PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

Elaboración de un Sistema de Recomendación de Publicaciones Científicas Nacionales de Acceso Abierto para los investigadores calificados del SINACYT

Tesis Para optar por el Título de Ingeniera Informática que presenta la bachillera:

Elizabeth Jenisse Vereau Zagastizábal

Asesor: Mg. César Augusto Olivares Poggi

Lima, junio de 2018

A mis padres y hermana por su apoyo incondicional.

A Gladys, Gilmer y Blanca por depositar su confianza en mi.

Agradecimientos

A mi asesor, por su apoyo y orientación brindados en el desarrollo de mi tesis.

A todos los profesores que contribuyeron a mi formación académica.

Resumen

Actualmente existe un crecimiento sostenido sobre la producción científica mundial. Esta producción científica es preservada a través de repositorios de acceso abierto digitales, los cuales se crean como herramientas de apoyo para el desarrollo de producción científica. Sin embargo, existen deficiencias en la funcionalidad de los mismos como herramientas de apoyo para el aumento de la visibilidad, uso e impacto de la producción científica que albergan.

El Perú, no es ajeno al crecimiento de la producción científica mundial. Con el avance del mismo, se implementaron nuevas plataformas (ALICIA y DINA) de difusión y promoción del intercambio de información entre las distintas instituciones y universidades locales. No obstante, estas plataformas se muestran como plataformas aisladas dentro del sistema científico-investigador, ya que no se encuentran integradas con las herramientas y procesos de los investigadores.

El objetivo de este Proyecto es el de presentar una alternativa de solución para la resolución del problema de carencia de mecanismos adecuados para la visualización de la producción científica peruana a través de la implementación de un Sistema de Recomendación de Publicaciones Científicas Nacionales de Acceso Abierto para los investigadores calificados del SINACYT.

Esta alternativa se basa en la generación de recomendaciones personalizadas de publicaciones en ALICIA, a través del uso del filtrado basado en contenido tomando en cuenta un perfil de investigador. Este perfil se construyó a partir de la información relevante sobre su producción científica publicada en Scopus y Orcid. La generación de recomendaciones se basó en la técnica de LSA (Latent Semantic Analysis), para descubrir estructuras semánticas escondidas sobre un conjunto de publicaciones científicas, y la técnica de Similitud Coseno, para encontrar aquellas publicaciones científicas con el mayor nivel de similitud.

Para el Proyecto, se implementaron los módulos de extracción, en donde se recoge la data de las publicaciones en ALICIA y las publicaciones en Scopus y Orcid para cada uno de los investigadores registrados en DINA a través de la técnica de extracción de datos de sitios web (web scrapping); de pre procesamiento, en donde se busca la mejora de la calidad de la data previamente extraída para su posterior uso en el modelo analítico dentro del marco de la minería de texto; de recomendación, en donde se capacita un modelo LSA y se generan recomendaciones sobre qué publicaciones científicas pueden interesar a los usuarios basado en sus publicaciones científicas en Scopus y Orcid; y de servicio, en donde se permite a otras aplicaciones consumir las recomendaciones generadas por el sistema.

Palabras Clave

Sistema de Recomendación; Producción Científica; Repositorios de Acceso Abierto, DINA; ALICIA; LSA; Similitud Coseno; Filtrado Basado en Contenido

Tema FCI

Tabla de Contenido

Índice de Ilustraciones

Índice de Tablas

Índice de Algoritmos

Capítulo 1.Generalidades

1.1 Problemática

La investigación es “un proceso que, mediante la aplicación del método científico, procura obtener información relevante y fidedigna, para entender, verificar, corregir y/o aplicar el conocimiento” (Tamayo, 2004). Esta investigación deriva en la producción de literatura científica, la cual es clave en el proceso de desarrollo tecnológico, económico y social de una nación (Tamayo, 2004).

La producción científica es considerada como “la parte materializada del conocimiento generado, es más que un conjunto de documentos almacenados en una institución de información. Se considera también que contempla todas las actividades académicas y científicas de un investigador” («La producción científica», 2013). Constituyen como producción científica publicaciones de carácter científico; entre las que se destacan tesis, artículos, libros, reportes, revistas indizadas nacionales, datos de investigación, entre otros («La producción científica», 2013).

En los últimos años, la producción científica mundial se ha expandido notablemente. De acuerdo al SCImago Journal & Country Rank (SJR), entre 1996 y 2016, el volumen total de documentos científicos considerados son de 44.7 millones («Scimago Journal & Country Rank», 2017). Para la región de América Latina, este volumen total es de 1.5 millones, teniendo a Brasil como mayor productor de literatura científica de la región, y a Perú representando solo el 1.1% del total («Scimago Journal & Country Rank», 2017).

Hablando estrictamente del Perú, el SCImago Journal & Country Rank (SJR) muestra el incremento notable de la producción científica del país con 164 documentos indexados en 1996 a más de 17000 en 2016 («Scimago Journal & Country Rank», 2017). Todos los países producen más literatura científica que hace veinte años («Scimago Journal & Country Rank», 2017). Sin embargo, el índice de crecimiento es diferente para cada uno de ellos («Scimago Journal & Country Rank», 2017). Así, se expone que el promedio regional para América Latina del índice de crecimiento de la producción científica en el 2012 fue de 4.7 (4.7 veces la producción de 1996) («Scimago Journal & Country Rank», 2017). En el caso peruano, este índice fue de 7.3, cifra que se encuentra por encima del promedio regional y lo posiciona como el segundo país con mayor índice de crecimiento después de Colombia («Scimago Journal & Country Rank», 2007).

De esta manera, el continuo crecimiento de la producción científica en la región, supone también el uso de la tecnología en materia de colección y preservación de la producción científica (María Inés Bravo, Ken Norsworthy, & Paula Pardo Lorca, 2004). Esto se explica en la aparición de dos tendencias claves: la “tremenda expansión en la cantidad

de información y análisis producida y ‘publicada’, sobre todo electrónicamente” (María Inés Bravo et al., 2004), además de la “demanda creciente dentro de la región por la información publicada por otros investigadores latinoamericanos” (María Inés Bravo et al., 2004).

En este contexto surgen los repositorios de acceso abierto. El movimiento de acceso abierto cuenta en el mundo con una amplia comunidad, que trata de compartir todo el acervo científico de las organizaciones (María Inés Bravo et al., 2004). Sin embargo, existen deficiencias en la funcionalidad de los mismos como herramienta de apoyo al desarrollo de investigaciones científicas (María Inés Bravo et al., 2004).

En América Latina, iniciativas como Dspace, SciELO, E-Prints y Open Journal System son los más utilizados por la comunidad científica como herramienta para el aporte en el desarrollo de investigaciones (Córdoba, 2011). Así mismo, en la región, existe LA Referencia (La Red Federada de Repositorios Institucionales de Publicaciones Científicas), red latinoamericana de repositorios de acceso abierto, de la cual el Perú es un país miembro con el repositorio nacional ALICIA (Acceso Libre de la Información de Ciencia y tecnología). ALICIA contribuye a La Referencia, albergando cerca de 100,000 documentos científicos («ALICIA», 2017).

Los principales objetivos que abordan estas iniciativas son la de promover la preservación digital («DSpace: un manual específico para gestores de la información y la documentación», s. f.), y la de aumentar la visibilidad, uso e impacto de las publicaciones científicas que albergan (Packer, Cop, Luccisano, Ramalho, & Spinak, 2014). Bajo estos mismos lineamientos es que nacen ALICIA y DINA. ALICIA como plataforma de difusión y promoción del intercambio de información entre las instituciones y universidades del Perú en distintas áreas («ALICIA», 2017), esto mediante la integración de repositorios nacionales que contienen publicaciones científicas para la generación de nuevo conocimiento (Atamari-Anahui & Díaz-Vélez, 2015), y DINA como plataforma virtual del Directorio Nacional de Investigadores e Innovadores con el objetivo de “dar visibilidad a la labor de los investigadores e innovadores peruanos, así como a vincularlos con sus pares para que puedan generar múltiples oportunidades de potenciar sus redes de colaboración” («CONCYTEC pone a disposición nueva plataforma virtual DINA para investigadores, innovadores y profesionales», s. f.).

Sin embargo, actualmente, estas iniciativas se muestran como plataformas aisladas dentro del sistema científico-investigador, ya que no se encuentran integradas con las herramientas y procesos de los investigadores a partir de la falta de integración de los repositorios de acceso abierto dentro de lo que concierne la labor investigadora (Lorenzo Gil, Braña Ferreiro, & Nieto Caramés, 2015).

En el caso peruano, las plataformas DINA y ALICIA se muestran aisladas y no contribuyen de manera eficiente en cumplir con los objetivos para los que fueron creados. La falta de interoperabilidad entre estas plataformas no facilita el cumplimiento del objetivo de creación y fortalecimiento de redes de colaboración entre los investigadores nacionales, representado en la dificultad para la agrupación natural de investigadores con temas afines en las distintas instituciones/regiones del Perú.

Por otro lado, la visibilidad se establece como una característica importante de la producción científica almacenada en los repositorios de acceso abierto y se encuentra fuertemente relacionado con la interoperabilidad (Ferreras-Fernández, García-Peñalvo, & Merlo-Vega, 2015).

En el caso peruano, la falta de mecanismos de integración para la visualización de producción científica de ALICIA para usuarios de DINA, es una de las causas para el desconocimiento por parte de los investigadores peruanos, sobre la producción científica nacional existentes. La existencia de estos mecanismos promovería la obtención de información sobre aquellos investigadores con líneas de investigación afines, así como el conocimiento las variantes de temas afines a los investigadores.

Por último, ante todo lo anteriormente expuesto, el objetivo de este proyecto de fin de carrera es proponer una alternativa de solución para la resolución del problema de carencia de mecanismos adecuados para la visualización de la producción científica de acceso abierto para los investigadores calificados del SINACYT.

1.2 Objetivos

1.2.1 Objetivo general

Implementar un sistema de recomendación que permita sugerir publicaciones científicas relevantes del Repositorio Nacional de Acceso Abierto (ALICIA) para los investigadores calificados del Sistema Nacional de Ciencia y Tecnología (SINACYT) a partir de su producción científica en Scopus y Orcid.

1.2.2 Objetivos específicos

O 1. Extraer, pre-procesar y modelar de los datos correspondientes a los investigadores calificados del SINACYT, así como también para las publicaciones científicas almacenadas en ALICIA.

O 2. Desarrollar un modelo de recomendación basado en los enfoques investigados.

O 3. Implementar un servicio web para la publicación de las recomendaciones generadas por el sistema de recomendación construido a partir de los componentes de arquitectura definidos.

O 4. Evaluar las recomendaciones generadas bajo la técnica de evaluación offline utilizando la métrica de Precisión en N.

1.2.3 Resultados esperados

R 1. Modelo de datos que representa el perfil del usuario (investigadores calificados del SINACYT) y provee conocimiento de los intereses del mismo (O1).

R 2. Modelo de datos que representa el elemento a recomendar (publicaciones científicas almacenados en ALICIA) (O1).

R 3. Definición de la arquitectura del sistema de recomendación (O1).

R 4. Módulo relacionado a la Extracción de Datos de las plataformas web de ALICIA y DINA (O1).

R 5. Módulo relacionado al Pre procesamiento de Datos (O1).

R 6. Datos normalizados de las publicaciones científicas de ALICIA y de los investigadores calificados de DINA (O1)

R 7. GUI para la interacción con los módulos del sistema (O1).

R 8. Definición de los enfoques de recomendación que se implementaran (O2).

R 9. Módulo relacionado al Modelo de Recomendación (O2).

R 10. Definición de la arquitectura para el servicio web (O3).

R 11. Integración de los componentes de arquitectura (03)

R 12. Servicio web REST para la trasferencia de las recomendaciones obtenidas hacia una plataforma de visualización (O3).

R 13. Cuestionario para la obtención de información sobre la relevancia de las recomendaciones generadas

R 14. Definición de las métricas que deben satisfacerse en la evaluación del rendimiento del sistema de recomendación (O4).

R 15. Evaluación offline de los resultados obtenidos utilizando la métrica de Precisión en N(O4).

1.2.4 Mapeo de objetivos, resultados y verificación

En la Tabla 1 se muestra el mapeo de objetivos, resultados y verificación.

Objetivo: (O1) Extraer, pre-procesar y modelar de los datos correspondientes a los investigadores calificados del SINACYT, así como también para las publicaciones científicas almacenadas en ALICIA.

Resultado	**Meta física**	**Medio de verificación**
(R1) Modelo de datos que representa el perfil del usuario (investigadores calificados del SINACYT) y provee conocimiento de los intereses del mismo	Documento	- Diagrama de Clases
(R2) Modelo de datos que representa el elemento a recomendar (publicaciones científicas almacenados en ALICIA)	Documento	- Diagrama de Clases
(R3) Definición de la arquitectura del sistema de recomendación.	Documento	- Documento que describe la arquitectura del sistema de recomendación

Resultado	**Meta física**	**Medio de verificación**
(R4) Módulo relacionado a la Extracción de Datos de las plataformas web de ALICIA y DINA, Scopus y Orcid	Software	- Pruebas unitarias
(R5) Módulo relacionado al Procesamiento de Datos	Software	- Pruebas de consistencia - Pruebas unitarias
(R6) Datos normalizados de las publicaciones científicas de ALICIA y de los investigadores calificados de DINA	Data	- Pruebas de consistencia
(R7) GUI para la interacción con los módulos del sistema	Software	- Pruebas unitarias

Objetivo: (O2) Desarrollar un modelo de recomendación basado en los enfoques investigados

Resultado	**Meta física**	**Medio de verificación**
(R8) Definición de los enfoques de recomendación que se implementaran	Documento	- Documento que describe los enfoques seleccionados y que se adecuen a la recomendación de documentos científicos

Resultado	Meta física	Medio de verificación
(R9) Módulo relacionado al Modelo de Recomendación	Software	- Pruebas de consistencia - Pruebas unitarias

Objetivo: (O3) Implementar un servicio web para la publicación de las recomendaciones generadas por el sistema de recomendación construido a partir de los componentes de arquitectura definidos.

Resultado	Meta física	Medio de verificación
(R10) Definición de la arquitectura para el servicio web	Documento	- Documento que describe la arquitectura del sistema de recomendación
(R11) Integración de los componentes de arquitectura	Software	- Pruebas de consistencia - Pruebas unitarias
(R12) Servicio web REST para la trasferencia de las recomendaciones obtenidas hacia una plataforma de visualización	Software	- Reporte técnico que describe el funcionamiento del servicio web - pruebas unitarias realizadas sobre las recomendaciones generadas

Objetivo: (O4) Evaluar las recomendaciones generadas bajo la técnica de evaluación offline utilizando la métrica de Precisión.

Resultado	Meta física	Medio de verificación
(R13) Cuestionario para la obtención de información sobre la relevancia de las recomendaciones generadas	Documento	- Formulario de encuesta
(R14) Definición de las métricas que deben satisfacerse en la evaluación del rendimiento del sistema de recomendación	Documento	- Documento que describe las métricas seleccionadas para la evaluación del sistema
(R15) Evaluación offline de los resultados utilizando la métrica de Precisión en N	Datos	- Reporte de evaluación de métricas

Tabla 1. Mapeo de Objetivos y Resultados (Elaboración propia)

1.3 Herramienta, métodos y procedimientos

Esta sección tiene como finalidad el dar a conocer las herramientas, métodos y procedimientos definidos para el logro del objetivo general del presente proyecto de fin de carrera.

A continuación en la Tabla 2 se muestra un mapeo resumido las herramientas y métodos alineados a cada resultado esperado del presente proyecto de fin de carrera.

Resultados Esperados	Herramientas y Métodos
(R3) Componente relacionado a la Extracción de Datos de las	- Beautiful soap - Sickle - Pyscopus

Resultados Esperados	Herramientas y Métodos
plataformas web de ALICIA y DINA	
(R4) Componente relacionado al Procesamiento de Datos	- NLTK
(R5) Datos normalizados de las publicaciones científicas de ALICIA y de los investigadores calificados de DINA	- NLTK
(R7) Componente relacionado al Modelo de Recomendación	- TF-IDF - VSD - LSA - Similitud coseno - Gensim
(R9) Integración de los componentes de arquitectura	- Python - Luigi
(R10) Servicio web REST para la trasferencia de las recomendaciones obtenidas hacia una plataforma de visualización	- Python
(R12) Evaluación de los resultados obtenidos a partir de las	- Precisión

Resultados Esperados	Herramientas y Métodos
recomendaciones generadas utilizando las métricas previamente definidas	

Tabla 2. Mapeo de Resultados con Herramientas y Métodos utilizados (Elaboración propia)

1.3.1 Herramientas

1.3.1.1 **Python**

Python es un lenguaje de programación multiparadigma, el cual soporta la programación orientada a objetos, imperativa y funcional («Welcome to Python.org», s. f.). Así mismo, posee una gran cantidad de librerías las cuales proveen de herramientas facilitadoras en el proceso de desarrollo («Welcome to Python.org», s. f.). Por esta última característica, es que Python es tan utilizada en la programación dentro del campo de las ciencias de la computación («Python Data Science Handbook | Python Data Science Handbook», s. f.). La existencia de librerías para el Análisis de Datos y sus aplicaciones en el campo del Aprendizaje Maquina («Python Data Science Handbook | Python Data Science Handbook», s. f.).

Su elección para el desarrollo del sistema de recomendación, se basa principalmente en la utilización de las librerías relacionadas al tratamiento de los datos como NumPy, Pandas, Scikit-learn, entre otros (Dolgert, s. f.). Así mismo, herramientas como Surprise, Crab y RecSys son librerías que están directamente relacionadas a la construcción y el análisis de los sistemas de recomendación («Python Libraries For Building Recommender Systems», s. f.).

1.3.1.2 **Beautiful Soap**

Beautiful soap es una librería de Python que ayuda en la tarea de extracción de páginas web (web scrapping). Esta herramienta permite la extracción de data almacenada en

archivos HTML y XML, proporcionando al desarrollador de funciones simples y útiles en la navegación y búsqueda de elementos contenidos dentro de las estructuras de páginas web estáticas («Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation», s. f.).

Su elección como herramienta en el proceso de extracción de la data de los investigadores, se basa principalmente en la facilidad para la navegación y obtención de elementos HTML contenidos, en este caso, en cada una de las páginas web de los usuarios investigadores registrados en DINA («Web Scraping with Beautiful Soup», s. f.).

1.3.1.3 **Sickle**

Sickle es una librería de Python cuya funcionalidad principal es la de recuperar la metadata de publicaciones científicas almacenadas en repositorios de acceso abierto («Sickle: OAI-PMH for Humans — Sickle 0.6.2 documentation», s. f.). Sickle utiliza la interface OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), el cual es un mecanismo para la interoperabilidad de repositorios de acceso abierto («Sickle: OAI-PMH for Humans — Sickle 0.6.2 documentation», s. f.).

Su elección para la extracción de la información correspondiente a la producción científica almacenada en ALICIA, se basa en la fácil y limpia extracción de la metadata. Sickle facilita el acceso de esta metadata mediante la utilización de diccionarios, los cuales almacenan atributos como autor, descripción, fecha de publicación, y demás información específica para el modelo de metadatos Dublin Core.

1.3.1.4 **Pyscopus**

Pyscopus es una librería de Python que utiliza el API de Scopus para la ejecución de servicios relacionados con la obtención de información sobre los investigadores registrados y sus publicaciones científicas (Zuo, Zhao, & Eichmann, s. f.).

Su elección para la extracción de información contenida en Scopus, se basa principalmente, en la simplicidad en la obtención de las publicaciones (título y abstract) pertenecientes a un determinado usuario Scopus.

1.3.1.5 **NLTK**

NLTK (Natural Language Toolkit, por sus siglas en inglés), es un paquete de librerías para el procesamiento de lenguaje natural en Python. Esta herramienta fue creada con la intención de facilitar el procesamiento del lenguaje humano, ya que contiene librerías de procesamiento de texto para transformación de la data, así como también más de 50 conjuntos de textos (corpus) y recursos léxicos («Natural Language Toolkit — NLTK 3.3 documentation», s. f.).

Su elección como herramienta en el proceso de pre procesamiento de la data radica principalmente en el fácil uso de sus componentes para las tareas de reducción del ruido y normalización de palabras, que juntas permiten la obtención de características claves y fundamentales en el proceso de recomendación («Natural Language Toolkit — NLTK 3.3 documentation», s. f.).

1.3.1.6 **Luigi**

Es un marco de trabajo en Python puro que facilita la implementación de tuberías para procesos batch; así mismo, gestiona la resolución de dependencias, el flujo de procesos, visualización de procesos y fallas (*luigi*, 2012/2018).

La elección de Luigi como herramienta para la integración de los módulos como procesos interrelacionados, se basa en la buena abstracción que otorga esta librería para definir flujos de procesos en términos de entrada y salida además de resolver las dependencias necesarias (Marco, 2015).

1.3.1.7 **Gensim**

Gensim es un marco de software que facilita el procesamiento de lenguaje natural. Dentro de esta herramienta se implementan distintos algoritmos populares para la inferencia tópica, como el análisis semántico latente y el análisis latente Dirichlet, independientemente del tamaño del corpus de entrenamiento (Rehurek & Sojka, 2010).

Se eligió el uso de esta herramienta ya que facilita la implementación de distintos algoritmos que se utilizan en el análisis de lenguaje natural, así como su eficiencia en el

procesamiento de grandes corpus sin la necesidad de cargar toda la data a memoria (Rehurek & Sojka, 2010).

1.3.2 Métodos

1.3.2.1 **Term Frequency Inverse Document Frequency (TF-IDF)**

TF-IDF (Term frequency – Inverse document frequency, por sus siglas en inglés) es una medida la cual señala el nivel de relevancia de una palabra para un documento en un conjunto de los mismos (Bean, 2016). Es decir, las palabras que sean comunes para el conjunto de documentos en su totalidad tendrán un peso menor; mientras que, aquellas palabras que sean comunes solo en el ámbito del documento más no en el conjunto de estos tendrán un peso mayor.

El modelo TF-IDF se basa en el uso de una matriz de palabras vs. documentos donde se almacena el peso correspondiente a una determinada palabra en un documento del conjunto de documentos (Bean, 2016).

Este modelo TF-IDF se seleccionó por ser un modelo robusto y de fácil manejo para las tareas de recomendación bajo un enfoque de filtrado basado en contenido (Lops, de Gemmis, & Semeraro, 2011).

En la Ecuación 1, se muestra la ecuación matemática que representa el modelo TF-IDF

$$W_d = f_{w,d} \times \log(|D|/f_{w,D})$$

Ecuación 1 Descripción formal del modelo TF-IDF. Donde f w,d es el número de veces w aparece en d, |D| es el tamaño del corpus, y f w,D es el número de documentos en los que w aparece en D (Berger, Caruana, Cohn, Freitag, & Mittal, 2000; Salton & BuckLey, 1988).

1.3.2.2 **Latent Semantic Analysis (LSA)**

LSA se basa en el principio de que las palabras utilizadas dentro de los mismos contextos tienden a tener significados similares (Landauer, Foltz, & Laham, 1998).

Este método tiene como principal función el de extraer e inferir relaciones de las palabras bajo el uso contextual de los mismos, utilizando bases de conocimiento, redes semánticas, gramáticas, sintácticas, analizadores, morfologías, o demás herramientas que facilitan el descubrimiento de relaciones conceptuales entre palabras en un texto (Landauer et al., 1998). Cabe resaltar que, el método dentro del marco de la actividad de IR (Information Retrieval, por sus siglas en inglés) también es llamado Latent Semantic Indexing.

Este método utiliza la representación de texto como una matriz TF-IDF, donde cada una de las frecuencias es ponderada a razón de la importancia que la palabra tiene en el texto y el grado en el que este tipo de palabra posee información en el dominio del conjunto de documentos analizados (Landauer et al., 1998). Esta matriz, luego, pasa por un proceso de descomposición, llamado SVD (Singular Value Decomposition). Esta descomposición da lugar a la conservación de aquellas dimensiones con mayor importancia asociadas con los mayores valores singulares de la matriz de coocurrencia (Pilato & Vassallo, 2015).

La elección de este método para el modelado por características (topic modelling) se basa en la mejor adaptabilidad que posee en cuanto a la mejor selección de textos relacionados, lo que afecta en el proceso de recomendación de documentos. Su superioridad a menudo se refleja en su capacidad para hacer coincidir correctamente documentos conceptualmente similares pero con palabras distintas (Landauer et al., 1998). Además, LSA soluciona los típicos problemas de sinonimia, polisemia, palabras compuestas, entre otros.

1.3.2.3 **Singular Value Decomposition (SVD)**

La técnica de descomposición utilizada en el método de LSA es la de SVD donde luego de aplicada se obtienen 3 matrices: Matriz de Término por Dimensión, Matriz de Valor Singular (dimensión por dimensión) y Matriz de Documento por Dimensión (Kontostathis & Pottenger, 2006). La técnica de descomposición utilizada es la SVD (Singular Value Decomposition, por sus siglas en inglés) (Kontostathis & Pottenger, 2006). La matriz descompuesta se define en la Ecuación 2.

$$A_k = T_k \, S_k D$$

Ecuación 2 Ecuación para la Matriz Descompuesta, donde T y D tienen columnas orto normales y S es normal. La selección de la técnica de SVD es útil en muchas tareas. A partir de la descomposición del valor singular de A, podemos obtener la matriz B del rango k que mejor se aproxima a A, la cual es una matriz reducida.

1.3.2.4 **Similitud coseno**

La similitud coseno es la medida de similitud entre dos vectores derivados del coseno del ángulo entre ellos (Ye, 2011). Bajo la aplicación de los sistemas de recomendación, los ítems a recomendar son tratados como un vector en el espacio de los usuarios a los que se les da las recomendaciones. Es el coseno entre estos vectores lo que se obtiene como medida de similitud (Ye, 2011).

En la Ecuación 3 se muestra l función matemática que representa el cálculo de la medida de similitud coseno

$$sim(\vec{\imath}, \vec{\jmath}) = \cos(\vec{\imath}, \vec{\jmath}) = \frac{\vec{\imath} \cdot \vec{\jmath}}{\| \vec{\imath} \| \cdot \| \vec{\jmath} \|}$$

Ecuación 3 Ecuación para el cálculo de la medida de Similitud Coseno, donde i, j son los vectores a comparar. La selección de esta técnica se basa principalmente sobre su uso para identificar similitudes de documentos de texto y páginas web. Es una de las técnicas más populares y efectivas en la recuperación de información, la agrupación e incluso aplicada al reconocimiento de patrones y al diagnóstico médico (Ye, 2011).

1.3.2.5 **Precisión**

La medida más popular dentro el ámbito de la extracción de información es la precisión. La precisión se define por la fracción de documentos extraídos que son relevantes como resultado de una consulta a un sistema de recuperación de información (IR, Information Retrieval por sus siglas en ingles) (Gunawardana & Shani, s. f.).

En la Ecuación 3, se muestra la fórmula para el cálculo de la métrica de precisión.

$$precision = \frac{\#tp}{\#tp + \#fp}$$

Ecuación 4 Fórmula para el cálculo de la Precisión, donde #tp es el número de verdaderos positivos (número de elementos relevantes recuperados) y #fp es el número de falsos positivos (número de elementos no relevantes que fueron recuperados) (Gunawardana & Shani, s. f.).

1.3.3 Procedimientos

Para la culminación del presente Proyecto se ha seguido la siguiente estructura de trabajo:

1.3.3.1 **Estudio de la literatura**

Para el presente Proyecto, se realizó la búsqueda de literatura relacionada a los sistemas de recomendación de literatura científica, con el fin de obtener los conocimientos necesarios para el desarrollo de la solución al problema presentado. En el caso de este proyecto se realizó un estudio de cada uno de los métodos aplicados en la construcción de soluciones de recomendación de artículos científicos para usuarios investigadores.

1.3.3.2 **Adquisición de la base de datos de usuarios investigadores de DINA y documentos de investigación en ALICIA**

Previo al inicio del desarrollo del prototipo fue necesario considerar sobre qué base de datos se realizaría la implementación, ya que bajo esta data es que se tendría que definir los modelos de datos a utilizar.

1.3.3.3 **Desarrollo del sistema de recomendación**

La etapa central del proyecto es la del desarrollo del sistema de recomendación, el cual incluye las siguientes etapas:

1. Pre-procesamiento

La data obtenida correspondiente a los usuarios investigadores de DINA, así como la correspondiente a los artículos científicos almacenados en ALICIA, deben pasar por el

proceso de pre-procesamiento. Esta etapa es muy importante para la generación de recomendaciones relevantes, ya que un mal pre procesamiento o uno incompleto puede influir en la obtención de resultados con cierto grado de error.

2. Modelamiento de datos

Esta etapa refiere a la construcción de los modelos de datos para el usuario investigador y al artículo científico almacenado en ALICIA.

3. Recomendación

Esta etapa se inicia el proceso de recomendación propiamente dicho, en base a los modelos de datos generados en la etapa previa se realiza el procesamiento correspondiente para la obtención de la recomendación de artículos científicos relevantes para cada uno de los usuarios investigadores registrados en DINA.

4. Publicación

Esta etapa es la cual se hace utilización del servicio web para el envío de las recomendaciones generadas.

1.3.3.4 **Evaluación del sistema y conclusiones obtenidas**

Posterior a la ejecución del sistema para la obtención de sus resultados. Se llevara a cabo una evaluación cuantitativa sobre la precisión de los resultados generados por el sistema de recomendación. Esto permitirá medir la eficiencia del sistema de recomendación.

1.4 Alcance y limitaciones

1.4.1 Alcances

El presente Proyecto se encuentra orientado a la generación de recomendaciones de artículos científicos de ALICIA para usuarios investigadores de DINA. El universo de datos solo comprenderán estos dos conjuntos de datos, los cuales son de acceso público a través de los correspondientes sitios web a cargo del CONCYTEC. Así mismo, se ejecutó las pruebas de validación del sistema solo mediante la utilización de técnicas

offline, y se aplicó los métodos y métricas de evaluación propuestas en la sección 1.3 Herramientas, métodos y procedimientos.

Cabe destacar que la utilización de técnicas offline para la evaluación no comprende la interacción del usuario con las recomendaciones generadas por lo que no existe un proceso de evaluación de la recomendación por parte del usuario. Así tenemos que la captación y tratamiento de la interacción del usuario con la recomendación a través de puntuaciones, esta fuera del alcance del presente Proyecto.

Por otro lado se hace mención que el presente proyecto de fin de carrera se limita a la construcción de una solución para la recomendación de artículos científicos, mas no contempla la implantación del mismo en los ambientes de producción del CONCYTEC. Así se tiene que, la instalación y puesta en marcha del sistema de recomendación se encuentra fuera del alcance del Proyecto.

1.4.2 Limitaciones

El presente Proyecto posee una limitación que afecta al desarrollo del proyecto, la cual está relacionada a la completitud y exactitud de los datos recopilados. Esta limitación recae en el hecho que la data recopilada correspondiente a los usuarios investigadores de DINA puede ser incompleta lo que repercute en el diseño de los perfiles de usuarios utilizados por el sistema de recomendación y por ende en la generación de la recomendación. Así mismo, en base a esto se puede concluir que esta data no es representativa del universo real de datos de los investigadores del SINACYT.

1.5 Justificación y viabilidad

1.5.1 Justificación

Este Proyecto se realizó con el fin de facilitar el trabajo de los investigadores peruanos en la producción de literatura científica, a través de una herramienta de recomendación. Así, se permitirá reducir los efectos relacionados a los problemas identificados en la sección 1.1 Problemática del presente documento.

Finalmente, cabe destacar que la comunidad científica peruana se beneficiara del presente proyecto ya que influirá en la identificación de aquellos trabajos de

investigación afines al investigador y la agrupación de investigadores con líneas de investigación afines.

1.5.2 Viabilidad

Esta sección tiene como propósito mostrar la viabilidad del proyecto de fin de carrera planteado, en términos de ejecución, económicos, de tiempo y de acceso a la información necesaria para su realización y culminación plena.

En primer lugar, en relación a la viabilidad técnica, se puede mencionar que los métodos elegidos para el desarrollo del sistema de recomendación cuentan con el soporte de diversas investigaciones en el campo que validan sus usos en el proyecto, lo cual se ve reflejado en el Estado del Arte. Así mismo, cabe destacar, el conocimiento básico necesario que posee el autor sobre los distintos métodos elegidos.

El Proyecto es viable en términos económicos, debido a que el software necesario para la desarrollo del sistema de recomendación se pueden adquirir de forma gratuita y legal. Tanto Sublime Text como Python y sus librerías son software libre, por lo que su adquisición no incurre en costo alguno.

Con respecto a la viabilidad temporal del Proyecto (dos semestres académicos), esta se demuestra a partir de la identificación de los tiempos necesarios para el aprendizaje, desarrollo y pruebas del sistema de recomendación. Ver Anexo 5 – Cronograma de Proyecto.

Finalmente, en relación a la viabilidad del acceso a la información, se demuestra su viabilidad gracias al abundante material bibliográfico de fácil acceso sobre los métodos y procedimientos necesarios para la construcción del sistema de recomendación. Casi la totalidad de los recursos literarios relacionados al presente proyecto de fin de carrera se encuentran de forma virtual y libre.

Capítulo 2. Marco Conceptual y Legal

2.1 Marco Conceptual

Esta sección tiene como objetivo exponer los conceptos que permitirán el entendimiento del problema: cómo mejorar la visibilidad de la producción científica de acceso abierto almacenada en el repositorio nacional ALICIA para los investigadores calificados del SINACYT.

2.1.1 Producción Científica

La producción científica ha sido definida como "la creación (es decir: producción) propiamente de los aportes científicos (nuevas teorías, nuevos métodos y procedimientos de investigación, nuevos productos científicos, etc.) que logran en su quehacer científico, los que pueden generar uno o más artículos por cada uno de dichos aportes obtenidos" (Morales Morejón & Morales Aguilera, 1997).

Esta "producción científica" se inicia a partir de la concepción por parte de los investigadores para generar nuevos aportes a un conjunto de conocimientos sólidamente establecidos (Silva, Oliveira, & Filho, 2005), y concluye en la difusión de esos nuevos aportes científicos a través de publicaciones de carácter científico (Filho & Siqueira, 2008).

En este sentido, se define a la producción científica como "toda producción documental sobre un determinado asunto de interés de una comunidad científica específica que contribuya al desarrollo de la ciencia y para la apertura de nuevos horizontes de investigación" (Lourenço, 2005).

Moura, a su vez, señala a la producción científica como el producto final del proceso que recorre la generación de ideas, el desarrollo de la investigación y la comunicación que impulsa el desarrollo científico, tecnología y social del país (Moura, A. M. S, Mattos, C. V, & Silva, D. C, 2002).

Por otro lado, autores como Witter y Skeef señalan a la producción científica como medio de difusión y de posicionamiento de las universidades a la hora de hacer ciencia

a través de métodos y procedimientos científicos en aras de la superación de una sociedad (Skeef, 1997; Witter, 1997).

Finalmente, la producción científica es una forma de expresión de conocimiento (tesis, artículos, libros, reportes, revistas indizadas nacionales, datos de investigación, entre otros), la cual resulta del proceso de generación de nuevos aportes científicos en determinadas áreas de conocimiento, y que conlleva al desarrollo tecnológico, económico y social.

En la Ilustración 1 se muestra la producción científica peruana almacenada en el repositorio de acceso libre a la información Alicia.

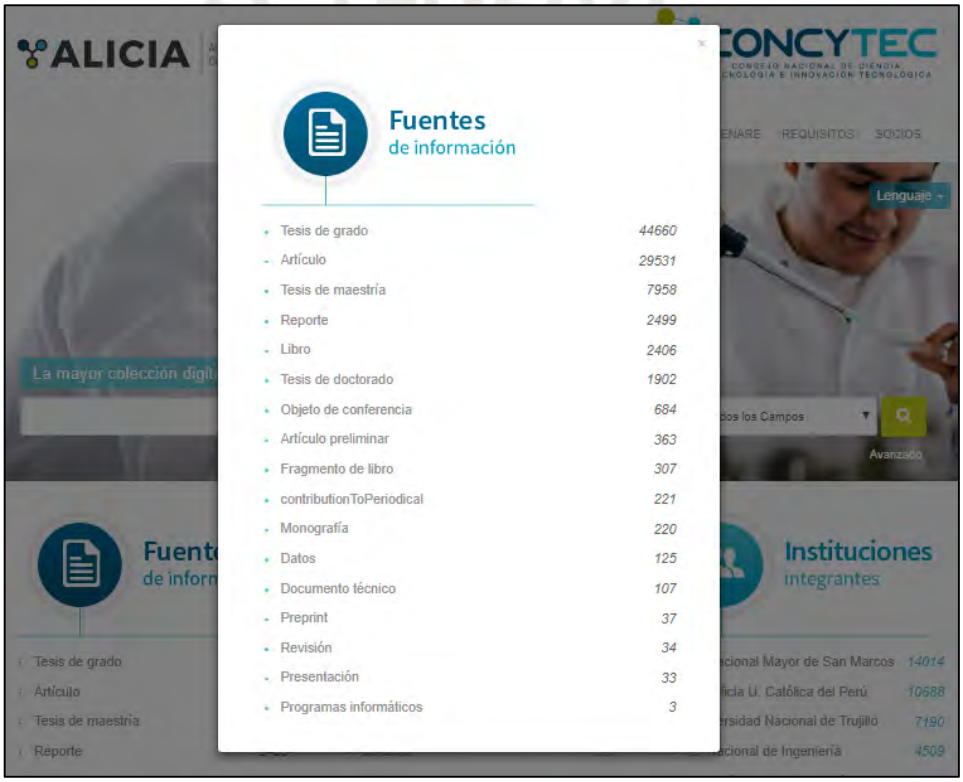

Ilustración 1. Producción Científica Peruana almacenada en el repositorio nacional ALICIA. Extraído de («Inicio de Búsqueda», s. f.).

2.1.2 Repositorio de Acceso Abierto

Se denomina repositorio a sistemas que almacenan recursos digitales (texto, imagen y sonido) de forma perpetua («What are Open Access repositories? - University of

Bradford», s. f.). Entre los recursos almacenados se tienen pre-publicaciones o post-publicaciones, ponencias de eventos, conferencias, informes de investigación, presentaciones a seminarios, tesis, textos de enseñanza y todo aquello que se defina como producción científica (Luque & M, 2009). De tal forma, se denomina acceso abierto a “la disponibilidad de contenido de forma gratuita, inmediata e irrestricta” (Pinfield, 2005). En consecuencia, se define como repositorio de acceso abierto a la colección de contenido en línea a través de la Internet donde su acceso es de forma gratuita, inmediata y sin ninguna restricción (Pinfield, 2005).

En la Ilustración 2 se muestra los porcentajes de uso para cada una de las tecnologías de Repositorios de Acceso Abierto.

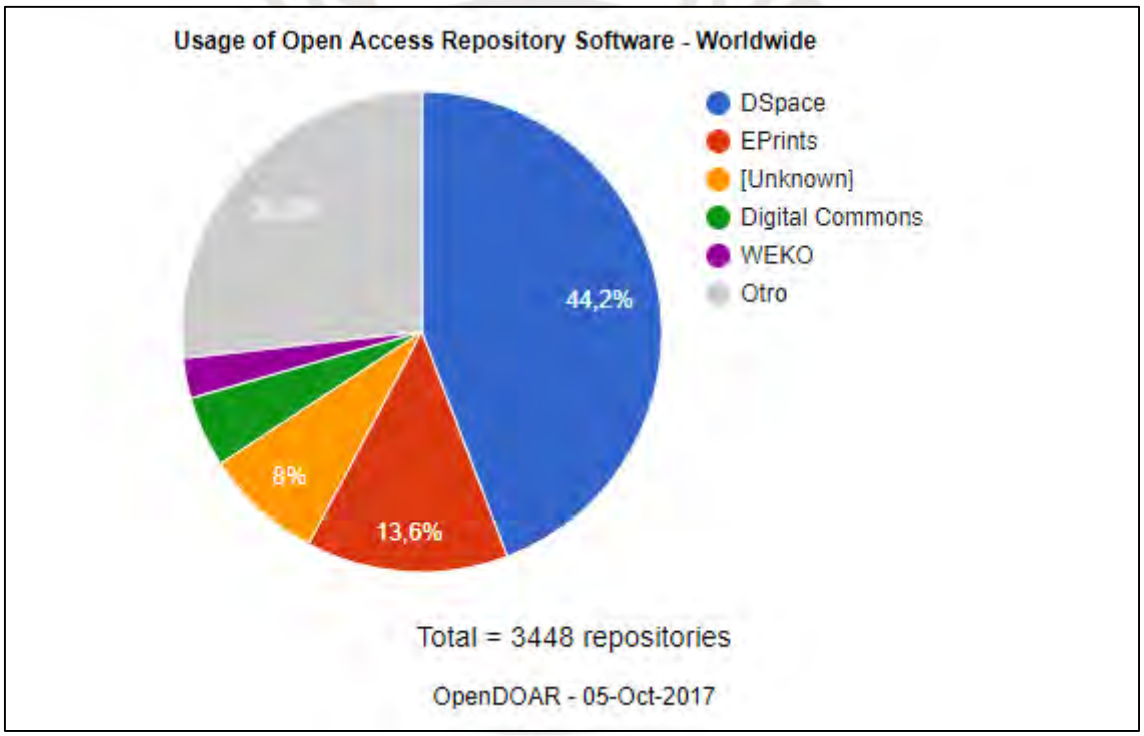

Ilustración 2. Porcentaje del uso de software de repositorios de acceso abierto en el mundo. Extraído de (Millington, 2006).

2.1.3 SINACYT

El Sistema Nacional de Ciencia, Tecnología e Innovación Tecnológica (SINACYT) son las instituciones y personas de origen peruano que se dedican a la Investigación,

Desarrollo e Innovación Tecnológica (I+D+I) en ciencia y tecnología así como a la promoción de la misma (Congreso de la República del Perú, 2004). Este sistema se rige bajo la Ley Marco de Ciencia, Tecnología e Innovación Tecnológica -N° 28303.

2.1.4 ALICIA

El Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto es una plataforma centralizada donde se almacena la información digital del resultado de la producción de ciencia tecnología e innovación, como lo son: libros, publicaciones, artículos de revistas especializada, trabajos técnico-científicos, programas informáticos, datos procesados y estadísticas de monitoreo, tesis académicas y similares. La información almacenada es “de acceso abierto, sin fines de lucro y sin requerimientos de registro, suscripción o pago alguno y está disponible para leer, descargar, reproducir, distribuir, imprimir, buscar o enlazar textos completos; considerando los derechos de autor” (Congreso de la República del Perú, 2013).

En base a la Ley N°30035 promulgada en junio de 2013, todas las instituciones que reciben financiamiento del Estado, tienen como obligatorio el uso de ALICIA, dando así a conocer el acceso libre a la información digital relacionada a la producción científica del país (Atamari-Anahui & Díaz-Vélez, 2015). ALICIA se rige bajo la Directiva 087-2016-CONCYTEC-P para su regulación.

.

2.1.5 REGINA

Se denomina REGINA al Registro Nacional de Investigadores en Ciencia y Tecnología, de la cual forman parte personas naturales que cumplan con las capacidades establecidas de acuerdo a una calificación, para el desarrollo de actividades de investigación científica y/o desarrollo tecnológico (Presidencia de la República del Perú, 2015a). En relación a la calificación, se tiene que la Dirección de Evaluación y Gestión del Conocimiento es la responsable de administrar los procedimientos relacionados al registro del Investigador, gestión de datos y las comunicaciones derivadas de los procedimientos establecidos en el Reglamento de Calificación y Registro de Investigadores en Ciencia y Tecnología del Sistema Nacional de Ciencia, Tecnología e

Innovación Tecnológica – SINACYT (Presidencia de la República del Perú, 2015a). Así mismo, en relación a la duración del registro del investigador en el REGINA, esta tiene una de dos (02) años, salvo se incurran en causales de expulsión (Presidencia de la República del Perú, 2015a).

2.1.6 DINA

El Directorio Nacional de Investigadores e Innovadores es una base de datos donde se registra y almacena la información de personas naturales que están relacionadas al campo de la ciencia, tecnología e innovación, tanto en el país como en el extranjero (Presidencia de la República del Perú, 2016b). Así mismo, cabe destacar que la relación existente entre el REGINA y DINA se basa en la utilización del segundo como fuente de información para el proceso calificador del primero. Así, REGINA se puede describir como la base de datos de miembros calificados de DINA.

2.2 Marco Legal

Esta sección tiene como objetivo exponer los fundamentos por los cuales se rigen ALICIA, DINA, REGINA y el SINACYT.

2.2.1 Ley N° 30035 que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto

Esta Ley tiene como objetivo establecer el marco normativo del Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto. Así mismo, define sus lineamientos rectores, los cuales son:

- Establecer y adoptar estrategias y políticas a fin de garantizar el acceso libre y abierto a la producción en ciencia, tecnología e innovación.
- Garantizar la gestión, divulgación y preservación a largo plazo de la información del Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto.

- Garantizar la seguridad y la calidad de la información y establecer las condiciones necesarias a fin de salvaguardar la propiedad intelectual.
- Fomentar el fortalecimiento de la red científica peruana.

Dentro de las disposiciones generales de esta Ley, se hace mención a la definición del Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, el cual fue descrito líneas arriba. Así mismo, se hace mención al ámbito en donde se aplica la Ley: entidades del sector público, entidades del sector privado o personas naturales que deseen compartir información bajo el marco del Reglamento de la presente Ley, entidades privadas y personas naturales que hayan obtenido financiación del Estado para su investigación, y personas y entidades que componen el SINACYT (Congreso de la República del Perú, 2013, p. 30)

2.2.2 Reglamento de la Ley N° 30035 que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto

Este Reglamento tiene como finalidad desarrollar la Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto. Así mismo, se hace mención al ámbito en donde se aplica el presente Reglamento: entidades públicas miembros o no del SINACYT, entidades privadas y personas naturales que hayan obtenido financiación del Estado para su investigación, personas de nacionalidad peruana y extranjeros cuya producción intelectual se haya realizado dentro del país y no se encuentren afiliados a una institución que cuente con algún repositorio y que cumplan con las disposiciones técnicas académicas reguladas para ALICIA (Presidencia de la República del Perú, 2015b).

2.2.3 Directiva Nº 004-2016-CONCYTEC-DEGC que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto

Esta directiva tiene como finalidad dar conocimiento sobre la regulación del Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, también denominado “Acceso Libre a la Información Científica para la Innovación” – ALICIA. Así

mismo, esta Directiva es de aplicación para el Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica – CONCYTEC y para todas aquellas personas naturales y jurídicas, públicas o privadas mencionadas descritas en el Artículo 3º del Decreto Supremo Nº 006-2015-PCM, Reglamento de la Ley Nº 30035, cuyas obras, datos procesados o estadísticas de monitoreo, se incorporen o deban incorporarse a ALICIA en el marco de la Ley, su Reglamento y su Directiva. De tal forma, la presente Directiva define disposiciones específicas que se relacionan con la producción científica no susceptible a ser incorporada a ALICIA, como lo son toda información calificada como secreta, confidencial, reservada, con carácter de inteligencia y contrainteligencia, además de disposiciones relacionadas al aseguramiento de la calidad de la información almacenada en ALICIA. Así mismo, la presente Directiva también define los procedimientos para: la adhesión de repositorios institucionales, la incorporación de resultados de investigación de personas naturales, la solicitud de orientación técnica del CONCYTEC, el monitoreo y evaluación para la verificación de los metadatos, la exclusión de ALICIA y la postergación de una obra, resultado de una investigación o aquella que requiera de un periodo de exclusividad. De la misma manera, la Directiva define a las entidades públicas y privadas como responsables del cumplimiento de las disposiciones contenidas en la Ley , su Reglamento, la presente Directiva y los anexos de la misma (Presidencia de la República del Perú, 2016a).

2.2.4 Reglamento del Registro Nacional Científico, Tecnológico y de Innovación Tecnológica - RENACYT

El presente Reglamento tiene como finalidad regular el proceso de inscripción en el Registro Nacional Científico, Tecnológico y de Innovación Tecnológica – RENACYT, de las personas naturales y jurídicas relacionadas con la ciencia, tecnología o innovación tecnológica (CTI), dentro del territorio nacional, así como de nacionales residentes en territorio extranjero. El Reglamento también describe la relación que el Directorio Nacional de Investigadores e Innovadores (DINA) y el Directorio Nacional de Instituciones en CTI (DANI) tienen con el RENACYT, a partir de las dos categorías de inscripción existentes: personas naturales y jurídicas. La información consignada por las personas naturales y definidas en el presente Reglamento deben ser registradas en el

DINA, así mismo, la información consignada por las personas jurídicas definidas en el Reglamento deben ser registradas en el DANI. En el caso del registro de personas naturales que soliciten la calificación de Investigador en Ciencia y Tecnología del SINACYT, estas deberán cumplir con los criterios establecidos en el Reglamento de Calificación y Registro de Investigadores en Ciencia y Tecnología del Sistema Nacional de Ciencia, Tecnología e Innovación Científica – SINACYT, denominado REGINA. De tal forma, las personas que soliciten la calificación Evaluador en Ciencia y Tecnología y/o Evaluador en Innovación y Financiamiento de Proyectos deberán cumplir con los criterios establecidos en la Directiva que regula el Directorio Nacional de Evaluadores en Ciencia, Tecnología e Innovación – EVA (Presidencia de la República del Perú, 2016b).

2.2.5 Reglamento de Calificación y Registro de Investigadores en Ciencia y Tecnología del Sistema Nacional de Ciencia, Tecnología e Innovación Científica - SINACYT

El presente Reglamento tiene por objetivo regular el procedimiento para calificar y registrar como investigadores en Ciencia y Tecnología en el Perú a personas naturales que realizan labores de investigación. Este Reglamento describe los lineamientos en relación a los criterios de calificación, procedimientos para la calificación, así como también las características propias del registro de investigadores peruanos. El presente Reglamento fue publicado en Febrero de 2017 por Resolución de Presidencia Nº 023-2017-CONCYTEC-P y es un sustituto al publicado en Diciembre de 2017 a través de la Resolución de Presidencia Nº 184 -2015-CONCYTEC-P. Este Reglamento describe los lineamientos en relación a los criterios de calificación, procedimientos para la calificación, así como también las características propias del registro de investigadores peruanos. Cabe resaltar que el presente Reglamento define ocho criterios para la calificación de Investigador en Ciencia y Tecnología del SINACYT, los cuales son: grado de Bachiller, Maestro o Doctor, o título profesional, publicaciones en revistas científicas indexadas, publicación de libros y/o capítulos de libros o edición de libros de su especialidad, registro de propiedad intelectual como patentes u otras modalidades de protección de invenciones o nuevas tecnologías, asesoramiento de tesis sustentadas

de bachillerato, título profesional, maestría y/o doctorado, valor del índice h de Scopus, experiencia en proyectos de investigación científica y/o desarrollo tecnológico, y ponencias en congresos, seminarios u otros eventos de su especialidad a nivel nacional y/o internacional (Presidencia de la República del Perú, 2015a).

.

Capítulo 3.Estado del Arte

3.1 Método usado en la revisión

Para la revisión del Estado del Arte se utilizó la revisión sistemática. Se realizaron búsquedas a través de Google Scholar, buscador de Google enfocado en documentos académicos y científicos.

3.1.1 Formulación de la pregunta

Para realizar la búsqueda, se formuló la siguiente pregunta: ¿Qué tipos de sistemas de recomendación son más apropiados para la recomendación de publicaciones científicas? Para la resolución de la pregunta se utilizaron los siguientes términos: “scientific”, “article”, “articles”, “recommender”, “recommending”, “academics”, “cientific”, “journal”, “articles”, “article”, “papers”, “paper”, “publication”, “publications”.

3.1.2 Selección de las Fuentes

A partir de los términos anteriormente mencionados, se construyó una cadena de búsqueda utilizando los conectores lógicos AND y OR. Esta cadena de búsqueda fue la siguiente:

	Cadena de búsqueda generada
1	(“recommending” OR “recommender”) AND (“academic” OR “scientific” OR “journal”) AND (“articles” OR “article” OR “papers” OR “paper” OR “publication” OR “publications”)

Tabla 3. Cadena General de Búsqueda (Elaboración propia)

Para la selección de la información, se tomó en cuenta la fecha de publicación del documento encontrado. Fue así que, solo se seleccionó información con fecha de publicación desde el año 2008 hasta la actualidad. Además, solo se seleccionó los documentos que contengan la cadena de búsqueda en sus títulos o resúmenes.

3.2 Síntesis del Estado del Arte

En esta Sección se presenta un resumen sobre los distintos conceptos encontrados luego de la revisión de las fuentes de investigación mencionadas en el Anexo 1.

3.2.1 Filtrado basado en estereotipos

Esta clase se basa en la generación de recomendaciones a partir de características propias de los usuarios, las cuales a su vez se relacionan con algún estereotipo existente (Beel, Dinesh, Mayr, Carevic, & Raghvendra, 2017). No obstante, dos de los principales problemas con este enfoque son el estricto encasillamiento de los usuarios dentro de estereotipos definidos, así como la tarea exhaustiva de construcción de estereotipos (cada ítem a recomendar debe ser relacionado manualmente con algún estereotipo existente) (Barla, 2010). Por otro lado, cabe destacar, que los alcances basados en estereotipos no necesitan grandes capacidades de procesamiento para la obtención de sus resultados pero sí un esfuerzo manual para la creación de los estereotipos (Beel, Gipp, Langer, & Breitinger, 2015).

3.2.2 Filtrado basado en contenido

Esta clase es la más utilizada para la implementación de sistemas de recomendación (Ricci, Rokach, Shapira, & Kantor, 2010). Se base en la inferencia de los intereses de un determinado usuario a través de su interacción con los elementos de repositorios digitales (Seroussi, 2010). El modelo de datos del usuario comprende las características de los elementos seleccionados por el usuario; de manera que, las recomendaciones nacen a partir de la comparación del modelo de datos y las recomendaciones candidatas son comparadas, mediante la utilización, por ejemplo, de un modelo de espacio vectorial y el coseno del coeficiente de similitud (Beel et al., 2015). Así mismo, cabe resaltar, la ventaja que este enfoque posee sobre el de estereotipos permite una personalización basada en el usuario para que el sistema de recomendación pueda determinar las mejores recomendaciones para cada usuario individualmente, en lugar de limitarse a estereotipos (Beel et al., 2015). Sin embargo, algunas de las desventajas de este enfoque es: su baja serendipidad y sobre especialización (Ricci et al., 2010), lo que da lugar a la recomendación de elementos muy parecidos a los que el usuario ya conoce,

además de, ignorar características como la calidad y la popularidad de los elementos (Dong, Tokarchuk, & Ma, 2009). Ejemplo de la implementación de esta clase para la recomendación de elementos, dentro del dominio de los sistemas de recomendación de publicaciones científicas, es Docear (Beel, Langer, Genzmehr, & Nürnberger, 2013).

3.2.3 Filtrado colaborativo

Esta clase se basa en la teoría de recomendación a partir de la semejanza en la interacción de usuarios con los elementos de los repositorios digitales. La semejanza en la interacción de usuarios con los elementos se explica en cómo estos califican a los elementos del repositorio, esto a su vez da lugar a la identificación de usuarios con gustos similares (Beel et al., 2015). Así, aquellos elementos calificados positivamente por algún usuario, serán recomendados a otros con gustos similares. El filtrado colaborativo posee tres ventajas:

- Es independiente del contenido (Palopoli, Rosaci, & Sarné, 2013)
- Toma en consideración evaluaciones de calidad reales (Dong et al., 2009)
- Genera recomendaciones serendipitarias al no basarse en la similitud de los elementos sino en la similitud de los usuarios (Palopoli et al., 2013)

Por otro lado, algunos de los principales problemas del filtrado colaborativo recae en:

- El nivel de participación de los usuarios para calificar los elementos (Yang, Wei, wu, Zhang, & Zhang, 2009).
- El alto grado de dispersión dentro del dominio de los sistemas de recomendación de publicaciones científicas causado por la diferente proporción entre usuarios (menos científicos) y elementos (más publicaciones científicas), lo cual dificulta la búsqueda de usuarios afines (Vellino, 2013).
- Menor escalabilidad y mayor procesamiento de datos fuera de línea que el filtrado basado en contenido (Sosnovsky & Dicheva, 2010).

3.2.4 Co-ocurrencia

Esta clase se basa en la relatividad que tienen entre si los elementos; es decir, el enfoque de las recomendaciones generadas están basadas en las relaciones que los elementos tienen entre sí, mas no en su semejanza (Beel et al., 2015). Esta clase genera recomendaciones serendipitarias lo que la hace comparable con la clase de filtrado colaborativo (Sugiyama & Kan, 2011). Por otro lado, algunas de sus desventajas son la generación de recomendaciones que no son altamente personalizadas y que los elementos pueden ser recomendados solo cuando haya surgido, al menos una vez, una co-ocurrencia con alguno de ellos (Beel et al., 2015). Ejemplos de la implementación de esta clase para la recomendación de elementos, dentro del dominio de los sistemas de recomendación de publicaciones científicas, son los recomendores bX and BibTip, las cuales generan millones de recomendaciones mensuales (Mönnich & Spiering, 2008).

3.2.5 Basado en grafos

Esta clase se basa en la construcción de redes de grafos a partir de conexiones inherentes que, en el caso de publicaciones científicas, muestran como están conectadas a través de citas (Liang, Li, & Qian, 2011), lugar de publicación, autor, relación entre genes , entre otros (Beel et al., 2015).

3.2.6 Enfoques de recomendaciones híbridas

Esta clase se basa en la combinación de dos o más clases puras, las cuales resultan en la obtención de mejor rendimiento con un menor número de inconvenientes causados por la utilización de una sola clase (Beel et al., 2015). Generalmente, el filtrado colaborativo se combina con otra clase en un intento de reducir inconvenientes que puedan presentarse (Çano & Morisio, 2017). Ejemplos de la implementación de esta clase para la recomendación de elementos, dentro del dominio de los sistemas de recomendación de publicaciones científicas, son ARSYS (Bancu et al., 2012), Papyres (Naak, 2009) y Scienstein (Gipp, Beel, & Hentschel, 2009).

3.3 Conclusiones sobre el Estado del Arte

A partir de la pregunta planteada inicialmente y la revisión realizada a los documentos seleccionados, se puede concluir que, las investigaciones realizadas en el campo de los sistemas de recomendación relacionados a artículos de investigación centran su atención en la combinación de distintos métodos o técnicas para la resolución de los diversos problemas que se presentan en la construcción de recomendadores de artículos de investigación. Sin embargo, todas las investigaciones convergen en un mismo objetivo: proponer sistemas híbridos para la mejora en la generación de recomendaciones de artículos de investigación.

Además, la investigación “Sistemas de recomendación de documentos de investigación: una encuesta bibliográfica” recopila información importante sobre los enfoques aplicados en la construcción de los sistemas de recomendación de artículos científicos, y muestra los desafíos que se pueden encontrar en este campo. Así mismo la investigación expone los distintos métodos de evaluación para los sistemas de recomendación de artículos de investigación, concluyendo que no solo la métrica de exactitud refleja la satisfacción del usuario. La investigación también concluye y resalta la importancia del uso de un modelo de datos con el objetivo de identificar las necesidades de información del usuario y así generar recomendaciones relevantes para el mismo.

Por otro lado, en los últimos años, productos relacionados a la gestión de documentos de investigación han incluido funcionalidades para la recomendación de estos a sus usuarios investigadores. Así, Docear implementó la función de recomendación de documentos científicos a sus usuarios, diferenciándose así de sus competidores Zotero y Mendeley. De tal manera, Scienstein se presenta como una alternativa a los motores de búsqueda académicos, mejorando el enfoque de la tradicional técnica de búsqueda por palabra bajo un esquema hibrido de recomendación. Así mismo, ARSYS y Papyres implementan enfoques híbridos compuestos por filtrado baso en contenido y colaborativo. Cabe resaltar que, ambas soluciones utilizan la combinación de dos o más técnicas para la generación de recomendaciones a sus usuarios.

Finalmente, se concluye que, tanto las investigaciones como productos relacionados al tema del presente proyecto de fin de carrera, sugieren la aplicación de más de un enfoque para el desarrollo de sistemas de recomendación de artículos de investigación. De esta forma, se concluye que los sistemas con aplicación de enfoques híbridos son los más apropiados para la recomendación de publicaciones científicas.

Capítulo 4. Recolección y Pre procesamiento de la información

Este capítulo muestra cómo se desarrolló el primer objetivo específico, el cual consiste en el proceso de extracción, pre procesamiento y modelamiento de la información correspondiente a las publicaciones científicas almacenadas en ALICIA y los datos pertenecientes a los investigadores calificados del SINACYT registrados en el DINA.

Tres fueron las etapas que se ejecutaron para la obtención de los resultados:

- Modelamiento
- Extracción
- Pre procesamiento.

Los resultados esperados obtenidos, luego del desarrollo de cada una las etapas definidas, son: los modelos de datos que representan tanto el perfil del investigador y las publicaciones científicas de ALICIA, y los componentes Extractor y de Pre procesamiento.

4.1 Modelamiento de los datos

Para el caso de las publicaciones científicas, la data a modelar fue obtenida de las fuentes de acceso abierto de los repositorios pertenecientes a ALICIA. Por otro lado, para el caso de los investigadores calificados del SINACYT, la data a modelar fue obtenida del DINA, Scopus y Orcid. La información contenida dentro de estas plataformas web era extensa por lo que para el análisis solo se utilizó aquella que fuera relevante dentro del proceso de recomendación de publicaciones.

En la ilustración 3 se muestra el modelo de datos que se desarrolló como parte de la ejecución del primer objetivo específico del presente Proyecto. Para la generación de los modelos de datos se tomó en cuenta la relevancia de cada uno de los atributos dentro del proceso de recomendación.

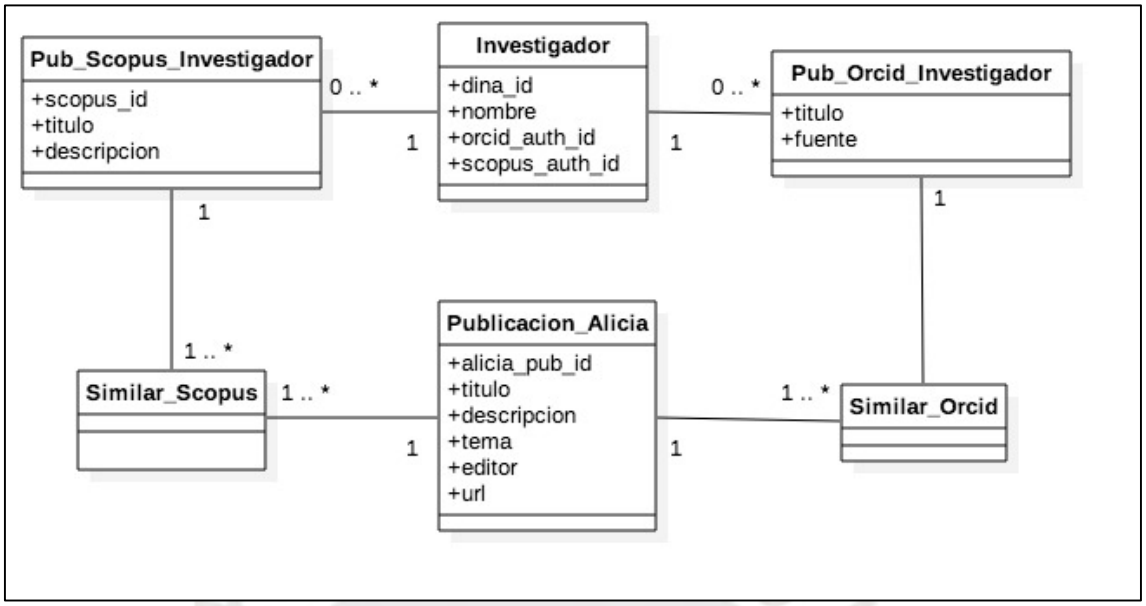

Ilustración 3 Diagrama de Clases para los modelos de datos correspondientes a las publicaciones de ALICIA y el perfil del investigador del SINACYT. El perfil del investigador consta también de las clases Pub_Scopus_Investigador y Pub_Orcid_Investigador que representan las publicaciones del autor contenidas en dichas plataformas. (Elaboración propia)

El modelo de datos del investigador está definido por tres estructuras:

- La estructura Investigador presenta los atributos básicos de los investigadores como su identificador DINA, nombre e identificadores Orcid y Scopus.
- La estructura Pub_Scopus_Investigador contiene los atributos relacionados a las publicaciones en Scopus del investigador. Los atributos relevantes son el identificador Scopus de la publicación, así como título y descripción.
- La estructura Pub_Orcid_Investigadores presenta los atributos relacionados a las publicaciones en Orcid del investigador. En este caso el título y la fuente son relevantes dentro del modelado.

El modelo de datos de Publicaciones está definido por una estructura:

- La estructura Publicacion_Alicia contiene atributos como identificador de la publicación, título, descripción, tema, editor y dirección web (url) de la publicación científica, los cuales son elementos relevantes dentro del proceso de recomendación.

Asimismo, cabe destacar la existencia de las estructuras Similares_Scopus y Similares_Orcid donde se relacionará cada publicación de Scopus u Orcid con todas aquellas recomendaciones de publicaciones de ALICIA.

4.2 Arquitectura del sistema

Para el presente Proyecto se identificó como módulos del sistema a cada uno de los procesos que intervienen para la obtención del resultado final. De esta manera, las tareas comprendidas dentro de los procesos de Extracción y Preprocesamiento pertenecen a los módulos del mismo nombre. Así mismo, se definieron los módulos asociados al proceso de generación del modelo de recomendación, a la provisión del servicio y a la interfaz gráfica. En la Ilustración 4 se muestra el diseño de la arquitectura a alto nivel.

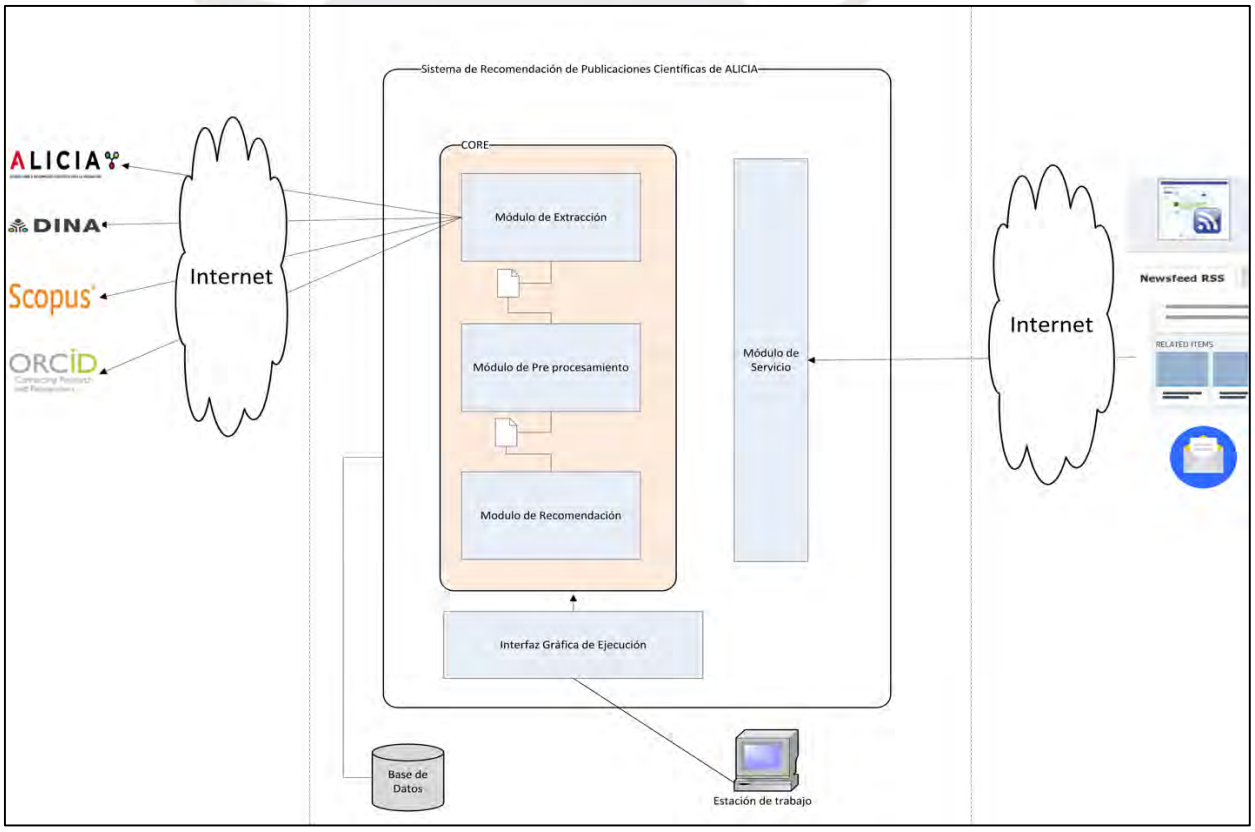

Ilustración 4 Arquitectura del sistema (Elaboración propia)

El diseño de la arquitectura se basa en las etapas de construcción del sistema de recomendación como solución integral, es así que está compuesto de 5 módulos:

- Módulo de Extracción: recoge la data correspondiente a las publicaciones científicas almacenadas en ALICIA, como también los datos sobre los investigadores registrados en la plataforma web del DINA junto con los datos concernientes a las publicaciones de estos en las plataformas Scopus y Orcid. Por tal motivo, este componente cuenta con la implementación de funciones que ayudan en el proceso de extracción utilizando técnicas extracción de datos de sitios web (web scrapping). Finalmente, este módulo ejecuta el almacenamiento de la data extraída siguiendo la estructura de los modelos de datos definidos tanto para las publicaciones científicas y los investigadores.
- Módulo de Pre procesamiento: se encarga de la transformación de la data en crudo extraída previamente. Esta transformación tiene como finalidad la mejora de la calidad de los datos de entrada de un modelo analítico dentro del marco de actividades de la minería de texto y/o el procesamiento de lenguaje natural. El resultado se almacena en forma de documentos donde un documento es una lista de términos pre procesados y que se relaciona con cada uno de las publicaciones científicas en ALICIA, así como las publicaciones de Scopus y Orcid de los investigadores.
- Módulo de Recomendación: se encarga de la implementación del modelo de recomendación. Es aquí donde los datos previamente extraídos y pre procesados son ingresados de manera que, se capacita a un modelo y se generan recomendaciones sobre qué publicaciones científicas pueden interesar a los usuarios basado en sus publicaciones científicas en Scopus y Orcid.
- Módulo de Servicio: capa de integración que permite a otras aplicaciones externas al sistema consumir la información generada por el sistema.
- Interfaz Gráfica de Ejecución: capa de interacción entre el sistema y el usuario del mismo. Permite la ejecución a demanda de los módulos de extracción y pre procesamiento.

4.3 Extracción de los datos

El proceso de extracción de datos implicó la creación del módulo de extracción de información de las plataformas web de ALICIA, DINA, Scopus y Orcid (Ver Ilustración 7).

Para la extracción de los metadatos de las publicaciones científicas en ALICIA se utilizó el protocolo OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). Se utilizó la librería de Python Sickle para la implementación de un cliente OAI-PMH y así obtener las listas de registros de publicaciones para cada uno de los repositorios dentro de la comunidad de ALICIA.

De tal manera, para la extracción de la data correspondiente a los investigadores registrados en DINA, se utilizó la técnica de extracción de datos de sitios web (web scrapping) para la obtención de sus identificadores Scopus y Orcid. Para la ejecución del web scrapping, se utilizó la librería Beautiful Soap para las páginas web de DINA y Orcid, mientras que para la extracción de información sobre las publicaciones registradas en Scopus por los investigadores nacionales se utilizó la librería Pyscopus.

Los pasos que se deben seguir para el proceso de extracción son los siguientes:

Para la extracción de Publicaciones:

1- Lectura del archivo que contiene las url OAI-PMH de cada uno de los repositorios asociados a ALICIA.
2- Extracción de la meta data de los registros existentes en los repositorios (ID publicación, titulo, descripción, tema, editor, url del sitio web).
3- Almacenamiento de los registros de publicaciones ALICIA en Base de datos y en archivos planos (Ver Ilustración 5).

Publicaciones científicas de los repositorios de ALICIA

identifier	title	description	subject	publisher	identifier-metadata
oai:dspace:U	Calidad De Vida Despu	La litiasis vesicular es una er	Calidad de vida Colecistec	Universidad Naci	http://dspace.unitru.edu.p
oai:dspace:U	Factores De Riesgo Pa	Se realizÃ³ un estudio retro	Factores de riesgo Morta	Universidad Naci	http://dspace.unitru.edu.p

Ilustración 5 Registros de publicaciones científicas con los atributos identificador, titulo, descripción, tema, editor y dirección web almacenados en un archivo plano (Elaboración propia).

Para la extracción de Investigadores:

1- Lectura del archivo que contiene las direcciones web DINA de cada uno de los investigadores calificados del SINACYT.
2- Extracción de los datos de los investigadores (nombre, ID scopus, ID orcid) publicados en los perfiles DINA.
3- Extracción de la información de las publicaciones (ID scopus, titulo, descripción) encontradas en los perfiles de Scopus de los investigadores.
4- Extracción de la información de las publicaciones (título, fuente) encontradas en los perfiles de Orcid de los investigadores.
5- Almacenamiento de los registros datos de investigador, publicaciones orcid, publicaciones scopus en Base de Datos y en archivos planos (Ver Ilustración 6).

Datos del investigador

link_dina	name	link_orcid	scopus_id
https://dina.concytec.gob.pe/appDirectorioCTI/VerDatosInvestigador.do?id_investigador=36887	ALVAÑA DE LA CRUZA ALDO ALFONSO	https://orcid.org/0000-0002-1190-0042	56196070900
https://dina.concytec.gob.pe/appDirectorioCTI/VerDatosInvestigador.do?id_investigador=52155	SALDAÑ'AÑ JIMENEZA MIGUEL JORGE	https://orcid.org/0000-0002-9164-7568	

Publicaciones extraídas de Scopus de los investigadores

source	title	link_dina
Aldo AlvÃ¡n	GeologÃ-a de los CuadrÃngulos de La Yarada, Tac	https://dina.concytec.gob.pe/appDirectorioCTI/VerDatosInvestigador.do?id_investigador=36887
Aldo AlvÃ¡n	Os fundamentos da fÃ-sica dos fÃ©rmions de Dir	https://dina.concytec.gob.pe/appDirectorioCTI/VerDatosInvestigador.do?id_investigador=52155
Aldo AlvÃ¡n	The lâ€•V characteristics of a graphene tunnel dic	https://dina.concytec.gob.pe/appDirectorioCTI/VerDatosInvestigador.do?id_investigador=52155

Publicaciones extraídas de Orcid de los investigadores

abstract	link_dina	scopus_id	title
Â© Springer Internati	https://dina.con	85040223939	A perception study of a new set of usability heuristics for transactional web sites
Â© 2017 Association f	https://dina.con	85033432256	Applying a user-centered design methodology to develop usable interfaces for an au

Ilustración 6 Registros de publicaciones científicas extraídas de Orcid y Scopus perteneciente a los investigadores (Elaboración propia).

4.4 Pre procesamiento de los datos

Las técnicas de pre procesamiento de datos dentro de la minería de textos, son empleadas para el tratamiento de data incompleta, irrelevante, redundante e inconsistente (Tyagi, Solanki, & Tyagi, s. f.). Estas características de inconsistencia, incompletitud, irrelevancia y redundancia son latentes en cualquier conjunto de datos reales y dificultan el proceso de extracción de conocimiento durante la etapa de entrenamiento del modelo de recomendación (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

De esta manera, la etapa de pre procesamiento es de suma importancia y critica, dentro de las actividades de minería de textos, procesamiento de lenguaje natural, recuperación de información y cualquier otra actividad relacionada con el análisis de datos, ya que influye directamente en la calidad de los resultados (Kotsiantis et al., 2006).

Para el presente Proyecto, el pre procesamiento de los datos obtenidos tras la ejecución de las funciones definidas en el componente extractor, está definido bajo el siguiente marco de trabajo (Ver Ilustración 7)

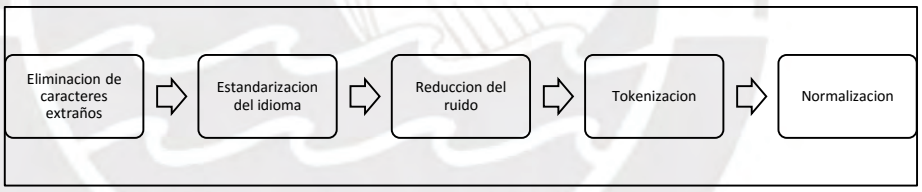

Ilustración 7 Marco de trabajo para el pre procesamiento de textos. (Elaboración propia)

La ejecución de todas las fases del ciclo de pre procesamiento, finaliza con la obtención de la data normalizada y lista para ser utilizada como dato de entrada para el componente de recomendación.

Así mismo, cabe destacar que la data a pre procesar la componen los valores de los atributos título y descripción de las publicaciones de ALICIA y las publicaciones del investigador en Scopus y Orcid.

Finalmente, luego de ejecutado el pre procesamiento de la data, el resultado del mismo se almacenó en un archivo plano (Ver Ilustración 8).

bag_of_words	identifier
pres trabajo analiz evolucin calidad servicio ref continuidad suministro empresa distribuidora pas brindado aplicacir	oai:cybertesis.uni.edu.pe:uni/10899
empres cemento pacasmayo saa finalidad envi seal alarma medida solicitada coessinac comit operacin econmic sist	oai:cybertesis.uni.edu.pe:uni/10900

Ilustración 8 Texto pre procesado para cada uno de las publicaciones científicas (Elaboración propia).

4.4.1 Eliminación de caracteres extraños

Para los procesos de minería de textos, los valores de entrada son típicamente bytes en un archivo o en un servidor web (Manning, Raghavan, & Schütze, 2008). Estas entradas al ser convertidas a secuencias lineales de caracteres, a menudo presentan distintos esquemas de codificación como Unicode UTF-8 entre otros. Es así que, caracteres como ®, ©, tildes y otros similares al pertenecer a un esquema de codificación distinto al ASCII, suelen generar ruido dentro de la data.

En este Proyecto la data predominante son textos en español, lo que implica la existencia de caracteres que no pertenecen al esquema de codificación ASCII. En la Ilustración 9, se puede ver como el valor del atributo descripción de una de las publicaciones extraídas contiene tanto caracteres ASCII y otros perteneciente al esquema de codificación UTF-8.

Para la eliminación de caracteres bajo un esquema de codificación diferente al ASCII, se utilizó la normalización de caracteres, mediante el uso de la librería unicodedata. Esta librería permite hacer uso de las varias formas de normalización (NF) basándose en la definición de equivalencia canónica y equivalencia de compatibilidad («7.9. unicodedata — Unicode Database — Python 2.7.15 documentation», 2018). Las formas normales son (Moran & Cysouw, 2018):

- La forma normal D (NFD) también se conoce como descomposición canónica, y traduce cada carácter en su forma descompuesta.
- La forma normal C (NFC) primero aplica una descomposición canónica y luego vuelve a componer caracteres pre combinados.

- La forma normal KD (NFKD) aplicará la descomposición de compatibilidad, es decir, reemplazará todos los caracteres de compatibilidad con sus equivalentes.

Esta normalización y posterior codificación y decodificación permite la transformación de letras tildadas a solo su equivalente sin tilde. Así se tiene:

á, é, í, ó, ú, ñ -> a, e, i, o, u, n

El Algoritmo 1 busca obtener el texto limpio de aquellos caracteres que no formen parte del esquema ASCII (caracteres extraños como ®, ©, tildes, y demás). Para esto, se normalizó y codificó cada palabra contenida en cada uno de los registros de ALICIA y DINA obtenidos en la etapa de extracción. La normalización a la forma NFKD implicó la descomposición de los caracteres y su conversión en caracteres compatibles. Luego, la codificación permitió obtener la versión codificada, bajo el esquema ASCII, del texto analizado.

```
1- Para cada palabra del texto
    2- Normalizar palabra a la forma NFKD
    3- Codificar al esquema ASCII
4- Fin para
```

Algoritmo 1 Algoritmo para la normalización de caracteres de la forma NFKD (Elaboración propia).

En la Ilustración 9 se muestra los caracteres extraños presentes en la data extraída tanto en las publicaciones ALICIA, ORCID y Scopus, los cuales fueron eliminados luego de ejecutado el proceso descrito anteriormente.

```
df_alicia['description'][0]

'La litiasis vesicular es una enfermedad frecuente, siendo el tratamiento de elecci\xc3\xb3n la Colecistectom\xc3\xad
a, ya sea v\xc3\xada laparotom\xc3\xada o laparosc\xc3\xb3pica. Actualmente diversos estudios muestran muchas ventaja
s a favor de la t\xc3\xa9cnica laparosc\xc3\xb3pica; sin embargo, muy pocos comparan estas dos t\xc3\xa9cnicas con re
laci\xc3\xb3n a la calidad de vida de los pacientes. Objetivo: Comparar los resultados de la percepci\xc3\xb3n de la
calidad de vida tras la colecistectom\xc3\xada laparosc\xc3\xb3pica y abierta. Material y M\xc3\xa9todos: Se realiz\x
c3\xb3 un estudio descriptivo, retrospectivo y transversal para determinar la calidad de vida posterior a una colecis
tectom\xc3\xada laparosc\xc3\xb3pica y abierta en adultos operados en el Hospital Regional Docente de Trujillo durant
e el a\xc3\xb1o 2006. Se realiz\xc3\xb3 entrevistas domiciliariamente para el llenado de los cuestionarios de calidad
de vida SF-36 y GIQLI. Resultados: Se incluyeron 23 pacientes en cada grupo de colecistectom\xc3\xada; 91.3% fueron m
ujeres siendo su edad promedio 41.91 a\xc3\xb1os en el grupo de colecistectom\xc3\xada laparosc\xc3\xb3pica y el 78.2
6% fueron mujeres con una edad promedio de 47.09 a\xc3\xb1os en la colecistectom\xc3\xada abierta. Se encontr\xc3\xb3
diferencia estad\xc3\xadsticamente significativa en los dominios funci\xc3\xb3n f\xc3\xadsica y rol f\xc3\xadsico del
SF-36 en favor de la colecistectom\xc3\xada laparosc\xc3\xb3pica. No se encontr\xc3\xb3 diferencia estad\xc3\xadstica
mente significativa en ninguno de los dominios del GIQLI. Conclusiones: La calidad de vida de los pacientes fueron si
milares entre los dos grupos de t\xc3\xa9cnicas quir\xc3\xbargicas despu\xc3\xa9s de haber sido realizadas, pero la c
olecistectom\xc3\xada laparosc\xc3\xb3pica fue significativamente superior a la t\xc3\xa9cnica abierta con respecto a
los dominios funci\xc3\xb3n f\xc3\xadsica y rol f\xc3\xadsico del SF-36. Palabras clave: Calidad de vida, Colecistect
om\xc3\xada Abierta, Colecistectom\xc3\xada Laparosc\xc3\xb3pica.'
```

Ilustración 9 Valor del atributo descripción para una de las publicaciones extraídas. Se visualiza la existencia de caracteres extraños (ej. caracteres con tildes) (Elaboración propia).

4.4.2 Estandarización del idioma

Para el presente Proyecto, se tuvo que tomar como consideración el manejo de distintos idiomas, ya que se contaba tanto con data en inglés así como en español. (Ver Ilustración 10).

Sin embargo, la data en español fue la más predominante, es por eso que se decidió estandarizar el texto al idioma español mediante el uso de funciones de la librería de Python Googletrans, la cual es un cliente Google Traductor, así como langdetect para la detección del idioma. Fue así que se pudo obtener toda la data extraída en un mismo idioma.

```
1- Si texto no es vacío
        2- Detectar idioma
        3- Si idioma es diferente al español
                4- Traducir texto con cliente Google Traductor
        5- Fin si
6- Fin si
```

Algoritmo 2 Algoritmo para la estandarización de textos al idioma español (Elaboración propia).

El algoritmo 2 busca realizar la estandarización del idioma de los textos extraídos. Para esto, primero se identificó el idioma del texto mediante el uso de la función detect de la librería de langdetect de Python. Si esta función no devolvía como resultado el valor ‘es’ (valor referido al español) pero sí otros valores como ‘en’ (referido al ingles), ‘fr’ (referido al frances), ‘de’ (referido al alemán), se procedía a su traducción mediante el uso de la función translate de la librería googletrans.

```
preprocesador2.df_scopus_rese['title'][0]
'A perception study of a new set of usability heuristics for transactional web sites'
```

Ilustración 10 Valor del atributo título para uno de los registros Scopus que pertenecen a un determinado investigador. El registro se encuentra en el idioma inglés (Elaboración propia).

4.4.3 Reducción del ruido

La reducción de elementos ruidosos es clave en la limpieza de datos, ya que el ruido dificulta la mayoría de los tipos de análisis de datos (Bidgoli, 2010). La presencia de ruido hace que la dimensionalidad del problema sea alta y, por lo tanto, la clasificación más difícil ya que cada palabra en el texto se trata como una dimensión (Haddi, Liu, & Shi, 2013).

El proceso de reducción de elementos con ruidos, se basa en la eliminación de aquellas palabras que no juegan ningún papel importante para la extracción de información dentro del proceso de recomendación ya que las mismas no aportan significado a los documentos (publicaciones de ALICIA, ORCID y Scopus) y por el contrario dificultan la tarea de recomendación a partir de conceptos (Vijayarani & Ilamathi, s. f.). En la Ilustración 11 se muestran la lista de palabras definidas como palabras vacías.

Ilustración 11 Lista de palabras vacías definidas para el filtrado sobre la data correspondiente a las publicaciones de ALICIA, y las publicaciones ORCID y Scopus de los investigadores (Elaboración propia).

Por otro lado, el proceso de reducción de elementos con ruidos, implicó también la evaluación de cada una de las palabras que no fueron filtradas como palabras vacías, de manera que se pudiera identificar si estas contenían caracteres diferentes a las letras del vocabulario español. Estos caracteres son considerados como ruidos dentro del procesamiento de lenguaje natural ya que no aportan ningún significado al texto analizado.

Para el presente Proyecto, la reducción de ruido se basó principalmente en la eliminación de caracteres que no conformen palabras, además de la eliminación de palabras vacías (stop words). El Algoritmo 3 define los pasos necesarios para la obtención de datos limpios.

Primero, se evaluó cada una de las palabras contenidas en el texto. La evaluación consistió en identificar si la palabra pertenecía a la lista de palabras vacías definidas (ver Ilustración 11). En caso, la palabra estuviera dentro de esta lista, se eliminaba. Caso contrario, se verificaba cada uno de los caracteres que la componían eliminando todos aquellos que no se encuentren dentro de los valores ASCII para los caracteres: A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z y sus equivalentes en minúsculas.

```
1- Para cada palabra del texto
    2- Si palabra esta en lista de palabras vacías entonces
        3- Eliminar palabra
    4– Sino
        5- Para cada carácter de la palabra
            6- Si (el código en ASCII del carácter > 90  y < 65) o
            (el código en ASCII del carácter > 122  y < 97)
            entonces
                7- Eliminar caracter
            8- Fin si
        9- Fin para
    10- Fin si
```

Algoritmo 3 Algoritmo para la reducción del ruido (Elaboración propia).

4.4.4 Tokenización y Normalización

La tokenización es un paso que divide cadenas de texto más largas en piezas más pequeñas o tokens. También se conoce como segmentación de texto o análisis léxico a nivel de palabras (Karthikeyan & Aruna, 2013).

Para la data analizada se utilizó la librería NLTK donde se encuentra implementada un tokenizador para texto natural para el idioma español (Ver Ilustración 12).

```
preprocesador.df_alicia=preprocesador.df_alicia[cols_alicia].applymap(preprocesador.tokenize_and_normalize)
preprocesador.df_alicia['description'][0]
```

```
['litias',
 'vesicul',
 'enfermedad',
 'frecu',
 'siendo',
 'tratamiento',
 'elecc',
 'colecistectom',
```

Ilustración 12 Separación de las palabras y normalización de las mismas en un texto, utilizando la técnica de Porter Stemming (Elaboración propia).

La normalización de textos se basa en la transformación de palabras en una forma base de manera que palabras que contengan esta forma base puedan coincidir en similitud (Toman, Tesar, & Jezek, s. f.). Así mismo, este proceso tiene como objetivo la estandarización y representación de rasgos semánticos existentes dentro de textos analizados (Rölleke, Tsikrika, & Kazai, 2006). Existen muchas enfoques para el proceso de normalización, siendo los más populares los de stemming y lemmatizing.

La técnica de stemming se relaciona con el proceso heurístico de cortar los extremos de las palabras, lo que incluye las derivaciones inflexionales; por otro lado, la técnica de lemmatizing se refiere a la normalización a través del uso un análisis morfológico y un vocabulario con la misma finalidad de eliminar las derivaciones inflexionales (Jivani, 2011).

La técnica de normalización utilizada en el Proyecto es la de Porter Stemming, la cual tiene como función la reducción de una determinada palabra a su raíz léxica (lexema), mediante la eliminación de terminaciones morfológicas e inflexionales más comunes en las palabras (Jivani, 2011). Esta técnica da la oportunidad de obtención de mejores valores de exhaustividad, medida sobre el número de documentos que se pueden encontrar con una consulta (Kraaij & Pohlmann, 1996).

La elección de la técnica de stemming, se fundamentó en la gran reducción en el almacenamiento requerida por un diccionario de (Bell & Jones, 2018) y al aumento en el rendimiento debido al uso de variantes de palabras (Hull, 2018). Así mismo, estudios comprueban la mejora, de hasta un 10%, en la precisión promedio mediante el uso de la técnica de stemming para el español; mientras que para la técnica de lemmatizing, las mejoras no son muy significativas (Hollink, Kamps, Monz, & de Rijke, 2004).

El algoritmo de Porter Stemming (Algoritmo 4) se basa principalmente en la aplicación de reglas de transformación de palabras para la remoción de sufijos definidos, a través de 5 pasos: el primero trabaja sobre los sufijos inflexionales, los cuatro siguientes trabajan sobre los sufijos derivacionales (Willett, 2006). La aplicación del algoritmo da como resultado la obtención de los denominados stems, los cuales representan las raíces léxicas de las palabras analizadas.

```
1- Para cada palabra en el texto
        2- Para cada paso en la lista de pasos de transformación
                3- Para cada patrón de transformación en el paso
                        4- Si patrón coincide entonces
                                5-Transforma palabra
                        6- Fin si
                7- Fin para
        8- Fin para
9- Fin para
```

Algoritmo 4 Algoritmo Porter Stemming (Elaboración propia).

Dentro de los patrones de transformación se tiene 5 pasos que agrupan reglas relacionadas con la transformación de sufijos y prefijos de palabras.

Así mismo, las reglas de transformación se verifican sobre las siguientes regiones (Barrenechea, 2006):

- R1: región que va desde la primera no vocal precedida por una vocal hasta el final de la palabra.
- R2: región que va desde la primera no vocal precedida por una vocal en R1.
- RV: Si la segunda letra es una consonante, RV es la región que va desde la siguiente vocal hasta el final de la palabra. Si las dos primeras letras son vocales, RV es la región que va desde la siguiente consonante hasta el final de la palabra. Para el caso contrario (caso consonante-vocal), RV es la región que va desde la tercera letra hasta el final de la palabra.

Con respecto a los pasos que se deben seguir para el proceso de Porter Stemming, estos se muestran en los Algoritmos 5, 6, 7 y 8 (Barrenechea, 2006):

1- Obtener el sufijo de mayor tamaño entre los sufijos: me, se, sela, selo, selas, selos, la, le, lo, las, les, los, nos, te, telo, melo, telos, melos, tela, mela, telas, melas

2- Si sufijo encontrado en RV entonces

3- Si sufijo se encuentra después de: iéndo, ándo, ár, ér, ír, ando, iendo, ar, er, ir y (u)yendo en R1entonces

4- Eliminar sufijo

5- Fin si

6- Fin si

Algoritmo 5 Primer paso del Algoritmo de Porter Stemming para la transformación de sufijos (Elaboración propia).

1- Obtener el sufijo de mayor tamaño entre los sufijos: anza, anzas, ico, ica, icos, icas, ismo, ismos, able ables, ible, ibles, ista, istas, oso, osa, osos, osas, amiento, amientos, imiento, imientos, icadora, icador, icacion, icadoras, icadores, icaciones, icante, icantes, icancia, icancias, adora, ador, acion, adoras, adores, aciones, ante, antes, ancia, ancias, logía, logías, ución, uciones, encia, encías, ativamente, ivamente, osamente, icamente, adamente, amente, antemente, ablemente, iblemente, mente, abilidad, abilidades, icidad, icidades, ividad, ividades, idad, idades, ativa, ativo, ativas, ativos, iva, ivo, ivas, ivos.

2- Si sufijo encontrado entonces

3- Si sufijo encontrado se encuentra en R2 entonces

4- Si sufijo eliminado es igual a logía o logías entonces

5- Reemplazar prefijo con log

6- De lo contrario si sufijo eliminado es igual a ución o uciones entonces

7- Reemplazar prefijo con u

8- De lo contrario si sufijo eliminado es igual a encia o encías entonces

9- Reemplazar prefijo con ente

10- Sino

11- Eliminar sufijo

Algoritmo 6 Segundo paso del Algoritmo de Porter Stemming para la transformación de sufijos (Elaboración propia).

1- Si en el Paso 2 no se realizó modificaciones en RV entonces

2- Obtener el sufijo de mayor tamaño entre los sufijos: ya, ye, yan, yen, yeron, yendo, yo, yo, yas, yes, yais, yamos

3- Si sufijo encontrado entonces

4- Si sufijo se encuentra después de: u entonces

5- Eliminar sufijo

6- Fin si

7- Fin si

8- Si no se eliminó sufijo entonces

9- Obtener el sufijo de mayor tamaño entre los sufijos: arıan, arias, aran, aras, arıais, aria, areis, arıamos, aremos, ara, are, erıan, erıas, eran, eras, erıais, erıa, ereis, erıamos, eremos, era, ere, irıan, irıas, iran, iras, irıais, irıa, ireis, irıamos, iremos, ira, ire, aba, ada, ida, ıa, ara, iera, ad, ed, id, ase, iese, aste, iste, an, aban, ıan, aran, ieran, asen, iesen, aron, ieron, ado, ido, ando, iendo, io, ar, er, ir, as, abas, adas, idas, ıas, aras, ieras, ases, ieses, ıs, ais, abais, ıais, arias, ierais, aseis, ieseis, asteis, isteis, ados, idos, amos, abamos, ıamos, imos, aramos, ieramos, iesemos, asemos, en, es, ´eis, emos, guen, gues, gueis, guemos

10- Si sufijo encontrado en RV entonces

11- Si sufijo encontrado es guen o gues o gueis o guemos entonces

12- Reemplazar sufijo con g

13- Sino

14- Eliminar sufijo

15- Fin si

16- Fin si

Algoritmo 5 Tercer paso del Algoritmo de Porter Stemming para la transformación de sufijos (Elaboración propia).

1- Obtener el sufijo de mayor tamaño entre los sufijos: os , a , o , á , í , ó , e , é , ue , ué

2- Si sufijo encontrado en RV entonces

3- Si sufijo encontrados es igual a e o é entonces

4- Si sufijo se encuentra después de: gu y la u en RV entonces

5- Eliminar u

6- Fin si

7- Sino

8- Eliminar sufijo

Algoritmo 6 Cuarto paso del Algoritmo de Porter Stemming para la transformación de sufijos (Elaboración propia).

4.4.5 Filtrado de stems con baja frecuencia

Como paso final dentro de la etapa de pre procesamiento y como técnica adicional para la reducción de la dimensionalidad de los datos, se realizó un proceso de filtrado sobre los stems obtenidos del proceso de tokenización y normalización. Esto permitió la eliminación de aquellos stems con poca representatividad dentro del conjunto de datos. La representatividad se midió a través de la frecuencia de aparición de un stem dentro del conjunto de publicaciones ALICIA.

Es así que, se calculó el promedio de veces que un stem apareció a lo largo del conjunto de publicaciones ALICIA, y fue a partir de esto que se eliminó aquellos stems con menor frecuencia a la del promedio.

La aplicación de esta técnica de filtrado adicional permitió la reducción de aquellos términos muy específicos y que no ayudan dentro del proceso de conceptualización en el proceso de recomendación.

```
1- Inicialización de la variable frecuencia del tipo Diccionario
2- Para cada uno de las publicaciones ALICIA:
    3- Para cada token de la publicación ALICIA:
        4- Incrementar la frecuencia del token en 1
    5- Fin para
6- Fin para
7- Obtener promedio de frecuencias
8- Para cada uno de las publicaciones ALICIA:
    9- Para cada token de la publicacion ALICIA:
        10- Si frecuencia del token < promedio:
                11- Eliminar token
        12- Fin si
    13- Fin para
```

Algoritmo 7 Algoritmo para el filtrado de Stems de poca frecuencia (Elaboración propia).

4.5 Implementación de los módulos de Extracción y de Pre procesamiento

El proceso de pre procesamiento se muestra como un proceso dependiente del proceso de extracción, así mismo las tareas ejecutadas en cada uno de los módulos describen un flujo de trabajo continuo donde el resultado final será un registro (publicación científica de Alicia, publicación Scopus u Orcid de investigador) pre procesado y listo para ser usado en el entrenamiento del modelo de recomendación.

Como parte de un flujo de trabajo ágil en el campo del análisis de datos, el uso de tuberías para la automatización, manejo de errores, parametrización, y ejecución de

pruebas aparece como una tecnología fácil de implementar («Using Luigi Pipelines in a Data Science Workflow», s. f.).

Para el Proyecto, se hizo uso de la librería Luigi que permite construir tuberías de procesos, manejando resolución de dependencias, flujo de trabajo y errores. Fue así que, se definieron 2 tareas, una para la extracción del registro y otra para el procesamiento de ese registro. Estas tareas corresponden a los módulos tanto de Extracción como el de Pre procesamiento.

Por otro lado, en relación a la ejecución del sistema, la implementación de un patrón de diseño basado en tareas permite que el sistema de recomendación pueda ser ejecutado manualmente a demanda a través de la interacción con una Interfaz Gráfica de Ejecución, o también programado para su ejecución automática como proceso batch.

4.5.1 Interfaz Gráfica de Ejecución

Como parte del desarrollo del sistema de recomendación, se desarrolló una Interfaz Gráfica de Ejecución, de manera que la ejecución del sistema pueda ser a demanda.

Esta interfaz simula el comportamiento de una consola de log, con el fin de permitir el seguimiento en tiempo real del estado de los procesos que se están ejecutando. En la Ilustración 13 se muestra la interfaz implementada. Esta cuenta con tres botones:

- Comenzar: este botón da inicio a la ejecución del sistema de recomendación.
- Limpiar Log: este botón limpia la consola, borra el contenido que se muestra.
- Parar/Empezar Log: este botón detiene la visualización del log en la pantalla. Así mismo, el botón play permite retomar la visualización del log en la pantalla.

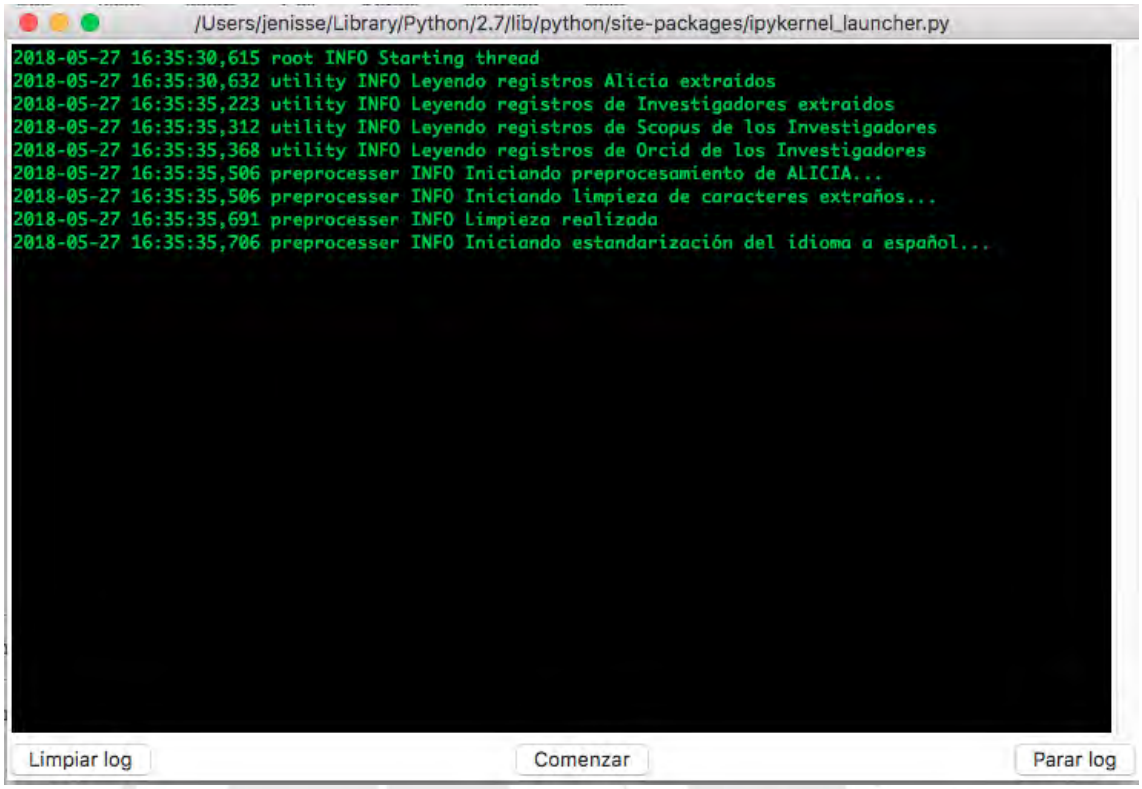

Ilustración 13 Interfaz Gráfica de Ejecución del sistema de recomendación. Pantalla que permite la interacción con el usuario para el inicio de ejecución del sistema (Elaboración propia).

Capítulo 5. Modelo de Recomendación

En este capítulo se muestra cómo se desarrolló el segundo objetivo específico, el cual consiste en la implementación del modelo de recomendación de las publicaciones científicas de ALICIA para los investigadores calificados del SINACYT.

5.1 Selección del enfoque de recomendación

Para la selección del enfoque adecuado se tuvo en cuenta las siguientes consideraciones:

- ¿Qué enfoque se adecua mejor en la recomendación de elementos cuyo contenido es principalmente texto?
- ¿Qué enfoque se adecua mejor en la recomendación de elementos si no se tiene la relación de puntuación-ítem-usuario?

Bajo estas consideraciones se pudo seleccionar al Filtrado Basado en Contenido, como el enfoque que mejor se adapta dentro del contexto de recomendación de literatura científica.

5.2 Modelo de recomendación

Ya definido el enfoque adecuado para el sistema de recomendación, se continúa con la implementación del mismo.

Previo a la implementación del enfoque de recomendación para el componente del mismo nombre, se utilizó la técnica de LSA (Latent Semantic Analysis) con el propósito de poder descubrir estructuras semánticas escondidas sobre un conjunto de documentos.

Para la obtención del LSA, se tuvo que crear un corpus a partir de cada uno de los registros de las publicaciones científicas, y a su vez una matriz TF-IDF.

Finalmente, se construyó la matriz de similitud utilizando la técnica similitud coseno y el modelo LSA obtenido previamente.

En la Ilustración 14, se muestra gráficamente el proceso anteriormente mencionado, con mayor detalle:

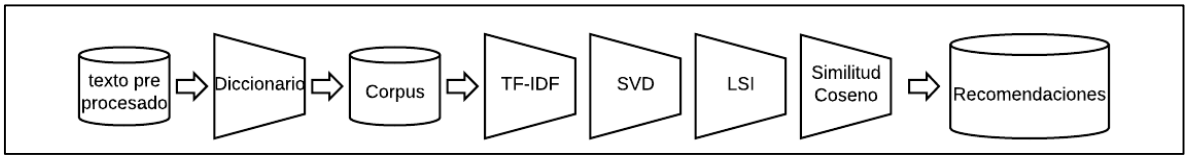

Ilustración 14 Etapas dentro del modelo de recomendación. (Elaboración propia)

5.3 Implementación del módulo de recomendación

El módulo de recomendación implementa las funciones necesarias para la ejecución de cada una de las etapas para la obtención del modelo de recomendación propuesto. Las funciones que se implementaron fueron las siguientes:

1- Reducción de dimensionalidad

2- Obtención del Corpus

3- Obtención del Corpus bajo TF-IDF

4- Obtención del Modelo LSA

5- Obtención de la Matriz de Similitud

6- Calculo de Recomendaciones

Estas funciones utilizan algunos de los métodos implementados en la librería Gensim.

Por otro, lado, la ejecución del módulo de recomendación depende de la finalización exitosa del flujo de trabajo integral de la tubería compuesta por los módulos de Extracción y Pre procesamiento. La razón de esta dependencia es la necesidad de utilizar toda la data pre procesada en la creación del modelo de recomendación. Este módulo toma como datos de entrada el archivo generado por el módulo Batch.

Finalmente, este módulo termina su ejecución con el almacenamiento de las recomendaciones generadas en la base de datos utilizada por el sistema.

5.3.1 Reducción de dimensionalidad

Como paso previo a la construcción del corpus, se aplicó un método de reducción de la dimensión de la data pre procesado. Esto, con el fin de eliminar aquellos términos muy específicos y particulares.

La poca representatividad del término sobre el conjunto de los mismos se relaciona con su característica de particularidad, la cual no influye en el proceso de obtención de dimensiones conceptuales.

Para el proyecto se definió eliminar aquellos términos que solo aparecieran una vez en el conjunto de términos. En el Algoritmo 10 se definen los pasos para la reducción de dimensionalidad.

```
1- Inicializar variable frecuencia de tokens
2- Para documento en lista de documentos
        3- Para token en lista de tokens
                4- Incrementar frecuencia de token en 1
        5- Fin para
6- Fin para
7- Para documento en lista de documentos
        8- Si frecuencia de token < 2 hacer
                9 Eliminar token
        10- Fin si
11- Fin para
```

Algoritmo 8 Algoritmo para la reducción de dimensionalidad (Elaboración propia).

5.3.2 Obtención del Corpus

El corpus se define como una colección de textos producidos dentro de contextos reales de aplicación de la lengua, los cuales son seleccionados y ordenados bajo una serie de criterios lingüísticos, de forma que aseguren su utilización como muestra representativa de una lengua determinada.(«What is a corpus?», 2018).

Para el Proyecto, el corpus fue obtenido a partir del uso de funciones de la librería Gensim de Python. Primero, se construyó el diccionario de palabras que consiste en el

conjunto de palabras únicas existentes que representan los datos a analizar, en este caso los títulos y descripciones de las publicaciones de ALICIA y de los investigadores.

En la Ilustración 15 se muestra algunos de los términos pertenecientes al diccionario de datos creado.

```
list(itertools.islice(dictionary.token2id.items(), 121, 127))
```

```
[(u'desinfecc', 6728),
 (u'termodinamic', 8559),
 (u'manteng', 6183),
 (u'adquirir', 5693),
 (u'catast', 9449),
 (u'infus', 3777)]
```

Ilustración 15 Algunos de los términos (pre procesados) que conforman el diccionario de datos generado. Se muestra el término junto con su identificador único (Elaboración propia).

A partir del diccionario, se entrenó un corpus, el cual es la colección de los documentos obtenidos luego de ejecutado el pre procesamiento. Este corpus contiene los títulos y descripciones representadas como vectores dispersos, con la información sobre la palabra y el número de veces que esta aparece dentro del documento al que pertenece.

En la Ilustración 16 se muestra algunos de los elementos correspondientes al corpus generado.

```
1 corpus[0]

[(0, 7),
 (1, 1),
 (2, 1),
 (3, 3),
 (4, 1),
 (5, 8),
 (6, 1),
 (7, 14),
 (8, 1),
 (9, 1),
```

```
3 dictionary[7]

'colecistectom'
```

Ilustración 16 Corpus correspondiente a una publicación ALICIA. El término 'colecistectom' aparece 14 veces para la primera publicación ALICIA del corpus (Elaboración propia).

5.3.3 Obtención del corpus bajo TF-IDF

El uso de TF-IDF para el presente Proyecto tiene como finalidad la de reflejar la importancia de los términos en el corpus. La idea en la aplicación de esta técnica es la de otorgar una medida de relevancia a diferentes términos. Así, el TF-IDF asocia un peso bajo a los términos que aparecen con frecuencia en el corpus y aumenta el peso de los términos que rara vez aparecen.

Bajo la aplicación de este método, el corpus entrenado define características existentes en las publicaciones científicas (términos con alto grado de relevancia). Se utilizó la función "TfidfModel" de la librería Gensim, la cual reemplaza el valor del contador asociado a cada término por el peso TFIDF obtenido luego de aplicada la técnica.

En la Ilustración 17, se muestra el resultado de los cálculos TF-IDF sobre el corpus previamente generado.

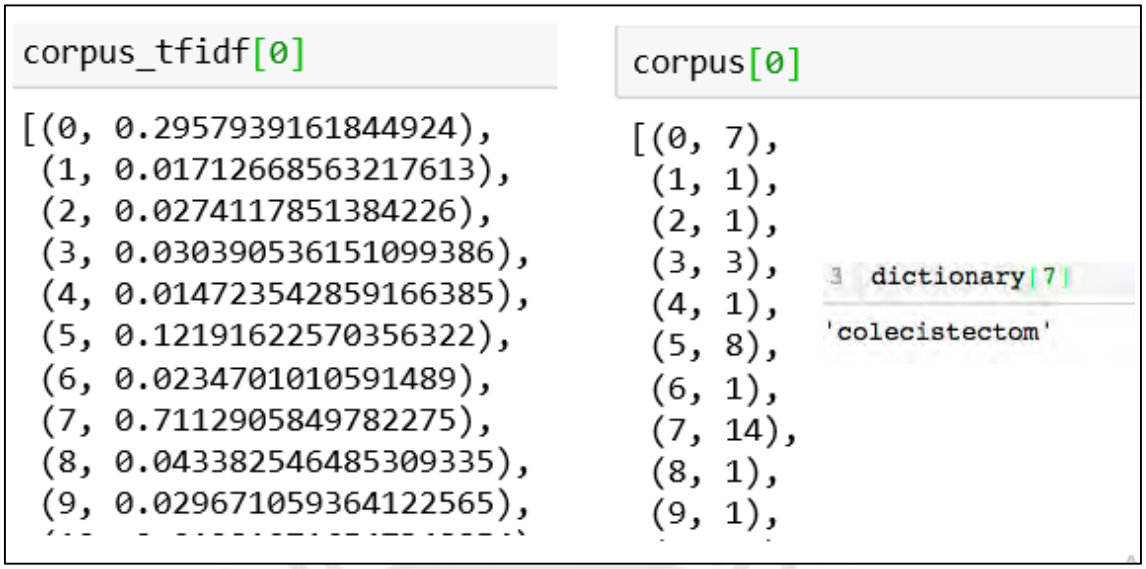

Ilustración 17 Corpus bajo TF-IDF vs Corpus simple correspondiente a una publicación ALICIA. El término 'colecistectom' ahora está representado por el valor numérico obtenido luego del cálculo de TF-ID (0.71129). Específicamente, para este ejemplo se puede ver como el término 'colecistectom' tiene mayor peso sobre los demás y posee un alto grado de relevancia para la primera publicación ALICIA (Elaboración propia).

5.3.4 Obtención del modelo LSA

La aplicación del método de Descomposición de Valores Singulares (SVD) tiene como finalidad la de encontrar una representación dimensional reducida de la matriz del corpus, enfatizando las relaciones más fuertes y desechando el ruido. En consecuencia, se efectúa mejor la reconstrucción de la matriz con la menor información posible (X. Zhang, Tang, Zhang, & Ji, 2016).

Así mismo, el modelo LSA busca identificar un conjunto de temas relacionados con las representaciones de las publicaciones científicas. El número de estos temas es igual a la dimensión de la matriz de aproximación resultante de la técnica de SVD. Este valor se obtiene de la selección de los N mayores valores singulares de la matriz del corpus (número de dimensiones). Así, LSA genera un espacio de rango reducido donde se pueden realizar comparaciones en distintos niveles conceptuales.

Para la construcción del modelo de dimensiones latentes, se utilizó la función LsiModel de la librería Gensim, la cual transforma al corpus TF-IDF en un espacio latente de un determinado número de dimensiones.

En relación al número de dimensiones, se tiene conocimiento que a 300 dimensiones las correlaciones entre el análisis semántico latente y los juicios de similitud de texto por parte de humanos, son empíricamente más altas (Landauer, Laham, & Derr, 2004). Se pueden usar menos dimensiones para comparaciones amplias (más conceptuales), mientras que utilizando una mayor cantidad de dimensiones para comparaciones específicas (más literales).

En la Ilustración 18 se muestran 5 de los temas generados por el modelo LSA.

```
Dimension #0
0.147*"gest" + 0.140*"estudy" + 0.130*"empres" + 0.116*"nivel" + 0.111*"aprendizas" + 0.104*"proceso" + 0.104*"lab" +
0.102*"instituc" + 0.100*"calidad" + 0.098*"organizac"
Dimension #1
-0.258*"estudy" + -0.248*"aprendizas" + 0.211*"empres" + -0.171*"doc" + -0.161*"educativ" + -0.149*"educac" + -0.145*
"instituc" + 0.143*"sistem" + -0.129*"secundar" + 0.125*"gest"
Dimension #2
-0.288*"aprendizas" + -0.213*"curso" + 0.196*"pacy" + -0.176*"estudy" + 0.173*"lab" + 0.164*"satisfacc" + -0.160*"mat
ematic" + 0.150*"salud" + 0.147*"hospit" + 0.124*"enfermer"
Dimension #3
-0.399*"lab" + -0.339*"clim" + -0.297*"organizac" + -0.227*"desempeno" + -0.219*"satisfacc" + -0.165*"trabajad" + -0.
144*"doc" + 0.144*"pacy" + 0.142*"nino" + -0.117*"empres"
Dimension #4
0.527*"curso" + 0.259*"silabo" + 0.257*"silab" + 0.215*"derecho" + -0.158*"aprendizas" + 0.148*"naturalez" + 0.127*"t
ema" + 0.118*"conocimiento" + 0.117*"lab" + 0.105*"profes"
Dimension #5
0.223*"pacy" + -0.199*"derecho" + 0.194*"enfermer" + 0.173*"satisfacc" + 0.173*"gest" + 0.167*"servicio" + 0.166*"apr
endizas" + 0.154*"calidad" + 0.148*"hospit" + 0.147*"curso"
```

Ilustración 18 Temas generados por el modelo LSA. Solo se muestran 5 de los 300 temas generados. Cada tema cuenta con tuplas de valor numérico-termino. Los valores numéricos representan el aporte de similitud que el término tiene sobre la dimensión (tema). Los valores negativos señalan disimilitud (la ocurrencia del concepto semántico acompaña la ausencia de la palabra dentro de la dimensión (Elaboración propia).

5.3.5 Calculo de Similitud Coseno

El algoritmo de la similitud coseno se implementa a través del uso de la función MatrixSimilarity de la librería Gensim, la cual calcula la similitud coseno contra un corpus de documentos almacenando la matriz de índice en memoria.

Esta matriz será utilizada para el posterior cálculo de recomendaciones.

5.3.6 Cálculo de Recomendaciones

Para la generación de recomendaciones, se implementó una función donde se evalúa cada una de las publicaciones de los investigadores tanto en Orcid como en Scopus. Cada una de estas publicaciones (en formato documento) se las vectoriza dentro del espacio LSI para luego aplicar un vector de similitud con relación al corpus. Luego, se seleccionan aquellos documentos con un nivel de similitud mayor o igual al definido en la función. En el Algoritmo 9 se muestran los pasos para el cálculo de recomendaciones.

```
1- Para cada investigador en investigadores
    2- Para cada publicación orcid en publicaciones orcid del investigador
        3- Convertir la publicación orcid al espacio LSA
        4- Realizar la comparación (similitud) de la publicación orcid en
        espacio LSA con el corpus para obtener similitudes
        5- Inicializar lista de recomendaciones vacía
        6- Para cada publicación Alicia en publicaciones Alicia
            7- Si grado de similitud de publicación orcid con publicación
            Alicia > NIVEL
                8- Agregar publicación Alicia a lista de recomendación
            9- Fin si
        10- Fin para
        11- Guardar lista de recomendación para publicación orcid
    11- Fin para
    12- Para cada publicación scopus en publicaciones scopus del investigador
        13- Convertir la publicación scopus al espacio LSA
        14- Realizar la comparación (similitud) de la publicación scopus en
        espacio LSA con el corpus para obtener similitudes
        15- Inicializar lista de recomendaciones vacía
        16- Para cada publicación Alicia en publicaciones Alicia
            17- Si grado de similitud de publicación scopus con publicación
            Alicia > NIVEL
                18- Agregar publicación Alicia a lista de recomendación
            19- Fin si
        20- Fin para
```

Algoritmo 9 Algoritmo para el cálculo de recomendaciones (Elaboración propia).

Finalmente, las recomendaciones generadas son guardadas en la Base de Datos del sistema.

Capítulo 6. Implementación del Servicio de Recomendación

En este capítulo se muestra cómo se desarrolló el tercer objetivo específico del presente Proyecto, el cual consiste en la implementación del servicio web para la publicación de las recomendaciones a partir de una arquitectura definida para la solución de recomendación.

6.1 Servicio Web REST

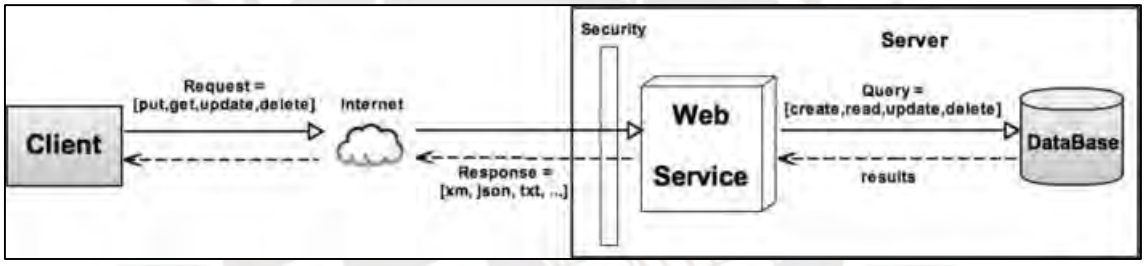

Ilustración 19 Esquema de la arquitectura de un servicio web REST (P Waller, Dresselhaus, & Yang, 2013)

El servicio web REST es un estilo arquitectónico, donde los datos o los componentes estructurales de un sistema se describen en forma de URI (identificador uniforme de recursos) y los comportamientos se describen en términos de métodos (Potti, 2011). Los recursos se pueden manipular usando operaciones CRUD (Crear, Leer, Actualizar y Eliminar). En la ilustración 19 se muestra la arquitectura seguida para el despliegue del servicio.

Para el Proyecto, se definieron los siguientes recursos:

- GET/recommendation/{#id_investigador} → La cual devolverá el detalle de alguna recomendación (publicación de ALICIA) generada para el investigador que posea el id especificado. La elección de la recomendación se realiza de manera aleatoria.

- GET/recommendations/{#id_investigador} → La cual devolverá la lista de recomendaciones (publicaciones de ALICIA) generadas para el investigador que posea el id especificado.

De esta forma los clientes que consuman el servicio estarán interactuando con el sistema de recomendación a través de la consulta de las recomendaciones almacenadas en la base de datos del sistema.

Ilustración 20 Ejemplo de un cliente android, que hace uso del servicio de recomendación aleatoria. Se debe especificar el id del investigador para hacer uso del servicio (Elaboración propia).

Como parte de la validación del módulo de servicio, este fue desplegado en un servidor de aplicaciones utilizando el marco de software Flask, así como sqalchemy para la

conexión y consultas a la base de datos del sistema, la cual almacena las recomendaciones para cada una de las publicaciones científicas de los investigadores existentes.

De tal manera, se desarrolló un cliente Android para que consumiera el servicio publicado. En la Ilustración 20 se muestra la pantalla con la respuesta enviada por el servicio ante el requerimiento del cliente.

Capítulo 7. Evaluación de Resultados

En este capítulo se muestra cómo se desarrolló el cuarto objetivo específico del presente Proyecto, el cual consiste en la evaluación offline de los resultados (recomendaciones) generados a partir del modelo de recomendación seleccionado.

En el escenario donde a un usuario se le proporciona una lista de recomendaciones en las que puede evaluar los elementos como afines o no, las métricas utilizadas en la recuperación de información como Precisión y Exhaustividad son útiles para evaluar la calidad de un método de recomendación (Sarwar, Karypis, Konstan, & Riedl, 2000).

Por tal motivo, la evaluación se basó en el análisis de las métricas de precisión para las N mejores recomendaciones generadas. Para esto, se realizó una encuesta sobre una pequeña muestra de investigadores con el fin de que puedan calificar a las recomendaciones como afines o no, en relación al ámbito de conocimiento en donde se desenvuelven.

La pregunta fue formulada de manera que pueda calificarse el grado de afinidad del contenido que se recomienda, mas no directamente la utilidad subjetiva percibida por el investigador. Esto debido a que el sistema de recomendaciones propuesto genera recomendaciones a partir de las publicaciones producidas por el investigador y no partir de sus preferencias subjetivas.

7.1 Muestra de la población

Se utilizó el método de muestreo conveniente donde se seleccionaron a 3 investigadores como población para la encuesta. Los investigadores que se seleccionaron forman parte de la comunidad de la Pontificia Universidad Católica del Perú.

Este método de muestreo no probabilístico fue seleccionado ya que se utiliza a menudo durante los esfuerzos de investigación preliminares de manera que se pueda obtener una estimación bruta de los resultados, sin incurrir en el costo o el tiempo requerido para seleccionar una muestra aleatoria (X. Wang, 2010).

7.2 Desarrollo del cuestionario

El cuestionario contó con cuatro claros elementos:

Título: Cuestionario para Calificación de Afinidad de Recomendaciones de Publicaciones Científicas.

Objetivo: Obtener información cuantitativa sobre las recomendaciones generadas por el sistema de recomendación de publicaciones científicas de ALICIA a investigadores calificados del SINACYT.

Supuesto: Afinidad de las recomendaciones en relación a su pertenencia al ámbito de conocimiento en donde Ud., (investigador) se desenvuelve.

Pregunta: ¿Cuán afín son las siguientes 10 publicaciones al ámbito de investigación en la cual Ud. se desenvuelve?

Opciones de Respuesta: Completamente afín - Muy afín - Medianamente afín - Poco afín - Nada afín. Estas opciones fueron seleccionadas a partir de la escala de Lickert (Norman, 2010).

Así mismo, junto con el cuestionario se preparó un Protocolo de Consentimiento informado para la Calificación de Recomendaciones de Publicaciones Científicas, la cual tenía como finalidad brindar a los participantes de la investigación una explicación clara de la naturaleza de la misma, así como el rol que tienen en ella. Este protocolo siguen los lineamientos propuestos en el formulario de C.I. del Comité de Ética del Departamento de Psicología de la PUCP (Espinoza & Alberto, 2018).

En el Anexo 2, se muestra el modelo de Protocolo de Consentimiento Informado que se utilizó para la investigación. Mientras que en el Anexo 3 se muestra el modelo de cuestionario que se utilizó para la calificación de las recomendaciones generadas por el sistema de recomendación propuesto.

7.3 Implementación de la encuesta

Para la implementación de la encuesta se adoptó el Método de diseño a medida (Faubion & Andrew, 2001). Este método muestra alta tasas de respuestas (X. Wang, 2010).

Para el proceso de implementación de la encuesta se siguieron los siguientes pasos:

1- El primer contacto fue un correo electrónico de aviso señalando el desarrollo de la encuesta y la selección de la persona como parte de la muestra de la encuesta.

2- Dos días después de enviado el aviso, se envió el cuestionario explicando el porqué de su utilidad.

3- Una semana después de enviado el cuestionario se envió un correo electrónico de recordatorio para el llenado de la encuesta.

7.4 Modelo de medición y análisis de resultados

En base al cuestionario realizado se pudo obtener los datos para la obtención de las métricas de:

- Precision@N: la proporción de las N primeras recomendaciones afines en relación al ámbito de conocimiento del investigador

  Donde N = 5,10

Para la medición de las métricas de precisión se siguió una lógica en específico, la cual se muestra en el Algoritmo 10.

1- Para N en lista de Ns para evaluar

2- Inicializar acumulador en 0

2- Para investigador en lista de investigadores

3- Obtener lista de publicaciones relevantes para el investigador (resultado de la encuesta)

4- Obtener lista de recomendación para el investigador a través del uso del modelo

5- Obtener cantidad de publicaciones que se encuentran tanto en la lista de recomendación así como en la lista de publicaciones relevantes (verdaderos positivos)

6 – Obtener cantidad de publicaciones de la lista de recomendación que no se encuentran en las publicaciones relevantes (falsos positivos)

7- Obtener precisión (verdaderos positivos) / ((verdaderos positivos + falsos positivos))

8- Actualizar acumulador precisión (acumulador <- acumulador + precisión)

9- Fin para

Algoritmo 10 Algoritmo para la obtención de la métrica de Precisión en N (Elaboración propia).

Donde los falsos positivos están representados por las calificaciones: Poco afín y Nada afín. Mientras que los verdaderos positivos son aquellos resultados con las calificaciones: Completamente afín, Muy afín y Medianamente afín.

De la evaluación (Ver Anexo 4) se obtuvo que para la precisión en N = 5, se obtuvieron un total de 3 recomendaciones sin afinidad al área de investigación en la que se desenvuelven y 12 recomendaciones afines. Así mismo, los resultados para la precisión en N=5, fueron de 8 recomendaciones sin afinidad a su área de investigación y 22 recomendaciones afines. El cálculo de la precisión para valores de N iguales a 5 y 10 se muestra en la Ilustración 21.

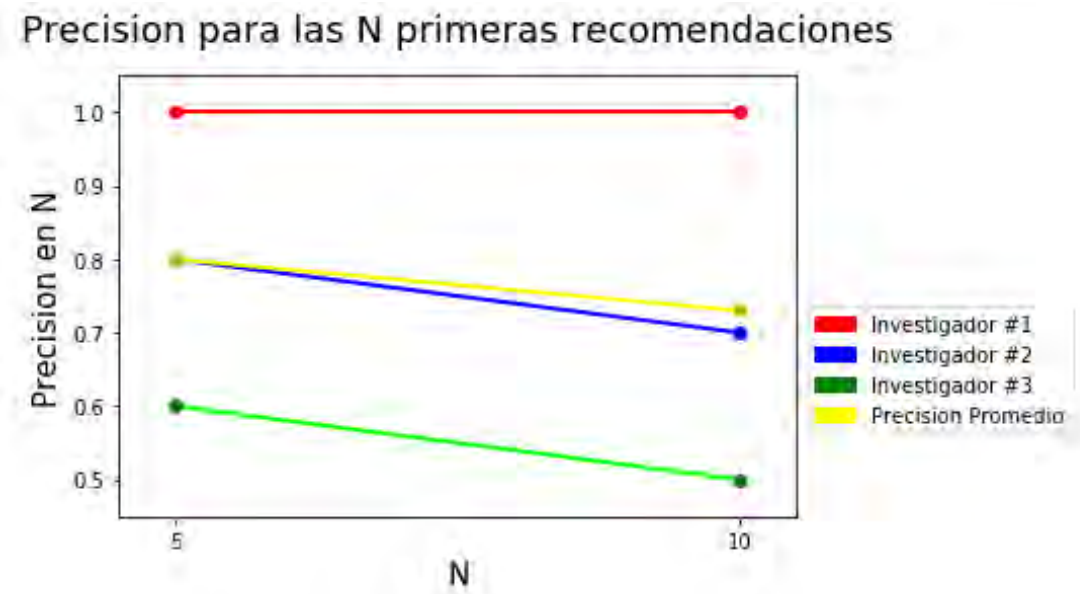

Ilustración 21 Cuadro de Precisión en N. Se muestra la precisión obtenida para cada investigador para N=5 y N=10, así como, la precisión promedio (Elaboración propia).

Se puede ver que el incremento en N impacta en la disminución de la precisión promedio. A mayor N, el valor de la precisión en N irá en descenso, mientras que el valor de la exhaustividad en N irá en aumento. Así se demuestran las características de proporcionalidad de N con las métricas de precisión y exhaustividad (Bondi, 2018).

Capítulo 8.Conclusiones y trabajos futuros

8.1 Conclusiones

En este apartado se presentan las conclusiones más significativas halladas a lo largo del desarrollo de presente Proyecto. A continuación, se muestran estas conclusiones:

1) Se logró desarrollar una solución que permita la visibilidad de las publicaciones científicas de ALICIA para con los investigadores, correspondientes a la misma área conceptual de su producción científica. Como consecuencia, se podrá ayudar a dar a conocer variantes de temas dentro de líneas de investigación.
2) Se construyó el perfil del investigador a través de la identificación de su producción científica en plataformas como ORCID y Scopus, permitiendo mayor información sobre el dominio de investigación al que pertenece.

3) Se logró la generación de recomendaciones personalizadas teniendo en cuenta la similitud (conceptual) entre la producción científica en plataformas como ORCID y Scopus correspondiente a los investigadores y las publicaciones científicas almacenadas en ALICIA. El uso del Modelo de Análisis Semántico Latente permitió la obtención de resultados que no se sesgaban en la simple similitud de las publicaciones científicas, sino a su similitud con los temas (conceptos) identificados luego de entrenado el modelo.
4) Se implementó una arquitectura basada en el modelo de programación de tuberías, lo que represento una reducción en el tiempo de procesamiento mediante la ejecución de varios procesos en paralelo. La gran cantidad tanto de publicaciones científicas de ALICIA como de publicaciones ORCID y Scopus correspondientes a los investigadores, justificó la implementación de un proceso batch junto con el uso de una tubería que relacionaba el proceso de extracción con el proceso de pre procesamiento.
5) Se logró tener una precisión promedio aceptable, lo que representa la correcta generación de recomendaciones teniendo en cuenta la afinidad de estas con la línea de investigación de los investigadores.

8.2 Trabajos futuros

Como trabajo futuro posible se propone la implementación de mecanismos para la interacción de los usuarios con las recomendaciones de publicaciones científicas generadas por el sistema de recomendación propuesto. De esta manera, se podría recabar información sobre las preferencias del investigador y/o patrones de comportamiento del mismo. Esto implicaría el aprendizaje del modelo de recomendación no solo por la descripción de las publicaciones a recomendar, sino también por la interacción de los usuarios con las publicaciones recomendadas.

Como mecanismos de interacción se proponen, la creación de un sistema web donde cada uno de los usuarios (investigadores) cuente con una cuenta para que así el sistema de recomendación pueda obtener la información generada directamente por el usuario (calificación online de las recomendaciones o CTR: número de clicks realizados sobre una recomendación o tiempo de lectura de la recomendación) y la publicación de más

servicios relacionados al envío de la información generada en las sesiones iniciadas por el usuario.

Relacionado a la propuesta de implementación de un módulo web de interacción de usuarios investigadores, se propone también, la implementación de un enfoque de recomendación complementario al ya propuesto en el Proyecto. Esto con la finalidad de generar recomendaciones con un alto grado de personalización, tomando en cuenta los gustos e intereses específicos de usuarios. Por ejemplo, la implementación de un enfoque de filtrado colaborativo basado en la interacción de usuario, de manera que se pueda aprovechar la información obtenida a través de los mecanismos de interacción mencionados anteriormente.

Asi mismo, a partir de la implementación de mecanismos de interacción, se propone también, la creación de un módulo de evaluación online adicional a los módulos de extracción, pre procesamiento, recomendación y de servicios del sistema de recomendación propuesto en el Proyecto.

Finalmente, se propone la ampliación del dominio de publicaciones científicas a recomendar. Esto es, evaluar también publicaciones científicas de distintos repositorios a nivel mundial, con la intención de incentivar la creación de redes de investigación internacionales.

Referencias

7.9. unicodedata — Unicode Database — Python 2.7.15 documentation. (2018). Recuperado 28 de mayo de 2018, de https://docs.python.org/2/library/unicodedata.html#unicodedata.normalize

ALICIA. (2017, septiembre 11). Recuperado 11 de septiembre de 2017, de https://portal.concytec.gob.pe/index.php/informacion-cti/alicia

Atamari-Anahui, N., & Díaz-Vélez, C. (2015). Repositorio Nacional Digital de Acceso Libre (ALICIA): oportunidad para el acceso a la información científica en el Perú.

Anales de la Facultad de Medicina, *76*(1), 81-82. https://doi.org/10.15381/anales.v76i1.11081

Bancu, C., Dagadita, M., Dascalu, M., Dobre, C., Trausan-Matu, S., & Florea, A. M. (2012). ARSYS – Article Recommender System. *2012 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 349-355. https://doi.org/10.1109/SYNASC.2012.38

Barla, M. (2010). *Towards Social-based User Modeling and Personalization*.

Barrenechea, D. D. P. (2006). *A Spanish Stemming Algorithm Implementation in PROLOG and C#*.

Bean, M. (2016). A Framework for Evaluating Recommender Systems. *All Theses and Dissertations*. Recuperado de https://scholarsarchive.byu.edu/etd/6195

Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation. (s. f.). Recuperado 27 de mayo de 2018, de https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Beel, J., Dinesh, S., Mayr, P., Carevic, Z., & Raghvendra, J. (2017, abril 1). *Stereotype and Most-Popular Recommendations in the Digital Library Sowiport*.

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2015). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, 1-34. https://doi.org/10.1007/s00799-015-0156-0

Beel, J., Langer, S., Genzmehr, M., & Nürnberger, A. (2013). Introducing Docear's research paper recommender system. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, 459-460. Indianapolis, Indiana, USA: ACM.

Bell, C., & Jones, K. P. (2018). Towards everyday language information retrieval systems via minicomputers. *Journal of the American Society for Information Science*, *30*(6), 334-339. https://doi.org/10.1002/asi.4630300606

Bidgoli, H. (2010). *The Handbook of Technology Management, Supply Chain Management, Marketing and Advertising, and Global Management*. John Wiley & Sons.

Bondi, L. (2018). *Information Retrieval Evaluation of IR systems*. 30.

Çano, E., & Morisio, M. (2017). *Hybrid Recommender Systems: A Systematic Literature Review* (Vol. 21). https://doi.org/10.3233/IDA-163209

CONCYTEC pone a disposición nueva plataforma virtual DINA para investigadores, innovadores y profesionales. (s. f.). Recuperado 17 de junio de 2018, de https://portal.concytec.gob.pe/index.php/noticias/289-concytec-pone-a-disposicion-nueva-plataforma-virtual-dina-para-investigadores-innovadores-y-profesionales

Congreso de la República del Perú. *Ley Marco de Ciencia, Tecnología e Innovación Tecnológica*. , Pub. L. No. Ley N° 28303 (2004).

Congreso de la República del Perú. *Ley que Regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto*. , Pub. L. No. Ley N° 30035 (2013).

Córdoba, S. (2011, febrero 16). La UCR y el movimiento de acceso abierto. Recuperado 27 de mayo de 2018, de Semanario Universidad website: https://semanariouniversidad.com/opinion/la-ucr-y-el-movimiento-de-acceso-abierto/

Dolgert, D. (s. f.). *Scripting for Data Analysis*. 26.

Dong, R., Tokarchuk, L., & Ma, A. (2009). *Digging Friendship: Paper Recommendation in Social Network*. 7.

DSpace: un manual específico para gestores de la información y la documentación. (s. f.). Recuperado 17 de junio de 2018, de http://bid.ub.edu/20rodri2.htm

Espinoza, P., & Alberto, F. (2018). Método para la evaluación de usabilidad de sitios web transaccionales basado en el proceso de inspección heurística. *Pontificia Universidad Católica del Perú*. Recuperado de http://tesis.pucp.edu.pe/repositorio//handle/123456789/9903

Faubion, C. W., & Andrew, J. D. (2001). Book Review: Dillman, D. A. (2000). Mail and Internet Surveys: The Tailored Design Method (2nd ed.). New York: Wiley 464 pp., $47.50 (hardcover). *Rehabilitation Counseling Bulletin*, *44*(3), 178-180. https://doi.org/10.1177/003435520104400309

Ferrara, F., Pudota, N., & Tasso, C. (2011). A Keyphrase-Based Paper Recommender System. *Digital Libraries and Archives*, 14-25. https://doi.org/10.1007/978-3-642-27302-5_2

Ferreras-Fernández, T., García-Peñalvo, F. J., & Merlo-Vega, J. A. (2015). Open Access Repositories As Channel of Publication Scientific Grey Literature. *Proceedings of the 3rd International Conference on Technological Ecosystems for Enhancing Multiculturality*, 419–426. https://doi.org/10.1145/2808580.2808643

Filho, G. A. L., & Siqueira, R. L. (2008). REVISTA CONTABILIDADE & FINANÇAS USP: UMA ANÁLISE BIBLIOMÉTRICA DE 1999 A 2006. *Revista de Informação Contábil*, *1*(2). Recuperado de https://periodicos.ufpe.br/revistas/ricontabeis/article/view/7736

Gipp, B., Beel, J., & Hentschel, C. (2009). *Scienstein: A Research Paper Recommender System*.

Gori, M., & Pucci, A. (2006). Research Paper Recommender Systems: A Random-Walk Based Approach. *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 778-781. IEEE Computer Society.

Gunawardana, A., & Shani, G. (s. f.). *A Survey of Accuracy Evaluation Metrics of Recommendation Tasks*. 28.

Haddi, E., Liu, X., & Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *Procedia Computer Science*, *17*, 26-32. https://doi.org/10.1016/j.procs.2013.05.005

Hollink, V., Kamps, J., Monz, C., & de Rijke, M. (2004). Monolingual Document Retrieval for European Languages. *Information Retrieval*, *7*(1/2), 33-52. https://doi.org/10.1023/B:INRT.0000009439.19151.4c

Hull, D. A. (2018). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, *47*(1), 70-84. https://doi.org/10.1002/(SICI)1097-4571(199601)47:1<70::AID-ASI7>3.0.CO;2-#

Inicio de Búsqueda. (s. f.). Recuperado 5 de octubre de 2017, de http://alicia.concytec.gob.pe/vufind/

Jiang, Y., Jia, A., Feng, Y., & Zhao, D. (2012). Recommending Academic Papers via Users' Reading Purposes. *Proceedings of the Sixth ACM Conference on Recommender Systems*, 241–244. https://doi.org/10.1145/2365952.2366004

Jivani, A. G. (2011). *A Comparative Study of Stemming Algorithms Ms .*

Karthikeyan, M., & Aruna, P. (2013). Probability based document clustering and image clustering using content-based image retrieval. *Applied Soft Computing*, *13*, 959–966. https://doi.org/10.1016/j.asoc.2012.09.013

Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management*, *42*(1), 56-73. https://doi.org/10.1016/j.ipm.2004.11.007

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). *Data Preprocessing for Supervised Leaning*. *1*(2), 7.

Kraaij, W., & Pohlmann, R. (1996). Viewing Stemming as Recall Enhancement. *In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 40–48.

La producción científica: un reto en Enfermería. (2013). *Revista Cubana de Enfermería*, *29*(1), 3-4.

Lakiotaki, K., Delias, P., Sakkalis, V., & Matsatsinis, N. F. (2009). User profiling based on multi-criteria analysis: the role of utility functions. *Operational Research*, *9*(1), 3-16. https://doi.org/10.1007/s12351-008-0024-4

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259-284. https://doi.org/10.1080/01638539809545028

Landauer, T. K., Laham, D., & Derr, M. (2004). From paragraph to graph: latent semantic analysis for information visualization. *Proceedings of the National Academy of*

Sciences of the United States of America, *101 Suppl 1*, 5214-5219. https://doi.org/10.1073/pnas.0400341101

Liang, Y., Li, Q., & Qian, T. (2011). Finding Relevant Papers Based on Citation Relations. En H. Wang, S. Li, S. Oyama, X. Hu, & T. Qian (Eds.), *Web-Age Information Management* (pp. 403-414). Springer Berlin Heidelberg.

Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based Recommender Systems: State of the Art and Trends. En F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender Systems Handbook* (pp. 73-105). https://doi.org/10.1007/978-0-387-85820-3_3

Lorenzo Gil, E., Braña Ferreiro, E., & Nieto Caramés, S. (2015). *Estudio de la integración de repositorios en el sistema científico-investigador: alternativas y estado actual*. Recuperado de http://helvia.uco.es/xmlui/handle/10396/12631

Lourenço, C. de A. (2005). Automação de Bibliotecas: Análise da Produção via Biblioinfo (1986-1994) p. 51-63. *Revista ACB*, *2*(2), 51-63. (Levantamento de dados).

luigi: Luigi is a Python module that helps you build complex pipelines of batch jobs. It handles dependency resolution, workflow management, visualization etc. It also comes with Hadoop support bu.. [Python]. (2018). Recuperado de https://github.com/spotify/luigi (Original work published 2012)

Luque, P., & M, A. (2009). Preservación documental en repositorios institucionales. *Investigación bibliotecológica*, *23*(49), 241-257.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Marco. (2015, octubre 24). Building Data Pipelines with Python and Luigi. Recuperado 27 de mayo de 2018, de Marco Bonzanini website: https://marcobonzanini.com/2015/10/24/building-data-pipelines-with-python-and-luigi/

María Inés Bravo, Ken Norsworthy, & Paula Pardo Lorca. (2004, octubre). *Bibliotecas Digitales Latinoamericanas en el Marco de OAI-PMH*.

Millington, P. (2006, septiembre 6). OpenDOAR - Home Page - Directory of Open Access Repositories. Recuperado 11 de septiembre de 2017, de http://www.opendoar.org/

Mönnich, M., & Spiering, M. (2008). Adding Value to the Library Catalog by Implementing a Recommendation System. *D-Lib Magazine*, *14*(5/6). https://doi.org/10.1045/may2008-monnich

Morales Morejón, M., & Morales Aguilera, M. (1997). La informetría y las fuentes de información personales e institucionales: su importancia en relación con la información de inteligencia. *Ciencias de la información*, *28*(3), 207-217.

Moran, S., & Cysouw, M. (2018). *The Unicode cookbook for linguists: Managing writing systems using orthography profiles*. Language Science Press.

Moura, A. M. S, Mattos, C. V, & Silva, D. C. (2002). *Acesso e recuperação da produção científica pela biblioteca universitária: os anais de eventos*. (Moura, A. M. S, Mattos, C. V, Silva, D. C. (2002). Acesso e recuperação da produção científica pela biblioteca universitária: os anais de eventos. Anais do Seminário Nacional de Bibliotecas Universitárias, Rio de Janeiro, RJ, Brasil, 12.), 12.

Naak, A. (2009). *Papyres : un système de gestion et de recommandation d'articles de recherche*. Recuperado de https://papyrus.bib.umontreal.ca/xmlui/handle/1866/3270

Natural Language Toolkit — NLTK 3.3 documentation. (s. f.). Recuperado 27 de mayo de 2018, de https://www.nltk.org/

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625-632. https://doi.org/10.1007/s10459-010-9222-y

P Waller, M., Dresselhaus, T., & Yang, J. (2013). JACOB: An Enterprise Framework for Computational Chemistry. *Journal of computational chemistry*, *34*. https://doi.org/10.1002/jcc.23272

Packer, A. L., Cop, N., Luccisano, A., Ramalho, A., & Spinak, E. (2014). *SciELO – 15 Años de Acceso Abierto: un estudio analítico sobre Acceso Abierto y comunicación científica*. https://doi.org/10.7476/9789233012370

Palopoli, L., Rosaci, D., & Sarné, G. M. L. (2013). A Multi-tiered Recommender System Architecture for Supporting E-Commerce. En G. Fortino, C. Badica, M. Malgeri, & R. Unland (Eds.), *Intelligent Distributed Computing VI* (pp. 71-81). Springer Berlin Heidelberg.

Pilato, G., & Vassallo, G. (2015). TSVD as a Statistical Estimator in the Latent Semantic Analysis Paradigm. *IEEE Transactions on Emerging Topics in Computing*, *3*(2), 185-192. https://doi.org/10.1109/TETC.2014.2385594

Pinfield, S. (2005). A mandate to self archive? The role of open access institutional repositories. *Serials*, *18*(1), 30-34.

Potti, P. K. (2011). *On the Design of Web Services: SOAP vs. REST*. 106.

Presidencia de la República del Perú. *Reglamento de Calificación y Registro de Investigadores en Ciencia y Tecnología del Sistema Nacional de Ciencia, Tecnología e Innovación Tecnológica - SINACYT*. , RESOLUCION-N° 023-2017-CONCYTEC-P § (2015).

Presidencia de la República del Perú. *Reglamento de la Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto*. , DECRETO SUPREMO N° 006-2015-PCM § (2015).

Presidencia de la República del Perú. *Directiva que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto*. , Directiva Nº 004-2016-CONCYTEC-DEGC § (2016).

Presidencia de la República del Perú. *Reglamento del Registro Nacional Científico, Tecnológico y de Innovación Tecnológica*. , RENACYT-RESOLUCION-Nº 045-2016-CONCYTEC-P § (2016).

Python Data Science Handbook | Python Data Science Handbook. (s. f.). Recuperado 27 de mayo de 2018, de https://jakevdp.github.io/PythonDataScienceHandbook/

Python Libraries For Building Recommender Systems. (s. f.). Recuperado 27 de mayo de 2018, de http://www.faroba.com/2015/12/03/a-python-libraries-for-building-recommender-systems/

Rehurek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks*, 45–50.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2010). *Recommender Systems Handbook* (1st ed.). Berlin, Heidelberg: Springer-Verlag.

Rölleke, T., Tsikrika, T., & Kazai, G. (2006). A general matrix framework for modelling Information Retrieval. *Information Processing & Management*, *42*(1), 4-30. https://doi.org/10.1016/j.ipm.2004.11.006

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Analysis of Recommendation Algorithms for e-Commerce. *Proceedings of the 2Nd ACM Conference on Electronic Commerce*, 158–167. https://doi.org/10.1145/352871.352887

Scimago Journal & Country Rank. (2007). Recuperado 9 de diciembre de 2017, de http://www.scimagojr.com/

Scimago Journal & Country Rank. (2017, septiembre 11). Recuperado 11 de septiembre de 2017, de http://www.scimagojr.com/

Seroussi, Y. (2010). Utilising User Texts to Improve Recommendations. En P. De Bra, A. Kobsa, & D. Chin (Eds.), *User Modeling, Adaptation, and Personalization* (pp. 403-406). Springer Berlin Heidelberg.

Sickle: OAI-PMH for Humans — Sickle 0.6.2 documentation. (s. f.). Recuperado 27 de mayo de 2018, de http://sickle.readthedocs.io/en/latest/

Silva, A. C. B. da, Oliveira, E. C. de, & Filho, J. F. R. (2005). Revista Contabilidade & Finanças - USP: uma comparação entre os períodos 1989/2001 e 2001/2004. *Revista Contabilidade & Finanças*, *16*(39), 20-32. https://doi.org/10.1590/S1519-70772005000300003

Skeef. (1997). *Citado por Mollo Pécora, Gláncia, 3 Actividades Académicas de pesquisador. En: Porto Witter, G. Produçao Científica, Campinas, SP: Editora Átomo, 1997, p. 158*.

Sosnovsky, S., & Dicheva, D. (2010). Ontological Technologies for User Modelling. *Int. J. Metadata Semant. Ontologies*, *5*(1), 32–71. https://doi.org/10.1504/IJMSO.2010.032649

Sugiyama, K., & Kan, M.-Y. (2011). Serendipitous Recommendation for Scholarly Papers Considering Relations Among Researchers. *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, 307–310. https://doi.org/10.1145/1998076.1998133

Tamayo, M. T. y. (2004). *El proceso de la investigación científica*. Editorial Limusa.

Toman, M., Tesar, R., & Jezek, K. (s. f.). *Influence of Word Normalization on Text Classification*. 5.

Tyagi, N. K., Solanki, A. K., & Tyagi, S. (s. f.). *AN ALGORITHMIC APPROACH TO DATA PREPROCESSING IN WEB USAGE MINING*. 5.

Using Luigi Pipelines in a Data Science Workflow. (s. f.). Recuperado 27 de mayo de 2018, de Pivotal Engineering Journal website: http://engineering.pivotal.io/post/luigi-data-science/

Vellino, A. (2013). *Usage-based vs. Citation-based Methods for Recommending Scholarly Research Articles*. Recuperado de https://arxiv.org/abs/1303.7149v2

Vijayarani, D. S., & Ilamathi, J. (s. f.). *Preprocessing Techniques for Text Mining - An Overview*. *5*, 10.

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 448-456. San Diego, California, USA: ACM.

Wang, X. (2010). An Empirical Investigation of Personal and Social Factors on Knowledge Sharing in China [Info:eu-repo/semantics/masterThesis]. Recuperado 27 de mayo de 2018, de http://essay.utwente.nl/60181/

Web Scraping with Beautiful Soup. (s. f.). Recuperado 27 de mayo de 2018, de http://web.stanford.edu/~zlotnick/TextAsData/Web_Scraping_with_Beautiful_Soup.html

Welcome to Python.org. (s. f.). Recuperado 27 de mayo de 2018, de Python.org website: https://www.python.org/

What are Open Access repositories? - University of Bradford. (s. f.). Recuperado 27 de mayo de 2018, de https://www.bradford.ac.uk/library/resources/open-access-publishing/what-are-open-access-repositories/

What is a corpus? (2018, abril 15). Recuperado 15 de abril de 2018, de http://www.ilc.cnr.it/EAGLES/corpintr/node13.html

Willett, P. (2006). The Porter stemming algorithm: then and now. *Program*, *40*(3), 219-223. https://doi.org/10.1108/00330330610681295

Witter, G. P. (1997). *Produção científica*. Editora Atomo.

Xia, F., Liu, H., Lee, I., & Cao, L. (2016). *Scientific Article Recommendation: Exploiting Common Author Relations and Historical Preferences* (Vol. 2). https://doi.org/10.1109/TBDATA.2016.2555318

Yang, C., Wei, B., wu, J., Zhang, Y., & Zhang, L. (2009, enero 1). *CARES: a ranking-oriented CADAL recommender system.* 203-212. https://doi.org/10.1145/1555400.1555432

Ye, J. (2011). Ye, J.: Cosine similarity measures for intuitionistic fuzzy sets and their applications. Mathematical and Computer Modelling 53, 91-97. *Mathematical and Computer Modelling*, *53*, 91-97. https://doi.org/10.1016/j.mcm.2010.07.022

Zhang, X., Tang, J., Zhang, M., & Ji, Q. (2016). Noise subspaces subtraction in SVD based on the difference of variance values [Research article]. https://doi.org/10.21595/jve.2016.16745

Zhang, Z., & Li, L. (2010). *A research paper recommender system based on spreading activation model*. https://doi.org/10.1109/ICISE.2010.5689417

Zuo, Z., Zhao, K., & Eichmann, D. (s. f.). The state and evolution of U.S. iSchools: From talent acquisitions to research outcome. *Journal of the Association for Information Science and Technology*, *68*(5), 1266-1277. https://doi.org/10.1002/asi.23751

Anexos

Anexo 1: Revisión de Fuentes: investigaciones y productos

Esta sección tiene como objetivo presentar y describir las investigaciones realizadas en los últimos años, así como los productos desarrollados relacionados al tema del presente proyecto de fin de carrera.

Investigaciones Primarias

Modelamiento de temas colaborativos para la recomendación de artículos científicos

Esta investigación se basa en la construcción de un algoritmo basado en aprendizaje máquina para la recomendación de artículos científicos a usuarios dentro de una comunidad científica virtual. Este algoritmo utiliza dos tipos de datos para la generación de recomendaciones: la librería de artículos científicos de otros usuarios y el contenido de estos artículos. Además, combina ideas de técnicas de filtrado colaborativo basado en el modelado de factores latentes y técnicas de análisis de contenidos basadas en el modelado de temas probabilísticos. Así, el algoritmo propuesto muestra recomendaciones donde se pueden encontrar artículos antiguos pero relevantes para otros usuarios con perfiles similares, como también artículos nuevos con contenido que satisfagan los intereses específicos de cada usuario (C. Wang & Blei, 2011).

Sistemas de recomendación de documentos de investigación: un enfoque basado en Caminos Aleatorios

Esta investigación tiene como objetivo proponer un algoritmo para la recomendación de artículos de investigación basado en grafos y propiedades de caminos aleatorios. Se plantea el problema a partir de la utilización de puntajes de

relevancia para cada documento en el cual el usuario esté trabajando. De esta forma, a más alto el puntaje de relevancia que se obtenga al analizar los documentos dentro de un repositorio en línea, más alta debe ser su relevancia con respecto al tema del documento en el cual se esté trabajando. Así mismo, también es posible la utilización de estructuras de grafos no dirigidos, donde cada nodo representa un documento dentro de un repositorio en línea. A esta estructura de grafos se le denomina Grafos de Citaciones. Por otro lado, en esta investigación también se hace presente la utilización del algoritmo PaperRank de Google, de manera que ayude en el proceso de filtrado de documentos almacenados en un repositorio. Así, se generarán recomendaciones de documentos basados en la bibliografía del documento en el cual se esté trabajando (Gori & Pucci, 2006).

Un sistema de recomendación de documentos de investigación basado en el modelo de activación por propagación

Esta investigación tiene como objetivo proponer un nuevo sistema de recomendación híbrido donde se utiliza un modelo de activación por propagación para la búsqueda de usuarios con intereses similares. Así mismo, la investigación describe el uso de árboles para la representación del perfil de cada usuario, lo que permite obtener la correlación entre usuarios a partir del TED (Tree Edit Distance) (Z. Zhang & Li, 2010).

Sistema basado en el análisis de múltiples criterios para la recomendación de artículos científicos

El objetivo de esta investigación es la de proponer un nuevo enfoque para la implementación de los sistemas de recomendación. De esta forma, se hace mención a las metodologías de Ayuda en la Toma de Decisiones basado en Múltiples Criterios, MCDA por sus siglas en ingles. El paradigma de análisis de múltiples criterios define los métodos y modelos para la toma de decisión en relación a la generación de una recomendación. A partir de este paradigma, se pueden distinguir 4 niveles, donde el segundo nivel se basa en el modelado de criterios, los cuales corresponden al año de publicación, el nivel de relevancia de las palabras claves, el factor de impacto de revistas (JIF), autor, reconocimientos, índice de citas e institución de proveniencia. Por otro lado,

el tercer nivel describe la construcción del modelo de preferencia a partir del enfoque de disgregación-agregación, el cual tiene como objetivo analizar el comportamiento del usuario (Lakiotaki, Delias, Sakkalis, & Matsatsinis, 2009).

Recomendación de artículos académicos a través de los intereses de lectura de usuarios

Esta investigación tiene como objetivo presentar una solución a la de generación de recomendaciones relacionadas al problema y/o solución presentados en los artículos científicos. La solución se base en satisfacer el interés específico de lectura del usuario a partir de la recomendación de artículos científicos en dos listas: recomendaciones de artículos más relevantes en relación al problema de investigación y en relación a la solución propuesta. La investigación también describe los componentes utilizados para el sistema de recomendación propuesto, donde se hacen uso de tres modelos de semejanza: modelo TF*IDF, modelo en base a tema, modelo en base a conceptos – con uso de LDA- (Jiang, Jia, Feng, & Zhao, 2012).

Sistema de Recomendación de artículos científicos basado en frases claves

El objetivo de esta investigación es de proponer un enfoque basado en contenido para la recomendación de artículos científicos dentro de repositorios digitales. El enfoque propuesto se basa en la extracción de palabras claves. El método de extracción de palabras clave trabaja en dos fases: la de identificación de candidato, donde se identifican todas las posibles frases de un artículo, y la de selección, donde se selecciona solo algunas frases candidatas como frases claves. Así mismo, la investigación describe la utilización de etiquetas de palabras, la utilización de uni-gramas, bi-gramas y tri-gramas para la extracción de las posibles frases candidatas, el uso de propiedades estadísticas y lingüísticas para la caracterización de las frases candidatas y el uso de puntajes para la elección de frases claves que representen apropiadamente el contenido del artículo científico. De esta manera, la investigación muestra también la creación de perfiles de usuarios a partir de frases claves, las cuales serán procesadas junto con las frases extraídas para cada uno de los artículos

científicos almacenados para conocer la semejanza entre ambas y así generar la recomendación (Ferrara, Pudota, & Tasso, 2011).

Recomendación de artículos científicos: Explotación de relaciones comunes de los autores y preferencias históricas

Esta investigación tiene como objetivo proponer un nuevo método para la recomendación de artículos de investigación. El método presentado se denomina CARE (Common Author Relation-Based Recommendation) por sus siglas en ingles. Este método se basa en la construcción de relaciones entre los artículos de investigación únicamente a través de la información del autor. Es así que, estas relaciones en conjunto con las preferencias históricas del investigador son usadas para la búsqueda de investigadores objetivo relevantes (Xia, Liu, Lee, & Cao, 2016).

Investigaciones Secundarias

Sistemas de recomendación de documentos de investigación: una encuesta bibliográfica

Esta artículo de investigación tiene como objetivo examinar el campo de los sistemas de recomendación de artículos de investigación, permitiendo así a los investigadores y desarrolladores a conocer el contexto actual de los sistemas de recomendación de artículos de investigación, identificar prometedores áreas de investigación y motivar a la comunidad a resolver los problemas más urgentes que vienen obstaculizando el uso efectivo de los sistemas de recomendación de artículos de investigación. Así mismo, la investigación señala que más de la mitad de los sistemas de recomendación analizados aplicaron el enfoque de filtrado basado en contenido (55%). Mientras que el enfoque de filtrado colaborativo fue aplicado por sólo el 18% y las recomendaciones basadas en grafos un 16%. Otros enfoques de recomendación que se encontraron tras el análisis fueron: estereotipos, recomendaciones centradas en características y recomendaciones híbridas. Por otro lado la investigación también describe los distintos métodos de evaluación para los enfoques analizados. Entre estos métodos están:

- Estudios de usuarios: suelen medir la satisfacción del usuario a través de clasificaciones explícitas. Los usuarios reciben recomendaciones generadas por

diferentes enfoques de recomendación, los usuarios valoran las recomendaciones y el enfoque con la calificación promedio más alta se considera más efectivo.

- Evaluaciones online: mide las tasas de aceptación de las recomendaciones en los sistemas de recomendación en el mundo real. Las tasas de aceptación se miden normalmente mediante la tasa de clics (CTR), es decir, la proporción entre las recomendaciones a las que se han hecho clic y las recomendaciones mostradas.
- Evaluaciones offline: miden la exactitud de un sistema de recomendación basado en la realidad. Algunas de las métricas de evaluación son: tasa de impacto, valor F, la media del rango recíproco (MRR por sus siglas en ingles), la ganancia descontada acumulada normalizada (nDCG), el error absoluto medio y el error cuadrático medio de la raíz.
- La perspectiva del operador: un sistema de recomendación efectivo puede ser uno que pueda ser desarrollado, operado y mantenido a bajo costo.
- Cobertura: describe cuántos documentos de los que figuran en la base de datos podrían ser potencialmente recomendados

Así mismo, la investigación describe algunas deficiencias y desafíos en el campo de los sistemas de recomendación de artículos de investigación. Algunos desafíos y deficiencias se muestran a continuación:

- Modelo de datos: parte fundamental de la generación de recomendaciones es el proceso de modelado del usuario que identifica las necesidades de información del usuario
- Exactitud: en el campo de sistemas de recomendación de artículos científicos, se hace mucho énfasis a la métrica de exactitud como indicador de la satisfacción del usuario. Sin embargo, no siempre la exactitud es sinónimo de satisfacción, otros factores como tareas del usuario, diversidad, diseño, características del usuario, tiempo de uso y retroalimentación del perfil del usuario.

- Implantación de los sistemas de recomendación en la vida real: la mayoría de los sistemas de recomendación aplican simples enfoques los cuales no se basan en últimas investigaciones. Por tal motivo, se puede concluir que la traducción de la investigación a la práctica es un desafío actual en la comunidad de sistemas de recomendación de artículos científicos.
- Persistencia y autoridades: de los 276 autores de los 185 artículos citados en esta investigación, el 73% publicaron un solo artículo, el 5% publicaron cinco o más artículos, pero de estos autores, varios coautores publicaron los mismos artículos. Esto significa que sólo hay unos pocos grupos que publican sistemáticamente investigaciones en el campo de sistemas de recomendación de artículos científicos.
- Dispersión de la información: este desafío significa, los problemas que conllevan la escasez en la publicación de información relacionada a los enfoques propuestos, como por ejemplo la dificultad de otros investigadores para implementar de nuevo el enfoque propuesto. Esto podría causar problemas en replicar las evaluaciones y reproducir los resultados de la investigación y obstaculizar la re-implementación y aplicación de enfoques prometedores en los sistemas de reconocimiento de palabras reales (Beel et al., 2015).

Productos desarrollados

Scienstein: Un Sistema de Recomendación de Documentos de Investigación

Scienstein es el primer sistema de recomendación híbrido que utiliza técnicas de filtrado basado en contenido y filtrado colaborativo. De esta manera, Scienstein combina análisis de citas, puntuación implícita y explicita, análisis de autor y análisis de origen para la recomendación de artículos científicos de manera holística. La combinación de estos enfoques es crítica ya que cada uno de ellos posee desventajas que solo pueden ser superadas a través de la combinación de los enfoques anteriormente mencionados (Gipp et al., 2009).

Papyres: Sistema de Administración de Artículos de Investigaciones

El sistema de recomendación implementado para Papyres se basa en la utilización de un enfoque híbrido a partir de los enfoques de filtrado basado en contenido y filtrado colaborativo. En Papyres, el filtrado basado en contenido tiene como objetivo el construir una lista de artículos científicos relevantes que representen el contexto de interés del usuario. Esta lista generada, luego se utiliza dentro del proceso de filtrado colaborativo multi-criterio. Este último enfoque es crítico en la recomendación de artículos científicos de calidad, la cual es relativa y no se refleja necesariamente en un factor global. Así mismo, el cálculo de la semejanza se describe como un proceso critico dentro del filtrado colaborativo. Por otro lado, Papyres también toma en consideración ciertos criterios de evaluación para la construcción de las recomendaciones. Estos se dividen en:

- Criterios generales: por presentación, por orientación técnica, por nivel técnico y por clasificación
- Criterios específicos dentro de las secciones del artículo científico: por calidad de la introducción, por estado del arte, por metodología, por experimentación y validación y por trabajos futuros
- Criterio global: evaluación global (Naak, 2009).

Sistema de Recomendación de BibTip

El sistema de recomendación BibTip se basa en los patrones de comportamiento de los usuarios a través de su interacción con un catálogo de la biblioteca. Este servicio de recomendación "implícita" se basa en la observación de patrones de usuario y en la evaluación estadística del uso de datos. Todos los datos almacenados y procesados son anónimos (números de identificación e identificadores de sesión). A nivel técnico, la arquitectura BibTip puede ser vista como una arquitectura de agente que involucra a tres agentes de software: el agente de observación, el agente de agregación y el agente de recomendación. El primer agente observa la selección de títulos dentro de las sesiones definidas. Estos datos se transfieren al agente de agregación, para realizar cálculos sobre el material estadístico para la construcción de una lista de recomendaciones. Por último, el agente de recomendación se encarga de presentar la lista de recomendaciones al usuario (Mönnich & Spiering, 2008).

ARSYS: Sistema de Recomendación de Artículos Científicos

ARSYS se basa en la aplicación de un enfoque híbrido para la construcción de un sistema de recomendación. Los enfoques utilizados son los de Filtrado basado en Contenido y Filtrado Colaborativo. Para el primer enfoque se utilizaron redes semánticas, las cuales representan los conceptos relacionados en una estructura específica para un dominio específico. Esta estructura semántica ayuda a la generación de recomendaciones de nuevos artículos científicos. Por otro lado, para el enfoque de filtrado colaborativo se implementó una red punto a punto donde la información es distribuida uniformemente a través de todos los usuario (Bancu et al., 2012).

Introducción al Sistema de Recomendación de Documentos de Investigación de Docear

Docear es un software completo que permite buscar, organizar y crear artículos de investigación. Además, posee un sistema de recomendación para los artículos de investigación que gestiona. Este sistema de recomendación se basa en la utilización de mapas mentales donde se gestiona la data del usuario –artículos, referencias, anotaciones, etc.-. A partir de estos mapas mentales se crea un modelo de datos del usuario y este a su vez es comparado con los más de 1.8 millones de artículos de investigación que gestiona la Biblioteca Digital de Docear. Finalmente el resultado será la recomendación de 10 artículos, generados a través de la técnica de filtrado basado en contenido (Beel et al., 2013).

Anexo 2: Protocolo de Consentimiento Informado

Comité de ética para la investigación con seres humanos y animales – CEI(sha)
Vicerrectorado de Investigación - PUCP

PROTOCOLO DE CONSENTIMIENTO INFORMADO PARA CALIFICACIÓN DE RECOMENDACIONES DE PUBLICACIONES CIENTÍFICAS

El propósito de este protocolo es de brindar a los participantes en esta investigación, una explicación clara de la naturaleza de la misma, así como del rol que tienen en ella.

La presente investigación es conducida por la estudiante del programa de pre grado en Ingeniería Informática de la Pontificia Universidad Católica del Perú, Elizabeth Jenisse Vereau Zagastizábal, asesorada por el Ing. Cesar Olivares Poggi. La meta de este estudio es la de evaluar el sistema de recomendación de publicaciones científicas de ALICIA propuesto en el Proyecto de Fin de Carrera de la estudiante, como parte de los cursos de INF391 Proyecto de Tesis1 e INF392 Proyecto de Tesis 2.

Si usted accede a participar en este estudio, se le pedirá participar de respondiendo una encuesta, lo cual le tomará 15 minutos de su tiempo. Las encuestas realizadas serán recopiladas por el investigador con el propósito de analizar los resultados y las ideas que usted haya expresado. Una vez finalizado el estudio todos los documentos serán destruidos.

Su participación será voluntaria. La información que se recoja será estrictamente confidencial y no se podrá utilizar para ningún otro propósito que no esté contemplado en esta investigación.

En principio, la encuesta resuelta por usted será anónima, por ello será codificada utilizando un número de identificación. Si la naturaleza del estudio requiriera su identificación, ello solo será posible si es que usted da su consentimiento expreso para proceder de esa manera.

Si tuviera alguna duda con relación al desarrollo del proyecto, usted es libre de formular las preguntas que considere pertinentes. Además, puede finalizar su participación en cualquier momento del estudio sin que esto represente algún perjuicio para usted. Si se sintiera incómoda o incómodo, frente a alguna de las preguntas, puede ponerlo en conocimiento de la persona a cargo de la investigación y abstenerse de responder.

Muchas gracias por su participación.

Yo, _____ doy mi consentimiento para participar en el estudio y soy consciente de que mi participación es enteramente voluntaria.

He recibido información en forma verbal sobre el estudio mencionado anteriormente y he leído la información escrita adjunta. He tenido la oportunidad de discutir sobre el estudio y hacer preguntas. Entiendo que el experimento tiene por objetivo evaluar las recomendaciones generadas por un sistema de software, NO mis capacidades /habilidades /conocimientos.

Al firmar este protocolo estoy de acuerdo con que mis datos personales, incluyendo datos relacionados a mi salud física y mental o condición, y raza u origen étnico, podrían ser usados según lo descrito en la hoja de información que detalla la investigación en la que estoy participando.

Entiendo que puedo finalizar mi participación en el estudio en cualquier momento, sin que esto represente algún perjuicio para mí. Entiendo que puedo comunicar al supervisor, en cualquier momento, sobre algún malestar, molestia o inconformidad que pueda sentir durante el desarrollo de las actividades, y que, por tal motivo, puedo abandonar las actividades en cualquier momento.

Entiendo que recibiré una copia de este formulario de consentimiento e información del estudio y que puedo pedir información sobre los resultados de este estudio cuando éste haya concluido. Para esto, puedo comunicarme con la estudiante Elizabeth Jenisse Vereau Zagastizábal al correo jenisse.vereau@pucp.pe o al teléfono 981546409, o de la misma forma, con el asesor Ing. Cesar Olivares Poggi al correo cesar.olivares@pucp.pe.

_____ _____ _____

_Nombre completo del (de la) participante Firma Fecha

ELIZABETH JENISSE VEREAU ZAGASTIZÁBAL

Para la elaboración de este protocolo se ha tenido en cuenta el formulario de C.I. del Comité de Ética del Departamento de Psicología de la PUCP

Anexo 3: Formato del Cuestionario para Calificación de Recomendaciones de Publicaciones Científicas

Cuestionario para Calificación de Recomendaciones de Publicaciones Científicas

Estimado(a) investigador(a):

Gracias por participar y completar esta encuesta, que tiene por objeto obtener información cuantitativa sobre recomendaciones generadas por el sistema de recomendación de publicaciones científicas de ALICIA a investigadores calificados del SINACYT.

1- ¿Cuán afines son las siguientes 10 publicaciones al ámbito de investigación en la cual Ud.se desenvuelve?

Publicación #1

Título:	Identificación y control de un gasificador de lecho fluidizado
Abstract:	El objetivo de la tesis es la identificación de la planta experimental del gasificador de lecho fluidizado instalado. Esto abarca la teoría matemática empleada para su obtención, clasificación y elección del modelo que mejor se ajuste a lo requerido para posteriormente aplicar control. El trabajo se enfoca en lo fundamental de la instalación del gasificador, su protocolo de puesta en marcha y los resultados de las pruebas realizadas. Seguidamente, se explica a profundidad los pasos a seguir para identificar, la clasificación de los modelos, en qué consiste el cálculo de un modelo paramétrico y la aproximación de los pasos que sigue el software MATLAB para calcular modelos. Por último, se trata el tema del control del modelo obtenido en la identificación.

Completamente afín( )	Muy afín( )	Medianamente afín( )	Poco afín( )	Nada afín( )

Publicación #2

Título:	Comparación de modelos hidrológicos precipitación-escorrentía determinísticos conceptuales con y sin uso de modelo estocástico
Abstract:	La tesis presenta la aplicación de tres modelos hidrológicos precipitación-escorrentía: TANQUE, NAM, SMA trabajados en la cuenca del río Cañete durante el periodo del primero de agosto de 1973 al primero de abril de 1976; los cuales fueron calibrados (del 01/08/1973 al 31/07/1974), simulados (de

	l 01/08/1974 al 31/07/1975) y validados (del 01/08/1975 al 01/04/1976), posteriormente a estos tres modelos se les agr ego modelos estocástico para mejorar la aproximación, es as í que primero se agregó un modelo autoregresivo AR(1), y de spués se aplicó el filtro de Kalman. La metodología para ev aluar una mejor aproximación, con los datos reales y calcul ados por las diferentes variantes de los modelos, fue el us o de índices, para la tesis estos son la raíz del error cua drático normalizado, el índice de eficiencia y el error med io normalizado, obteniéndose resultados satisfactorios para los modelos TANQUE y NAM.

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #3

Título:	Fundamentos de econometría : teoría y problemas
Abstract:	Contenido: 1.- REGRESIÓN LINEAL SIMPPLE. 1.1.- Introducción a la regresión lineal simple. 1.2.- Modelo clásico de regresión lineal: Recta de regresión simple muestral. 1.3.- Método de estimación de mínimos cuadrados ordinarios (MCO). 1.4.- Propiedades de los estimadores MCO. 1.5.- Cálculos adicionales sobre los estimadores sobre los estimadores MCO y la varianza del error. 1.6. Medidas de bondad de ajuste. 1.7. Pruebas de hipótesis. 2.- MODELO REGRESIÓN MULTIPLE. 2.1.- Función de regresión poblacional. 2.2.- Función de regresión muestral. 2.3.- Supuestos del modelo clásico de regresión lineal. 2.4.- Estigmación MCO. 2.5.- Propiedades de los estimadores MCO. 2.6.- Medidas de bondad de ajuste. 2.7.- Pruebas de hipótesis. 2.8.- Una visión matricial. 3.- MULTICOLINEALIDAD. 3.1.- Definición. 3.2.- Causas. 3.3.- Consecuencias. 3.4.- Detección. 3.5. Corrección. 4. HETEROCEDASTICIDAD. 4.1.- Definición de Heteroscedasticidad. 4.2.- Causas de la heteroscedasticidad. 4.3.- Consecuencias de utilizar MCO en presencia de heteroscedasticidad. 4.4.- Test de heteroscedasticidad Park. 4.5. Test de heteroscedasticidad de Glejser. 4.6.- Test de heteroscedasticidad Goldfeld-Quandt. 4.7.- Test de heteroscedasticidad de Breusch-Pagan-Godfrey. 4.8.- Test de heteroscedasticidad de White. 4.9.- Medidas correctivas cuando se conoce: Método de mínimos cuadrados ponderados. 4.10. Medidad correctivas cuando no se conoce. 5.- AUTOCORRELACIÓN. 5.1.- Definición. 5.2.- Modelo autorregresivo (AR). 5.3.- Causas. 5.4.- Consecuencias. 5.5.- Detección. 5.6.- Correción. 6.- VARIABLES DUMMY. 6.1.- Definición. 6.2.- Modelos econométricos con variables Dummy. 7.- PRUEBAS DE DIAGNÓSTICO Y SELECCIÓN DE MODELOS. 7.1.- Introducción. 7.2.- Pruebas de diagnóstico. 7.3.- Criterios de selección del modelo. 8.- MODELOS DE REGRESIÓN NO LINEALES. 8.1.- Definición. 8.2.- Estimación. 9.- MODELOS DE RESPUESTA CUALITATIVA

	. 9.1.- Introducción. 9.2.- Modelo lineal de probabilidad (MLP). 9.3.- Logit. 9.4.- Probit. 10.- DATA PANEL. 10.1.- De finición de modelos de regresión con datos de panel. 10.2.- Ventajas. 10.3.- Tipos. 10.4.- Técnicas de estimación con D ata Panel. 10.5.- Prueba de Hausman. 10.6.- Propiedades est adísticas de los estimadores. 10.7.- Comparación entre el m odelo de efectos fijos (MEF) y el modelo de efectos aleator ios (MCE). 11.- MODELOS DINÁMICOS AUTORREGRESIVOS Y DE REZA GOS DISTRIBUIDOS. 11.1.- Modelos econométricos de rezagos d istribuidos de Koyck. 11.2.- Modelo econométrico de expecta tivas adaptativas. 11.3.- Modelo de ajuste parcial. 11.4.- Modelo econométrico de rezagos distribuidos de Almon. 11.5 .- Causalidad de series de tiempo. 11.6.- Test de causalida d de Granger. 12.- MODELOS ECONOMÉTRICOS DE ECUACIONES SIMU LTÁNEAS. 12.1.- Introducción: Álgebra de sistemas de ecuac iones simultáneas. 13.-SERIES DE TIEMPO: ESTACIONARIEDAD, R AÍZ UNITARIA Y CONTEGRACIÓN. 13.1.- Definiciones. 13.2.- Es tacionariedad de un proceso estocástico. 14.- MODELOS ARIMA . 14.1.- Creación de modelos econométricas para series de t iempo: Ar, Ma, Arima. 14.2.- Metodología de Box-Jenkins (BJ). 14.3.- Identificación. 14.4.- Estimación. 14.5.- Estimac ión. 14.6.- Pronóstico.

Completamente afín( )	Muy afín( )	Medianamente afín( )	Poco afín( )	Nada afín( )

Publicación #4

Título:	Consorcio Web : Formulación e implementación de un marco co nceptual de integración eficiente de sitios web dentro de u na comunidad en internet
Abstrac t:	La realización de este trabajo plantea implementar una comu nidad vía Internet conformada por diversos sitios web auto-sostenidos, denominado, para efectos del estudio, como cons orcio web, demostrar que la implementación de dicho esquema es factible mediante un análisis de caso, y que eventualmen te permitiría al usuario la obtención de ingresos; la innov ación propuesta radica en la utilización de diferentes cate gorías de sitios web que se complementan mutuamente, confor mando un esquema simbiótico que contribuye a la sinergia de l modelo total. El problema identificado fue la no existencia de un marco c onceptual capaz de integrar de manera eficiente a los diversos sitios web. Adicionalmente, se busca identificar las ventajas y desvent ajas que este esquema de gestión implicaría, identificar los riesgos de su implementación y formalizar la metodología utilizada tanto a

	fin de hacerla repetible en el tiempo como de facilitar su transmisión a otras personas.

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #5

Título:	Diseño del controlador basado en un observador de estado re alimentado desde un controlador clásico aplicado a un manip ulador robótico con una articulación
Abstrac t:	El control automático puede ser fragmentado, comprimido en diferentes alternativas que den una solución al problema de l diseño de control. Las nuevas teorías de control y los co nceptos modernos son atractivos y pueden hacer que de algun a forma nos olvidemos del problema del diseño de un control mediante las técnicas clásicas. Si tenemos dos o más acerca mientos diferentes que proporcionan una solución buena al p roblema de control, entonces debe de existir una conexión f uerte entre ellos que por diferentes métodos solucionen el problema. Si podemos establecer tales conexiones, esto debe ría ser de gran ayuda a los investigadores a que puedan ent ender conceptos subyacentes que involucran al problema de d iseño del control. La transición de la realimentación de estados estimados en el control clásico es bien conocida; sin embargo, al conoci miento la transición inversa requiere ser articulada previa mente para un caso general. De ese modo en el presente info rme de suficiencia, consideramos al sistema en tiempo conti nuo y exploramos las conexiones de los sistemas single-inpu t, single-ouput (SISO) para los controladores clásicos line ales (es decir, aquellos definidos por la función de transf erencia). El acercamiento al diseño del control se basa en el diseño de un sistema de control usando la técnica del ob servador de estado de orden reducido realimentado desde un control clásico para el manipulador robótico con una articu lación.

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #6

Título:	Diseño de un algoritmo PID sintonizado mediante lógica fuzz y para controlar un brazo robótico
Abstrac t:	En los últimos años el desarrollo del control moderno ha ev olucionado considerablemente mostrando diferentes técnicas

	basadas en el control difuso, la aplicación de estas técnic as a sistemas reales es cada vez una tendencia con mejores posibilidades de control ofreciendo resultados en un menor tiempo de respuesta comparado a un clásico controlador PID. El control PID es confiable pero con las limitaciones enten didas por su diseño para sistemas no lineales se presenta e n el presente trabajo una sintonización en base a lógica di fusa de los parámetros proporcional, derivativa, integrativ a del controlador PID para controlar la posición de un robo t de tres grados de libertad. El Robot de tres grados de libertad que se formula la cinem ática para diseñar las medidas de los eslabones como de los componentes, desarrollados en software Solidwork, también l a dinámica para conocer el modelado del Robot para diseñar el controlador PID en software Matlab. El control propuesto es denominado Fuzzy PID usa la impreci sión del lenguaje difuso para la toma de decisiones en base a valores intermedios que pertenecen a dos conjuntos para s intonizar el control PID, las entradas de la función de mem bresía Fuzzy son el error y derivada del error estas ingres an a la función del tipo triangular con reglas Mandanni que explican la experiencia del experto y desarrolla una salida para los parámetros del controlador PID. El controlador Fuzzy-PID para un control de posición del Ro bot se confronta con una segunda alternativa de controlador denominado Fuzzy PID2 en pruebas de movimientos curvos o li neales concluyendo cual ofrece un mejor comportamiento y li mitación. Se busca tener una alternativa con el controlador Fuzzy PID para contribuir a la aplicación de nuevas técnica s de control de forma combinada con un control PID en un ni vel educativo.

Completamente afín( )	Muy afín( )	Medianamente afín( )	Poco afín( )	Nada afín( )

Publicación #7

Título:	Modelación hidrológica distribuida espacializada usando HEC – HMS para la represa Chirimayuni – Moquegua
Abstrac t:	La necesidad de contar con procedimientos más precisos que la metodología clásica, hizo que se plantee efectuar la mod elación hidrológica bajo un sistema distribuido espacial us ando software libre HEC-HMS para la simulación de la repres a Chirimayuni en la Región Moquegua. Este planteamiento se sustenta en aspectos básicos como identificar los elementos del sistema hidrológico que forman parte del sistema de la Represa Chirimayuni, asimismo efectuar la modelación distri buida espacial de las subcuencas concurrentes y finalmente

	efectuar la simulación hidrológica del embalse, para determ inar su comportamiento hidrológico. En tal sentido logramos elaborar el modelo espacializado co n detalle adecuado en la zona de estudio, lo que nos permit ió efectuar las simulaciones hidrológicas en las que podemo s notar diferencias en los resultados obtenidos respecto al modelo clásico elaborado para la misma Región hidrológica. En cuanto a las descargas máximas de ingreso al embalse ten emos que el modelo clásico arroja como resultado un valor d e 17,7 m3/s, mientras que el modelo espacializado un valor de 16,4 m3/s, el cual es ligeramente menor que el modelo cl ásico. Asimismo, en cuanto a las descargas pico evacuadas p or el vertedero de excedencias tenemos que el modelo clásic o arroja como resultado un valor de 6,8 m3/s, mientras que el modelo espacializado un valor de 5,9 m3/s, el cual es li geramente menor que el modelo clásico. En cuanto al volumen de almacenamiento para los modelos clá sico y espacializado ha sido establecido en 5,56 Hm3 y 5,53 Hm3 respectivamente. Luego del proceso de simulación y resu ltados obtenidos podemos observar diferencias más o menos s ignificativas en los valores calculados, siendo el modelo h idrológico espacializado el de mayor precisión, en virtud a que la información para la composición de este modelo es mu cho más detallada que el modelo clásico.

Completamente afín( )	Muy afín( )	Medianamente afín( )	Poco afín( )	Nada afín( )

Publicación #8

Título:	Estudio de la lógica borrosa en la regulación de sistemas c onmutados DC/DC
Abstrac t:	El trabajo que se ha desarrollado trata del estudio de los aportes que puede ofrecer la Lógica Borrosa en el campo de los Sistemas Conmutados DC/DC. En el caso del control basado en Lógica Borrosa no se prete nde estudiar el control directo de sistemas conmutados DC/DC si no los aportes de este tipo de control aplicado a sistemas que permitan su futura implementación con dispositivos estándar es usando los conceptos usados en: El Control en Modo Deslizante Para mejorar su respuesta y funcionamiento cuando este es c ontrolado en modo de corriente.

Completamente afín( )	Muy afín( )	Medianamente afín( )	Poco afín( )	Nada afín( )

Publicación #9

Título:	Diseño y simulación de un sistema de control no lineal multivariable por lógica difusa aplicado a un manipulador robótico translacional de 2DOF
Abstract:	La presente tesis trata sobre el control de trayectoria de un manipulador robótica traslacional multivariable de 2DOF (Two-Degrees-of-Freedoms) que consta de un móvil accionado por una polea y un eslabón articulado en el CG (centro de gravedad) de dicho móvil. Este proceso será controlado mediante la técnica de control fuzzy. La acción de control está orientada a controlar el movimiento traslacional del móvil y el movimiento angular del brazo que es libre de girar en ambas direcciones. Las metas impuestas en la tesis son: diseño, modelado, y simulación del sistema: controlado con la ley de control fuzzy. Los sistemas convencionales de control son diseñados tradicionalmente usando modelos matemáticos de sistemas físicos para posteriormente aplicar técnicas de diseño para obtener controladores apropiados para el sistema. Sin embargo, en la realidad, el modelo y sus parámetros son con frecuencia desconocidos. Esto se debe a cambios en el ambiente de trabajo, dinámicas no modeladas; y la presencia de no linealidades e incertidumbre que son difícil de tratar con los controladores convencionales los cuales no siempre son capaces de aprender o de adaptarse a nuevas situaciones. A continuación, se plantea este problema y la solución del mismo utilizando la técnica de lógica difusa. Actualmente las técnicas avanzadas de control basadas en modelos, como son el control predictivo, el control por modos deslizantes, el control adaptable, entre otras, siendo combinadas con técnicas de control basadas en los sistemas difusos.

Completamente afín( )	Muy afín( )	Medianamente afín( )	Poco afín( )	Nada afín( )

Publicación #10

Título:	Asimetrías forward-backward y left-right en el modelo 3-3-1
Abstract:	Aproximadamente para el 2030 se piensa inaugurar el ILC (International Linear Collider) en el cual se harán colisionar haces de electrones contra positrones con energías en el C.M. entre 0.5 y 3 TeV. Uno de los propósitos de éste futuro colisionador será evidenciar la existencia del bosón exótico Z' , el cual aparece en modelos que van más allá del Modelo Estándar (ME) tales como el Modelo SU (3)c x SU(3)L x U(l)y (Modelo 3-3-1). El presente trabajo tiene como principal objetivo desarrollar el sector electrodébil del 3-3-1 y calcular las Asimetrías Forward-Backward y Left-Right en función de la

	energía en el C.M. (vía el proceso e-e+ -+ μ-μ+), para pos teriormente comparar las predicciones obtenidas con las del ME y otros modelos. Palabras Claves: Asimetría Izquierda-Derecha y Atrás-Adelant e, Futuro colisionador e+e-, Modelo SU(3) x U (l).

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Anexo 4: Protocolos de Consentimiento Informado y Cuestionarios de la Evaluación.

Comité de ética para la investigación con seres humanos y animales – CEI(sha)
Vicerrectorado de Investigación - PUCP

PROTOCOLO DE CONSENTIMIENTO INFORMADO PARA CALIFICACIÓN DE RECOMENDACIONES DE PUBLICACIONES CIENTÍFICAS

El propósito de este protocolo es de brindar a los participantes en esta investigación, una explicación clara de la naturaleza de la misma, así como del rol que tienen en ella.

La presente investigación es conducida por la estudiante del programa de pre grado en Ingeniería Informática de la Pontificia Universidad Católica del Perú, Elizabeth Jenisse Vereau Zagastizábal, asesorada por el Ing. Cesar Olivares Poggi. La meta de este estudio es la de evaluar el sistema de recomendación de publicaciones científicas de ALICIA propuesto en el Proyecto de Fin de Carrera de la estudiante, como parte de los cursos de INF391 Proyecto de Tesis1 e INF392 Proyecto de Tesis 2.

Si usted accede a participar en este estudio, se le pedirá participar de respondiendo una encuesta, lo cual le tomará 15 minutos de su tiempo. Las encuestas realizadas serán recopiladas por el investigador con el propósito de analizar los resultados y las ideas que usted haya expresado. Una vez finalizado el estudio todos los documentos serán destruidos.

Su participación será voluntaria. La información que se recoja será estrictamente confidencial y no se podrá utilizar para ningún otro propósito que no esté contemplado en esta investigación.

En principio, la encuesta resuelta por usted será anónima, por ello será codificada utilizando un número de identificación. Si la naturaleza del estudio requiriera su identificación, ello solo será posible si es que usted da su consentimiento expreso para proceder de esa manera.

Si tuviera alguna duda con relación al desarrollo del proyecto, usted es libre de formular las preguntas que considere pertinentes. Además, puede finalizar su participación en cualquier momento del estudio sin que esto represente algún perjuicio para usted. Si se sintiera incómoda o incómodo, frente a alguna de las preguntas, puede ponerlo en conocimiento de la persona a cargo de la investigación y abstenerse de responder.

Muchas gracias por su participación.

Yo, [redacted] doy mi consentimiento para participar en el estudio y soy consciente de que mi participación es enteramente voluntaria.

He recibido información en forma verbal sobre el estudio mencionado anteriormente y he leído la información escrita adjunta. He tenido la oportunidad de discutir sobre el estudio y hacer preguntas. Entiendo que el experimento tiene por objetivo evaluar las recomendaciones generadas por un sistema de software, NO mis capacidades /habilidades /conocimientos.

Al firmar este protocolo estoy de acuerdo con que mis datos personales, incluyendo datos relacionados a mi salud física y mental o condición, y raza u origen étnico, podrían ser usados según lo descrito en la hoja de información que detalla la investigación en la que estoy participando.

Entiendo que puedo finalizar mi participación en el estudio en cualquier momento, sin que esto represente algún perjuicio para mí. Entiendo que puedo comunicar al supervisor, en cualquier momento, sobre algún malestar, molestia o inconformidad que pueda sentir durante el desarrollo de las actividades, y que, por tal motivo, puedo abandonar las actividades en cualquier momento.

Entiendo que recibiré una copia de este formulario de consentimiento e información del estudio y que puedo pedir información sobre los resultados de este estudio cuando éste haya concluido. Para esto, puedo comunicarme con la estudiante Elizabeth Jenisse Vereau Zagastizábal al correo jenisse.vereau@pucp.pe o al teléfono 981546409, o de la misma forma, con el asesor Ing. Cesar Olivares Poggi al correo cesar.olivares@pucp.pe.

[redacted]	[redacted]	19/06/18
Nombre completo del (de la) participante	Firma	Fecha

ELIZABETH JENISSE VEREAU ZAGASTIZÁBAL

Para la elaboración de este protocolo se ha tenido en cuenta el formulario de C.I. del Comité de Ética del Departamento de Psicología de la PUCP

Cuestionario para Calificación de Recomendaciones de Publicaciones Científicas

Estimado(a) investigador(a):

Gracias por participar y completar esta encuesta, que tiene por objeto obtener información cuantitativa sobre recomendaciones generadas por el sistema de recomendación de publicaciones científicas de ALICIA a investigadores calificados del SINACYT.

1- ¿Cuán afines son las siguientes 10 publicaciones al ámbito de investigación en la cual Ud.se desenvuelve?

Publicación #1

Título:	Estudio de usabilidad en la una propuesta de sitio Web basado en el diseño de la experiencia del´usuario
Abstract:	El propósito de esta investigación fue determinar que, en comparación al Sitio Web original, la aplicación del Diseño de la Experiencia del Usuario optimiza la Usabilidad en la Propuesta de Sitio Web. Para ello, se estudiaron los fundamentos, elementos y características del Diseño de la Experiencia de Usuario, y se aplicaron una serie de metodologías de desarrollo, propias de esta disciplina, para definir las necesidades, tanto de la institución propietaria del sitio web como de sus usuarios. Así se establecieron los requerimientos de la Propuesta de Sitio Web y posteriormente se creó un prototipo como Mínimo Producto Viable de la Propuesta. Para poder comparar la usabilidad en ambos sitios web, estos se evaluaron por medio de un instrumento de investigación basado en el Estudio de Usabilidad. Concluyendo, de esta forma, que la aplicación del Diseño de la Experiencia del Usuario permitió que la Propuesta de Sitio Web obtenga, tanto una mejor Usabilidad como una mejor retroalimentación de parte de los usuarios finales.

Completamente afín( ) Muy afín(X) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #2

Título:	Reconocimiento de patrones de comportamiento de usuarios en el portal web usando web Mining
Abstract:	La siguiente investigación tiene la finalidad de encontrar los patrones de comportamiento de los usuarios de un Portal Web, para esto se utiliza la metodología de Web Mining, la técnica del Clustering y el algoritmo de K-Means: esta metodología permite formar grupos con características iguales o similares (preferencias y cantidad de visitas a uno o más servicios). Una vez que se tenga el conocimiento de los grupos se personaliza el Portal Web para cada uno estos grupos sobre la base de sus preferencias. Con esto se logra un menor tiempo de acceso al servicio deseado y un mayor grado de satisfacción de los usuarios. Los datos a analizar se extraen de los registros de acceso de usuarios hacia el Portal Web (archivos Logs) en los cuales se registran los datos de navegación de los usuarios, y éstos se almacenan en el servidor web.

Completamente afín( ) Muy afín( ) Medianamente afín(X) Poco afín( ) Nada afín( )

Publicación #3

Título:	Implementación de estrategias de gamificacion y buenas prácticas para el sitio web del congreso internacional de ingeniería de la UPAO
Abstract:	La gamificación es un conjunto de técnicas de juego utilizadas en entornos no lúdicos, como por ejemplo el ambiente laboral, selección de personal, sitios web. La gamificación en un Sitio Web busca que los usuarios se vuelvan fieles al Sitio, premiándolos por realizar acciones como comentar, publicar artículos, entre otras. Para plantear todas estas acciones existen plataformas como "Captain Up", la cual le permite a los usuarios ganar puntos, medallas, subir de nivel (Principiante, Avanzado) realizando algunas acciones como compartir información en sus perfiles sociales, ver vídeos, twittear, dar Me Gusta, etc. A medida que los usuarios realicen una, dos, más veces las acciones mencionadas, se volverán fieles al Sitio Web. Esta fidelidad se mide a través de las visitas recurrentes que el Sitio recibe. Existen muchas herramientas que permiten saber la cantidad de visitantes, clasificada por país, ciudad, sistema operativo, navegador utilizado, redes sociales. La más usada y recomendada es "Google Analytics". Para que la experiencia del usuario sea la mejor, el Sitio Web debe contar con un Diseño Interactivo. Las tecnologías que permiten realizar este tipo de Diseño son: HTML5, CSS3, jQuery, Bootstrap, Responsive Web Design.

Completamente afín( ) Muy afín( ) Medianamente afín(X) Poco afín( ) Nada afín( )

Publicación #4

Título:	Consorcio Web : Formulación e implementación de un marco conceptual de integración eficiente de sitios web dentro de una comunidad en internet
Abstract:	La realización de este trabajo plantea implementar una comunidad vía Internet conformada por diversos sitios web autosostenidos, denominado, para efectos del estudio, como consorcio web, demostrar que la implementación de dicho esquema es factible mediante un análisis de caso, y que eventualmente permitiría al usuario la obtención de ingresos; la innovación propuesta radica en la utilización de diferentes categorías de sitios web que se complementan mutuamente, conformando un esquema simbiótico que contribuye a la sinergia del modelo total. El problema identificado fue la no existencia de un marco conceptual capaz de integrar de manera eficiente a los diversos sitios web. Adicionalmente, se busca identificar las ventajas y desventajas que este esquema de gestión implicaría, identificar los riesgos de su implementación y formalizar la metodología utilizada tanto a fin de hacerla repetible en el tiempo como de facilitar su transmisión a otras personas.

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín(X) Nada afín( )

Publicación #5

Título:	Sistema informático web basado en la tecnología RIA"S para el control de personal de una empresa de seguridad y vigilancia privada
Abstract:	Este trabajo de investigación presenta un estudio y posterior desarrollo de una aplicación Web siguiendo los conceptos de RIA - Rich Internet Application, que pretende obtener un mejor rendimiento de la tecnología del lado cliente a fin de proveer nuevos tipos de sitios Web, más interactivos, que se caracterizan por la sofisticación de aplicativos de escritorio, pero sin comprometer la facilidad de desarrollo y gerencia de los aplicativos Web. Para el desarrollo se utilizó la plataforma apache flex, una nueva tecnología que permite construir aplicaciones para Internet de manera más rápida, inteligente, y fluida. El foco principal de este estudio es demostrar las ventajas de este nuevo concepto de desarrollo de aplicativos Web y, consecuentemente, de la plataforma apache flex.

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín(X) Nada afín( )

Publicación #6

Título:	Transmisión de datos vía TCP/IP en tiempo real
Abstract:	La automatización de procesos no ha sido una tarea sencilla de solucionar. Mas aún si se trata de automatizar procesos de transferencia de información, es decir, tratar de lograr que el usuario obtenga de manera automática la información y que la pueda visualizar. Actualmente esta automatización se trata de solventar con aplicaciones Web; pero este tipo de aplicaciones tiene limitaciones al momento de actualizar la de manera automática en el Navegador Web. Existen componentes que son utilizados dentro de las páginas Web para mostrar ese comportamiento dinámico que requiere la actualización automática de (a información; esos componentes se denominan Applets. pero estos componentes por lo general consumen una regular cantidad de recursos de memoria, además son peligrosos en cuanto a la vulnerabilidad en la seguridad de la PC del Cliente.

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín(X)

Publicación #7

Título:	Desarrollo de un portal web que permita mejorar la gestión y el acceso a la información para la Escuela de Informática de la Universidad Nacional de Trujillo
Abstract:	Due to the deficit with which account the current website of Computing National University of Trujillo, in a matter of informing and providing academic students of Computer Science, National University of Trujillo services; and building on existing information technologies, the web portal for the School of Computing, National University of Trujillo was

	developed., which improved the management and access to information. Throughout the development of this thesis, we have applied the RUP; and all artifacts, UML diagrams and modelamientos necessary for the further development of the Web Portal, was developed the website using the Joomla CMS. Furthermore, the design of tests were conducted to verify the functionality of the web portal, which were satisfactory. For the proof of the thesis, two surveys were applied to the end users of the web portal, the first survey was conducted before the implementation of the proposed web portal, where users had the old website; The second survey was conducted after completion of the proposed web portal; all this in order to demonstrate the improved management and access to information web portal. The obtained and subsequently discussed final results, they conclude that the web portal developed able to improve the management and access to information for School of Computing National University of Trujillo

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #8

Título:	Desarrollo de un portal web para la Facultad de Ciencias Físicas y Matemáticas de la UNT
Abstract:	Considering that the growth of technology and the need to stay connected globally with the world, it is feasible to have a dynamic and operative web portal that allow attend requirements of faculty of Physics and Mathematics of the National University of Trujillo about services of broadcast news, events and academic information, also allows us to create a competitive image within the university environment. The website is aimed at authorities, administrative officials, students, teachers and the general public, allows the dissemination of activities in an efficient, effective and fast results in the delivery of information to the user. The portal will be a tool that gives users a single point of access to certain services (search of documents, corporate information, class schedules, information on academic events, news, information adjusted to user profile, photo gallery, etc.). The purpose of this project is therefore to design and implement a website using a simplified version of the RUP, modeling language UML and tool CMS JOOMLA, to which are installed and configured in the portal extensions that give solution all functional requirements that are demanded, which will be fully integrated into a single web environment

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #9

Título:	La especificación de requisitos para la usabilidad del sitio Web de la empresa Crecer Jugando E.I.R.L., en la ciudad de Arequipa, año 2011
Abstract:	El trabajo busca analizar hasta qué punto la especificación de requisitos influye en la usabilidad del sitio web de la

	empresa Crecer Jugando EIRL, en la ciudad de Arequipa, año 2011. La investigación estuvo enmarcada dentro de un estudio de campo con diseño no experimental, método descriptivo, de carácter transaccional, para lo cual se ha empleado como instrumento de recolección de datos, la encuesta, mediante un cuestionario. Entre las conclusiones más destacadas podemos mencionar que la correcta especificación de requisitos en la etapa inicial del desarrollo de un sitio web influye al momento de realizar el análisis de usabilidad a dicho sitio web, detectando errores y anomalías que se pueden corregir modificando los requisitos. El presente trabajo ayuda también a detectar problemas con el hardware del usuario que pocas veces se toma en cuenta. Cabe resaltar que una correcta especificación de requisitos nos ayudará a elaborar un plan de desarrollo más eficiente ganando tiempo para tener un resultado óptimo acorde con las especificaciones del usuario.

Completamente afín( ) Muy afín(X) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #10

Título:	Evaluación de la usabilidad en las interfaces de usuario de las aplicaciones web mediante normas de calidad
Abstract:	Hoy en día las páginas web se han convertido en una medio de comunicación muy útil para cualquier tipo de organización, En la actualidad una página web se encuentra en diversos campos de la actividad humana, por lo que es muy importante que reúna ciertos criterios de calidad para satisfacer las necesidades de los usuarios como es la Usabilidad ya que aquellos sitios web que ofrezcan a los usuarios información útil, bien organizada y navegable dentro de un diseño bien estructurado, tienen más probabilidades de retener a los usuarios, Y de ello puede depender el éxito o fracaso de la misma. El presente trabajo de investigación proporciona un enfoque para la evaluación de la usabilidad de aplicaciones web. Para ello evaluamos la aplicación de gestión de ventas SAAS de la empresa EKAMPERU S.A.C.; con la ayuda de la norma ISO 9241 – 11. Para realizar dicha evaluación se realizó una guía de observación la cual contiene 21 actividades que el usuario realizó y también elegimos un cuestionario SUS (Escala para la usabilidad de los sistemas). Una vez obtenidos y analizados los resultados se diseñó un plan de mejoras, en donde se propone varias acciones que se deberían de realizar para optimizar la interacción de los usuarios con la aplicación evaluada. Finalmente, como resultado general de toda la investigación, se puede concluir que se han alcanzado todos los objetivos propuestos al inicio de la misma.

Completamente afín(X) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Comité de ética para la investigación con seres humanos y animales – CEI(sha)
Vicerrectorado de Investigación - PUCP

PROTOCOLO DE CONSENTIMIENTO INFORMADO PARA CALIFICACIÓN DE RECOMENDACIONES DE PUBLICACIONES CIENTÍFICAS

El propósito de este protocolo es de brindar a los participantes en esta investigación, una explicación clara de la naturaleza de la misma, así como del rol que tienen en ella.

La presente investigación es conducida por la estudiante del programa de pre grado en Ingeniería Informática de la Pontificia Universidad Católica del Perú, Elizabeth Jenisse Vereau Zagastizábal, asesorada por el Ing. Cesar Olivares Poggi. La meta de este estudio es la de evaluar el sistema de recomendación de publicaciones científicas de ALICIA propuesto en el Proyecto de Fin de Carrera de la estudiante, como parte de los cursos de INF391 Proyecto de Tesis1 e INF392 Proyecto de Tesis 2.

Si usted accede a participar en este estudio, se le pedirá participar de respondiendo una encuesta, lo cual le tomará 15 minutos de su tiempo. Las encuestas realizadas serán recopiladas por el investigador con el propósito de analizar los resultados y las ideas que usted haya expresado. Una vez finalizado el estudio todos los documentos serán destruidos.

Su participación será voluntaria. La información que se recoja será estrictamente confidencial y no se podrá utilizar para ningún otro propósito que no esté contemplado en esta investigación.

En principio, la encuesta resuelta por usted será anónima, por ello será codificada utilizando un número de identificación. Si la naturaleza del estudio requiriera su identificación, ello solo será posible si es que usted da su consentimiento expreso para proceder de esa manera.

Si tuviera alguna duda con relación al desarrollo del proyecto, usted es libre de formular las preguntas que considere pertinentes. Además, puede finalizar su participación en cualquier momento del estudio sin que esto represente algún perjuicio para usted. Si se sintiera incómoda o incómodo, frente a alguna de las preguntas, puede ponerlo en conocimiento de la persona a cargo de la investigación y abstenerse de responder.

Muchas gracias por su participación.

Yo, [redacted] doy mi consentimiento para participar en el estudio y soy consciente de que mi participación es enteramente voluntaria.

He recibido información en forma verbal sobre el estudio mencionado anteriormente y he leído la información escrita adjunta. He tenido la oportunidad de discutir sobre el estudio y hacer preguntas. Entiendo que el experimento tiene por objetivo evaluar las recomendaciones generadas por un sistema de software, NO mis capacidades /habilidades /conocimientos.

Al firmar este protocolo estoy de acuerdo con que mis datos personales, incluyendo datos relacionados a mi salud física y mental o condición, y raza u origen étnico, podrían ser usados según lo descrito en la hoja de información que detalla la investigación en la que estoy participando.

Entiendo que puedo finalizar mi participación en el estudio en cualquier momento, sin que esto represente algún perjuicio para mí. Entiendo que puedo comunicar al supervisor, en cualquier momento, sobre algún malestar, molestia o inconformidad que pueda sentir durante el desarrollo de las actividades, y que, por tal motivo, puedo abandonar las actividades en cualquier momento.

Entiendo que recibiré una copia de este formulario de consentimiento e información del estudio y que puedo pedir información sobre los resultados de este estudio cuando éste haya concluido. Para esto, puedo comunicarme con la estudiante Elizabeth Jenisse Vereau Zagastizábal al correo jenisse.vereau@pucp.pe o al teléfono 981546409, o de la misma forma, con el asesor Ing. Cesar Olivares Poggi al correo cesar.olivares@pucp.pe.

[redacted] [redacted] 19/06/18

Nombre completo del (de la) participante | Firma | Fecha

ELIZABETH JENISSE VEREAU ZAGASTIZÁBAL

Para la elaboración de este protocolo se ha tenido en cuenta el formulario de C.I. del Comité de Ética del Departamento de Psicología de la PUCP

Cuestionario para Calificación de Recomendaciones de Publicaciones Científicas

Estimado(a) investigador(a):

Gracias por participar y completar esta encuesta, que tiene por objeto obtener información cuantitativa sobre recomendaciones generadas por el sistema de recomendación de publicaciones científicas de ALICIA a investigadores calificados del SINACYT.

1- ¿Cuán afines son las siguientes 10 publicaciones al ámbito de investigación en la cual Ud.se desenvuelve?

Publicación #1

Título:	Estudio de usabilidad en la una propuesta de sitio Web basado en el diseño de la experiencia del´usuario
Abstract:	El propósito de esta investigación fue determinar que, en comparación al Sitio Web original, la aplicación del Diseño de la Experiencia del Usuario optimiza la Usabilidad en la Propuesta de Sitio Web. Para ello, se estudiaron los fundamentos, elementos y características del Diseño de la Experiencia de Usuario, y se aplicaron una serie de metodologías de desarrollo, propias de esta disciplina, para definir las necesidades, tanto de la institución propietaria del sitio web como de sus usuarios. Así se establecieron los requerimientos de la Propuesta de Sitio Web y posteriormente se creó un prototipo como Mínimo Producto Viable de la Propuesta. Para poder comparar la usabilidad en ambos sitios web, estos se evaluaron por medio de un instrumento de investigación basado en el Estudio de Usabilidad. Concluyendo, de esta forma, que la aplicación del Diseño de la Experiencia del Usuario permitió que la Propuesta de Sitio Web obtenga, tanto una mejor Usabilidad como una mejor retroalimentación de parte de los usuarios finales.

Completamente afín( ) Muy afín(X) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #2

Título:	Reconocimiento de patrones de comportamiento de usuarios en el portal web usando web Mining
Abstract:	La siguiente investigación tiene la finalidad de encontrar los patrones de comportamiento de los usuarios de un Portal Web, para esto se utiliza la metodología de Web Mining, la técnica del Clustering y el algoritmo de K-Means: esta metodología permite formar grupos con características iguales o similares (preferencias y cantidad de visitas a uno o más servicios). Una vez que se tenga el conocimiento de los grupos se personaliza el Portal Web para cada uno estos grupos sobre la base de sus preferencias. Con esto se logra un menor tiempo de acceso al servicio deseado y un mayor grado de satisfacción de los usuarios. Los datos a analizar se extraen de los registros de acceso de usuarios hacia el Portal Web (archivos Logs) en los cuales se registran los datos de navegación de los usuarios, y éstos se almacenan en el servidor web.

Completamente afín( ) Muy afín( ) Medianamente afín(X) Poco afín( ) Nada afín( )

Publicación #3

Título:	Consorcio Web : Formulación e implementación de un marco conceptual de integración eficiente de sitios web dentro de una comunidad en internet
Abstract:	La realización de este trabajo plantea implementar una comunidad vía Internet conformada por diversos sitios web auto-sostenidos, denominado, para efectos del estudio, como consorcio web, demostrar que la implementación de dicho esquema es factible mediante un análisis de caso, y que eventualmente permitiría al usuario la obtención de ingresos; la innovación propuesta radica en la utilización de diferentes categorías de sitios web que se complementan mutuamente, conformando un esquema simbiótico que contribuye a la sinergia del modelo total. El problema identificado fue la no existencia de un marco conceptual capaz de integrar de manera eficiente a los diversos sitios web. Adicionalmente, se busca identificar las ventajas y desventajas que este esquema de gestión implicaría, identificar los riesgos de su implementación y formalizar la metodología utilizada tanto a fin de hacerla repetible en el tiempo como o de facilitar su transmisión a otras personas.

Completamente afín( ) Muy afín( ) Medianamente afín(X) Poco afín( ) Nada afín( )

Publicación #4

Título:	Metodología para la evaluación de la usabilidad de sitios web aplicada a una empresa exportadora de productos naturales
Abstract:	La evolución tecnológica propone nuevos desafíos para asegurar la "Accesibilidad Universal." Los diferentes tipos de usuarios de la Web, se enfrentan con numerosas barreras de accesibilidad cuando interactúan con los diferentes tipos de sitios y aplicaciones que coexisten hoy en la Web 2.0, desde las páginas web, las aplicaciones móviles, sistemas de información y dispositivos electrónicos como un Smart TV. Necesitan metodologías que permitan evaluar la evolución del equipo en términos de "que tan útil es" con la finalidad de retroalimentarse en esa iteración y nutrirse de mejoras necesarias antes de su lanzamiento a producción. El presente proyecto de investigación formula una metodología que permite evaluar en términos de usabilidad las páginas web. Con la finalidad de comprobar su aplicabilidad se escogió una empresa del sector industrial dedicada a la comercialización y exportación de bebidas basadas en productos naturales, en donde al aplicar la metodología permitió clasificar su página web como página usable, identificando los puntos a mejorar.

Completamente afín(X) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #5

Título:	Desarrollo de un portal web que permita mejorar la gestión y el acceso a la información para la Escuela de Informática de la Universidad Nacional de Trujillo
Abstract:	Debido al déficit con el que cuenta el sitio web actual de la Universidad Nacional de Computación de Trujillo, en cuestión de informar y proporcionar estudiantes académicos de Ciencias de la Computación, servicios de la Universidad Nacional de Trujillo; y sobre la base de las tecnologías de la información existentes, se desarrolló el portal web de la Facultad de Informática de la Universidad Nacional de Trujillo, que mejoró la gestión y el acceso a la información. A lo largo del desarrollo de esta tesis, hemos aplicado el RUP; y todos los artefactos, diagramas UML y modelamientos necesarios para el posterior desarrollo del Portal Web, se desarrolló el sitio web utilizando el Joomla CMS. Además, el diseño de las pruebas se realizó para verificar la funcionalidad del portal web, que fueron satisfactorias. Para la prueba de la tesis, se aplicaron dos encuestas a los usuarios finales del portal web, la primera encuesta se realizó antes de la implementación del portal web propuesto, donde los usuarios tenían el sitio web anterior; La segunda encuesta se realizó después de la finalización del portal web propuesto; todo esto para demostrar el mejor portal de administración y acceso a la información. Los resultados finales obtenidos y posteriormente discutidos, concluyen que el portal web desarrollado permitió mejorar la gestión y acceso a la información para la Facultad de Informática de la Universidad Nacional de Trujillo

Completamente afín( ) Muy afín( ) Medianamente afín(X) Poco afín( ) Nada afín( )

Publicación #6

Título:	Herramientas 3.0 para la calidad educativa
Abstract:	La Web 3.0 es un término que no termina de tener un significado ya que varios expertos han intentado dar definiciones que no concuerdan o encajan la una con la otra pero que, en definitiva, va unida a veces con la Web Semántica. En lo que a su aspecto semántico se refiere, la Web 3.0 es una extensión del World Wide Web en la que se puede expresar no sólo lenguaje natural, también se puede utilizar un lenguaje que se puede entender, interpretar utilizar por agentes software, permitiendo de este modo encontrar, compartir e integrar la información más fácilmente.

Completamente afín( ) Muy afín( ) Medianamente afín(X) Poco afín( ) Nada afín( )

Publicación #7

Título:	Implementación de estrategias de gamificacion y buenas prácticas para el sitio web del congreso internacional de ingeniería de la UPAO
Abstract:	La gamificación es un conjunto de técnicas de juego utilizadas en entornos no lúdicos, como por ejemplo el ambiente laboral, selección de personal, sitios web. La

	gamificación en un Sitio Web busca que los usuarios se vuel van fieles al Sitio, premiándolos por realizar acciones com o comentar, publicar artículos, entre otras. Para plantear todas estas acciones existen plataformas como "Captain Up", la cual le permite a los usuarios ganar puntos, medallas, s ubir de nivel (Principiante, Avanzado) realizando algunas a cciones como compartir información en sus perfiles sociales , ver vídeos, twittear, dar Me Gusta, etc. A medida que los usuarios realicen una, dos, más veces las acciones menciona das, se volverán fieles al Sitio Web. Esta fidelidad se mid e a través de las visitas recurrentes que el Sitio recibe. Existen muchas herramientas que permiten saber la cantidad de visitantes, clasificada por país, ciudad, sistema operat ivo, navegador utilizado, redes sociales. La más usada y re comendada es "Google Analytics". Para que la experiencia de l usuario sea la mejor, el Sitio Web debe contar con un Diseño Interactivo. Las tecnologías que permiten realizar e ste tipo de Diseño son: HTML5, CSS3, jQuery, Bootstrap, Re sponsive Web Design.

Completamente afín( ) Muy afín(X) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #8

Título:	La especificación de requisitos para la usabilidad del siti o Web de la empresa Crecer Jugando E.I.R.L., en la ciudad d e Arequipa, año 2011
Abstract:	El trabajo busca analizar hasta qué punto la especificación de requisitos influye en la usabilidad del sitio web de la empresa Crecer Jugando EIRL, en la ciudad de Arequipa, año 2011. La investigación estuvo enmarcada dentro de un estudi o de campo con diseño no experimental, método descriptivo, de carácter transaccional, para lo cual se ha empleado como instrumento de recolección de datos, la encuesta, mediante un cuestionario. Entre las conclusiones más destacadas pode mos mencionar que la correcta especificación de requisitos en la etapa inicial del desarrollo de un sitio web influye al momento de realizar el análisis de usabilidad a dicho si tio web, detectando errores y anomalías que se pueden corre gir modificando los requisitos. El presente trabajo ayuda t ambién a detectar problemas con el hardware del usuario que pocas veces se toma en cuenta. Cabe resaltar que una correc ta especificación de requisitos nos ayudará a elaborar un p lan de desarrollo más eficiente ganando tiempo para tener u n resultado óptimo acorde con las especificaciones del usua rio.

Completamente afín(X) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #9

Título:	Portales de la Presidencia del Consejo de Ministros en el A ño 2014" "ESTÁNDAR Wcag 2.0 y Su Accesibilidad Web
Abstract:	En el presente trabajo de investigación, se hace un estudio sobre la accesibilidad en los sitios web de la Presidencia de Consejo de Ministros (PCM), con el objetivo de determina r la medida en que se cumplen los principios de accesibilid ad web según el estándar WCAG 2.0 en los portales de la PCM en el año 2014. El trabajo es de tipo: Observacional, descr

	iptivo y transversal. La población de estudio estuvo constituida por 1184 páginas web y una muestra de 290. Los datos se han recolectado con un cuestionario de 61 preguntas el cual se ha divido en 4 partes: Perceptibilidad, Operabilidad, Comprensibilidad y Robustez. El nivel de accesibilidad se categorizó A, AA, AAA y No se cumple el estándar. Los datos fueron analizados, tratados mediante estadística descriptiva y llegando a la conclusión de que no se cumplen los principios de accesibilidad web según el estándar WCAG 2.0 en mayor proporción en los portales de la Presidencia del Consejo de Ministros en el año 2014

Completamente afín(X) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #10

Título:	Evaluación de la usabilidad en las interfaces de usuario de las aplicaciones web mediante normas de calidad
Abstract:	Hoy en día las páginas web se han convertido en una medio de comunicación muy útil para cualquier tipo de organización, En la actualidad una página web se encuentra en diversos campos de la actividad humana, por lo que es muy importante que reúna ciertos criterios de calidad para satisfacer las necesidades de los usuarios como es la Usabilidad ya que aquellos sitios web que ofrezcan a los usuarios información útil, bien organizada y navegable dentro de un diseño bien estructurado, tienen más probabilidades de retener a los usuarios, Y de ello puede depender el éxito o fracaso de la misma. El presente trabajo de investigación proporciona un enfoque para la evaluación de la usabilidad de aplicaciones web. Para ello evaluamos la aplicación de gestión de ventas SAAS de la empresa EKAMPERU S.A.C.; con la ayuda de la norma ISO 9241 - 11. Para realizar dicha evaluación se realizó una guía de observación la cual contiene 21 actividades que el usuario realizó y también elegimos un cuestionario SUS (Escala para la usabilidad de los sistemas). Una vez obtenidos y analizados los resultados se diseñó un plan de mejoras, en donde se propone varias acciones que se deberían de realizar para optimizar la interacción de los usuarios con la aplicación evaluada. Finalmente, como resultado general de toda la investigación, se puede concluir que se han alcanzado todos los objetivos propuestos al inicio de la misma.

Completamente afín(X) Muy afín( ) Medianamente afín( ) Poco afín( ) Nada afín( )

Comité de ética para la investigación con seres humanos y animales – CEI(sha)
Vicerrectorado de Investigación - PUCP

PROTOCOLO DE CONSENTIMIENTO INFORMADO PARA CALIFICACIÓN DE RECOMENDACIONES DE PUBLICACIONES CIENTÍFICAS

El propósito de este protocolo es de brindar a los participantes en esta investigación, una explicación clara de la naturaleza de la misma, así como del rol que tienen en ella.

La presente investigación es conducida por la estudiante del programa de pre grado en Ingeniería Informática de la Pontificia Universidad Católica del Perú, Elizabeth Jenisse Vereau Zagastizábal, asesorada por el Ing. Cesar Olivares Poggi. La meta de este estudio es la de evaluar el sistema de recomendación de publicaciones científicas de ALICIA propuesto en el Proyecto de Fin de Carrera de la estudiante, como parte de los cursos de INF391 Proyecto de Tesis1 e INF392 Proyecto de Tesis 2.

Si usted accede a participar en este estudio, se le pedirá participar de respondiendo una encuesta, lo cual le tomará 15 minutos de su tiempo. Las encuestas realizadas serán recopiladas por el investigador con el propósito de analizar los resultados y las ideas que usted haya expresado. Una vez finalizado el estudio todos los documentos serán destruidos.

Su participación será voluntaria. La información que se recoja será estrictamente confidencial y no se podrá utilizar para ningún otro propósito que no esté contemplado en esta investigación.

En principio, la encuesta resuelta por usted será anónima, por ello será codificada utilizando un número de identificación. Si la naturaleza del estudio requiriera su identificación, ello solo será posible si es que usted da su consentimiento expreso para proceder de esa manera.

Si tuviera alguna duda con relación al desarrollo del proyecto, usted es libre de formular las preguntas que considere pertinentes. Además, puede finalizar su participación en cualquier momento del estudio sin que esto represente algún perjuicio para usted. Si se sintiera incómoda o incómodo, frente a alguna de las preguntas, puede ponerlo en conocimiento de la persona a cargo de la investigación y abstenerse de responder.

Muchas gracias por su participación.

Yo, [redacted] doy mi consentimiento para participar en el estudio y soy consciente de que mi participación es enteramente voluntaria.

He recibido información en forma verbal sobre el estudio mencionado anteriormente y he leído la información escrita adjunta. He tenido la oportunidad de discutir sobre el estudio y hacer preguntas. Entiendo que el experimento tiene por objetivo evaluar las recomendaciones generadas por un sistema de software, NO mis capacidades /habilidades /conocimientos.

Al firmar este protocolo estoy de acuerdo con que mis datos personales, incluyendo datos relacionados a mi salud física y mental o condición, y raza u origen étnico, podrían ser usados según lo descrito en la hoja de información que detalla la investigación en la que estoy participando.

Entiendo que puedo finalizar mi participación en el estudio en cualquier momento, sin que esto represente algún perjuicio para mí. Entiendo que puedo comunicar al supervisor, en cualquier momento, sobre algún malestar, molestia o inconformidad que pueda sentir durante el desarrollo de las actividades, y que, por tal motivo, puedo abandonar las actividades en cualquier momento.

Entiendo que recibiré una copia de este formulario de consentimiento e información del estudio y que puedo pedir información sobre los resultados de este estudio cuando éste haya concluido. Para esto, puedo comunicarme con la estudiante Elizabeth Jenisse Vereau Zagastizábal al correo jenisse.vereau@pucp.pe o al teléfono 981546409, o de la misma forma, con el asesor Ing. Cesar Olivares Poggi al correo cesar.olivares@pucp.pe.

[redacted] [redacted] 20/06/18

Nombre completo del (de la) participante | Firma | Fecha

ELIZABETH JENISSE VEREAU ZAGASTIZÁBAL

Para la elaboración de este protocolo se ha tenido en cuenta el formulario de C.I. del Comité de Ética del Departamento de Psicología de la PUCP

Cuestionario para Calificación de Recomendaciones de Publicaciones Científicas

Estimado(a) investigador(a):

Gracias por participar y completar esta encuesta, que tiene por objeto obtener información cuantitativa sobre recomendaciones generadas por el sistema de recomendación de publicaciones científicas de ALICIA a investigadores calificados del SINACYT.

1- ¿Cuán afines son las siguientes 10 publicaciones al ámbito de investigación en la cual Ud.se desenvuelve?

Publicación #1

Título:	Análisis del actual algoritmo de asignación de ancho de banda de la red Ethernet óptica pasiva y del proceso de acceso al medio de la Red Óptica conmutada en ráfagas para la integración entre estas dos redes
Abstract:	En la presente Tesis, basándose en una red todo óptico (red de acceso óptico y red de core/metro óptico) en base a soportar los nuevos servicios que mejoran y aportan una calidad de vida a la sociedad como el e-learning, e-business, e-health y el e-government; se plantea el uso de la red Ethernet Óptica Pasiva (EPON) como red de acceso y la red óptica de conmutación de Burst (OBS) como red de core/metro. Se analizan las técnicas de acceso al medio de transmisión de las redes EPON y OBS pasando por la descripción de cada una de estas redes, la forma en que asignan el ancho de banda, analizando el rendimiento de la red en términos de retardo y overhead de la red conformada por la integración de EPON y OBS y finalmente planteando la necesidad del desarrollo de un nuevo algoritmo de asignación de ancho de banda dinámico para la red EPON en concordancia de integración con la red OBS.

Completamente afín( ) Muy afín(X) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #2

Título:	Diseño para la interconexión de redes bancarias compartiendo una Arquitectura Ethernet - ATM
Abstract:	Describir una Solución AM/Ethernet para la Interconexión de Redes Bancaria con los proveedores de Información que tengan en común, logrando generar una Red flexible, funcional y escalable. Adicionalmente, lograr compartir los costos en de dicha red disminuyendo la cantidad de enlace y estandarizando la manera de entregarlo

Completamente afín( ) Muy afín(X) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #3

Título:	MODELO DE IMPLEMENTACIÓN JERÁRQUICO DE LA RED DE DATOS IP D E LA UNIVERSIDAD CATÓLICA DE SANTA MARÍA UTILIZANDO LA METO DOLOGÍA TOP DOWN Y EL ESTÁNDAR IEEE 802.3AE, IEEE 802.3AN Y IEEE 802.3AB
Abstract:	MODELO OSI MODELO DE REDES JERÁRQUICAS CAPAS DEL MODELO JER ÁRQUICO BENEFICIOS PRINCIPIOS IEEE TECNOLOGÍAS LAN GIGABIT ETHERNET 10 GIGABIT ETHERNET REDES CONVERGENTES REDES MÚLTI PLES DE MÚLTIPLES SERVICIOS REDES CONVERGENTES REDES DE INF ORMACIÓN INTELIGENTES CALIDAD DE SERVICIO QOS QUE ES QOS FI NALIDAD DE QOS FACTORES QUE AFECTAN LA QOS PRINCIPIOS DE QO S ARQUITECTURA DE QOS HERRAMIENTAS DE CLASIFICACIÓN Y MARCA DO DE PAQUETES METODOLOGÍA DE IMPLEMENTACIÓN DE RED DE DATO S DE CAMPUS Y ENTERPRISE OBJETIVO NECESIDAD DE UN MODELO DE IMPLEMENTACIÓN ESTRUCTURA DEL MODELO IDENTIFICAR LAS NECESI DADES Y LOS OBJETIVOS ANALIZAR LAS METAS Y RESTRICCIONES DE LA ORGANIZACIÓN UTILIZAR LA METODOLOGÍA TOP DOWN ANALIZAR L AS METAS DE LA ORGANIZACIÓN ANALIZAR METAS Y RESTRICCIONES ANALIZAR LAS METAS TÉCNICAS Y COMPENSACIONES ESCALABILIDAD DISPONIBILIDAD PERFORMANCE DE LA RED SEGURIDAD CAPACIDAD DE ADMINISTRACIÓN USABILIDAD ADAPTABILIDAD ASEQUIBILIDAD HACER COMPENSACIONES EN EL DISEÑO DE RED REVISIÓN DE METAS TÉCNIC AS IDENTIFICAR LA RED EXISTENTE REVISAR LA SALUD DE LA RED ACTUAL LISTA DE REVISIÓN DE LA SALUD DE LA RED IDENTIFICAR EL TRAFICO DE RED IDENTIFICAR EL FLUJO DE TRAFICO IDENTIFIC AR LA CARGA DE TRAFICO IDENTIFICAR EL COMPORTAMIENTO DEL TR AFICO IDENTIFICAR REQUERIMIENTOS DE CALIDAD DE SERVICIO LIS TA DE REVISIÓN DE TRAFICO DE RED DISEÑO DE RED LÓGICO DISEÑ AR LA TOPOLOGÍA DE RED DISEÑAR MODELOS PARA DIRECCIONAMIENT O Y NUMERA MIENTO SELECCIONAR PROTOCOLOS DE SWITCHING Y ROU TING DESARROLLAR ESTRATEGIAS DE SEGURIDAD DE LA RED DESARRO LLAR ESTRATEGIAS DE ADMINISTRACIÓN DE LA RED DISEÑO DE RED FÍSICO SELECCIONAR TECNOLOGÍAS Y DISPOSITIVOS PARA REDES DE CAMPUS SELECCIONAR TECNOLOGÍAS Y DISPOSITIVOS PARA REDES EN TERPRISE PRUEBAS, OPTIMIZACIÓN, DOCUMENTACIÓN DEL DISEÑO PR OBAR EL DISEÑO DE RED OPTIMIZAR EL DISEÑO DE RED DOCUMENTAR EL DISEÑO DE RED IMPLEMENTACIÓN DE LA RED DE CAMPUS DE LA U CSM RESUMEN DEL PROYECTO METAS DEL PROYECTO ÁMBITO REQUERIM IENTOS DE DISEÑO ESTADO ACTUAL DE LA RED - MAYO 2011 DISEÑO LÓGICO PROPUESTO DISEÑO FÍSICO PROPUESTO RESULTADOS DE LAS PRUEBAS VALIDACIONES COMPROBACIÓN DE LA HIPÓTESIS VALIDACIÓ N DE INDICADORES

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín(X) Nada afín( )

Publicación #4

Título:	Diseño de Red Power Over Ethernet con Categoría 6A para Apl icación en Data Center
Abstract:	La presente tesis describe el Proyecto de Diseño de una Red Over Ethernet sobre una red f?sica DE 10 Gigabit Ethernet u tilizando categor?a 6A, demostrando la convergencia de esta s dos tecnolog?as logrando el mayor rendimiento y cumpliend o con todas las normas internacionales y nacionales aplicad os en un Data Center.

Completamente afín( ) Muy afín(X) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #5

Título:	Convergencia de redes en el servicio de larga distancia en el Perú
Abstract:	Con la llegada de las Redes de Nueva Generación (NGN), se h ace inevitable realizar un proyecto que involucre el "inter working" entre la red mencionada líneas arriba, la Red de T elefonía Pública Conmutada (PSTN) y la Red de Datos. Es por eso por lo que el presente Informe de Suficiencia utiliza e stas tres redes con el fin de crear un servicio que permite al público usuario, con acceso a Internet por Banda Ancha, realizar llamadas de larga distancia nacional e internacion al. La utilización de la Red de Telefonía Pública Conmutada hace que no se tenga que invertir grandes cantidades de din ero en una reconversión tecnológica hasta el usuario final y además aprovechar la calidad de servicio innata en la red PSTN. Utilizar la Red de Datos, en este caso se utiliza la red IP, en el transporte de voz tiene como objetivo reducir los costos operativos en el transporte de dicho tráfico, as í como aprovechar todos los beneficios que nos brinda dicha red. Para una fácil comunicación entre ambas redes se hace uso de los elementos que conforman una red NGN. Todo esto a punta a una reducción de tarifas en las llamadas de larga d istancia en beneficio del usuario final.

Completamente afín( ) Muy afín(X) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #6

Título:	Análisis y actualización del diseño de la red de datos de l a clínica adventista Ana Stahl, Iquitos-Perú
Abstract:	El presente trabajo de análisis y actualización de la red d e datos de la Clínica Adventista Ana Stahl (CAAS), se hizo para solucionar la problemática encontrada en relación a su s redes de datos. Esta problemática consistía en la falta d e control del crecimiento de la red. Actualmente la CAAS cu enta con más de ochenta (80) usuarios con estaciones de cóm puto y algunas impresoras de red. La distribución de los eq uipos se h,¡¡ce sin un previo análisis y según la necesidad del momento. No se proyecta a futuros usuarios ni un porcen taje decrecimiento por áreas. Por este motivo se desarrolla un diseño que tiene como objetivo la adecuada distribución de los dispositivos de red, así como de los puntos de red d e cada área ubicada en el primer piso y las viviendas ubica das en la parte posterior de las oficinas administrativas q ue constituyen el CAMPUS. Asimismo, se propone un diseño fí sico y lógico de toda la red de datos de la CAAS. Y finalme nte se documento la red existente mediante un diagrama lógi co y estadísticas que muestran el tipo de tráfico predomina nte. Una de las alternativas propuestas para mejorar la seg uridad y control en los switches, es el de ubicarlos en gab inetes. A nivel lógico, bloquear los puertos no usados y cr ear VLAN's mejora el rendimiento de la red aumentando la se guridad en los switches y disminuyendo el tráfico broadcast que constituía más del 50% de todo el tráfico de paquetes. Para el desarrollo de este trabajo se empleó la metodología Descendente (Top Down) con la que se lograra una red óptima de datos, mejorando el flujo de información y la comunicaci ón con los servidores, basándonos en las ventajas que ofrec e el uso de dispositivos de red, como son los equipos Switc h administrables de capa 2 y cable UTP categoría 5e; esto p

	ermitirá a esta red LAN cumplir con los principales benefic ios que representa este tipo de redes que son: fiabilidad, confiabilidad, escalabilidad, seguridad, control de acceso en el área de cobertura de todo el campus de la Clínica y s obre todo la posibilidad de crecimiento de interconectivida d hacia otras áreas y así poder tener una mejor administrac ión de la red.

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín(X) Nada afín( )

Publicación #7

Título:	Segmentación de redes LAN mediante VLANs
Abstract:	Sabemos que la mayoría de las redes LAN implementadas en nu estro país están basadas en Ethernet, estas redes LAN han c recido en número de computadores en los últimos años. El tr áfico de datos en las LAN se ha incrementado como consecuen cia de nuevos servicios y más computadores en la LAN. En varias empresas estatales se tienen un buen número de co mputadores todos ellos en un solo dominio de difusión, con problemas de congestión debidos al alto tráfico de difusión . El presente trabajo, intenta mostrar cómo podemos mejorar e l rendimiento de la LAN mediante el uso de redes LAN virtua les (VLAN). En los primeros capítulos se describe en qué co nsisten las VLANs, los tipos de VLAN, los protocolos utiliz ados. Más adelante se muestran como se configuran las VLAN en con mutadores de la marca Cisco, luego se tiene una aplicación de un proyecto de VLAN en una empresa local. Finalmente se describe las ventajas que se tienen al implem entar VLAN, que mejoras se logran en la red.

Completamente afín( ) Muy afín( ) Medianamente afín( ) Poco afín(X) Nada afín( )

Publicación #8

Título:	Adecuación de una red de acceso 2G PDH, SDH para evolución a red multiservicio 3,5G
Abstract:	El presente informe de ingeniería tiene como objetivo expon er los criterios para realizar la adecuación de una red de acceso 2G PDH, SDH de un operador de servicios de telefonía móvil para su evolución a una red multiservicios 3.5G, para esto se integra y aplica los conocimientos adquiridos duran te el ejercicio de la carrera. Para el desarrollo del infor me se considera sólo un segmento de la red de acceso de Lim a Metropolitana, entendiéndose que el análisis realizado se hace extensivo a toda la red de acceso. El proyecto se inic ia con la revisión de la topología actual de la red de acce so PDH, SDH 2G, verificando las capacidades de los enlaces de microondas y los recursos de red utilizados, seguidament e se realizan los informes del estado de la infraestructura (TSS:Technical site survay) en las estaciones que conforman la red de acceso. Respecto a la ingeniería del proyecto, se analizan las diferentes alternativas y variantes para la im plementación de la red multiservicios, escogiéndose la red basada en Ethernet que es la que ofrece mayor escalabilidad

	y rendimiento económico. El análisis de la estructura de co stos, basado en precios acorde al mercado, nos permite calc ular el monto de las inversiones, el cual es analizado y po steriormente aprobado por el departamento de finanzas de la operadora móvil, confirmando de esta manera la factibilidad de la implementación del proyecto al área de red.

Completamente afín( ) Muy afín( ) Medianamente afín(X) Poco afín( ) Nada afín( )

Publicación #9

Título:	Migración de una red ATM hacia una red MPLS para el mejoram iento e implementación de servicios de una empresa
Abstract:	EL presente Documento tiene como finalidad explicar el Proc eso de Migración de todos los Servicios de Comunicaciones d e la Empresa DELOSI S.A. que se encuentran soportados bajo la Red ATM (Modo de Transferencia Asíncrona) IP DATA (Servi cio de CLARO basado en su Red ATM) hacia la Nueva Red RPV (Red Privada Virtual) IP/MPLS+Metro de CLARO PERÚ, el cual s e detalla de la siguiente manera: En el Capítulo I se explicará cuál es la problemática del p royecto describiendo el funcionamiento de la Red ATM de CLA RO y cuáles son las carencias de esta Red. Podremos ver cuá les son los objetivos del trabajo, sus limitaciones y una p equeña reseña del trabajo que vamos a realizar. En el Capítulo II se explicará los inicios de la Red ATM en el Perú. Una descripción teórica detallada de los puntos má s importantes de las Redes ATM y MPLS (Conmutación Multi-Pr otocolo mediante Etiquetas). En el Capítulo III se explicará cuáles son las alternativas de solución. Luego una descripción completa de la solución escogida para este proyecto, explicando el funcionamiento d e la Red MPLS en la Red de CLARO y el Proceso de Migración de la Red ATM hacia la Red MPLS. En el Capítulo IV se explicará todo el análisis de los resu ltados obtenidos en el Proceso de Migración. Adicionalmente se detallarán los costos y el cronograma utilizado para est e Proyecto.

Completamente afín( ) Muy afín(X) Medianamente afín( ) Poco afín( ) Nada afín( )

Publicación #10

Título:	Diseño de una red de acceso óptica para la siguiente generac ión
Abstract:	Las redes de acceso actuales en el Perú para el servicio de banda ancha fija tienen limitaciones técnicas debido al uso del cobre como medio de transmisión. Estos inconvenientes se reflejan en el servicio que se brinda a los usuarios debido a que la demanda de ancho de banda es cada vez mayor. El objetivo principal del trabajo es diseñar una red de acce so para la siguiente generación con elementos ópticos pasivo s. Para tal fin, se brinda una metodología basada en un enfo que de "arriba hacia abajo" donde se toma como referencia la s capas del modelo OSI, es decir, se inicia en la capa de ap licación y se termina en la capa física. Se empieza la metodología con la preparación de la red, en l a que se analiza los requerimientos organizacionales y técni cos de una empresa operadora de cable. Este operador tiene u

	na red HFC desplegada en la zona norte de Lima. En esta fase , se abarca lo correspondiente a la capa de aplicación y se evalúa las diferentes alternativas de solución, donde se eli ge y fundamenta una opción tecnológica. En la segunda fase se realiza la planificación, donde se car acteriza la red existente del operador y se elige una zona d e la cobertura para el diseño de un piloto de red. En esta e tapa se obtiene un mayor conocimiento de la red y se puede v isualizar los cambios necesarios para cumplir con todos los requerimientos actuales y futuros. Esta evaluación se realiz a en base a la opción tecnológica elegida. Finalmente, en el diseño de la red se estudian las diferente s arquitecturas que se utilizan tanto en capa 2 como en capa 1. La capa 3 no se analiza ya que involucra a los equipos de la red de núcleo. Adicionalmente, se muestran las proyeccion es de la red a futuro de acuerdo a la solución planteada.

Completamente afín( ) Muy afín( ) Medianamente afín(X) Poco afín( ) Nada afín( )

Anexo 5: Cronograma de Proyecto

Id	Modo de tarea	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras	Nombres de los recursos
1		**Sistema de Recomendación de Publicaciones Cientificas Nacionales a Investigadores del SINACYT**	**75 días**	**lun 15/01/18**	**vie 27/04/18**		
2		**Manejo de la documentacion del proyecto**	**60 días**	**lun 15/01/18**	**vie 06/04/18**		
3		Elaboracion de la documentacion requerida	60 días	lun 15/01/18	vie 06/04/18		
4		**Preparacion de la Base de Datos de articulos cientificos de ALICIA e investigadores registrados en DINA**	**5 días**	**lun 15/01/18**	**vie 19/01/18**		
5		Extraccion de datos ALICIA	2 días	lun 15/01/18	mar 16/01/18		
6		Extraccion de datos DINA	2 días	mié 17/01/18	jue 18/01/18	5	
7		Normalizacion de data extraida	1 día	vie 19/01/18	vie 19/01/18	6	
8		**Implementacion de modelos de datos**	**15 días**	**lun 22/01/18**	**vie 09/02/18**		
9		Construccion de modelos de datos de usuarios de DINA	5 días	lun 22/01/18	vie 26/01/18	7	
10		Construccion de modelos de datos de publicaciones cientificas en ALICIA	5 días	lun 29/01/18	vie 02/02/18	9	
11		Construccion de modelo de recomendación	5 días	lun 05/02/18	vie 09/02/18	10	
12		**Desarrollo de sistema de recomendación**	**30 días**	**lun 12/02/18**	**vie 23/03/18**		
13		Desarrollo del componente extractor de datos de ALICIA y	5 días	lun 12/02/18	vie 16/02/18	11	
14		Desarrollo del componente de recoleccion y procesamiento de datos	6 días	lun 19/02/18	lun 26/02/18	13	
15		Desarrollo del componente recomendador	10 días	mar 27/02/18	lun 12/03/18	14	
16		Integracion de los componentes	7 días	mar 13/03/18	mié 21/03/18	15	
17		Evaluacion de las recomendaciones	3 días	jue 22/03/18	lun 26/03/18	16	
18		Mejoras y Correcciones	15 días	jue 22/03/18	mié 11/04/18	16	
19		**Implementacion del servicio de recomendación**	**10 días**	**jue 12/04/18**	**mié 25/04/18**		
20		Implementacion de servicio web REST para la posterior publicacion de recomendaciones	10 días	jue 12/04/18	mié 25/04/18	18