conseguir el formato de respuestas requerido, son costosas en términos computacionales en el caso de querer mejorarlos y usarlos en el desarrollo, generalizan demasiado o simplemente ofrecen una estadística de palabras separadas que no puede ofrecer una buena comprensión del contexto.

En el presente trabajo se ha logrado analizar de manera automática un conjunto de respuestas de los encuestados, se ha detectado los deseos, problemáticas, “dolores” del auditorio. Las frases extraídas podrían ser consideradas como ejemplos de expresiones de los encuestados.

Por otro lado, se prevén varias posibilidades de mejorar, por ejemplo vía una corrección de errores avanzada previa al embedding, experimentos con diferentes algoritmos de procesamiento, utilización de otras arquitecturas de redes de aprendizaje profundo para NLP fuera del embudo utilizado(BERTOPIC).

CÓDIGO: GP_NLP_TETIANA_IANIUK.ipynb

(https://colab.research.google.com/drive/1tAKBaR1lCE_BScXKCav3xtehyenIBHmN?usp=sharing)

ENLACE VIDEO PRESENTACIÓN:
https://www.flexclip.com/es/share/4101809e5134725bcd7ce1f7f618062161a1a9a.html

INTRODUCCIÓN

La introducción debe contener el detalle sobre:

- Cual es problema detectado por el alumno y cómo lo ha identificado?
- ¿Cómo se ha solucionado este problema anteriormente?
- ¿Cuál es la solución planteada por el alumno?

- ¿Por qué esta solución es adecuada e innovadora?
- Cual ha sido el procedimiento seguido para lograr la solución planteada.
- Resumen brevísimo de los resultados obtenidos.
- Resumen de la estructura de este documento.

Extensión máxima: 2000 palabras.

¿Cuál es problema detectado por el alumno y cómo lo ha identificado?

En el mundo de hoy hay una constante carrera de detectar los deseos del usuario, sus miedos, penas, molestias, opiniones, etc. Las empresas que quieren competir tienen que estar siempre al tanto del comportamiento de sus consumidores, sea una empresa con productos digitales, físicos o servicios.

Las empresas analizan las opiniones públicas, textos de redes sociales, realizan encuestas y son las que más datos tienen para realizar en el idioma español, muchas veces tienen que resolver tareas no triviales para lograr análisis estos datos y poder aprovechar los resultados innovando y superando la calidad esperada por el público.

Las características de los textos normalmente son:

- faltas de ortografía y puntuación son muy habituales
- lenguaje coloquial
- idioma español latino
- comentarios cortos
- pueden llegar contener varias frases claves importantes en un comentario/post

Este trabajo es una investigación de algoritmos contemporáneos ejecutada sin fines lucrativos para una empresa que vende cursos online. Suele preguntar a su auditorio sobre las razones de haber contratado el curso, si les ha gustado o no y por qué, qué necesidades había o queda sin resolver, etc. La empresa tiene alrededor de 120 000 suscriptores que contestan sus encuestas.

Para experimentar y tratar de conseguir las metas propuestas en el presente trabajo se han usado 1495 registros de comentarios de los encuestados, que contestaron a la siguiente pregunta ***"8. Si estuvieras compartiendo un café con un terapeuta angelical, ¿Cuál sería la pregunta más importantes que le pedirías que te respondiera? - 1. Input - textarea".***

Actualmente, es imposible para la empresa procesar esta información completa e íntegra de manera manual.

El interés especial de la empresa en el análisis de respuestas es:

- Procesar gran cantidad de respuestas sin gastos de recursos adicionales
- Resumir las ideas principales de cada encuesta.
- Conseguir ideas nuevas
- Comunicarse con sus clientes en su lenguaje

La meta principal del trabajo fue extraer una lista de frases claves más cruciales de respuestas a preguntas abiertas en unas encuestas digitales. Las frases de interés fueron relacionadas con las dificultades, miedos, "barreras", deseos de las personas, problemas que les gustaría resolver con el producto comercial de la empresa proveedora de datos para el presente proyecto.

¿Cómo se ha solucionado este problema anteriormente?

Actualmente, hay varias soluciones con las que un emprendedor podría analizar un conjunto de respuestas o comentarios de sus clientes potenciales. Sin embargo, estas herramientas no son ideales ni muy específicas. La empresa en cuestión ha probado algunas herramientas disponibles actualmente en el mercado para realizar un análisis de este tipo y ninguno ha dado los resultados esperados:

- Open AI Playground - tiene un límite de **4500 tokens (4000 palabras** aproximadamente) dependiendo del plan contratado. Da los mejores resultados de resumen de patrones en el texto de todas las herramientas vistas por la empresa de cursos, pero el inconveniente es que hay saber formular bien las preguntas, tiene límites de input, no muestra gráficos visuales de estadística, como una herramienta generativa, el usuario no sabe en qué medida ha alucinado en su respuesta aun si le ha seteado la temperatura de creatividad al mínimo.
- Insightbase.ai - sugiere que preguntas puedes hacer a tu base de datos y no requiere uso de SQL, muestra gráficos informativos. Pero no da respuestas relevantes sobre contenido de texto, no está en español.
- WordClouds.com - extrae una nube de palabras más utilizadas, pero solo es palabra por palabra y no frases, y toma en cuenta palabras sin sentido, lo que tendría más sentido para el análisis de opiniones, tampoco muestra una estadística exacta.

- ¿Cuál es la solución planteada por el alumno?

La idea del proyecto es aplicar algoritmos avanzados de Machine Learning adaptado a los datos, sin necesitar formular preguntas a la base de datos, no a un chat, devolver al usuario principales temas de los que hablan los clientes potenciales en formato de una lista, mantener al máximo el lenguaje de clientes, recomendar las temáticas nuevas que pueden ser transformados a los nichos de temas nuevos de cursos y clusterizar a los clientes potenciales uno por uno, para que el usuario pueda obtener etiquetas de clases de intereses.

- ¿Por qué esta solución es adecuada e innovadora?

La innovación es encontrar una solución que utilice los últimos logros de algoritmos NLP, logrando los requerimientos esperados y considerando limitaciones del dataset y los recursos, según se describe a continuación.

Estos planteamientos junto con otras limitaciones han formado una serie de condiciones para cumplir este proyecto. Los cuales se identificaron como siguientes:

- Encontrar un modelo para trabajar con textos en español latino, lenguaje coloquial, etc.
- El modelo debe ser el más avanzado y potente posible para conseguir mejores resultados
- Debe ser posible descargar y ejecutar en el Google Collab gratuito (CPU/TPU/ T4 GPU).
- El algoritmo debe mantener el estilo del lenguaje del texto original

- Ausencia de posibilidades para anotación (entrenamiento supervisado excluido)
- Debe extraer una lista de frases claves, no un párrafo de texto generalizado, ni palabras claves.
- Debe ser relativamente fácil en ejecución debido a la alta complejidad y variedad de las opciones de biblioteca, modelos, librerías, pipelines, versiones disponibles.

- ¿Cuál ha sido el procedimiento seguido para lograr la solución planteada?

Vectorizar frases enteras en las respuestas para captar su sentido general, usando modelos más avanzados disponibles para idioma español, correr el algoritmo de reducción de dimensionalidad, clasterizarlos, seleccionar las palabr/frases más representativas cada cluster mediante una técnica estadística c-TF-IDF (una adaptación de TF-IDF para el pipeline utilizado). Luego, con KeyBertInspired (algoritmo inspirado en KeyBert) calculamos la similitud de coseno entre las frases claves candidatas y las frases de su clúster.

Experimente aplicando RAKE (Rapid Automatic Keyword Extraction) para rankear las frases de temáticas más importantes. Es una técnica simple a base de grafo, calcula la coocurrencia de palabras de siguiente manera: tokeniza las frases por signos de puntuación y las stopwords, luego calcula la frecuencia de palabras, “degree of word” (sumas de cuantas veces cada palabra ha ocurrido con cada otra), luego normaliza dividiendo este valor a la frecuencia de ocurrencia general. Al final a cada frase clave se le asigna un score que es la suma de degree of word normalizado de cada palabra que la compone.

- Resumen brevísimo de los resultados obtenidos.

Se han detectado temas principales, muy parecidos a los que ha encontrado el chatGPT. Aunque el último devuelve frases que son más fácilmente entendibles para el ser humano. Las frases claves resultantes describen los deseos, problemáticas, “dolores” del auditorio. A la vez, un punto de mejora será reducir la cantidad de temas por cluster y seleccionarlos o rankear de una mejor manera para lograr establecer prioridad y mejorar comprensibilidad.

Se prevén varias posibilidades de mejorar, por ejemplo vía corrección de errores avanzada, extracción de frases clave mediante RAKE (Rapid Automatic Keyword Extraction - una técnica a base de grafo o algunos autores como Sharma y Li (2019) en su trabajo Self-Supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling lo clasifican como un algoritmo estadístico, creo que depende tan solo de la interpretación preferida, calcula la coocurrencia de palabras (Barrera, 2018) y luego pasarla misma combinación de algoritmos, uso de los transformers sec2sec para mejorar las frases detectadas, además de la utilización de los modelos radicalmente distintos.

ESTADO DEL ARTE

En este apartado el estudiante deberá reportar aproximadamente 10-15 proyectos, o bibliografía disponible en buscadores científicos que describan cómo se ha afrontado el problema que él está tratando de solucionar, justificando con esta literatura la relevancia de la solución que ha planteado.

Un proyecto de máster implica que el estudiante se acerque al planteamiento de un problema poco explorado en la literatura y que ofrezca una solución innovadora al mismo.

Por tanto, el estado del arte debe tener un sub-apartado de conclusiones en el cual el estudiante justifique el carácter innovador y de nuevo conocimiento de su Global Project.

Extensión máxima: 2000 palabras.

Escribe aquí…

Análisis de modelos usados en los trabajos recientes

NER

La misma tarea se podría tratar como una tarea de reconocimiento de entidades nombradas (NER de sus siglas en inglés) . Esta tarea implica identificar y clasificar entidades como nombres de personas, ubicaciones, fechas y entidades específicas del dominio, como términos médicos, en un texto dado.

Se usa frecuentemente en el ámbito de la medicina. Una de aplicaciones es el procesamiento de las hojas de interconsultas médicas una por una y extracción de ellas datos sobre medicamentos, enfermedades, síntomas, etc. relacionadas, para su posterior clasificación por las vías de asistencia médica más adecuada (Báez et al., 2020).

En el caso nuestro de querer procesar de datos de encuestas, se necesitaría una etapa posterior después de la NER , la que consistirá en clusterizar y rankear las entidades reconocidas y devolver al usuario una N cantidad de las clases/clusters más frecuentes.

Las ventajas potenciales de este método comparando con los resultados de modelos generativos: que el output del modelo se basaría únicamente en el input sin lugar a generación de contenido que no existía en nuestro input. Es beneficioso para el negocio debido a que permite conocer lenguaje original de los consumidores y utilizarlas en campañas futuras; seguridad de las estadísticas obtenidas no han sido distorsionadas; se controla la variedad y granularidad del output mediante parámetros de clusterización.

Desventajas que se prevén en la etapa de diseño:

- La necesidad de anotación manual del texto - requiere un esfuerzo y tiempo importante.
- Necesidad de una aplicación de una capa de otra red neuronal de clusterización, que aumentaría margen de error al ejecutarse arriba de los datos suplidos por otra red neuronal que va a tener su propio nivel de error probablemente alto.

BERT

BERT y sus modificaciones. Estos modelos se basan en la codificación de relaciones entre palabras y frases, extrayéndolas de manera bidireccional de las frases en corpus. Son originalmente de multitarea y para conseguir el mejor rendimiento posible se usan capas de output dependiendo de la tarea a resolver. El entrenamiento se basa en maskear palabras en una frase y tratar de predecirlas (Devlin et al., 2018).

Las ventajas son:

- No es necesaria anotación previa. Es muy costoso en general y no es posible para nuestro caso en particular.
- Existe una gran cantidad de diferentes tipos de modelos BERT especializados en tareas de PLN ligeramente diferentes. Se puede seleccionar una que sería más cercana a nuestra tarea.
- Hay modelos BERT preentrenados en idiomas diferentes y se puede conseguir utilizar el modelo preentrenado en un corpus en español - una solución “out of the box” (igual que en el caso de GPT).

Desventajas:

- BERT se especializa en modelar la relación entre palabras y entre frases, pero no puede tener en cuenta las partes que se repiten en el resumen final en el caso del output generativo. Además, las frases en el resumen pueden depender unas de otras en términos de contenido y significado. Algo que tratan combatir las redes diseñadas específicamente para Extractive Summarization (resumen extractivo - en español) (Jia et al., 2020).
- BERT considera hasta 512 tokens, - si hay una secuencia de texto larga que debe dividirse en múltiples secuencias de texto cortas de 512 tokens (*Fine-tuning BERT with sequences longer than 512 tokens*, 2021).

MODELOS UNIVERSALES DE LENGUAJE sec2sec

GPT

Uno de los modelos más potentes que existen hasta ahora para NLP generativo, permitirá recibir respuestas de una vez (zero-shot) o mostrando al modelo ejemplos de respuestas (few-shot). Será fácil hacer el siguiente paso del análisis, que espera el negocio - generar recomendaciones de temáticas y/o necesidades a cubrir relacionados con los detectados de los textos procesados.

La desventaja del método (para su utilización directa por el negocio) es que el tamaño del input es limitado a 4087 tokens , dependiendo del modelo GPT a usar. El límite de tokens para gpt-35-turbo es 4096 tokens, mientras que los límites de token para gpt-4 y gpt-4-32k son 8192 y

32768 respectivamente (Mrbullwinkle, 2023). Estos límites incluyen tanto el número de tokens en los mensajes enviados a lo largo de una "conversación", como los de las respuestas del modelo.

El set de datos (resultados de encuesta) más pequeño a procesar contiene 1100 palabras de 84 respuestas, el de 1500 respuestas: 14500 + y el set de 3200 respuestas: 33999 palabras. Por tanto, sería deseable desarrollar un algoritmo que pueda analizar el input de todos los datos a la vez de un solo cuestionario del caso contrario debería ser procesado por partes.

Otra desventaja ya para su uso en este proyecto es que es de pago, lo que es una limitación en la presente etapa.

LLAMA

LLAMA - un modelo sec2sec inicialmente preentrenado para tareas universales. Los autores comentan que la versión LLaMA-13B supera al GPT-3 (175B) en la mayoría de las pruebas, y en su versión máxima LLaMA-65B compite con los mejores modelos, Chinchilla-70B y PaLM-540B. Otra ventaja importante es que el modelo fue entrenado en un corpus en varios idiomas, incluido el español. Entre las fuentes hay archivos de Wikipedia correspondientes al período de junio a agosto de 2022 (Touvron et al., 2023). ¡Además, la licencia, que permite su libre utilización para fines comerciales, está disponible en el Hugging Face Hub! Ya está publicada la segunda versión del modelo - Llama2.

Una de las desventajas de este punto puede ser un bajo porcentaje de textos en idioma español en el corpus - una parte indeterminada de 4,5% pertenecientes a los archivos de Wikipedia en 20 idiomas (Touvron et al., 2023).

Otra desventaja común de todos los modelos generativos para extraer las palabras claves de las personas - no se va a saber con seguridad cual es el porcentaje de extracción y generación hay en el resumen generado por estas redes. Aunque al final se trató de hacer un experimento con este modelo. Encontré una versión preentrenada en un set de datos en español, pero no he podido descargar la versión de 7B parametros (github: https://github.com/Garrachonr/LlamaDos) (Garrachonr, s. f.) ni con un método de compresión de doble cuantización de 4 bits.

ELECTRA

ELECTRA - modelo pequeño y no enmascara, sino que sustituye palabras para aprendizaje no supervisado descrita en el artículo por Clark et al. (2020).

RigoBerta

La búsqueda de los modelos preentrenados en español me crucé con el artículo RigoBERTa: A State-of-the-Art Language Model For Spanish escrito por Vaca et al.(2022). La cual trata de competir con otros modelos recientes basados en Bert (Maria, RoBerta). El modelo está basado en la arquitectura de DeBERTa , un BERT con mejoras en el encoder y en el decoder para que pueda generalizar y predecir con mejor precisión. El modelo DeBERTa lideraba el benchmark SuperGLUE para el 6 de enero de 2021, superando el umbral humano por un margen considerable (90.3 frente a 89.8) , según los autores de DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION (He et al., 2021).

Los autores el artículo RigoBERTa: A State-of-the-Art Language Model For Spanish mencionan que a pesar de que el español es una de las lenguas más habladas, hay pocos modelos de lenguaje para este idioma. Esto se debe a la falta de corpus de alta calidad en español y a los costos significativos de entrenar modelos grandes, que son más asequibles para empresas multinacionales. La mayoría de los modelos en otros idiomas son desarrollados en entornos académicos (Vaca et al., 2022).

MODELOS PARA RESUMEN EXTRACTIVO DE TEXTO

Hay dos lógicas principales de trabajo para resumir textos con redes neuronales: abstractivo y extractivo. El paradigma abstractivo se centra en generar un resumen palabra por palabra después de codificar el documento completo. El enfoque extractivo selecciona directamente frases del documento para ensamblar un resumen.

Como se ha mencionado anteriormente, para el presente trabajo es importante extraer el lenguaje que usan personas encuestadas tanto para su análisis como para su uso en campañas de marketing. La investigación reciente sobre resumen extractivo abarca una amplia gama de enfoques. Estos trabajos suelen instanciar su arquitectura codificador-decodificador, eligiendo RNN, Transformer o Hierarchical GNN como codificador, decodificadores autorregresivos o no autorregresivos. En su trabajo Jia et al. (2020b) los investigadores presentan su modelo HAHSum (Hierarchical Attentive Heterogeneous Graph for Text Summarization). El modelo está basado en grafos y maneja diferentes jerarquías de información, incluyendo palabras y oraciones, y destaca las dependencias de redundancia entre las oraciones. Según los autores, HAHSum demuestra un rendimiento excepcional en pruebas a gran escala.

En el mismo trabajo se mencionan otros modelos que han sido populares para sumarizar los textos. Aunque me he concentrado en los modelos más mainstream, los menciono aquí brevemente para cualquier referencia de los lectores en un futuro:

- El Oracle es el modelo resumen extraído de acuerdo a las etiquetas de la verdad absoluta.
- Lead es un método base para la extracción de resúmenes de texto que elige las primeras frases como un resumen.
- SummaRuNNer tiene en cuenta el contenido, la relevancia, la novedad y la posición de cada oración al decidir si debe incluirse en el resumen extractivo.
- PNBERT intenta emplear el conocimiento transferible no supervisado.
- BERTSUMEXT aplica BERT preentrenado en la sumarización de texto y propone un marco general tanto para modelos extractivos como abstractivos.
- MATCHSUM es un método de dos etapas para extraer y luego comparar, y la primera etapa es BERTSUMEXT.

En el trabajo Analyzing Social Media Data: A mixed-methods framework combining computational and qualitative text analysis Andreotta et al. (2019) el equipo de investigadores busca

una solución para detectar temas comunes de los tweets sobre el cambio climático. Usaron factorización interconjunta de matrices no negativas (NMijF) de Nugroho et al., (2017) para agrupar tweets por temáticas. Este proceso utiliza tanto los patrones de co-ocurrencia de términos entre tweets, como la relación socio-temporal entre los tweets para derivar temas. La granularidad de tópicos detectados de manera automática ha sido evaluada manualmente por los investigadores y ayudantes involucrados. Posteriormente, se realizó un análisis temático cualitativo (TA; Braun y Clarke, 2006) ligeramente adaptado para utilizarlo en frases cortas como Tweets. Este análisis agrupó cada tweet en 5 temas distintos.

En el trabajo de tesis Aplicación de procesamiento de lenguaje natural sobre una encuesta de satisfacción de Álvarez y Pontificia Universidad Católica de Chile. Escuela de Ingeniería (2021) , un trabajo extenso y muy bien explicado, se desea clasificar los comentarios de acuerdo a los ejes de negocio de los que tratan (labels conocidos) y si es bueno o malo respecto del rendimiento esperado por cada uno de ellos.

Se realizó un proceso de limpieza y normalización del texto: Se eliminaron caracteres especiales que no corresponden al español y se corrigieron las faltas de ortografía. Luego se emplea la tokenización por palabra y se juntan los vectores de un comentario para alimentar el modelo de vectorización FastText que calcula los vectores prediciendo las palabras que rodean a una palabra objetivo. Luego se extraen los pesos y se forma una representación vectorial de un comentario como el promedio de todos los vectores de palabras que lo componen.

También se utilizó el modelo BERT fine-tuned sobre el corpus normalizado y creado con ortografía revisada para experimentar con la tokenización y vectorización. Los autores explican que la vectorización previa a Bert fue necesaria, ya que es un modelo preentrenado en sus propios tokens que no coincidían con los tokens del corpus en cuestión, se usa el tokenizador implementado en la librería Transformers.

Para la clasificación se emplea la lógica de multi-label clasificación para poder asignar varias o zero etiquetas a un comentario. Los autores aplican tres diferentes tipos de algoritmos para conseguir este tipo de clasificación: Multi Layer Perceptron Classifier, Classifier Chains y Random K-Labelsets. Se aprovecha que BERT capta relación no solo entre palabras en una frase , sino también entre las frases, pudiendo distinguir si una frase es continuación de la otra.

Para el resumen de texto en el trabajo de tesis analizado se utiliza el algoritmo de clusterización K-means para agrupar las representaciones vectoriales de las frases del texto en clusters y seleccionar las oraciones más cercanas al centroide de cada uno.

Los trabajos que menciona la tesis analizada Towards automatic extractive text summarization of A-133 Single Audit reports with machine learning Chou et al. (2019) y Leveraging BERT for Extractive Text Summarization on Lectures D. Miller (2019) también usan estos algoritmos para el resumen extractivo de textos, como se explicará a continuación.

El trabajo de Chou et al. (2019) presenta un enfoque detallado sobre cómo resume auditorías de instituciones federales de Estados Unidos. Al igual que la tesis de Álvarez y Pontificia Universidad Católica de Chile. Escuela de Ingeniería (2021), utilizan el algoritmo K-means para agrupar las oraciones de las auditorías en clusters. Para vectorización se utilizó embedding GloVe el número de clusters se determina mediante el método del codo, se elige el valor de *k* tiene el más conveniente balance entre la que tan grandes son las distancias entre puntos incluidos dentro de cada cluster hasta su centro y el valor *k*. Posteriormente, los autores revisan las oraciones de los clusters definidos para identificar temas y descartan las oraciones no informativas. La cantidad de

oraciones en el resumen se define mediante una heurística que varía según la longitud del texto original: para textos pequeños se escogen más oraciones para formar el resumen y para los grandes - se escoge un porcentaje más pequeño (5%). Como métrica de distancia se utilizan las distancias de coseno.

Por otro lado, en el enfoque de D. Miller (2019), se centra en la generación de resúmenes de clases académicas para estudiantes. El texto se preprocesa, se divide en oraciones y se vectoriza utilizando el modelo BERT preentrenado. La cantidad de oraciones en el resumen se deja a elección del usuario a través del hiperparámetro *k*. Para la formación del resumen, se elige una oración de cada uno de los *k* clusters, basándose en la distancia euclidiana al centroide. Los investigadores admiten que los algoritmos seleccionados tenían unas desventajas que pudieron haber afectado a los resultados de manera negativa, como por ejemplo, el límite de procesamiento de BERT que fue de 512 tokens por batch, el corpus general no relacionado específicamente a las clases académicas, además de que el lenguaje de corpus fue más propenso a ser colonial, algunas veces el algoritmo pudiera dar oraciones gramaticalmente incorrectas y confusas, por lo que sugieren pasar los resúmenes por un algoritmo que los haga parecer más al lenguaje humano.

Como precisa Miller (2019), debido a la ausencia de "gold standard" de resúmenes, se usó el feedback humano para la medición de la calidad de los resúmenes.

De los artículos analizados hasta ahora hago la conclusión que en el conjunto vectorización + k-means el componente del que depende la calidad de resultados en una gran medida es el algoritmo de vectorización seleccionado.

Conclusiones sobre el estado del arte

Actualmente, las tareas de resumen extractivo de texto se realizan a base de los modelos basados en Transformer(más bien encoders) de última generación. Los benchmarks más recientes y de autoridad para tareas de NLP en general es SuperGLUE.

Los tipos de modelos representativos serían:

- BERT y sus modificaciones se usan cambiando las capas de output dependiendo de la tarea o de tratamiento que se quiere aplicar.
- Transformers (encoder + decoder)
- NER modelos a entrenar para reconocimiento de miedos, barreras, deseos dentro del texto + cálculos estadísticos de las entidades reconocidas de la encuesta.
- Modelos basados en grafos.
- Combinación de modelos algebraicos (verctorizador + clusterizador).

De los artículos analizados hasta ahora hago la conclusión que en el conjunto vectorización + k-means el componente del que depende la calidad de resultados en una gran medida es el algoritmo de vectorización seleccionado.

OBJETIVOS

En este apartado el estudiante definirá el objetivo general de su Trabajo Final de Máster y máximo 4 objetivos específicos que describan cómo llevó a cabo el objetivo general.

El objetivo general debe dar solución a problema realmente retador en el área del BI o DS, y debe haber quedado bien justificado en el estado del arte.

Objetivo general

Objetivos específicos

1. Objetivo 1
2. Objetivo 2
3. Objetivo 3
4. Objetivo 4

Extensión máxima: 1 página.

Objetivo general

Extraer una lista de ideas claves expresadas por los encuestados en forma de texto libre escrito.

Objetivo 1

Utilizar el embudo modular BERTopic (Grootendorst, 2022) junto con modelos preentrenados en corpus en español para vectorización.

Objetivo 2

Generar una lista de frases claves en cantidad comprensible.

Objetivo 3

Evaluar mediante un feedback humano/comparacion propia subjetiva con chatGPT 3.5

SOLUCIÓN PLANTEADA

En esta sección el estudiante debe describir la solución planteada, comenzando por la metodología (pasos que siguió) de desarrollo. Posteriormente, la descripción del desarrollo de cada etapa seguida.

La metodología que debe utilizar el estudiante de máster debe estar validada por la comunidad científica y el estudiante debe justificarlo.

1. METODOLOGÍA
    - Etapa 1
    - Etapa 2
    - Etapa 3
    - Etapa 4
    - Etapa 5
2. DESARROLLO DE CADA ETAPA

Extensión máxima: 15 páginas.

Se recomienda utilizar anexos tras el final de la memoria, únicamente si se estima estrictamente necesario incluir una gran cantidad de material gráfico o tabular.

Escribe aquí…

Voy a empezar este capítulo acordando la lista de limitaciones y metas del proyecto.

Mi tarea resultó ser una tarea no trivial: se propuso analizar respuestas de encuestados en el texto escrito, contestando a preguntas abiertas para extraer los intereses, problemas, deseos y miedos del auditorio e innovar, además de la finalidad de usar las frases usadas por el auditorio en las campañas de marketing.

Las características de los textos normalmente son:

- faltas de ortografía y puntuación son muy habituales
- lenguaje coloquial
- idioma español latino
- comentarios cortos
- pueden llegar contener varias frases claves importantes en un comentario/post

En este trabajo se utilizó uno de los archivos provistos por la empresa y se eligió una columna con respuestas separadas por filas. En total fueron 1495 celdas correspondientes a comentarios de los encuestados. La pregunta a la que contestaban en esta columna era: **"*Si estuvieras compartiendo un café con un terapeuta angelical, ¿Cuál sería la pregunta más importantes que le pedirías que te respondiera?*"** (se conserva la gramática original).

El interés especial de la empresa en el análisis estaba en:

- Procesar gran cantidad de respuestas sin gastos de recursos adicionales
- Resumir las ideas principales.
- Conseguir ideas nuevas
- Comunicarse con sus clientes en su lenguaje

Considerando estas peculiaridades se ha formado una serie de condiciones para cumplir este proyecto. Los cuales se identificaron como siguientes:

- Modelo que pueda trabajar con textos en español latino, lenguaje coloquial, etc.
- Modelo debe ser el más avanzado y potente posible para conseguir resultados que tengan sentido.
- Debe ser posible descargar y ejecutar en el Google collab gratuito (CPU/TPU/ T4 GPU).
- El algoritmo debe mantener el estilo del lenguaje del texto original en la medida de lo posible.

- No puede haber anotación humana
- Debe extraer una lista de frases claves, no un párrafo de texto generalizado, ni palabras claves.
- Debe ser relativamente fácil en ejecución debido a la alta complejidad y variedad de las opciones de biblioteca, modelos, librerías, pipelines, versiones disponibles actualmente.

METODOLOGÍA Y PASOS

Al investigar varios modelos y algoritmos para resumen extractivo he decidido iniciar de los más sencillos. Me propuse experimentar utilizando siguiente metodología:

1. **Embedding** tipo encoder BERT (explicado al inicio en state-of-art apartado) para vectorización de texto.

2.**Reducción de dimensionalidad** con UMAP.

3. Pasar un algoritmo de **clusterización** (K-means y DBSCAN).

4.**Extraer las frases más representativas** de cada clúster, usando KeyBertInspired (algoritmo inspirado en KeyBert) basado en la similitud de coseno entre las frases claves candidatas y todas las frases de su clúster.

5. Evaluación

Debido a la ausencia de "gold standard" de resúmenes, Miller (2019) usaron el feedback humano para la medición de la calidad de los resúmenes extractivos. Yo tampoco tengo el "gold standard", por tanto, de manera similar utilizaré **feedback humano** para orientarme durante el desarrollo o a mi opinión subjetiva al final - al feedback de la empresa que originó los datos y sería el consumidor principal del resultado de este proyecto después de unas mejoras

Baselines

Decidí utilizar uno de los algoritmos más simples de extracción de frases claves, que es **RAKE** un modelo que extrae frases claves de un masivo textual con un score representatividad asignado. El último servirá de un tipo de baseline bajo. Un ejemplo de aplicación ha dado Barrera, M. C. (2018) en el artículo Aplicación del algoritmo RAKE en la indización de documentos digitales.

RAKE (Rapid Automatic Keyword Extraction) es una técnica a base de grafo o, como lo clasifica Sharma y Li (2019b), un algoritmo estadístico, creo que depende tan solo de la interpretación preferida, calcula la coocurrencia de palabras. La ventaja de este algoritmo es que atribuye un score a las frases claves del texto, permitiendo rankearlas en relación con que tanto representan el documento y que tan interconectados son. Además, esta técnica favorece a las frases claves largas, debido a que en la última etapa se suman las veces (normalizadas) de coocurrencia de palabras juntas.

Los resultados de **ChatGPT 3,5** (disponible en https://chat.openai.com/ (*ChatGPT*, s. f.)) - servirán de una orientativa de alta calidad en cuanto a la comprensión humana y generalización.

Definiciones

Un poco de definiciones que no han sonado antes.

UMAP (sus siglas en inglés de Uniform Manifold Approximation and Projection for Dimension Reduction) es una buena elección para posterior clusterización porque se construye a

partir de un marco teórico basado en la geometría de Riemann y la topología algebraica. Preserva parte de la estructura global, tiene un buen rendimiento en tiempo de ejecución y no tiene restricciones computacionales en la dimensión de incrustación (McInnes & Healy, 2018).

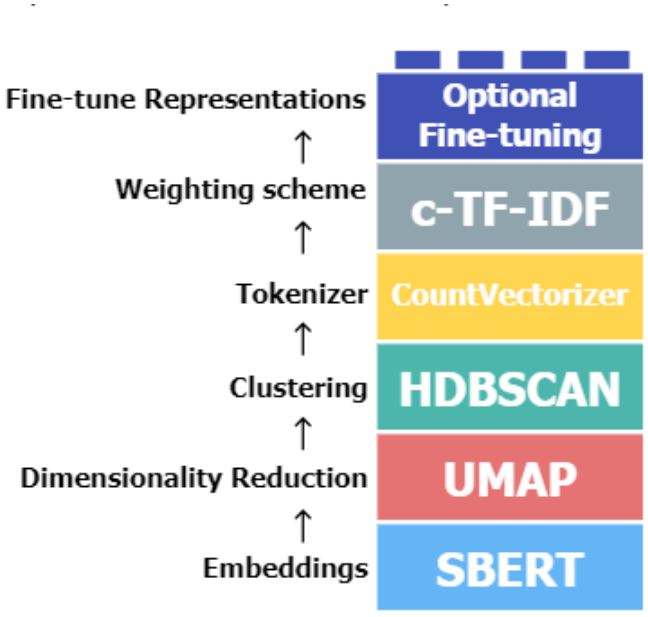
K-means es un algoritmo de agrupamiento que divide un conjunto de datos en k grupos, donde k es un número predefinido por el usuario(por esto conviene bien para uso en este proyecto - permitirá elegir el grado de granularidad/generalización de temáticas). Clusterizar las observaciones a base la distancia euclidiana a los centroides de los grupos y luego recalculando los centroides.

DBSCAN Funciona encontrando regiones densas de puntos cercanos en el espacio de características. DBSCAN es bueno para identificar grupos de datos de diferentes tamaños y formas en datos no lineales y puede detectar puntos atípicos o ruidos en el conjunto de datos. Como se basa en densidad de observaciones, es particularmente útil cuando los grupos pueden tener formas irregulares y tamaños variables.

KeyBert es una técnica que extrae los embeddings de documentos con el modelo BERT para obtener una representación a nivel de documento (en nuestro caso - a nivel de cluster). Luego, se extraen los embeddings de palabras a nivel de palabras/frases de N-gramas. Finalmente, utilizamos la similitud del coseno para encontrar las palabras/frases que son más similares al documento. Las palabras más similares a sus clusters respectivos podrían entonces identificarse como las palabras que mejor describen el documento completo. Descrito e implementado en la librería [GitHub - MaartenGr/KeyBERT: Minimal keyword extraction with BERT] (MaartenGr, n.d.)

HERRAMIENTAS

El mismo autor de la librería KeyBert, Maarten Grootendorst ofreció otra librería que básicamente es un pipeline que une modelos y algoritmos que podrían ser los descritos anteriormente como elegidos para este proyecto (MaartenGr, n.d.).



EJECUCIÓN

Embedding

A lo largo de trabajo experimenté ejecutando el embudo con dos modelos de embedding a base de encoders:

xlm-r-bert-base-nli-stsb-mean-tokens
(*Sentence-transformers/Xlm-r-100langs-bert-base-nli-stsb-mean-tokens · Hugging Face*, n.d.)

y sentence-transformers/paraphrase-multilingual-mpnet-base-v2
(*Sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 · Hugging Face*, n.d.)

El último de los dos, como así lo compartió un speaker de IBM en uno de sus eventos públicos que había atendido, se usó para producir embeddings en un proyecto comercial de servicios de IA chatbot realizados para una de las instituciones públicas de España. (ver tabla de comparación del portal sbert).

Secuencia de pasos que une BERTopic (MaartenGr, n.d.):

Model Name	Performance Sentence Embeddings (14 Datasets)	Performance Semantic Search (6 Datasets)	Avg. Performance	Speed	Model Size
all-mpnet-base-v2	69.57	57.02	63.30	2800	420 MB
multi-qa-mpnet-base-dot-v1	66.76	57.60	62.18	2800	420 MB
all-distilroberta-v1	68.73	50.94	59.84	4000	290 MB
all-MiniLM-L12-v2	68.70	50.82	59.76	7500	120 MB
multi-qa-distilbert-cos-v1	65.98	52.83	59.41	4000	250 MB
all-MiniLM-L6-v2	68.06	49.54	58.80	14200	80 MB
multi-qa-MiniLM-L6-cos-v1	64.33	51.83	58.08	14200	80 MB
paraphrase-multilingual-mpnet-base-v2	65.83	41.68	53.75	2500	970 MB
paraphrase-albert-small-v2	64.46	40.04	52.25	5000	43 MB
paraphrase-multilingual-MiniLM-L12-v2	64.25	39.19	51.72	7500	420 MB

paraphrase-multilingual-MiniLM-L12-v2

Base Model:	Teacher: paraphrase-MiniLM-L12-v2; Student: microsoft/Multilingual-MiniLM-L12-H384
Max Sequence Length:	128
Dimensions:	384
Normalized Embeddings:	false
Suitable Score Functions:	cosine-similarity (util.cos_sim)

Tabla de benchmark de sbert.net de modelos preentrenados para embeddings de oraciones y rendimiento en búsqueda semántica.

El primer modelo al final lo descarté debido a que me encontré una advertencia en Hugging Fase Hub de que este modelo tenía un problema en el tokenizador y desconocía si se había arreglado o no.

Sentence-BERT **(SBERT)** es una modificación de la red preentrenada BERT que obtiene embedding de oraciones, captando parte su sentido semántico y luego pueden ser comparadas utilizando la similitud de coseno (Reimers & Gurevych, 2019).

Clusterización

Dentro de cada modelo usado experimenté con los dos algoritmos de clusterización: K-means y HDBSCAN.

En K-means cambiaba la cantidad de clústeres, dejando el modelo con el último valor de k que parecía funcionar mejor y recular al final en frases claves que mejor reflejaban los comentarios representativos del cluster.

El HDBSCAN produce clusters con temáticas parecidas en la mayoría de los casos, pero no convenía que a veces a una tercera parte de la base de datos como outliers, atribuyéndole un cluster con el número -1. Por esta razón y porque da posibilidad de seleccionar el número de clusters he elegido el algoritmo K-menas para la etapa de clusterización.

Tokenización

En el modelo de vectorización que tokeniza el input a c-TF-IDF jugaba con cantidad de palabras en n-gramas. He detectado que el rango de 1 a 4 de palabras es el óptimo para formar frases claves representativas entendibles.

StopWords

En un momento dado quite los stop-words en español antes de la tokenización (pero los deje en el embedding inicial, porque contribuyen a mejor generalización del sentido de frases y textos por los encoders). Quitar los stopwords en esta etapa permitió extraer más sentido con la misma cantidad de tokens.

Fine-tuning

Llama2

Intente mejorar los resultados con uno de los modelos generativos de state-of-the-art Llama2 de 7B , entrenado en un corpus en español (github: https://github.com/Garrachonr/LlamaDos Garrachonr, n.d.), inicialmente elaborado por meta (Touvron et al., 2023b).

Lamentablemente aun aplicando una técnica de optimización no he podido descargarla al Google Collab conectado a los recursos gratuitos por falta de los mismos y fue descartada.

RAKE on top

En búsqueda de reducir las temáticas y rankearlas de alguna manera, use el algoritmo RAKE arriba del conjunto de frases claves encontradas en la última etapa del embudo Bertopic con KeyBertInspired.

Aun si se quitan las frases claves que se alejan por el score, no logramos hacer más comprensible esta lista resultante para el ser humano. Por tanto, se optó por mostrar los N top-scored frases, siendo N un número que elegiría el usuario del algoritmo, dependiendo de qué tan detallado quiere estudiar el tema.

Baselines

RAKE (Rapid Automatic Keyword Extraction) . RAKE calcula la coocurrencia de palabras de siguiente manera: tokeniza las frases por signos de puntuación y las stopwords, luego calcula la frecuencia de palabras, “degree of word” (sumas de cuantas veces cada palabra ha ocurrido con cada otra), luego normaliza dividiendo este valor a la frecuencia de ocurrencia general. Al final a cada frase clave se le asigna un score que es la suma de degree of word normalizado de cada palabra que

la compone. RAKE favorece, como se vio en práctica, a las frases claves largas, debido a que en la última etapa se suman las veces (normalizadas) de coocurrencia de palabras juntas.

Mientras hacía mi estudio de state-of-the art modelos y trabajos, me di cuenta de que en realidad hay muy buenos modelos de licencia libre, pero la mayoría no trabaja con texto en español. Y como los autores del artículo RigoBERTa: A State-of-the-Art Language Model For Spanish Vaca et al. (2022) mencionan, a pesar de que el español es una de las lenguas más habladas, hay pocos modelos de lenguaje para este idioma.

EVALUACIÓN

El estudiante en esta sección debe describir cómo realizó la evaluación de la solución planteada.

La evaluación de un Global Project debe basarse en una metodología validada y bien establecida que el estudiante deberá justificar.

Extensión máxima: 6 páginas.

Se recomienda utilizar anexos tras el final de la memoria, únicamente si se estima estrictamente necesario incluir una gran cantidad de material gráfico o tabular.

Escribe aquí…

Debido a la ausencia de “gold standard” de resúmenes, se usó el feedback humano para la medición de la calidad de los resúmenes extractivos(similar a D. Miller (2019) en Leveraging BERT for Extractive Text Summarization on Lectures).

Durante el desarrollo me apoyo con propia opinión subjetiva y al final del trabajo - al feedback de la empresa que originó los datos y sería el consumidor principal del resultado de este proyecto después de unas mejoras.

Aunque los **criterios de evaluación** son subjetivos y difícilmente cuantificables, es bueno definirlos:

- ☐ Si las frases generadas tienen sentido
- ☐ Si parecen representar ideas claves por cluster
- ☐ Qué tan homogéneos son los clusters en sus temáticas
- ☐ Cuantas temáticas se generó (escala: demasiadas-escasas y escala se entrecruzan a lo largo del documento)
- ☐ Representación semántica de los deseos, problemáticas, necesidades a satisfacer de los encuestados
- ☐ Contenido de temáticas poco frecuentes(ideas originales).

Se compararon los resultados de cuatro metodologías y se evaluó subjetivamente , considerando los criterios descritos arriba.

Las metodologías comparadas:

1. KeyBertInspired + Rake = ['KeyBERTRake']
2. baseline solo Rake = ['soloRake']
3. KeyBertInspired que recibió el texto como una lista sin separación por párrafos entre los comentarios. = ['KeyBERTList']
4. Baseline alto - chatGPT , análisis de los primeros 60 comentarios.

RESULTADOS

El estudiante en esta sección debe describir los resultados obtenidos en el proceso de evaluación de la solución planteada.

Debe usar notación estándar para presentación de resultados de carácter científico.

Extensión máxima: 6 páginas.

Se recomienda utilizar anexos tras el final de la memoria, únicamente si se estima estrictamente necesario incluir una gran cantidad de material gráfico o tabular.

Escribe aquí…

Se han detectado temas principales, que corresponden a tres de las cuatro metas principales propuestas al inicio:

- Se ha logrado analizar el conjunto entero de las respuestas en un documento
- Las frases claves detectadas describen los deseos, problemáticas, “dolores” del auditorio.
- En gran medida reflejan el estilo utilizado por los encuestados.

A la vez, un punto de mejora será reducir la cantidad de temas por cluster y seleccionarlos o rankear de una mejor manera para lograr establecer prioridad y mejorar comprensibilidad.

Los temas resultantes son muy parecidos a los que ha encontrado el chatGPT. Aunque el último devuelve frases que son más fácilmente entendibles para el ser humano

Las metodologías comparadas se mencionan en el apartado anterior.

EJEMPLOS DE RESULTADOS DE CADA METODOLOGÍA

['KeyBERTRake']: se obtuvieron 431 frases rankeadas. Esta combinación de algoritmos ofrece una lista de frases claves ranqueada, donde previamente se ha reducido la cantidad de temas mediante una clusterización. Y, por tanto, ofrece un formato de resultados más parecido a lo que se ha propuesto para conseguir.

```
res_1 #RAKE on top de KeyBertInspired, 431 frase clave
```

1 to 25 of 431 entries Filter

index	score	phrase
165	9.0	ven hablan visualizarlos
16	9.0	plenitud campo laboral
338	9.0	hablando creadores diosa
262	9.0	tantas empezar preguntas
351	9.0	tendria muchas preguntas
260	9.0	humanidad dios sienta
332	9.0	presentes despierten conciencia
51	9.0	controlar empiezo tantas
108	9.0	hijos mio mismo
416	9.0	primeros pasos puedes
253	9.0	humm realmente mos
329	9.0	integral dos hijos
318	9.0	perdoname pensado ello
310	9.0	informacion querido fallecido
66	9.0	pasado traumas pasado
246	9.0	gente conecten comprendan
459	9.0	perdin aprender dejar
261	9.0	gente honesta ayudarles
229	8.9375	queridos ubicarme proposito
69	8.875	obtener verdadero conocimiento
105	8.869565217391305	encomendado vida ayudas
273	8.869565217391305	percibir sucede vida
385	8.869565217391305	vida logras percibir
5	8.869565217391305	realidad continuidad vida
265	8.833333333333334	mundo cuesta sentirme

Baseline solo Rake - ['soloRake'], 3319 frases candidatas, de los cuales fueron seleccionadas solo con el score mayor a 100 (80 frases).

```
res_2 =phrase_df_new2.loc[phrase_df_new2.score>100].sort_values('score', ascending = False)
res_2 #solo RAKE
```

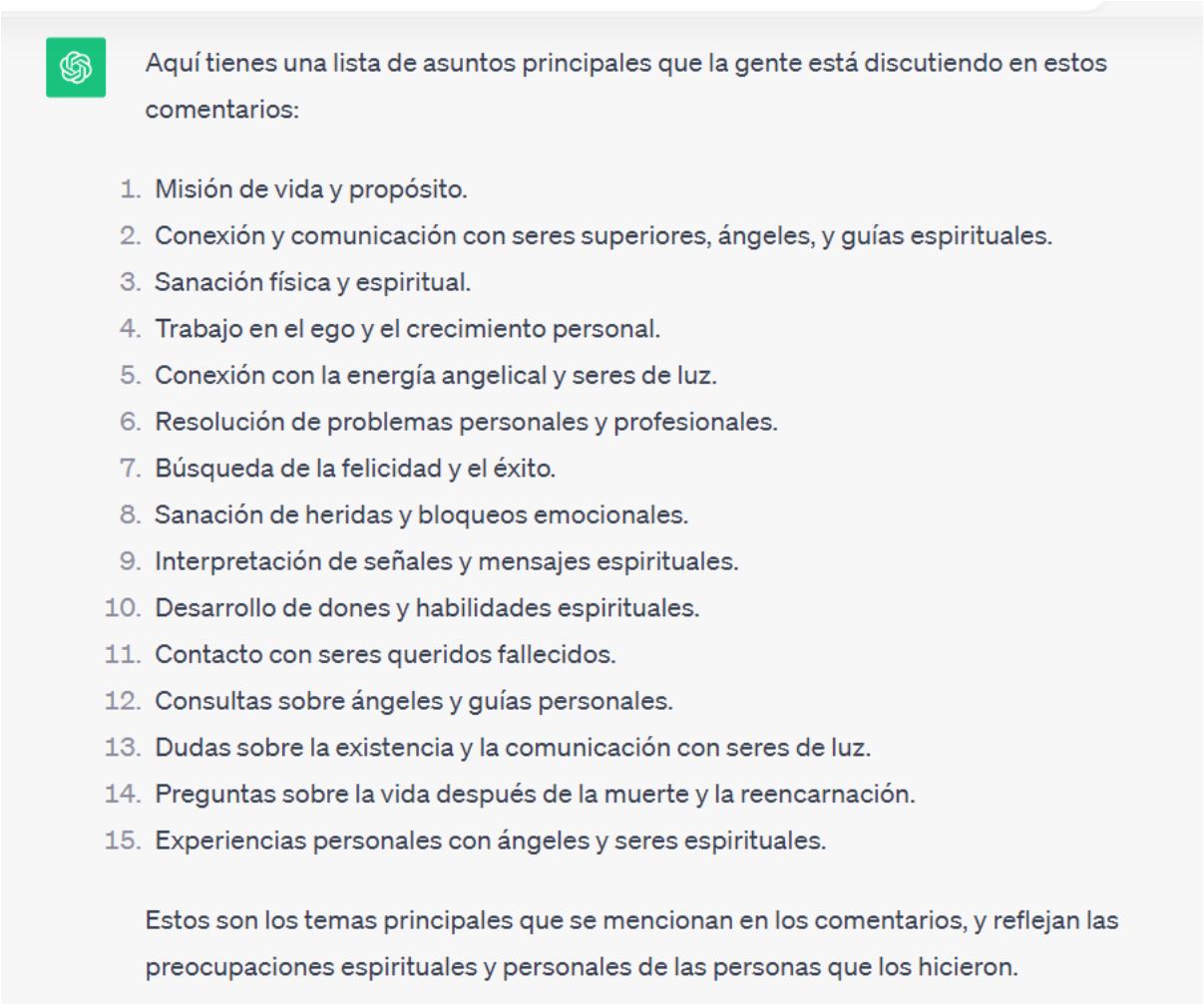
1 to 25 of 80 entries Filter

index	score	phrase
2072	232.70909423638705	como sabes que estas con una energia angelical en un mundo llego de angeles caidos sobre todo es estos tiempos de tantas pruebas
2531	215.9418283468679	en que estoy escaseando preguntaria acerca de mi hogar preguntaria acerca como hacer para que mi nina pequena aprenda de los angeles
1054	205.24718790161688	como minimizar mi ego por sobre todo para que mi vida espiritual sea pura e inspiradora en los mensajes del maestro jesus
2604	198.01670780342323	en cada una de las dimensiones materiales existentes tendremos que estar equilibrados con esas dimensiones para poder evolucionar
1676	196.01870374583447	le preguntaria el porque uno tarda tanto en darse cuenta que los angeles son quienes siempre nos ponen sobre aviso
2053	176.20412587662346	quiero dar consejos aunque siempre termino hablamdo de la respondabilidad del ser humano de lo que le sucede
1506	175.3898650555027	cual es la formula ideal para vivir en plenitud con las personas que amo disfrutando de una gran prosperidad
1693	173.25756850579168	le preguntaria como sanar mi arbol genealogico para cortar con enfermedades que se van presentando en la vida
2459	170.6354823390025	como sanar mis memorias inconscientes heredadas de mis ancestros que provocan problemas fisicos en mi vida presente
1420	166.91580768688837	le preguntaria cual es mi camino del alma si estoy en el correcto pero tambien tengo otra muy importante para mi
502	165.9583372413024	m8 mayor experiencia seria compartir con ella los poderes metafidicos q pofemos akcanzar con esfuerzo unido
2154	164.70048185226645	creo que ese cafe vendria muy bien acompanado con la historia de vida de mi terapeuta en relacion
1274	161.6358895362232	q hacee para q la humanidad cambie todo esto q estamos viviendo q ya parecemos sombies
758	157.33680532169822	cuando lo utilizo para la sanacion se efectivo bueno yo les pediria tantas cosas pero lo importante es que
1248	154.731460281696	si es verdad que dicen que los familiares que fallecen muchas veces se convierten en nuestros acompanantes
3019	152.93127520983572	sera posible que al momento que deba trascender del mundo fisico pueda sentir la presencia de mis angeles
1954	152.78957466803828	cual es la maejor manera para controlar el ego para vivir en paz en esta encarnacion como aprender
802	148.77924674351294	corporal que sea efectiva puntual en corto tiempo donde el amor cure lo mas pronto posible
1550	147.19597206307094	si se puede hacer algo al respecto con lo que ha pactado mi alma antes de nacer
165	145.4902057804945	sano mi karma familiar para poder recibir libremente los mensajes del cielo ya estando sanada
742	145.30967144192786	porque hay tantos que dicen que los angeles es un programa de la matrix para enganarnos
237	142.34680415278805	evolucionar en este mundo donde los seres de luz estan alli para llevarnos de la mano
652	141.92675766864573	le preguntaria como podria entender mas los mensajes por mas que les pido que sean mas sencillos aun

['KeyBERTList'], 50 clusters detectados y para cada cluster hay varias frases claves.

KeyBERTList
sanar otras personas,como sanar mi,como sanar,como lograr sanarme,vida como sanar,sanar como,lograr sanarme,para sanar,poder sanar,de salud como
angeles como comunicarme,los mensajes angelicales,mensajes angelicales,angeles arcangeles como,angelicales como,angeles los seres,angelicales,angeles en que,angeles como,arcangeles como
no se,no se en,no lo se,ocurre no se,lo se no,no se muy,ahora no se,no sabria,no se me,se no se
porque mi vida,mi vida desde,por que estoy,que me pasa,mi vida,por que he,porque todo me,una vida tan,porque he fracasado,mundo me aburro
como escucharlos como,de escucharlos,como escucharlos,me escuchan,me estan escuchando,escucharlos como,estan escuchando,para ver escuchar,escucharlos,escuchan
vida mi mision,mision de vida,es mi mision,mi mision en,proposito de vida,mi mision de,mi proposito,con mi mision,mi mision,es la mision
espiritualmente como puedo,ser mas espiritual,guias espirituales,mis guias espirituales,espiritualmente como,mi guia espiritual,guia espiritual,espiritual que me,espiritualmente,espiritual como
trabajar con energias,poder elevar la,mismo poder elevar,puedo vibrar adecuadamente,hacer para elevar,si podrian trabajar,mi vibracion alta,mi frecuencia vibratoria,mantener mi vibracion,para elevar mi
angeles que me,los angeles como,los angeles que,angeles como,con los angeles,angeles que,angeles si,seres angelicales,mis angeles,angeles
que hay mas,que encarne en,mas alla para,mas bien de,que encarne,para que encarne,si es con,mas alla,se,son mas
porque los angeles,los angeles que,que los angeles,angeles que,los angeles son,angeles,les pidamos,que les pidamos,los arcangeles,arcangeles
mi angel guardian,angel guardian como,mi angel como,sobre mi angel,es mi angel,mi arcangel,mi angel de,angel quien es,de mi angel,llama mi angel
seres de luz,ser de luz,de luz como,luz como,luz como es,la luz,luz como puedo,de luz,de luz porque,luz
comunicarme con ellos,contactar con ellos,contactarlos,comunicarme con,hablar con ellos,manera de comunicarme,contactar con,comunicarme,me comunico como,como me comunico
mis hijos por,mis hijos sobre,padres,nietos lograre ayudar,mis hijos,ninos sus hijos,mi familia,familia,sus hijos,mis padres como
saber como empezo,saber como comenzo,realizar esos trabajos,para realizar esos,lograrlo como,lograrlo como

Resultados obtenidos por el **ChatGPT** a partir de los primeros 60 comentarios de la base de datos.



Aquí tienes una lista de asuntos principales que la gente está discutiendo en estos comentarios:

1. Misión de vida y propósito.
2. Conexión y comunicación con seres superiores, ángeles, y guías espirituales.
3. Sanación física y espiritual.
4. Trabajo en el ego y el crecimiento personal.
5. Conexión con la energía angelical y seres de luz.
6. Resolución de problemas personales y profesionales.
7. Búsqueda de la felicidad y el éxito.
8. Sanación de heridas y bloqueos emocionales.
9. Interpretación de señales y mensajes espirituales.
10. Desarrollo de dones y habilidades espirituales.
11. Contacto con seres queridos fallecidos.
12. Consultas sobre ángeles y guías personales.
13. Dudas sobre la existencia y la comunicación con seres de luz.
14. Preguntas sobre la vida después de la muerte y la reencarnación.
15. Experiencias personales con ángeles y seres espirituales.

Estos son los temas principales que se mencionan en los comentarios, y reflejan las preocupaciones espirituales y personales de las personas que los hicieron.

A juicio subjetivo, son bastante parecidos a los que ha detectado el algoritmo usado en el proyecto, pero está formulado con un lenguaje más comprensible. Aunque la desventaja de no saber la escala de la distorsión permanece.

Igual para el futuro se prevé selección y uso de métricas cuantitativas para identificar el grado de similitud de los resultados y poder compararlos (por ejemplo, un embudo: embedding de frases + reducción de dimensionalidad + similitud de coseno).

Mientras tanto, de todas las combinaciones de algoritmos que se han utilizado en los experimentos dentro del ambiente Google Collab, con la de **['KeyBERTRake']** se consiguió el formato más próximo al esperado: una lista de frases representativas de las temáticas mencionadas en los comentarios. El potencial usuario del algoritmo podría elegir si ver solo las frases más frecuentemente utilizadas o las más raras, que pudieran ofrecer nuevas ideas para desarrollo de productos para el auditorio. En ambos casos van a ser frases que representan algunas temáticas expresadas por los encuestados, consecuencia de utilizar no solo Rake, sino Rake arriba de embedding de oraciones, clusterización, extracción de frases, que destacan cada cluster, selección de las mismas por su similitud con los comentarios en cada cluster.

CONCLUSIONES Y TRABAJOS FUTUROS

El estudiante plantea las conclusiones de su trabajo, y cómo considera que puede seguir avanzando en el mejoramiento de la solución planteada.

Extensión máxima: 1 página

Escribe aquí…

La meta de este trabajo resultó no ser fácil, considerando las limitaciones , como se ha detallado en la introducción.

La dificultad de la meta propuesta consiste en una necesidad práctica del negocio de analizar las fuentes textuales para resumir las ideas contenidas, como también captar el estilo original del auditorio, extraer insights.

Sin embargo, en este trabajo **se ha logrado** analizar de manera automática el conjunto entero de las respuestas en un documento, detectados los deseos, problemáticas, "dolores" del auditorio. Las frases extraídas podrían ser consideradas como ejemplos de expresiones de los encuestados. Estos resultados se acercan a las metas que se han propuesto al inicio del proyecto.

Como se explicó más detalladamente en el apartado anterior, de todas las combinaciones de algoritmos que se han utilizado en los experimentos, la de **embedding de oraciones + clusterización + extracción de frases claves por cluster + selección de las más similares con los comentarios en cada cluster + ranking con RAKE** se consiguió el formato más próximo al esperado: una lista de frases representativas rankeadas por frecuencia de uso, que representan las temáticas mencionadas en los comentarios.

Por otro lado, se prevén varias **posibilidades de mejorar,** por ejemplo vía corrección de errores avanzada, cambio de fases del embudo utilizado, como por ejemplo, procesamiento de los comentarios con el algoritmo Rake antes de pasar la misma combinación de algoritmos, uso de los transformers sec2sec en las últimas etapas del embudo para mejorar las frases detectadas, además de la utilización de los modelos de deep learning basados en técnicas distintas(NER, POS, grafos etc.).

Uno de los otros puntos que se podría mejorar fácilmente sería la ejecución de algoritmos usados en mis experimentos c-TF-IDF y KeyBertInspired. Estos se enfocan en extraer temáticas diferentes de cada cluster detectado, algo que me aleja de la meta de tratar todos los comentarios como un texto íntegro para poder manejar frases similares de cada cluster de mejor manera y no repetirlas. Jugar con la cantidad de clusters no ayuda a resolver este problema, porque en el caso de disminución de su cantidad, el algoritmo empieza a generalizar y perder temáticas. Para el futuro preveo experimentar realizando un parcing de oraciones y crear una lista de elementos donde cada oración sería un elemento aparte, lo que puede ayudar a una mejor clusterización. Y luego, como una etapa crucial para el experimento, sería tratar encontrar los mejores hiperparametros para la fase de selección de las frases representativas de cada cluster con c-TF-IDF y KeyBertInspired usados en la mayoría de mis experimentos en este trabajo.

También se plantea la formalización de medición cuantitativa de resultados para tener puntos referencia firmes a la hora de desarrollo e investigación. En este sentido se prevé la selección y uso de métricas cuantitativas para identificar el grado de similitud de los resultados y poder compararlos.

REFERENCIAS

Lista de referencias bibliográficas utilizadas, siguiendo las normas APA.

Referencias

Álvarez, C. & Pontificia Universidad Católica de Chile. Escuela de Ingeniería. (2021). *APLICACION DE PROCESAMIENTO DE LENGUAJE NATURAL SOBRE UNA ENCUESTA DE SATISFACCION* [Tesis de maestría, Pontificia Universidad Católica de Chile]. https://doi.org/10.7764/tesisUC/ING/62782

Araujo, V., Trusca, M., Tufiño, R., Moens, M., Leuven, K. (2023) Sequence-to-Sequence Spanish Pre-trained Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2309.11259

Andreotta, M., Nugroho, R., Hurlstone, M. J., Boschetti, F., Farrell, S., Walker, I., & Paris, C. (2019). Analyzing Social Media Data: A mixed-methods framework combining computational and qualitative text analysis. *Behavior Research Methods*, *51*(4), 1766-1781. https://doi.org/10.3758/s13428-019-01202-8

Báez, P., Villena, F., Rojas, M., Durán, M., & Dunstan, J. (2020). The Chilean Waiting List Corpus: a new resource for clinical Named Entity Recognition in Spanish. *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. https://doi.org/10.18653/v1/2020.clinicalnlp-1.32

Barrera, M. C. (2018). Aplicación del algoritmo RAKE en la indización de documentos digitales. *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*. https://doi.org/10.22201/iibi.24488321xe.2018.75.57951

ChatGPT. (s. f.). chat.openai. https://chat.openai.com/

Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv (Cornell University)*. https://arxiv.org/pdf/2003.10555.pdf

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv (Cornell University)*. https://arxiv.org/pdf/1810.04805v2

Fine-tuning BERT with sequences longer than 512 tokens. (2021, 9 diciembre). Hugging Face Forums. https://discuss.huggingface.co/t/fine-tuning-bert-with-sequences-longer-than-512-tokens/12652

Garrachonr. (s. f.-a). *GitHub - Garrachonr/LlamaDOs: Finetuning of a LLAMA2-7B to give it the ability of having a fluent conversation in Spanish.* GitHub. https://github.com/Garrachonr/LlamaDos

Garrachonr. (s. f.-b). *GitHub - Garrachonr/LlamaDos: Finetuning of a Llama2-7B to give it the ability of having a fluent conversation in Spanish.* GitHub. https://github.com/Garrachonr/LlamaDos

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2203.05794

He, P., Liu, X., Gao, J., & Chen, W. (2021). DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. *International Conference on Learning Representations*. https://openreview.net/pdf?id=XPZIaotutsD

Jia, R., Cao, Y., Tang, H., Fang, F., Cao, C., & Wang, S. (2020a). Neural Extractive Summarization with Hierarchical Attentive Heterogeneous Graph Network. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).Association for Computational Linguistics*, 291-300. https://doi.org/10.18653/v1/2020.emnlp-main.295

Jia, R., Cao, Y., Tang, H., Fang, F., Cao, C., & Wang, S. (2020b). Neural Extractive Summarization with Hierarchical Attentive Heterogeneous Graph Network. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics*, 3622-3631. https://doi.org/10.18653/v1/2020.emnlp-main.295

MaartenGr. (s. f.-a). *GitHub - MaartenGr/BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.* GitHub. https://github.com/MaartenGr/BERTopic

MaartenGr. (s. f.-b). *GitHub - MaartenGr/KeyBERT: Minimal keyword extraction with BERT*. GitHub. https://github.com/MaartenGr/KeyBERT

McInnes, L., & Healy, J. J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv (Cornell University)*. http://export.arxiv.org/pdf/1802.03426

Miller, D. M. (2019). Leveraging BERT for Extractive Text Summarization on Lectures. *arXiv (Cornell University)*. https://arxiv.org/pdf/1906.04165

Mrbullwinkle. (2023, 20 julio). *Cómo trabajar con los modelos GPT-35-Turbo y GPT-4 - Azure OpenAI Service*. Microsoft Learn. https://learn.microsoft.com/es-es/azure/ai-services/openai/how-to/chatgpt

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. *arXiv (Cornell University)*. https://arxiv.org/pdf/1908.10084.pdf

sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 · Hugging Face. (s. f.). https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens · Hugging Face. (s. f.). https://huggingface.co/sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens

Sharma, P., & Li, Y. (2019). Self-Supervised Contextual Keyword and Keyphrase Retrieval with Self-Labelling. *Preprints*. https://doi.org/10.20944/preprints201908.0073.v1

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, É., & Lample, G. (2023a). LLAMA: Open and Efficient Foundation Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.13971

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, É., & Lample, G. (2023b). LLAMA: Open and Efficient Foundation Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2302.13971

Vaca, A., Barbero, A., Guerrero, M., Moreno, A., Betancur, D., Samy, D., Aldama Garcia, N., Montoro Zamorano, H., & Garcia Subies, G. (2022). RigoBERTa: A State-of-the-Art Language Model For Spanish. *arXiv (Cornell University)*. https://doi.org/10.48550/arXiv.2205.10233

Wei, Y., & Ding, Y. (2023). Application of Text Rank Algorithm Fused With LDA in Information Extraction Model. *IEEE Access*, *11*, 84301-84312. https://doi.org/10.1109/access.2023.3296141