# Recomendador de recursos académicos.

## Propósito

Desarrollar una interfaz de programación de aplicaciones (API) que utilice técnicas de procesamiento de lenguaje natural para que, dada la consulta de un usuario sobre un tema, pueda presentar un conjunto de recursos, previamente indizados, que puedan ser académicamente relevantes para la ampliación y desarrollo de éste. Mostrar, además, esta interfaz aplicada en software.

## Objetivos

* Establecer un modelo para la jerarquización e indización de conceptos en documentos, además de la elección de las fuentes de documentos para el desarrollo y pruebas.
* Elegir las técnicas a utilizar para la indización de los documentos, siendo deseable una combinación de técnicas estadísticas y lingüísticas aptas para el tratamiento de los documentos. Y ponderar los métodos a usar para el tratamiento de las consultas del usuario, presentando éstas un mayor número de consideraciones de tratamiento lingüístico, dado su carácter informal.
* Desarrollar un módulo de indización, que permita la clasificación de documentos y un componente de búsqueda, que deje la posibilidad al usuario de sugerir nuevos recursos a indizar. Consolidar la interfaz e implementarla en una aplicación para las subsecuentes pruebas.
* Evaluar el rendimiento de la interfaz en un entorno multilingüe (tanto en documentos como en búsquedas).

## resultados esperados

* El desarrollo de una interfaz que permita la presentación a usuarios de documentos relevantes a la búsqueda de 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 futuras investigaciones de extracción de la información que generen jerarquías de conceptos.
* Contribuir a la investigación del procesamiento del lenguaje natural y a la comunidad de desarrolladores con una solución abierta al público.

## palabras clave

Extracción de la información, procesamiento de lenguaje natural, indización de documentos, recuperación de información.

## 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, la extracción de la información y la minería de textos y conceptos, han surgido para satisfacer hasta cierto punto esta necesidad en dominios específicos, y existen proyectos aún en desarrollo (como ser el buscador *Powerset)*  que llevarán este concepto al internet. Pero muy pocas aplicaciones públicas y orientadas a dominios de conocimiento específico han visto la luz, y es aquí donde se pretende contribuir: al campo de la recuperación de la información en un dominio basada en el procesamiento de lenguaje natural.

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, lingüistica básica y manipulación básica 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

[TODO]