# sistema Recomendador de documentos aplicado a un contexto académico

## Propósito

Desarrollar un servicio web que recomiende al usuario documentos relevantes a un tópico, ya sea en base a una descripción del mismo o a otros documentos que él proporcione. La recomendación consistirá en la búsqueda de información relacionada conceptualmente a la proporcionada por el usuario en un índice de documentos mantenido por el sistema. La retroalimentación del usuario y la consiguiente contribución al conocimiento del sistema sobre los temas permitirá la mejora de futuras recomendaciones.

## Objetivos

* Elegir las fuentes de documentos para el desarrollo y pruebas, implicando esto decidir el idioma principal a tratar y también la variedad temática del conjunto documental.
* Determinar las técnicas a utilizar para la clasificación conceptual de los documentos, siendo deseable una combinación de técnicas estadísticas y lingüísticas aptas para el tratamiento de los mismos. En base a estas técnicas, desarrollar un módulo para la clasificación de documentos, con el fin de obtener un índice conceptual.
* Encontrar métodos eficaces para el tratamiento de las consultas del usuario (explícitas o mediante documentos proporcionados por él respecto a un tema), presentando éstas un mayor número de consideraciones de tratamiento lingüístico, dado su carácter informal. Con el fundamento de éstos desarrollar un componente de búsqueda que deje la posibilidad al usuario de sugerir nuevos recursos y tópicos a clasificar, esto es, que el sistema pueda aprovechar la retroalimentación del usuario para la tarea de clasificación.

## resultados esperados

* El desarrollo de una interfaz que permita la presentación a usuarios de documentos relevantes a un tópico, aplicada al dominio conceptual del sector académico mediante la aplicación de algoritmos estadísticos y de procesamiento lingüístico.
* Sentar una base y un marco de prueba y aplicación para futuros proyectos de recuperación de la información orientados a conceptos que se valgan de técnicas de procesamiento de lenguaje natural.
* Contribuir a las áreas de investigación del procesamiento del lenguaje natural y la recuperación de la información, en general, y a la comunidad de desarrolladores con una solución abierta al público.

## palabras clave

Procesamiento de lenguaje natural, recuperación de información, web semántica, aprendizaje de máquina, clasificación de documentos.

## contexto e introducción

Un componente de la buena docencia, y el buen estudio, es la información: el disponer de recursos que permitan adquirir dominio sobre el tema. Ahora bien, la adquisición de los mismos no es tarea fácil y usualmente se requiere demasiado tiempo, y hasta creatividad, encontrar algo útil entre grandes colecciones de documentos, siendo el internet la elección antonomástica, en los últimos tiempos, para buscar tales recursos. Los motores de búsqueda nos han facilitado la vida, pero se limitan a encontrar información que se adhiere a las palabras que usamos en la búsqueda, dejándonos a nosotros la tarea de analizar los documentos para determinar si tienen, en realidad, relevancia para nuestros propósitos. Y, además del tiempo perdido en lecturas que se prueban fútiles al final, lo perdemos también en formular y reformular nuestras búsquedas para ver si, con suerte, algo que nos aporte al fin aparece.

La investigación en el área del procesamiento del lenguaje natural siempre ha querido establecer una comunicación más significativa entre el hombre y la máquina; y en las áreas que nacieron con la lingüística computacional, como responder preguntas hechas por usuarios humanos, se ve la necesidad que tenemos de que las computadoras entiendan lo que queremos, aunque no lo expresemos de una manera formal y estructurada.

En la recuperación de información, se han hecho investigaciones que exploran la potencialidad de aplicar el procesamiento del lenguaje natural a las búsquedas, para que el usuario encuentre documentos relevantes no en base a palabras clave y un poco de suerte, sino en base al *significado conceptual* de lo que busca. Productos comerciales que utilizan técnicas de otras áreas de la lingüística computacional, como ser la extracción de la información y la minería de textos y conceptos, han surgido para satisfacer hasta cierto punto esta necesidad. Incluso existen proyectos aún en desarrollo (como ser el buscador *Powerset*  o la base de datos de conocimiento en línea *Freebase)*  que llevarán este concepto al internet. En los últimos años muchos proyectos industriales y de investigación se han llevado a cabo para la recuperación de información orientada a conceptos, el campo es, sin embargo, joven, y, por tanto, está abierto a que se explore aún en él.

Elegí este proyecto debido, en primera instancia, al interés que tengo en explorar el campo del procesamiento del lenguaje natural como una oportunidad de conciliar mis mayores intereses académicos: las lenguas y las ciencias computacionales. Además, existen herramientas de código abierto que permiten concentrase en esta tarea sin preocuparse por cuestiones técnicas que no vienen al caso (como el pre-procesamiento de documentos, la implementación de análisis estadísticos, etc.).De hecho, una de las herramientas más útiles que encontré para la investigación y desarrollo en el área está implementada en un lenguaje que he llegado a dominar satisfactoriamente: python, de ahí que el desarrollo del software propuesto no se verá ralentizado por el aprendizaje de un lenguaje para la implementación. Para un sistema como el propuesto, no se necesitan habilidades muy expertas en el área, sino más bien dominio del lenguaje de programación, matemáticas, lingüística básica y manipulación de textos, todas habilidades en las cuales no me siento en desventaja, gracias a la heterogeneidad y calidad de la educación que he recibido en la universidad.

## investigaciones relacionadas

Como he mencionado, existen productos comerciales, como ser *Naveganza*, para el mercado hispanoparlante o *ConceptMine* , de la empresa Sciento, para el angloparlante. (Haav & Lubi, 2001) Mencionan 13 distintos productos, de los cuales 4 son resultado de investigación académica y los otros, comerciales. Distintas técnicas se discuten, pero se resalta el tema de la falta de automatización en la clasificación y la relativa novedad del campo (a 2001, herramientas de nivel comercial tenían apenas cinco años). En el instituto politécnico nacional de México, por ejemplo, (Gelbukh, Sidorov, & Guzmán-Arenas, 2005) investigaron el impacto del uso de una jerarquía ponderada en la recuperación de información en documentos clasificados mediante una taxonomía preestablecida. Incluso existen patentes para sistemas que hacen este tipo de búsquedas, por ejemplo (Lin, y otros, 2004). Además, iniciativas de organizar el conocimiento humano como *Freebase.com* pueden servir de base para investigaciones en este aspecto. En el campo teórico, avances tales como el aprendizaje de máquina no supervisado pueden servir para la creación de taxonomías u ontologías pudiendo incluso servir éstas para luego aplicar algoritmos de aprendizaje supervisado que sean útiles para la clasificación de textos; además, claro, de la lingüística computacional como tal, cuyos aportes, tales como los análisis semánticos o pragmáticos, hacen posible que investigaciones como ésta sean cada vez más factibles, pudiendo aspirar, así, a tecnologías de recuperación de la información que sean más naturales y , a la vez, precisas, dejando atrás los modelos simplificados del lenguaje que se usan en la actualidad.

## tipo de proyecto

De investigación y desarrollo.

## preguntas de investigación e hipótesis

* ¿Cómo se determinan los conceptos de un documento y una consulta escritos en un lenguaje natural?
* ¿Cómo se pueden recuperar documentos en base a modelos de lenguaje natural y conceptos?
* ¿Qué tan preciso y confiable es el uso de aprendizaje de máquina para la clasificación de textos?
* ¿Cómo afectan el tamaño y la heterogeneidad de un cuerpo de textos a un sistema que se base en búsqueda conceptual?
* ¿Cómo afecta el idioma elegido a un sistema basado en el procesamiento de lenguaje natural?
* ¿Es posible que un sistema basado en modelos de lenguaje pueda rendir en un entorno translingüistico?
* ¿Qué tanto aporta al campo de la recuperación de la información el tratamiento conceptual de las búsquedas y los documentos?

## métodos

El método de investigación consistirá, más que todo, en la recopilación y revisión de bibliografía. Se evaluarán experimentalmente, además, técnicas de clasificación y búsqueda de documentos, para elegir las más adecuadas para el desarrollo del software.

## Recursos requeridos

Respecto al software, pretendo utilizar herramientas de código abierto para el desarrollo; a la hora de escribir esta propuesta, dos conjuntos de herramientas de desarrollo de procesamiento de lenguaje natural y recuperación de la información han sido elegidas como los candidatos más prometedores: el *natural language toolkit (NLTK),* en pythony el *dragon toolkit*, en java. Para almacenamiento, he considerado MySql y PostgreSQL y para el desarrollo de una interfaz para el servicio, el *framework* de desarrollo web Django.

En cuanto al hardware, la mayor parte del desarrollo lo haré en mi computadora personal, pero para la adquisición de los textos que usaré para la clasificación y búsquedas y para la experimentación con los distintos algoritmos de clasificación y técnicas de búsqueda, sería idóneo contar con los servidores del laboratorio de sistemas de la universidad.

## alcance y restricciones

Se pretende investigar cómo se pueden recuperar documentos de un dominio específico del conocimiento basándose en conceptos: tanto en la consulta, consistente o en una pesquisa explícita o en un conjunto de documentos que el usuario estima relevantes a un tópico previamente declarado, como en los documentos que consistirán el espacio de búsqueda. Para este fin, técnicas lingüísticas y estadísticas del procesamiento del lenguaje natural se utilizarán.

Por cuestiones de complejidad, recursos y tiempo, no se explorarán técnicas novedosas o poco exploradas de clasificación de documentos o tratamiento del lenguaje natural, sino que se evaluarán y aplicarán en una aplicación orientada a una comunidad de usuarios que contribuyen, aquellas que hayan probado, de alguna manera, ser aptas para la tarea.

# Trabajos citados

Gelbukh, A., Sidorov, G., & Guzmán-Arenas, A. (2005). Document Indexing with a Concept Hierarchy. *Comunicación y Sistemas* , 281-292.

Haav, H.-M., & Lubi, T.-L. (2001). *A Survey of Concept-based Information Retrieval Tools on the Web.* Tallin: Institute of Cybernetics at Tallinn Technical University.

Lin, A. D., Graydon, P. J., Busch, J. E., Caudill, M., Chinchor, N. A., Tseng, J. C.-M., et al. (2004). *Patent No. 6675159.* Estados Unidos.