UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

"Evaluación de sistemas de extracción de frases clave"

Tesis
Para Obtener el Título de
Ingeniero en Software

Que Presenta

Jesús Ernesto Padilla Camacho

Directora:
Dra. Yulia Nikolaevna Ledeneva

TIANGUISTENCO, MÉX. Mayo 2016

Abstract

Nowadays the amount of digital information that found in internet has considerably increased that is why online search brings recovered documents. These results must be verified in order to know whether they contain the necessary information. A way to simplify an online search is using *keywords* or *keyphrases* since they act as filters within a search field. Keywords and keyphrases are used in many fields, for example, marketing and advertising, where the main objective is to catch public's attention, as well as to expose as many people as possible to a certain product or service. In text documents in both physical and electronic formats, keyphrases help readers to find certain information by displaying the main ideas in a specific text. This thesis presents the comparison of automatic keyphrase extraction systems based on a collection of scientific papers used in task 5: "Automatic keyphrase extraction from scientific articles" of SemEval-2010, with the objective of know systems that can find the keyphrases that were proposed by a human being for each paper. In the chapter of experiments, the results are presented among installable and online systems. Finally, the obtained results of the evaluation of task 5 of SemEval-2010 are compared.

Resumen

Hoy en día, la cantidad de información electrónica en forma de texto ha aumentado considerablemente por lo que una búsqueda de información puede traer consigo varios documentos recuperados. Posteriormente, los documentos recuperados se tienen que revisar para saber si contienen lo que realmente se busca. Una manera de simplificar una búsqueda es el empleo de palabras o frases clave ya que actúan como filtro en un campo de búsqueda. Las palabras o frases clave se utilizan en muchas áreas, por ejemplo, la mercadotecnia y publicidad, en donde el objetivo es capturar la atención del público. De igual forma, en todo aquello que se quiere dar a conocer hacia el público en general. Ya sean documentos de textos impresos o electrónicos, las frases clave ayudan al lector mostrándole las ideas principales del texto. En esta tesis, se comparan los sistemas de extracción automática de frases clave sobre un conjunto de artículos científicos utilizados en la tarea 5 del SemEval-2010, con el objetivo de conocer qué sistemas pueden encontrar las frases clave que fueron propuestas por un ser humano. En la experimentación se presentan los resultados de la comparación entre los sistemas instalables y en línea. Por último, los resultados de la evaluación se comparan con los de la tarea 5 del SemEval-2010.

Contenido

Página

Lista de figuras

Lista de tablas

CAPÍTULO 1.

Introducción

El manejo de información en la actualidad es un factor de gran importancia dentro de los sectores públicos y privados. Con el crecimiento constante de los volúmenes de información electrónica, ésta requiere ser organizada para su uso. Con la tecnología que se cuenta hoy en día, el manejo de la información se ha facilitado. Dentro del área de Procesamiento de Lenguaje Natural una de las disciplinas que la integran es la Recuperación de Información (RI). La RI es el proceso de encontrar en un repositorio grande de datos, material (usualmente documentos) de naturaleza no estructurada (usualmente texto) o semiestructurada (páginas Web) que satisfaga una necesidad de información [Manning 09]. Para lograr la RI, existen varios sistemas, estos reciben el nombre de sistemas de recuperación de información. Un Sistema de Recuperación de Información (SRI) consiste básicamente de un conjunto de procesos interrelacionados que permiten obtener información de interés, a partir de una determinada consulta [Jiménez 03].

Un ejemplo de un SRI es un buscador Web, ya que se ingresa una búsqueda y este devuelve los resultados que han coincidido con el texto de entrada.

El objetivo principal de un SRI es recuperar todos los documentos que son relevantes a una consulta del usuario mientras recupera el menor número de documentos no relevantes como sea posible [Baeza 99]. En la figura 1 se presenta el proceso interno que realiza un SRI [Baeza 99], en el cual una de las fases que la integran es la indexación, que según [Medelyan 09a] es: "*donde todas las palabras y frases, sin incluir a las stopwords, son extraídas de un documento*". Esto a su vez se relaciona con la extracción de frases clave ya que se extraen las frases más importantes de un texto y a partir de las cuales se presentan los resultados de una búsqueda de información.

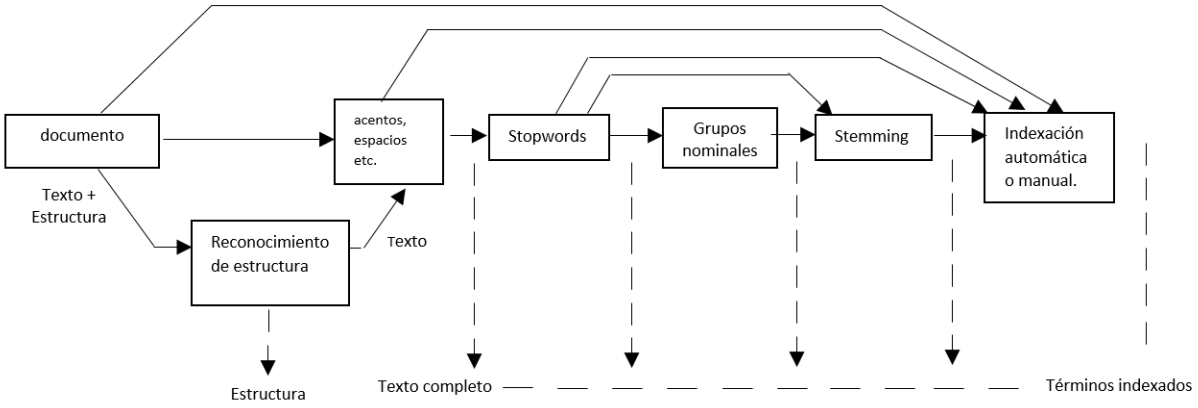


Figura 1. Vista lógica de un documento [Baeza 99].

Sin embargo, un SRI no siempre cumple con lo que realmente se necesita, ya que una búsqueda de información implica una gran cantidad de resultados, que posteriormente hay que revisar para verificar si contienen lo que se busca y en lo cual se debe invertir tiempo. Es aquí donde entra en juego las palabras o frases clave, para determinar qué información es de nuestro interés. Las frases clave influyen en la RI pues son de gran utilidad al momento de buscar información en un gran conjunto de datos, ya que actúan como filtro para presentar los temas destacados que son tocados por el autor en el contenido de un texto. Según [Witten99] las frases clave suelen ser elegidas de forma manual en muchos contextos académicos por los autores los cuales asignan palabras clave a los documentos que han escrito. Los indizadores profesionales suelen elegir frases de un "vocabulario controlado" predefinido relevante para el dominio en cuestión. Sin embargo, la gran mayoría de los documentos vienen sin frases clave por lo que asignarlas manualmente es un proceso tedioso ya que requiere el conocimiento sobre la materia. Es decir, la extracción automática de frases clave facilita la asignación de frases clave, ya que al realizar el proceso de manera automática reduce factores como el tiempo y el costo que implica la asignación manual.

En el presente trabajo se realiza la comparación del desempeño entre los sistemas de extracción automática de frases clave disponibles en 2016, para conocer cuáles devuelven los resultados más parecidos a los que propone un ser humano en un conjunto de artículos científicos.
Para comparar los sistemas se utiliza un conjunto de artículos científicos de la Biblioteca Digital ACM (ponencias y talleres), que fueron previamente tratados en el taller de evaluación Semántica SemEval-2010 "*Task 5: Automatic Keyphrase Extraction from Scientific Articles*", y que servirán de entrada en los sistemas a evaluar, para así poder caracterizar sus resultados con este conjunto de datos.

1.1 Antecedentes

Senseval es una organización internacional dedicada a la evaluación de sistemas de desambiguación de la palabra. El propósito de Senseval es evaluar las fortalezas y debilidades de este tipo de programas en relación con diferentes palabras, diferentes variedades de la lengua, y los diferentes idiomas [Senseval 15]. A partir del 2007, Senseval evolucionó a SemEval en donde se incluye la evaluación semántica dando lugar a más tareas de Procesamiento de Lenguaje Natural. En el 2010 se realiza el segundo taller SemEval-2010 (SemEval-2) en el cual se incluyeron 18 tareas dirigidas a la evaluación de los sistemas de análisis semánticos, en 2010 la tarea 5 fue: "*SemEval-2010 Task 5: Automatic Keyphrases Extraction from Scientific Articles*". Para esta tarea:

- Se compiló un conjunto de artículos científicos con asignación de frases clave por parte del: autor, lector y combinado (combinación entre autor y lector).
- La tarea consiste en desarrollar sistemas que producen frases claves automáticamente para cada artículo científico.
- Participaron oficialmente 19 sistemas de extracción automática de frases clave.
- Los sistemas participantes fueron evaluados de manera automática [Kim 10].

En la tabla 1, se muestran las áreas que han tenido lugar en los talleres de Senseval hasta SemEval y se marca (en letra resaltada e itálica) el área "*Extracción de frases clave (Extracción de información)*" donde se observa que desde el SemEval-2010 en adelante no se ha realizado esta área.

Areas de estudio	**S1**	**S2**	**S3**	**SE07**	**SE10**	**SE12**	**SE13**	**SE14**	**SE15**	**SE16**
Bioinformática / Análisis de texto clínico								✓	✓	✓
Razonamiento de sentido común (COPA)						✓				
Resolución de la correferencia					✓					
Compuestos nominales (Extracción de información)					✓		✓			
Ellipsis					✓					
Inducción gramática								✓		
Extracción de frases clave (Extracción de información)					✓					
Simplificación léxica						✓				
Sustitución léxica (Multilingual or Crosslingual)		✓		✓	✓					
Complejidad léxica										✓
Metonymy (Extracción de información)				✓	✓					
Paráfrasis						✓	✓		✓	✓
Búsqueda de Respuestas									✓	✓
Similitud relacional						✓	✓			
Análisis semántico						✓	✓	✓	✓	✓
Identificación de relación semántica				✓	✓					
Etiquetado de papel semántico			✓	✓	✓		✓			
Similitud semántica						✓	✓	✓	✓	✓
Similitud semántica (Crosslingual)										✓
Similitud semántica (Multilingual)								✓	✓	✓
Análisis de sentimiento				✓	✓		✓	✓	✓	✓
Etiquetado de papel espacial						✓	✓			
Taxonomía de inducción/Enriquecimiento									✓	✓
Pruebas de implicación					✓	✓	✓	✓		
Pruebas de implicación (Cross-lingual)							✓			
Anotación temporal				✓	✓		✓	✓	✓	✓
Análisis de Twitter							✓	✓	✓	✓
Desambiguación del sentido de la palabra (Muestra léxica)	✓	✓	✓	✓	✓					
Desambiguación del sentido de la palabra (Todas las palabras)		✓	✓	✓	✓		✓		✓	
Desambiguación del sentido de la palabra (Multilingual)							✓		✓	
Desambiguación del sentido de la palabra (Cross-lingual)							✓	✓		
Inducción del sentido de la palabra				✓	✓		✓			

Tabla 1. Áreas de estudio en los talleres desde Senseval a SemEval. [SemEval 12], [SemEval 13], [SemEval 14], [SemEval 15], [SemEval 16], [SemEval-Wikipedia 16].

1.2 Motivación de la tesis

Con el continuo crecimiento de la información, la cantidad de documentos que contienen texto han aumentado considerablemente y una búsqueda trae consigo varios resultados. HaCohen [HaCohen 03] menciona un aspecto en la búsqueda dentro de una gran cantidad de documentos: "*Con el crecimiento explosivo de la información en línea, más y más personas dependen de los resúmenes. La gente no tiene tiempo para leer todo el contenido del texto pues prefiere leer los resúmenes, antes de decidir si les gustaría leer todo el texto o no. Las palabras clave son consideradas como resúmenes muy cortos, que pueden ser de gran ayuda*". Es decir, las frases clave facilitan la búsqueda de información dentro de un gran conjunto de documentos pues presentan las ideas principales de un texto. Los sistemas de extracción automática de frases clave reducen el tiempo, personal y costo que implica la asignación manual de frases clave en uno o varios documentos.

Con la tarea 5 del SemEval-2010 se conoce el avance de los métodos de extracción automática de frases clave. Sin embargo no se conoce el desempeño de los sistemas de extracción de frases clave disponibles en 2016 por lo que es necesario determinar su calidad de extracción y con ello caracterizar dichos sistemas.

1.3 Planteamiento del problema

Hoy en día existen sistemas que realizan la tarea de extraer frases clave de textos, pero no todos usan el mismo procedimiento para realizar esa tarea ya que utilizan diversos enfoques, colecciones, métodos y algoritmos.

Con lo anteriormente mencionado surge la pregunta de investigación:

¿Cuál es el desempeño de los sistemas instalables y en línea disponibles en 2016, para la extracción automática de frases clave en comparación con las frases clave asignadas por un ser humano, sobre un conjunto de artículos científicos de la colección SemEval-2010?

1.4 Hipótesis

Si se comparan los sistemas de extracción automática de frases clave disponibles, será posible conocer los mejores sistemas actuales y con ello caracterizar los resultados que devuelven.

1.5 Objetivos

1.5.1 Objetivo general

Como respuesta a la pregunta de investigación, el objetivo general de esta tesis es:

Realizar la comparación de sistemas disponibles para conocer cuál devuelve los resultados más parecidos a los que propuso un ser humano, en cada uno de los artículos contenidos en el conjunto de datos SemEval-2010.

1.5.2 Objetivos específicos

Como tareas a cumplir los objetivos específicos son:

- Analizar y conocer los sistemas disponibles que extraen una lista de frases clave de textos.
- Describir los pasos que siguen los sistemas para hacer su trabajo de extracción.
- Poner a prueba los sistemas en la búsqueda de frases clave en referencia a los documentos contenidos del SemEval-2010.
- Evaluar a los sistemas de manera automática con la metodología propuesta en este trabajo.
- Analizar y comparar los resultados de la evaluación y con ello caracterizar a los sistemas de extracción automática de frases clave.
- Comparar los resultados de esta evaluación con los de la tarea 5 del SemEval-2010.

1.6 Delimitación del problema

- Solo se evalúan sistemas disponibles (comerciales y libres) para medir su desempeño.
- Se trabaja con el conjunto de datos que fue utilizado en la tarea 5 del SemEval-2010.
- La herramienta de evaluación ("performance.pl") es la misma que se utilizó en SemEval-2010 y que se utiliza en esta tesis.
- No se proponen correcciones a los sistemas, solo conocer cuál es su calidad de extracción.

1.7 Estructura de la tesis

El esquema del documento es el siguiente: En el capítulo 1 se presentó la orientación de este trabajo así como los objetivos a alcanzar. En el capítulo 2 se presentan los principales conceptos referentes a la extracción automática de frases clave. En el capítulo 3 se presentan trabajos relacionados que se han desarrollado en cuanto a la extracción automática de frases clave. En el capítulo 4 se presenta la metodología propuesta que se utiliza para evaluar a los sistemas. En el capítulo 5 se presenta la descripción del conjunto de datos, se muestran los resultados obtenidos en esta evaluación y se comparan con los resultados de la tarea 5 del SemEval-2010. En el capítulo 6 se presentan las conclusiones y se proponen direcciones futuras.

CAPÍTULO 2.

Marco Teórico

En primera parte, en este capítulo se abordan los conceptos sobre las frases clave. Después se presenta una tabla [Medelyan 09a] en la cual se muestran las tareas que se relacionan con la extracción automática de frases clave y el objetivo de cada una de ellas. Posteriormente, se presentan los enfoques que hay en la extracción automática de frases clave. Por último, se menciona la extracción de frases clave en artículos científicos, así como los factores que influyen en los conjuntos de datos haciendo que la extracción se realice con ventajas o desventajas.

.

2.1 Frases clave

Las frases clave son la unión de palabras que representan las ideas principales de un texto y proporcionan una percepción breve de su contenido [Kim 10], [Witten 99], [Medelyan 06], [Turney 99]. Es decir, son las frases que vuelven representativo a todo un texto, ya que con el simple hecho de que un ser humano observe una frase puede también deducir de qué trata el contenido o ligar esa frase con temas que son de su interés.

2.2 Extracción automática de frases clave

La extracción automática de frases clave es la selección automática de un conjunto de frases que mejor describen a un texto [Turney 99], [Mihalcea 04]. Es decir, la extracción automática de frases clave se encarga de realizar todo el proceso que lleva acabo un indexador profesional, ya que al realizar el proceso de manera automática reduce factores como el costo de contratar un experto en el tema y el tiempo que implica.

2.3 Asignación de frases clave y extracción de frases clave

Dentro de la extracción de frases clave [Witten 99], [Beliga 14], [Medelyan 06] [Frank 99] existen dos enfoques**:**

1. Asignación de frases clave
2. Extracción de frases clave

1.- **Asignación de frases clave:** en este enfoque las frases clave son elegidas de un vocabulario controlado, en donde el objetivo es encontrar un conjunto de términos que describan a un documento individual.

2.- **Extracción de frases clave:** en este enfoque se analizan las frases en el documento y se identifican como frases clave a las más representativas y por estar presentes en el documento, no es necesario contar con un vocabulario para poder seleccionarlas.

2.4 Diferencia entre extracción de terminología y extracción de frases clave

¿Qué diferencia hay entre la extracción de terminología y la extracción de frases clave? Para responder esta pregunta, primero debemos conocer la orientación de cada una:

- La extracción de terminología se encarga de extraer los términos que pertenecen a un dominio específico en un texto dado.
- La extracción de frases clave se encarga de extraer las frases que caracterizan a un texto (como se menciona en el apartado 2.3).

Park [Park 10] La extracción de frases clave, sin embargo, es más que la extracción de terminología o el análisis de la estructura del documento, Mientras que los términos son palabras que aparecen en contextos específicos y analizan estructuras conceptuales en los dominios de la actividad humana, las frases clave son palabras que captan la idea clave de los documentos. Además, mientras que los términos por lo general ocurren en el documento dado más a menudo de lo que se puede esperar que se produzca, las frases más importantes no necesariamente ocurren con frecuencia o frases clave no se producen en absoluto en el documento.

En los trabajos del estado del arte de extracción automática de frases clave comúnmente encontraremos términos parecidos a la extracción de frases clave los cuales podemos confundir o malinterpretar como sinónimos por ejemplo: categorización, extracción de términos, etiquetado, etc.

A partir de Medelyan [Medelyan 09a], se presentan las tareas relacionadas que hay en el tema de indexación (ver tabla 2).

Nombre de la tarea	Conocido como	Descripción
Categorización de texto	Clasificación de texto	Muy pocas categorías generales, como la política o las noticas, son asignadas usualmente a partir de vocabularios controlados pequeños.
Asignación de términos	Indexación de temas	Los principales temas se expresan utilizando términos de un vocabulario controlado grande. Eje: thesaurus.
Extracción de frases clave	***Extracción de palabras clave, Extracción de términos clave***	***Los temas principales son expresados usando las palabras y las frases más importantes en el documento.***
Extracción de terminología	Indexación	Todas las palabras y frases relevantes de un dominio se extraen de un documento.
Indexación de texto completo	Indexación completa, indexación de texto libre	Todas las palabras y frases, sin incluir a las stopwords, son extraídas de un documento.
Indexación de frases clave	Asignación de frases clave	Un término general que se refiere tanto a la asignación de términos y la extracción de frases clave.
Etiquetado	Etiquetado colaborativo, Etiquetado social, auto-etiquetado, etiquetado automático	El usuario define muchos temas como desea. Cualquier palabra o frase puede servir como etiqueta. Se aplica principalmente a los sitios web de colaboración.

Tabla 2. Tareas relacionadas en la indexación [Medelyan 09a].

En esta tesis, la tarea que se sigue es la extracción de frases clave, por lo cual se marca en la tabla 2 (en letra resaltada e itálica) la orientación de esta tarea y es la se pretende cumplir con los sistemas a evaluar en el conjunto de datos SemEval2010.

2.5 Enfoques en la extracción automática de frases clave

2.5.1 Enfoque supervisado

Hasan [Hasan 14] indica que el objetivo de un enfoque supervisado, es capacitar a un clasificador de documentos anotados con frases clave para determinar si una frase es una frase clave candidata.

2.5.2 Enfoque no supervisado

En este enfoque no se necesitan datos de entrenamiento, ya que son extraídas las características del texto analizado, para seleccionar las frases más importantes [Mihalcea 04], [Bordea 10].

2.5.3 Enfoque semi-supervisado

Decong Li [Decong Li 10] dice que el objetivo del aprendizaje Semi-supervisado es diseñar una función que sea suficientemente suave con respecto a la estructura intrínseca revelada por las frases de título y otras frases, partiendo del supuesto de que las frases relacionadas semánticamente son probables que tengan puntajes similares.

Beliga y Zahang [Beliga 14], [Zahang 08] coinciden en que la extracción automática de frases clave se ha desarrollado a partir de distintos enfoques, de la siguiente forma (ver las secciones 2.5.1 - 2.5.9):

2.5.4 Enfoque de estadística simple

Comprende métodos sencillos que no requieren los datos de entrenamiento. Además, los métodos son el lenguaje y en el dominio independiente. Las estadísticas de las palabras del documento se pueden usar para identificar las palabras clave: *n-gramas estáticos*, *frecuencia de la palabra*, TF-IDF, *co-ocurrencia de la palabra*, *PAT Tree* (*Patricia Tree* (del inglés), un árbol de sufijos o árbol de posición), etc.

2.5.5 Enfoque lingüístico

Utilizan la característica de la lingüística principalmente de las palabras, frases y documentos. Léxico, sintáctico, análisis semántico y el discurso son algunos de los análisis más comunes, pero complejos.

2.5.6 Enfoque de aprendizaje automático

Induce un modelo que está entrenado en un conjunto de palabras clave. Requieren una anotación manual en el conjunto de datos de aprendizaje que es muy tedioso e inconsistente. Por desgracia, los autores suelen asignar palabras clave a sus documentos sólo cuando se ven obligados a hacerlo. El modelo inducido se aplica para la extracción de palabras clave de un nuevo documento. Este enfoque incluye *Naïve Bayes, SVM, C4.5, Bagging* etc. Estos métodos requieren datos de entrenamiento, y a menudo dependen del dominio. El sistema tiene que volver a aprender y establecer el modelo cada vez que se cambia de dominio. El modelo de inducción puede ser muy exigente y requiere mucho tiempo en conjuntos de datos masivos.

2.5.7 Enfoque basado en grafos

Grafo es un modelo matemático que permite la exploración de las relaciones y la información estructural de manera muy eficaz. El documento son los modelos como gráfico donde los términos están representados por vértices y las relaciones entre términos están representados por las aristas. La relación de aristas entre dos términos se puede establecer en muchos principios que explotan diferente alcance o las relaciones de texto para la construcción del grafo:

- Palabras co-ocurrentes juntas en una oración, párrafo, sección o documento agregado al gráfico como un clique;
- La intersección de las palabras de una oración, párrafo, sección o documento;
- Palabras concurrentes dentro de la ventana fija en el texto;
- Las relaciones semánticas - palabras de conexión que tienen significado similar, palabras escritas de la misma manera, pero tienen diferente significado, sinónimos, antónimos, heterónimos, etc.

2.5.8 Otros enfoques

Para la extracción de palabras clave, en general, se combinan todos los métodos mencionados anteriormente. Además, a veces por la fusión que incorpora el conocimiento heurístico, tales como la posición, la longitud, las características de diseño de los términos, HTML, etiquetas similares, el formato de texto, etc.

2.5.9 Modelo de espacio vectorial

Este modelo es adecuado para capturar la frecuencia sencilla de la palabra, sin embargo la información estructural y semántica se suelen pasar por alto. Por lo tanto, debido a la simplicidad VSM tiene varias desventajas:

- El significado de un texto y la estructura no puede ser expresado.
- Cada palabra es independiente de la otra, la secuencia de aparición de la palabra u otras relaciones no se puede exigir.
- Si dos documentos tienen un significado similar, pero son de diferentes palabras, la similitud no puede ser calculada fácilmente.

En la figura 2, a partir de [Beliga 14] se muestran los principales métodos de extracción de frases clave.

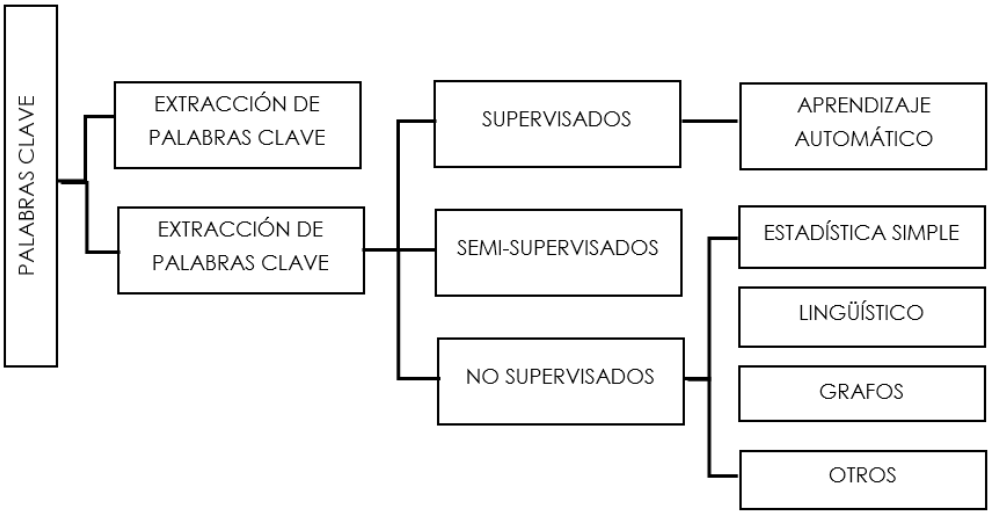


Figura 2. Clasificación de métodos para extracción de palabras clave [Beliga 14].

2.6 La extracción de frases clave en artículos científicos

La extracción automática de frases clave se emplea en diversos campos ya que es de gran utilidad para un área en específico y tiene un gran impacto hoy en día en la mercadotecnia. Dentro de los estudios en donde se han presentado métodos de extracción, los conjuntos de datos con los que se experimenta van desde noticias, artículos médicos, notas periodísticas, redes sociales, páginas web, emails, artículos científicos, etc. El conjunto de datos SemEval2010 es el utilizado en este trabajo, el cual está formado por artículos científicos. Nguyen [Nyugen 07] menciona que los artículos científicos se distinguen de los demás en función de su uso de lenguaje técnico, así como su rica estructura del documento.

A continuación se describen los factores que afectan la extracción de frases clave en los artículos científicos:

1.- Longitud

La dificultad de la tarea aumenta con la longitud del documento de entrada, documentos más largos producen más frases clave [Hasan 14].

Hasan [Hasan14] y [Hasan 10] indica: Cada resumen en el corpus Inspec (*conjunto de artículos científicos*) tiene un promedio de 10 frases clave autor y 34 frases clave candidatas. Por el contrario, un trabajo científico tiene típicamente al menos 10 frases clave y cientos de frases clave candidatas, dando un espacio de búsqueda mucho más grande (ver tabla 3).

	Inspec	**SemEval-2010**
Promedio frases clave	10	10
Candidatos	34	Varios

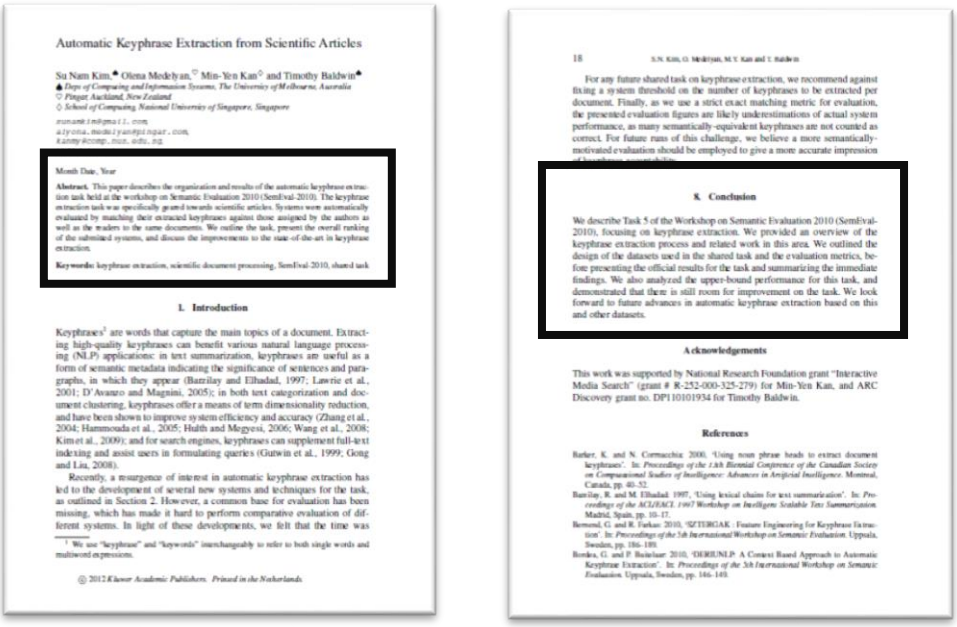
Tabla 3. Promedio de frases y candidatos en el corpus Inspec y SemEval-2010 [Hasan 14].

Es decir, dado que en un artículo científico contiene áreas de investigación y éstas a su vez tienen sub-áreas en los cuales se desarrolla la investigación, se presentan como candidatos a frases clave para cada una de ellas, con lo que las probabilidades de ser una frase clave aumentan.

2.- Consistencia estructural

En un documento estructurado, hay ciertos lugares donde es más probable que aparezca una frase clave. La mayoría de palabras clave de un artículo científico aparece en el resumen y la introducción [Hasan 14].

[Hasan 14] a partir de [Kim 13] indica que: La información estructural se ha explotado para extraer palabras clave de artículos científicos (por ejemplo, el título, la información de la sección ver figura 3).



Automatic Keyphrase Extraction from Scientific Articles

Su Nam Kim,♠ Olena Medelyan,♡ Min-Yen Kan◇ and Timothy Baldwin♠

♠ Dept. of Computing and Information Systems, The University of Melbourne, Australia
♡ Pingar Auckland, New Zealand
◇ School of Computing, National University of Singapore, Singapore

{snamkim@gmail.com, olena.medelyan@pingar.com, kanmy@comp.nus.edu.sg, tb@ldwin.net}

Month Date, Year

Abstract. This paper describes the organization and results of the automatic keyphrase extraction task held at the workshop on Semantic Evaluation 2010 (SemEval-2010). The keyphrase extraction task was specifically geared towards scientific articles. Systems were automatically evaluated by matching their extracted keyphrases against those assigned by the authors as well as the readers to the same documents. We outline the task, present the overall ranking of the submitted systems, and discuss the improvements to the state-of-the-art in keyphrase extraction.

Keywords: keyphrase extraction, scientific document processing, SemEval-2010, shared task

1. Introduction

Keyphrases[1] are words that capture the main topics of a document. Extracting high-quality keyphrases can benefit various natural language processing (NLP) applications: in text summarization, keyphrases are useful as a form of semantic metadata indicating the significance of sentences and paragraphs, in which they appear (Barzilay and Elhadad, 1997; Lawrie et al., 2001; D'Avanzo and Magnini, 2005); in both text categorization and document clustering, keyphrases offer a means of term dimensionality reduction, and have been shown to improve system efficiency and accuracy (Zhang et al., 2004; Hammouda et al., 2005; Hulth and Megyesi, 2006; Wang et al., 2008; Kim et al., 2009); and for search engines, keyphrases can supplement full-text indexing and assist users in formulating queries (Gutwin et al., 1999; Gong and Liu, 2008).

Recently, a resurgence of interest in automatic keyphrase extraction has led to the development of several new systems and techniques for the task, as outlined in Section 2. However, a common base for evaluation has been missing, which has made it hard to perform comparative evaluation of different systems. In light of these developments, we felt that the time was

[1] We use "keyphrase" and "keywords" interchangeably to refer to both single words and multiword expressions.

For any future shared task on keyphrase extraction, we recommend against fixing a system threshold on the number of keyphrases to be extracted per document. Finally, as we use a strict exact matching metric for evaluation, the presented evaluation figures are likely underestimations of actual system performance, as many semantically-equivalent keyphrases are not counted as correct. For future runs of this challenge, we believe a more semantically-motivated evaluation should be employed to give a more accurate impression

8. Conclusion

We describe Task 5 of the Workshop on Semantic Evaluation 2010 (SemEval-2010), focusing on keyphrase extraction. We provided an overview of the keyphrase extraction process and related work in this area. We outlined the design of the datasets used in the shared task and the evaluation metrics, before presenting the official results for the task and summarizing the immediate findings. We also analyzed the upper-bound performance for this task, and demonstrated that there is still room for improvement on the task. We look forward to future advances in automatic keyphrase extraction based on this and other datasets.

Acknowledgements

This work was supported by National Research Foundation grant "Interactive Media Search" (grant # R-252-000-325-279) for Min-Yen Kan, and ARC Discovery grant no. DP110101934 for Timothy Baldwin.

References

Barker, K. and N. Cornacchia: 2000, 'Using noun phrase heads to extract document keyphrases'. In: *Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*. Montreal, Canada, pp. 40–52.

Barzilay, R. and M. Elhadad: 1997, 'Using lexical chains for text summarization'. In: *Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarization*. Madrid, Spain, pp. 10–17.

Bernond, G. and R. Farkas: 2010, 'SZTERGAK : Feature Engineering for Keyphrase Extraction'. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden, pp. 186–189.

Bordea, G. and P. Buitelaar: 2010, 'DERIUNLP: A Context Based Approach to Automatic Keyphrase Extraction'. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala, Sweden, pp. 146–149.

Figura 3. Secciones de resumen y conclusiones de un artículo científico.

En conclusión con lo anterior, la mayoría de artículos científicos tienden a ser redactados mediante una estructura que debe de contener el artículo en la cual se desarrolla la investigación. Lo anterior es una ventaja ya que identificando las secciones en el papel se puede extraer las frases clave que suelen estar presentes en secciones como la introducción y al final en las conclusiones de artículo, pues en estas secciones es donde mayormente están presentes. Sin embargo representa un problema cuando no hay una estructura como lo es una página Web donde las frases clave pueden estar distribuidas por todas partes.

3.- Cambios de tema

Hasan [Hasan 14] a partir de [Medeylan 09] y [Witten 99] indica que: Una observación común explotada en la extracción de frases clave de artículos científicos y artículos de noticias es que las frases clave normalmente no aparecen sólo al principio, sino también al final de un documento.

Esta observación no necesariamente se mantiene para el texto conversacional (por ejemplo, reuniones, charlas), sin embargo la razón es simple: en una conversación, los temas (es decir, sus temas de conversación) el cambio como la interacción se mueve hacia adelante en el tiempo, y también lo hacen las palabras clave asociadas con un tema [Hasan 14].

Es decir, mientras que en un artículo las secciones ayudan a desarrollar una investigación y a seguir una línea de desarrollo y tiempo, esto no se aplica en pláticas informales como un chat por ejemplo, en donde nosotros podemos estar platicado con una persona de un tema de deportes, enseguida tocar un tema personal y después de un tiempo volver a hablar del tema de deportes. Haciendo que conforme a la fluidez de la plática las frases clave vayan acompañando al tema en cuestión.

4.- Correlación de tema

Hasan [Hasan 14] a partir de [Turney 03] y [Mihalcea 04] indica que: Otra observación comúnmente explotada en la extracción de frases clave de artículos científicos y artículos de noticias es que las frases clave en un documento normalmente se relacionan entre sí.

Sin embargo, esta observación no necesariamente se mantiene para el texto informal (por ejemplo, correos electrónicos, chats, reuniones informales, blogs personales), donde la gente puede hablar sobre cualquier número de temas potencialmente no correlacionados [Hasan 14]. En conclusión, con lo anterior dentro de la extracción de palabras se busca la relación entre palabras para poder realizar la búsqueda de los candidatos más probables a ser frases clave pero, como menciona Hasan, en pláticas informales esto no se mantiene, ya que las personas pueden hablar de temas que no tiene ninguna relación.

CAPÍTULO 3.

Estado del Arte

En este capítulo, se presenta trabajos relacionados a la extracción automática de frases clave. En primer lugar, se presenta una breve descripción de la tarea 5 del SemEVal-2010 que se llama "*Extracción automática de frases clave de artículos científicos*". Después se presentan los sistemas que ocuparon las primeras posiciones en la competencia. Posteriormente, en la segunda sección se describen los trabajos del estado de arte en la extracción automática de frases clave.

3.1 Tarea 5: Extracción automática de frases clave de artículos científicos

En el año 2010, [Kim 10] junto a otros organizadores realizó la tarea compartida "*Task 5: Automatic Keyphrases extraction from Scientific Articles*" que se incluyó en el SemEval-2010. El propósito fue desarrollar sistemas de extracción automática de frases clave de artículos científicos y comparar la lista de frases propuestas por cada sistema participante, con las frases clave que fueron asignadas por seres humanos a cada uno de los artículos científicos, evaluando los resultados de manera automática. El sistema que obtuvo el mejor resultado en la tarea fue HUMB [López 10] con F-medida de 27.5% (F-medida y otras medidas se explican brevemente en la página 37).

A continuación se presentan los sistemas que obtuvieron los primeros lugares en la tarea 5 del SemEval-2010:

3.1.1 HUMB: Extracción automática de términos clave de los artículos científicos en GROBID

[López 10] participa en la tarea 5 del SemEval-2010 con el sistema HUMB obtenido el primer lugar en la competencia. HUMB es de enfoque supervisado, para la extracción de frases clave analiza la estructura del documento (resumen, conclusión, referencias). La selección de candidatos que implementa es la extracción de n-gramas de hasta 5 palabras, eliminación de candidatos que empiezan o terminan con *stopwords,* filtrado de símbolos matemáticos. La clasificación de candidatos la realiza mediante un árbol de decisión, además de utilizar las bases de datos terminológicas de GRISP y Wikipedia. El resultado más alto que obtuvo de las tres categorías (autor, lector y combinado), fue en las frases "Combinado" con 27.5% de F-medida.

3.1.2 WINGNUS: Extracción de frases clave utilizando la estructura lógica del documento

[Nguyen 10] participa en la tarea 5 del SemEval-2010 con el sistema WINGNUS. WINGNUS es de enfoque supervisado, una de las características principales que maneja para la extracción de frases clave es la estructura lógica del documento, para hacer menos el texto a analizar e identificar las secciones en donde es más probable que se encuentren las frases clave. Para la clasificación de candidatos emplea 19 funciones sintácticas de las cuales obtiene su mejor resultado con las funciones: F1-F3: TF x IDF, frecuencia de los términos, frecuencia de sub-cadenas, F4: primera ocurrencia y F6: longitud de la frase en palabras.

3.1.3 KP-Miner

[El-Beltagy 10] participa en la tarea 5 del SemEva-2010 con el sistema KP-miner. KP-miner es de enfoque no supervisado que extrae frases clave a partir de textos en árabe e inglés. Su proceso es de tres pasos: 1.- Selección de candidatos donde filtra palabras que no estén separadas por signos de puntuación o *stopwords,* también se incluye la frecuencia de la frase y la primera aparición, 2.- Cálculo de pesos: peso del término, frecuencia del término, termino IDF, factor

de aumento y posición del término. 3.- Lista de refinamiento de candidato final: que es una característica opcional del sistema para refinar los candidatos.

3.1.4 SZTERGAK: función de ingeniería para extracción de frases clave

[Bernend 10] participa en la tarea 5 del SemEvla-2010 con el sistema SZTERGAK. SZTERGAK es un sistema de enfoque supervisado. La selección de candidatos que implementa es la extracción de n-gramas de hasta 4 palabras, sus características se agrupan en cuatro categorías: 1.-Nivel de frase (longitud de la palabra, POS *pattern*), 2.-Nivel de documento (características: *acronomity*, PMI, *Sintactic*), 3.-Nivel de corpus (TF x IDF, *Keyphraseness*) y 4.-Conocieminto externo: uso de Wikipedia como recurso externo.

3.1.5 KX: un sistema flexible para la extracción de frases clave

[Pianta 10] participa en la tarea 5 del SemEval-2010 con el sistema KX. KX es un sistema de enfoque no supervisado, en la selección de candidatos extrae n-gramas de hasta 4 palabas, KX emplea cuatro pasos: tres a nivel corpus y uno extrae la información específica de un documento. Para la clasificación de candidatos emplea las siguientes características: IDF, longitud de la frase, posición de la primera aparición, subsunción y *boosting*.

3.2 Sistemas de extracción automática de frases clave del estado del arte

3.2.1 Una comparación de los modelos supervisados de extracción de frases clave

En el año 2015 [Bulgarov 15] presenta la comparación de su modelo para extracción de frases clave "CeKE". El propósito de este trabajo fue comparar el sistema CeKE con sistemas de extracción automática de frases clave de enfoque supervisado tales como Maui, KEA y el de Hulth "n-gram con etiquetas". Para evaluar el desempeño de cada uno de estos sistemas utilizó un conjunto de datos formado de World Wide Web (WWW) y descubrimiento de conocimiento y minería de datos (KDD) con asignaciones de frases clave de autor, Bulgarov reporta que CeKE más la característica keyphraseness obtiene mayores resultados en Precisión y F-medida que los otros sistemas de extracción. En la tabla 4, [Bulgarov 15] presentan los resultados de CeKE más la característica keyphraseness (*la característica keyphraseness cuantifica con qué frecuencia una frase candidato aparece como etiqueta o frase clave en el entrenamiento* [Medelyan 09]) en la cual se marcan los resultados en Precisión, Recuerdo y F-medida.

	WWW			KDD		
Method	Precision	Recall	F1-score	Precision	Recall	F1-score
CeKE	0.228	0.386	0.285	0.213	0.413	0.280
Maui	0.120	**0.502**	0.193	0.104	**0.466**	0.170
Hulth - n-gram with tags	0.165	0.107	0.129	0.206	0.151	0.172
KEA	0.210	0.146	0.168	0.178	0.124	0.145
CeKE + keyphraseness - Na¨ıve Bayes	**0.251**	0.460	**0.322**	**0.254**	0.440	**0.321**

Tabla 4. Resultados de Ceke y los sistemas del estado del arte en la comparación de Bulgarov.

3.2.2 Extracción automática de frases clave: Un estudio al estado del arte

En el año 2014 [Hasan 14] presenta un estudio sobre la extracción automática de frases clave. En el cuál se mencionan aspectos que influyen en la extracción, tipos de corpus utilizados en trabajos del estado del arte, enfoques de la extracción, sistemas que han alcanzado las mejores puntuaciones en un conjunto de datos, análisis de errores y recomendaciones para mejorar la calidad de extracción. En la tabla 5, a partir de Hasan se muestra el corpus utilizado por el sistema, técnica de extracción y su puntuación obtenida en Precisión, Recuerdo y F-medida.

Dataset	**Approach and System [Supervised?]**	**Score**		
		P	**R**	**F**
Abstracts (*Inspec*)	Topic clustering (Liu et al., 2009b) [×]	35.0	66.0	45.7
Blogs	Topic community detection (Grineva et al., 2009) [×]	35.1	61.5	44.7
News (DUC -2001)	Graph-based ranking for extended neighborhood (Wan and Xiao, 2008b) [×]	28.8	35.4	31.7
Papers (SemEval -2010)	Statistical, semantic, and distributional features (Lopez and Romary, 2010) [✓]	27.2	27.8	27.5

Tabla 5. Mejores puntuaciones de sistemas en diferentes conjuntos de datos [Hasan 14].

3.2.3 Palabras clave y técnicas de extracción de frases clave: Una revisión a la literatura

En el año 2015 [Siddiqui 15] presenta un estudio cronológico al estado del arte en la extracción automática de frases clave, separa los trabajos de acuerdo al enfoque que pertenecen: estadístico, supervisados, sin supervisión y semi-supervisado. También presenta las características que se emplean para poder clasificar a los candidatos a frases clave dentro del texto analizado.

3.2.4 *Extracción de frases clave en publicaciones científicas*

En el año 2007 [Nguyen 07] presenta un algoritmo para la extracción de frases claves en artículos científicos, con base en la estructura de un artículo científico o académico de acuerdo a sus secciones (resumen, introducción, conclusiones). La identificación de estas secciones facilitan la identificación de candidatos a frases clave, ya que la mayor parte de las veces las frases clave de un artículo científico se ubican en secciones como introducción y conclusión.

3.2.5 *KEA: Práctica de Extracción automática de frases clave*

[Witten 99] crea junto con otros investigadores un algoritmo para la extracción de frases clave (KEA), el algoritmo es de enfoque supervisado utiliza la técnica Naive Bayes, el cual a partir de datos de entrenamiento crea un modelo de formación que puede extraer las frases clave de nuevos documentos. KEA emplea 2 características TF-IDF y primera aparición de una frase. En la figura 8, a partir de Witten presenta el funcionamiento interno que realiza KEA para cumplir con la extracción de frases clave, en donde teniendo un conjunto de documentos o un solo documento, se extraen los candidatos a frases clave, se aplican las técnicas de extracción. Posteriormente, se consulta si hay entrenamiento de datos para aplicar un modelo de formación y se extraen los candidatos con más probabilidad de ser una frase clave (ver Figura 4).

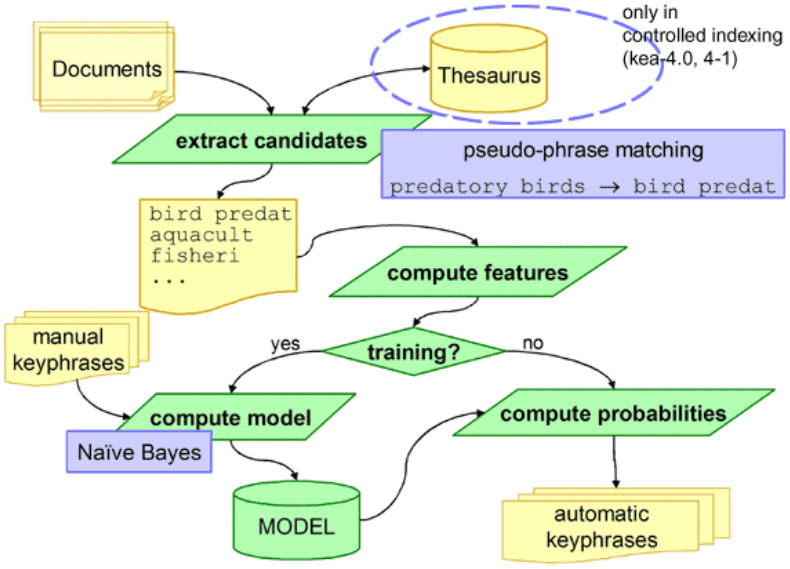


Figura 4. Proceso de extracción KEA [KEA 15].

3.2.6 *Genex*

[Turney 2000] presenta los resultados de una comparación entre un modelo de extracción basado en un algoritmo genético y una implementación de árboles de decisión C4.5. Turney informa que el algoritmo genético emite mejores palabras clave que los árboles de decisión.

3.2.7 Mejora de la extracción automática de frases clave dando conocimiento lingüístico

[Hulth 03] presenta un método basado en un algoritmo supervisado con la adición de conocimiento lingüístico, partir de 4 funciones: frecuencia dentro del documento, frecuencia de recogida, posición relativa de la primera aparición, secuencia de parte de la etiqueta de voz además de NP-chunker y Pos Tag, el uso de la etiqueta "Pos" como característica asignada a la selección de candidatos devuelve una mejor extracción de frases clave.

3.2.8 Etiquetado humano competitivo usando extracción automática de frases clave

[Medelyan 09] presenta Maui una variante de Kea. Maui es un algoritmo de enfoque supervisado para la indexación automática, utiliza información semántica extraída de Wikipedia con lo cual se manejan recursos externos para poder obtener una mejor extracción de frases clave basándose en los títulos de Wikipedia.

3.2.9 TextRank: Poniendo orden en textos

[Mihalcea 04] presenta un modelo de clasificación basado en un grafo no supervisado, que utiliza la co-ocurrencia y relación entre palabras que se añaden al grafo para posteriormente dar peso a los vértices. TextRank realiza dos tareas dentro de la recuperación de información que son: extracción de palabras clave y extracción de frases para resúmenes automáticos.

CAPÍTULO 4.

Metodología de trabajo

En este capítulo, se presenta la metodología propuesta para evaluar a los sistemas en esta tesis. Posteriormente se menciona cada etapa que integra a la metodología, así como el proceso que realiza cada una de estas, para llevar acabo la evaluación de los sistemas de extracción automática de frases clave.

4.1 Metodología de trabajo

La metodología de trabajo propuesta para evaluar a los sistemas de extracción de frases clave es la siguiente: Se trabaja el conjunto de datos "SemEval2010" específicamente con los artículos de la carpeta "TEST", a los cuales se les aplica pre-procesamiento. Después, se ingresa cada artículo en los sistemas a evaluar para comenzar con la extracción. Los parámetros que se ingresan para la extracción de frases es una configuración estándar para todos los sistemas. Al tener los resultados de los sistemas se les aplica el algoritmo "Porter stemmer" (stemming). Posteriormente se ingresan los resultados al formato de evaluación. La evaluación se realiza con el programa que evalúa los resultados, mismo que se utilizó en la tarea 5 del SemEval-2010. Finalmente, se clasifica a cada sistema con base en los resultados obtenidos.

Siguiendo la metodología propuesta se pueden evaluar todos los sistemas y conocer el desempeño de cada uno de ellos sobre el corpus SemEval2010. En la figura 5, se presenta la metodología a emplear de manera gráfica (ver figura 5).

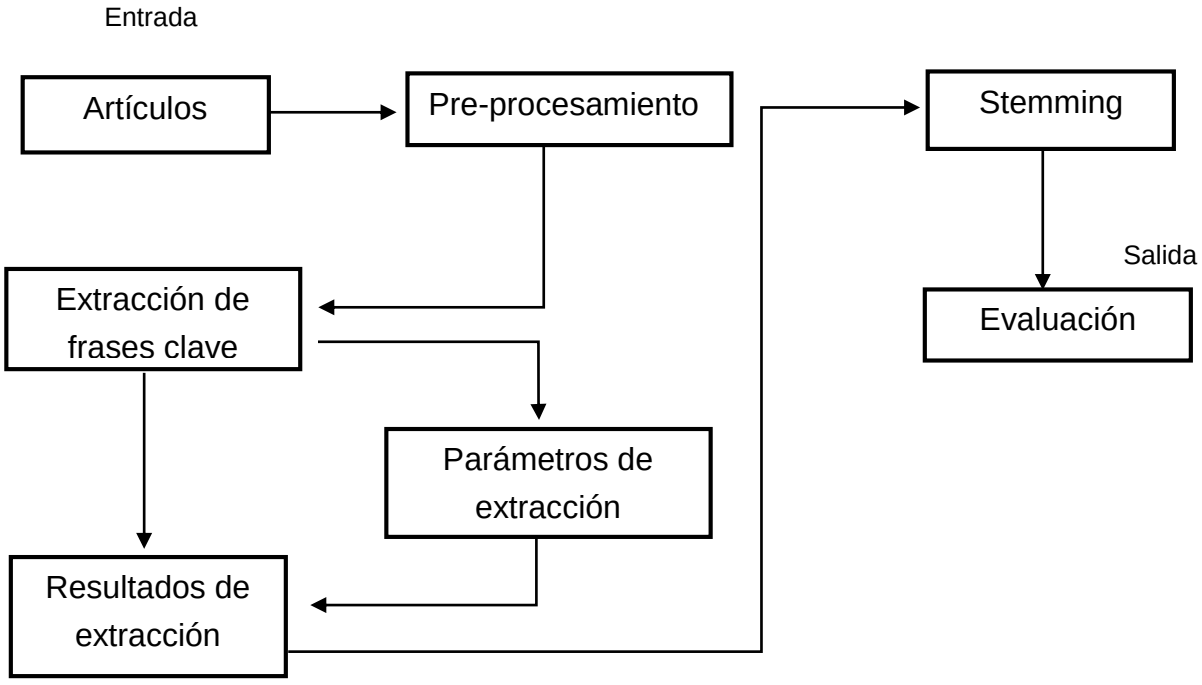


Figura 5. Metodología para evaluar a los sistemas 2016.

4.1.1 Entrada

La entrada para la generación de frases clave son los 100 artículos contenidos en la carpeta "TEST "del conjunto de datos SemEval2010.

4.1.2 Pre-procesamiento

Antes de extraer las frases clave con los sistemas evaluados se aplicó el pre-procesamiento a los artículos el cual consistió en: colocar todo el texto en una sola línea, remover fórmulas matemáticas y remover los caracteres que no fueran números ni letras a excepción de puntos, comas y guiones cortos.

En la etapa de colocar el texto en una sola línea se realizó con el fin de poder separar las palabras ya que al quitar el salto de línea se formaba una sola palabra, lo cual era incorrecto. A continuación se muestra una parte del texto original:

ABSTRACT
Efficient discovery of grid services is essential for the success of grid computing. The standardization of grids based on web services has resulted in the need for scalable web service.

Texto en una sola línea:

ABSTRACT Efficient discovery of grid services is essential for the success of grid computing. The standardization of grids based on web services has resulted in the need for scalable web service

En la etapa de eliminación, se quitaron las fórmulas matemáticas. [Kim 13] menciona que el formato original de los artículos era PDF pero se convirtieron a texto plano para uso en la tarea, sin embargo, las fórmulas matemáticas o tablas no se convirtieron en su totalidad, razón por la cual causa ruido en el texto. A continuación se muestra una fórmula matemática que se encuentra en el artículo C-14 del conjunto de datos SemEval2010 en formato de texto plano.

$$Dv(u) = Prob\ N\ i=1\ K\ ||u - si||k + Ni \geq \eta = Prob\ N\ i=1\ Ni \geq \eta - N\ i=1\ K\ ||u - si||k$$

Por último, se removieron los caracteres que no son números, ni letras a excepción de los puntos, comas y guiones. Ya que con la conversión de PDF a texto plano se sustituyeron caracteres que no fueron reconocidos en su totalidad. A continuación se muestra un ejemplo de una fórmula que se encuentra en el artículo C-17 del conjunto de datos SemEval2010 en formato PDF (ver figura 6):

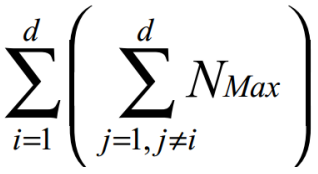


Figura 6. Fórmula matemática en el artículo C-17 en formato PDF.

Con la conversión de PDF a texto plano, la formula anterior se sustituyó por los siguientes caracteres (ver figura 7):

∑ ∑= ≠=□□□□□□□□

Figura 7. Fórmula matemática en el artículo C-17 en texto plano.

4.1.3 Sistemas a evaluar

Los sistemas evaluados son los siguientes (listados por orden alfabético):

- Alchemy (extracción de frases clave)
- Extractor (extracción de frases clave)
- Fivefilters (extracción de términos)
- Genia (extracción de términos)
- Kea (extracción de frases clave)
- Skyttle (extracción de frases clave)
- Tree tagger (extracción de términos)
- Texlexan (extracción de frases clave)
- Translatedlab(extracción de términos)
- Wordstat (extracción de frases clave)

Los sistemas evaluados se pueden dividir en dos categorías de la siguiente forma:
Sistemas libres: KEA, Fivefilters, Texlexan

Sistemas comerciales: Alchemy, Extractor, Genia, Skyttle, Tree Tagger, Translatedlab, Wordstat

Se eligieron estos sistemas de extracción por la disponibilidad que tienen, tanto los de licencia libre como los comerciales que ofrecen su versión de prueba. A continuación se presenta una tabla comparativa con las características (licencia, instalable, demo en línea, opciones disponibles, idioma, plataforma) de los sistemas que son evaluados (ver tabla 6).

Sistema KE	Licencia	Instalable	Demo En línea	Opciones disponibles	Idiomas	Plataforma
AlchemyAPI	Comercial		✓	-Análisis de sentimiento -Ayuda de idioma -Relevancia de clasificado -Formatos de respuesta	Inglés, Alemán, Francés, Italiano, Portugués, Ruso, Español y Sueco	Web (Demo)
Fivefilters 1.0	Libre		✓	-Máximo de resultados -Salida de resultado -Máximo de palabras por termino -Términos en minúscula	Inglés	Web
KEA	Libre	✓		-Longitud mínima de frase -Longitud máxima de frase -Mínimo de ocurrencia -Nombre vocabulario	Inglés, español y francés	Multiplataforma
NaCTeM	Comercial		✓	-Gennia Tagger 2.1 -Tree Tagger 3.1	Inglés	Web
TexLexAn	Libre	✓		Default	Inglés, español, francés, alemán, italiano	Linux
Wordstat 7	Comercial	✓		- Longitud mínima de frase - Longitud máxima de frase - Minino de casos	Multilenguaje	Windows, Mac, Linux
Extractor 7.2	Comercial	✓	✓	-Número de frases a extraer -Ingreso de Stopwords -Ir a Frases	Inglés, Español, Francés, Alemán, Japonés, Coreano	Windows, Solaris, Linux, MacOS y HP/UX
Translated	Comercial		✓	Default	Inglés, Italiano, Francés	Web(Demo)
Skyttle	Comercial		✓	-Extracción de frases clave -Análisis de sentimientos	Inglés	Web(Demo)

Tabla 6. Características funcionales de los sistemas evaluados.

4.1.4 Parámetros de extracción

Para realizar la extracción de frases clave algunos sistemas requieren parámetros. Los parámetros de extracción que un sistema solicita son los siguientes: longitud de una frase, frecuencia, número de frases a extraer, etc. Como es el caso de Wordstat 7 en donde se ingresa la longitud de una frase o en caso contrario estos no se solicitan como AchemyAPI, en donde lo único por hacer es pegar el texto. En tarea 5 del SemEval-2010, no se solicitó a los sistemas participantes una configuración estándar para la extracción de las frases clave ya que los participantes podían adaptar los parámetros de extracción a la configuración de su sistema [Medelyan 15].

Al darse a conocer los resultados de la tarea 5 del SemEval-2010, el sistema que alcanzó la mejor posición durante la competencia fue el sistema HUMB [López 10], dentro de las características que este sistema uso en el corpus SemEval2010 para las frases clave menciona las siguientes:

- Extracción todos los n-gramas de hasta 5 palabras
- Filtro de términos que contienen símbolos matemáticos
- Normalizar a los candidatos por minúscula utilizando el algoritmo de Porter (Stemming)

En esta tesis, para realizar la evaluación se usa una configuración estándar con el fin de medir el desempeño de los sistemas bajo una sola configuración. Con lo mencionado anteriormente, los parámetros de extracción en esta evaluación son:

Número de frases a extraer:
Extraer una lista de 15 frases clave por cada uno de los 100 artículos del SemEval2010.

Longitud mínima:
1 palabra ya que se representa como palabra clave.

Longitud máxima:
Con base en el sistema HUMB [López 10], el número de palabras que puede contener una frase es de 5. Se hace esto con el fin de abarcar la mayor cantidad de frases clave con 4 y 5 palabras, pues son las que mayormente se presentan en el corpus. También hay frases de mayor longitud; sin embargo contienen *stopwords* y los sistemas que se están evaluando filtran estas palabras.

Frecuencia:
En el caso de los sistemas que requieran este parámetro se dejará por defecto al que traigan consigo.
La tabla 7, presentan los parámetros que se usan en los sistemas evaluados en esta tesis.

Sistema de extracción de frases clave		Opción utilizada	Idioma
Instalable	**En línea**	**Características de extracción**	
	AlchemyAPI	Default	Inglés
	Fivefilters 1.0	Máximo de resultados: 15 Máximo de palabras por término: 5	Inglés
KEA		Número de frases a extraer: 15 Longitud máxima de frase: 5 Longitud mínima de frase: 1 Mínimo de ocurrencia: 3	Inglés
	NaCTem (Genia Tagger & Tree Tagger)	Default	Inglés
TexLexAn		Default	Inglés
Wordstat 7		Longitud máxima de frase: 5 Longitud mínima de frase: 2 Mínimo de ocurrencia: 3	Inglés
Extractor 7.2		Número de frases a extraer: 15	Inglés
	Translated	Default	Inglés
	Skyttle	Default	Inglés

Tabla 7. Opciones utilizadas en los sistemas para realizar la extracción automática de frases clave.

4.1.4 Stemming

El algoritmo de *Porter stemming* es un procedimiento para encontrar la raíz de una palabra. Se aplica a los resultados de los sistemas de igual forma que en SemEval-2010. A continuación se presenta un ejemplo de frases clave en su estado original y posteriormente con *stemming.*

Frases clave en su estado original:
uddi registries, uddi registry, dht, multiple uddi registries, local uddi registry, private uddi registries, uddi business registries, proxy registry, uddi key, unique uddi key, respective uddi registries, relevant uddi registries, dude distributed uddi, uddi local registry, multiple proxy uddi

Frases clave con stemming:
uddi registri, uddi registri, dht, multipl uddi registri, local uddi registri, privat uddi registri, uddi busi registri, proxi registri, uddi kei, uniqu uddi kei, respect uddi registri, relev uddi registri, dude distribut uddi, uddi local registri, multipl proxi uddi

4.1.5 Evaluación

La evaluación se realiza de manera automática con el programa ("performance.pl") que evalúa los resultados de los sistemas y que fue el mismo que se utilizó en la tarea 5 del SemEval-2010. La evaluación de los sistemas se lleva a cabo bajo lo siguiente: las frases propuestas por los sistemas se comparan contra las frases que asigno el autor y el lector para cada uno de los artículos [Kim 13]. Los archivos que contienes estas frases son:

- *Autor-asignado*: contiene las frases clave que propuso el autor del artículo.
- *Lector-asignado*: contiene las frases clave que propuso el lector artículo.
- *Combinado*: contiene las frases clave de los autores y lectores.

La evaluación se obtiene con las métricas de Precisión, Recuerdo y F-medida para los mejores 5, 10, y 15 frases clave.

Precisión

Expresa el número de coincidencias ("correctas") como una proporción de todos los temas del algoritmo [Medelyan 09a] (ver figura 8).

$$P = \frac{\#correct\ extracted\ topics}{\#all\ extracted\ topics}$$

Figura 8. Fórmula de Precisión.

Recuerdo

Es la proporción de los temas humanos que están cubiertos [Medelyan 09a] (ver figura 9).

$$R = \frac{\#correct\ extracted\ topics}{\#manually\ assigned\ topics}$$

Figura 9. Fórmula del Recuerdo.

F-medida

F-medida (también llamada *F-score*) combina la Precisión y el Recuerdo, en toda su generalidad implica un parámetro β que permite al evaluador dar más peso a Precisión o Recuerdo [Medelyan 09a] (ver figura 10).

$$F_\beta = \frac{(1 + \beta^2)\, PR}{\beta^2 P + R}$$

Figura 10. Fórmula del F-medida.

Para el posicionamiento de resultados, los sistemas se clasifican en el orden descendente de su F-medida en el top-15 para las frases clave que proponen, en el caso de empate en el F-medida, se sub-clasifica a los equipos con el F-medida descendente sobre el conjunto de datos completo [Kim 13].

CAPÍTULO 5.

Experimentación

En este capítulo, se describe el conjunto de datos SemEval2010. Posteriormente, se presentan los resultados de los sistemas instalables y en línea, posicionados conforme a su desempeño. Finalmente, los resultados de esta evaluación se comparan con los resultados de la tarea 5 del SemEval-2010.

5.1 Descripción del Corpus SemEval2010

El conjunto de datos se creó partir de artículos que originalmente tuvieran las frases clave de su autor. Sin embargo, se agregaron las frases clave propuestas por los lectores del artículo. [Kim 13] menciona que en la asignación manual de frases clave por parte de los lectores para los artículos consistía en:

- Extraer frase clave que realmente aparecieran en el texto.
- No crear frases semánticamente equivalentes.
- Extraer frases de cualquier parte del documento incluyendo encabezados y pies de imágenes.

Sin embargo [Kim 13] reporta que en las frases de lector-asignado un 15% no aparecen en el texto y en las de autor-asignado un 19%, por lo que el alcance máximo que los sistemas pueden lograr en estos documentos es el 85% y el 81% para las frases clave lector y autor. Es decir, los sistemas que fueron evaluados en 2010, como en este trabajo, no alcanzan un 100% en sus resultados, porque algunas frases con las que se hacen coincidir en los archivos que contienen los patrones oro no están presentes en el texto original (ver figura 11).

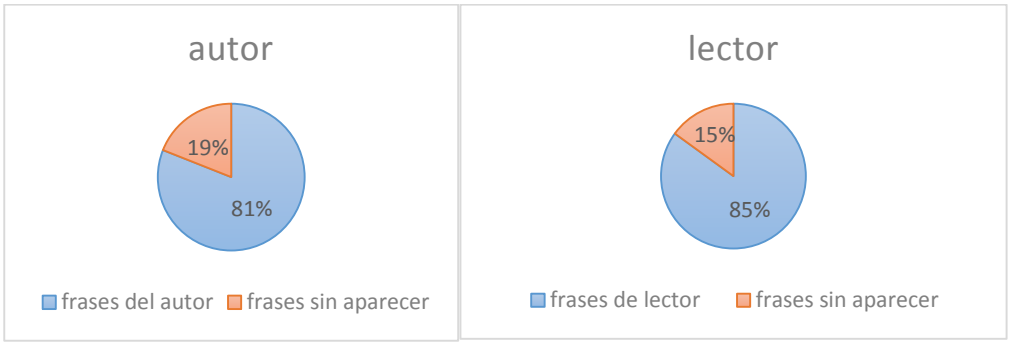


Figura 11. Porcentaje de frases que no aparecen en el texto original de los artículos del "Test" para el corpus SemEval-2010 en las frases del autor y lector.

El conjunto de datos utilizado en los experimentos es una colección de artículos científicos que fueron usados en SemEval-2010. Los artículos provienen de la Biblioteca Digital ACM (ponencias y talleres). La tabla 8 muestra la distribución de las cuatro categorías de los artículos donde:

- **C**: Sistemas Distribuidos
- **H**: Búsqueda de información y recuperación
- **I**: Inteligencia Artificial – Sistemas Multi-agente
- **J:** Ciencias Sociales del Comportamiento

Dataset	Total	C	H	I	J
Trial	40	10	10	10	10
Training	144	34	39	35	36
Test	100	100	25	25	25

Tabla 8. Distribución de las 4 áreas que integran el corpus SemEval-2010.

La longitud de los artículos es de 6 a 8 páginas, Originalmente el formato de los artículos estaba en PDF pero fue convertido a texto plano para uso de los participantes [Kim 13]. Para realizar los experimentos con los sistemas, se utilizan los archivos que contiene la carpeta "test" los cuales fueron utilizados en SemEval-2010. El conjunto de datos SemEval2010 está disponible gratuitamente, puede ser descargado desde el sitio https://github.com/snkim en donde también se encuentran otros corpus de artículos científicos.

5.2 Resultados obtenidos

A continuación se muestras los resultados de los sistemas de extracción automática de frases clave para conocer cuáles de estos presentan los mejores resultados en cada asignación, se clasifican por su resultado de F-medida en el top 15.

En las tablas siguientes:

- P se refiere a: Precisión
- R se refiere a: Recuerdo
- F se refiere a: F-medida

En las figuras se presentan los resultados de manera gráfica, la señalización roja indica la primera posición, mientras que el color azul del gráfico es para los sistemas instalables, el color verde lo es para los sistemas en línea.

En esta sección en particular se muestran los resultados del top 15, que son lo que se toman en cuenta en la tarea 5 del SemEval-2010 y en esta tesis, para clasificar a los sistemas de extracción automática de frases clave. Sin embargo las posiciones dentro del top 5 y top 10 varían. En el Anexo 1y 2 se puede encontrar la clasificación de los sistemas en el top 5 y 10.

La evaluación se realizó con una configuración de parámetros estándar en todos los sistemas (ver capítulo 4). A continuación se muestran los resultados de los sistemas "*instalables*" y "*en línea*" en una sola clasificación para cada una de las tres asignaciones: autor, lector y combinado sobre el corpus SemEval2010.

5.2.1 Resultados en las frases clave asignadas por el autor

En los resultados para las frases clave asignadas por el autor en el top 15, Extractor con Precisión 9.0%, Recuerdo 34.88% y F-medida de 14.31%, se posiciona en el primer lugar dentro de esta asignación. En la tabla 9, se muestran los resultados y se marca el valor más alto (ver tabla 9).

		Top 15		
Sistema	Rank	P	R	F
Extractor	1	**9**	**34,88**	**14,31**
Kea	2	8,87	34,37	14,1
Alchemy	3	8,47	32,82	13,47
Wordstat	4	7,67	29,72	12,19
Genia	5	6,87	26,61	10,92
Tree tagger	6	6,6	25,58	10,49
Skyttle	7	5,33	20,67	8,47
Five filters	8	5,27	20,41	8,38
Texlexan	9	4	15,5	6,36
Translatedlab	10	3,2	12,4	5,09

Tabla 9. Clasificación de los sistemas en las frases clave asignadas por el autor en el top 15.

En la figura 12, se muestran los resultados de manera gráfica del top 15, para las frases clave del autor y se señala el mayor resultado.

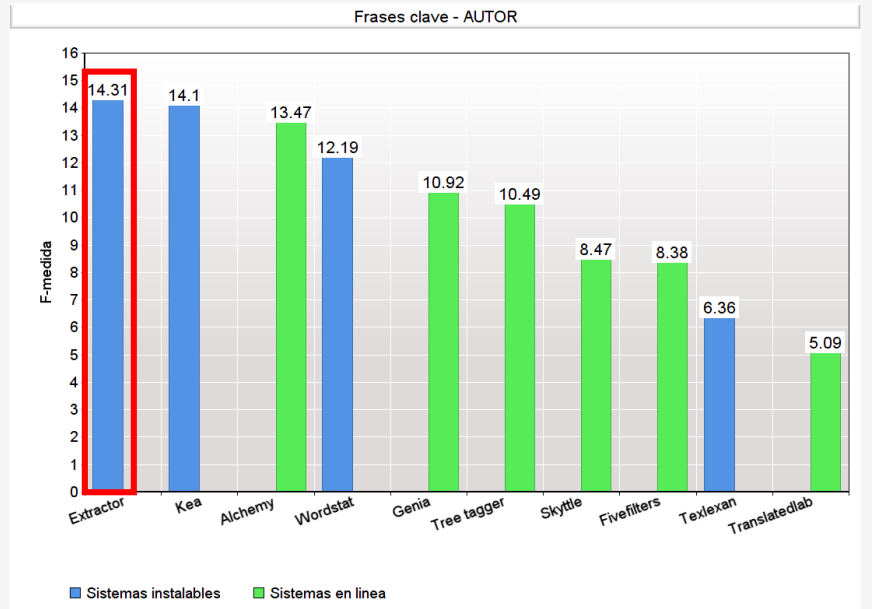


Figura 12. Desempeño de los sistemas sobre las frases clave asignadas por el autor en el top 15.

5.2.2 Resultados en las frases clave asignadas por el lector

En los resultados paras las frases clave asignadas por el lector en el top 15, Alchemy con Precisión 17.4%, Recuerdo 21.68% y F-medida de 19.31%, se ubica como el primer lugar dentro de esta asignación. En la tabla 10 se muestran los resultados y se marca el valor más alto (ver tabla 10).

		Top15		
Sistema	Rank	P	R	F
Alchemy	1	**17,4**	**21,68**	**19,31**
Wordstat	2	16,53	20,6	18,34
Extractor	3	16,33	20,35	18,12
Genia	4	15,53	19,35	17,23
Tree tagger	5	15,07	18,77	16,72
Kea	6	14,67	18,27	16,27
Fivefilters	7	10,87	13,54	12,06
Skyttle	8	10	12,46	11,1
Translatedlab	9	7,93	9,88	8,8
Texlexan	10	6,4	7,97	7,1

Tabla 10. Clasificación de los sistemas en las frases clave asignadas por el lector en el top 15.

En la figura 13, se muestran los resultados de manera gráfica del top 15, para las frases clave del lector y se señala el mayor resultado.

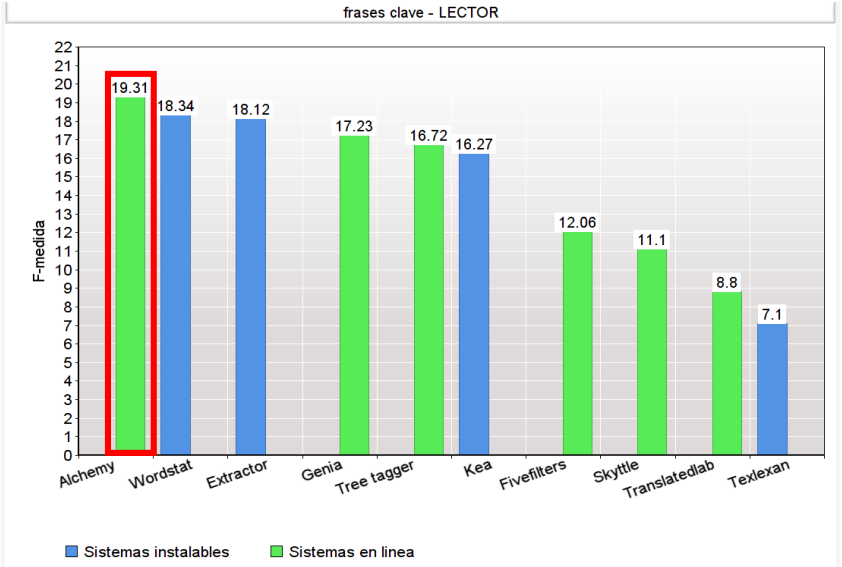


Figura 13. Desempeño de los sistemas sobre las frases clave asignadas por el lector en el top 15.

5.2.3 Resultados en las frases clave combinado

En los resultados para las frases clave combinado en el top 15, Alchemy con Precisión de 21.13%, Recuerdo 21.62% y F-medida 21.37%, se posiciona como el primer lugar dentro de esta asignación. En la tabla 11 se muestran los resultados y se marca el valor más alto (ver tabla 11).

Top15

Sistema	Rank	P	R	F
Alchemy	1	**21,13**	**21,62**	**21,37**
Extractor	2	20,8	21,28	21,04
Wordstat	3	20,27	20,74	20,5
Kea	4	19,33	19,78	19,55
Genia	5	18,6	19,03	18,81
Tree tagger	6	17,93	18,35	18,14
Fivefilters	7	13,07	13,37	13,22
Skyttle	8	12,67	12,96	12,81
Translatedlab	9	9,33	9,55	9,44
Texlexan	10	8,87	9,07	8,97

Tabla 11. Clasificación de los sistemas en las frases clave combinado en el top 15.

En la figura 14, se muestran los resultados de manera gráfica del top 15, para las frases clave combinado y se señala el mayor resultado.

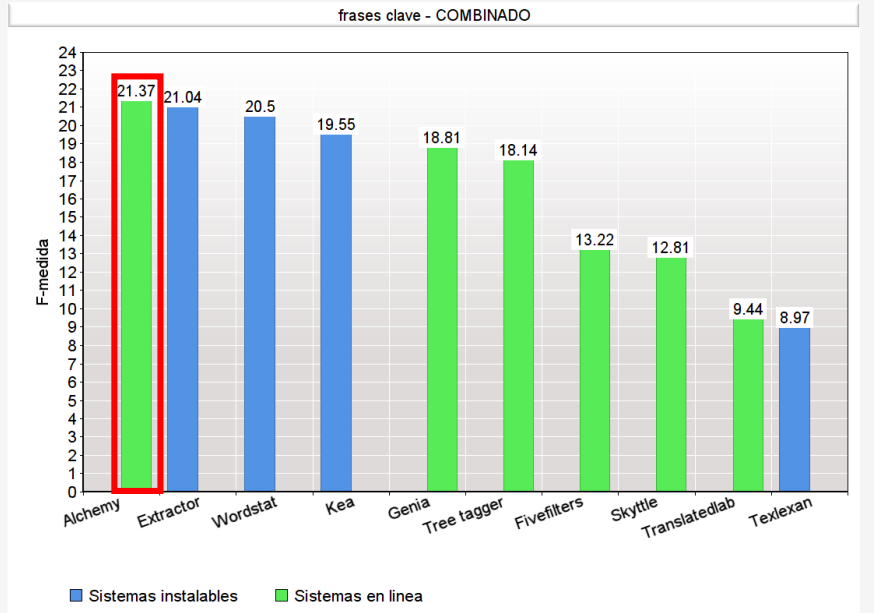


Figura 14. Desempeño de los sistemas sobre las frases clave combinado.

Como se puede observar, con los resultados obtenidos de los sistemas sobre las 3 asignaciones que contiene las frases clave oro, el sistema Alchemy quedó como el sistema con el mejor desempeño sobre las frases del lector y las frases "combinado", mientras que en las frases clave del autor lo fue el sistema Extractor.

5.4 Comparación de resultados de esta evaluación y de la tarea 5 del SemEval-2010.

A continuación se presentan la comparación de los resultados de esta evaluación y de la tarea 5 del SemEval-2010, con el objetivo de conocer si los sistemas disponibles en la actualidad presentan un mejor funcionamiento que los ya evaluados con anterioridad.

Comparación de resultados en las frases clave asignadas por el autor

En los resultados obtenidos sobre las frases clave del autor en esta tesis, el sistema Extractor con el resultado de 14.31%, tiene un resultado similar al de DERIUNLP 14.7% en SemEval-2010, mientras que el resultado más bajo en esta evaluación es de Translatedlab con 5.09%. En el top 15, en SemEval-2010 es el de UKP con 1.3%. En la figura 15, se señala el mejor resultado alcanzado en esta evaluación por Extractor 14.31% y en SemEval-2010 por HUMB 19.3%. Las barras de color azul pertenecen a los sistemas instalables y las barras verdes a los sistemas en línea que se comparan en esta evaluación mientras que las barras amarillas pertenecen a los sistemas que participaron en la tarea 5 del SemEval-2010 (ver figura 15).

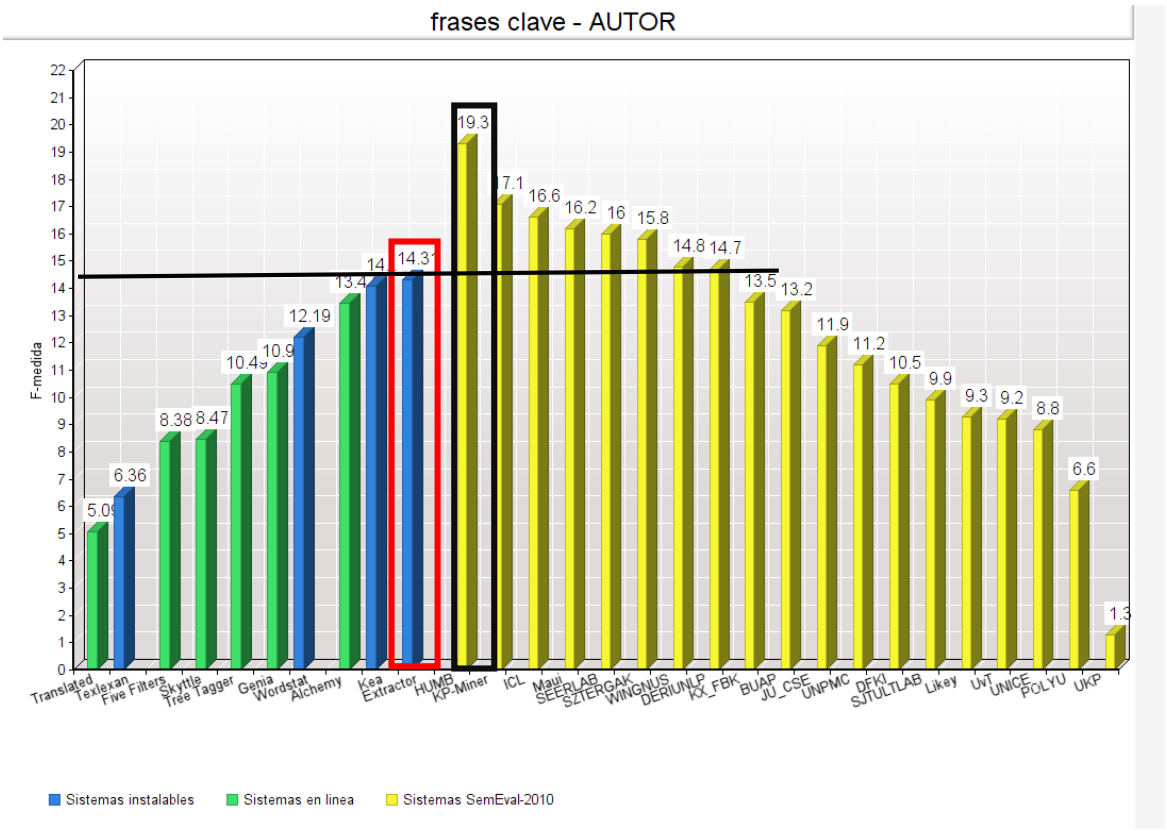


Figura 15. Desempeño de los sistemas evaluados en este trabajo y en SemEval-2010 sobre las frases clave asignadas por el autor.

Comparación de resultados en las frases clave asignadas por el lector

En las frases clave del lector, el sistema con el resultado más alto es Alchemy con 19.31%, su resultado es similar al de DERIUNLP con 19.5% y al de DFKI con 19.3% del SemEval-2010. El sistema con el resultado más bajo en esta evaluación es TexLexAn con 7.1%, mientras que en el SemEval-2010 es UKP con 5.2%. En la figura 16, se señala el mejor resultado alcanzado en esta evaluación por Alchemy con 19.31% y en SemEval-2010 por HUMB con 23.5%. Las barras de color azul pertenecen a los sistemas instalables y las barras verdes a los sistemas en línea que se comparan en esta evaluación mientras que las barras amarillas pertenecen a los sistemas que participaron en la tarea 5 del SemEval-2010 (ver figura 16).

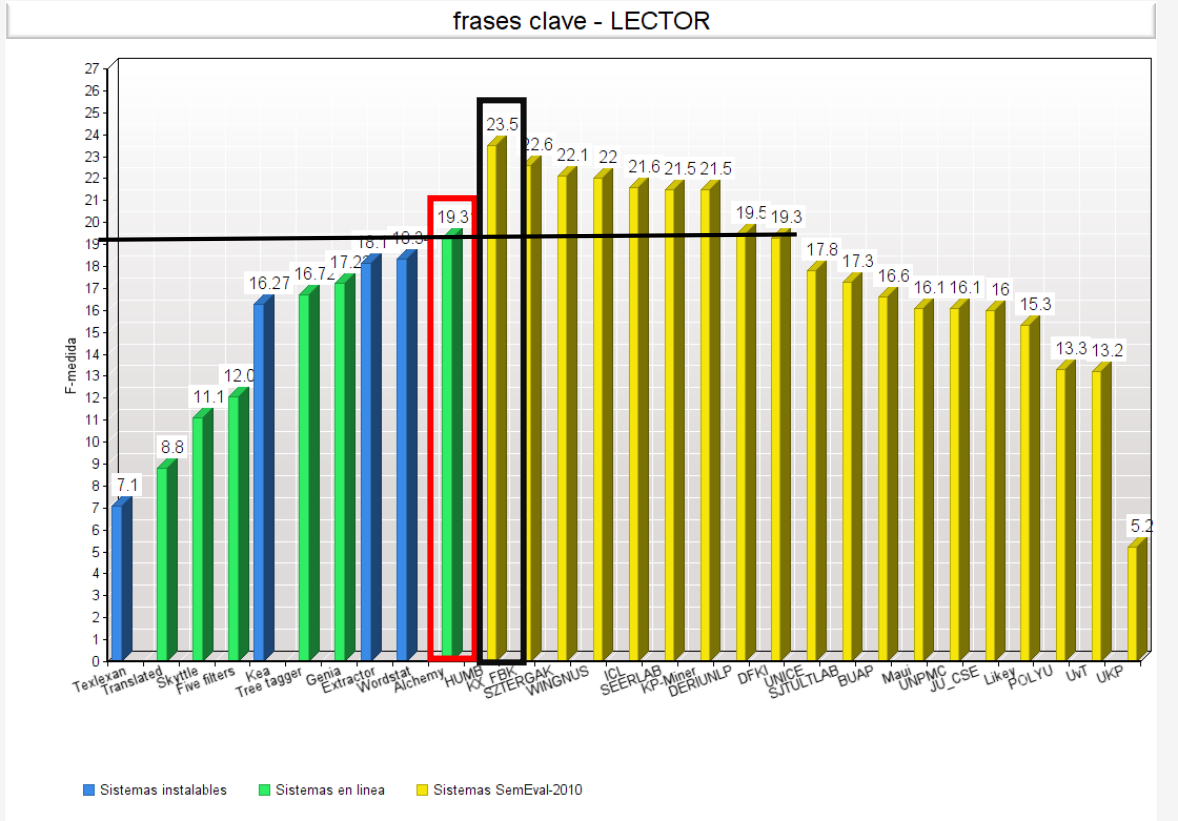


Figura 16. Desempeño de los sistemas evaluados en este trabajo y en SemEval-2010 sobre las frases clave asignadas por el lector.

Comparación de resultados en las frases clave combinado

En la característica de frases clave "combinado", el sistema con el resultado más alto en esta evaluación es Alchemy con 21.37%, su resultado está por debajo de DERIUNLP con 22.3% y por arriba de Maui con 20.6%. El sistema con el resultado más bajo en esta evaluación es TexLexAn, con 8.97% mientras que en SemEval-2010 es UKP con 5.3%. En la figura 17, se señala el resultado más alto en esta evaluación, que es Alchemy con 21.37% y en SemEval-2010 por HUMB con 27.5%. Las barras de color azul pertenecen a los sistemas instalables y las barras verdes a los sistemas en línea que se comparan en esta evaluación mientras que las barras amarillas pertenecen a los sistemas que participaron en la tarea 5 del SemEval-2010 (ver figura 17).

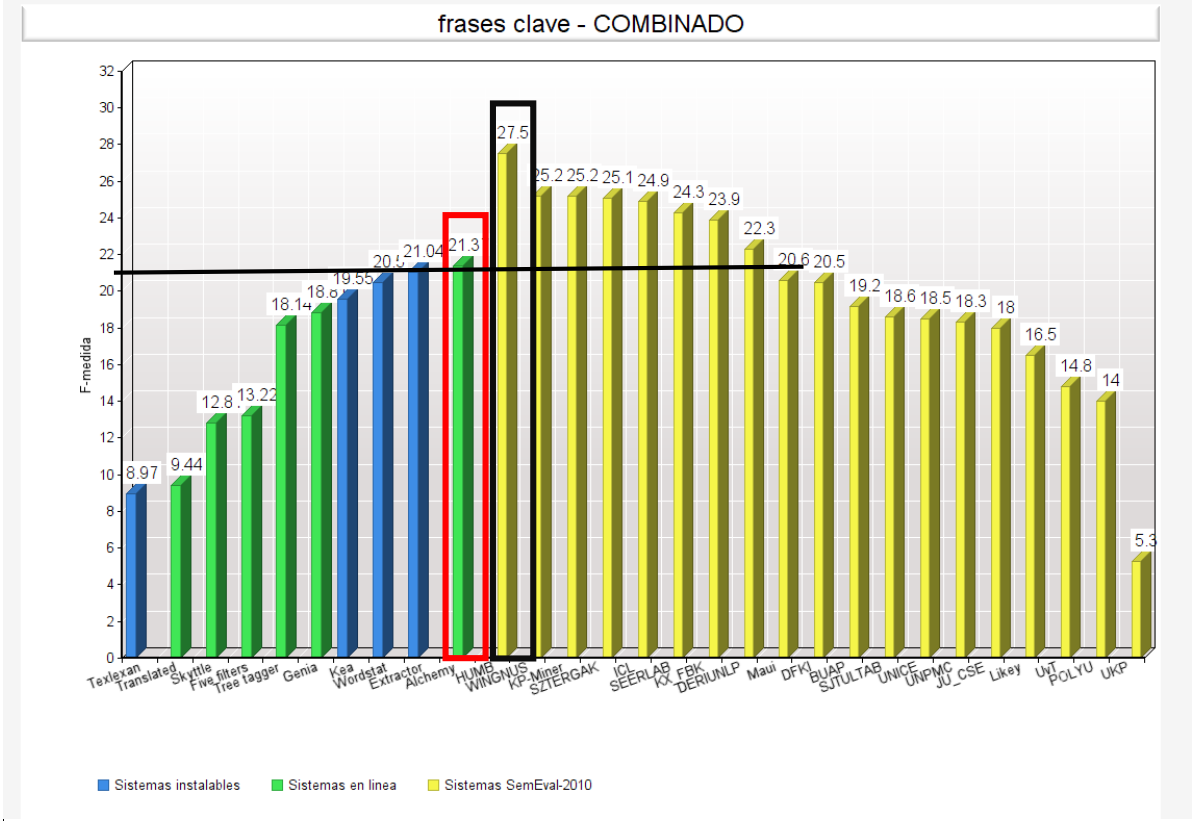


Figura 17. Desempeño de los sistemas evaluados en este trabajo y en SemEval-2010 sobre las frases clave combinado.

Con la comparación se puede observar que los resultados de la evaluación en la tarea 5 del 2010 son superiores en algunos casos y en otros se igualan a los resultados obtenidos en esta tesis. Se esperaba que los sistemas de esta evaluación presentaran un mejor desempeño que los ya evaluados con anterioridad, por el tiempo que ha transcurrido desde el 2010, lo cual no se logró en su totalidad.

Se usaron sistemas de extracción de términos pues en su descripción dicen encontrar las frases clave de un texto, no se esperaba que tuvieran un desempeño igual al de un sistema de extracción de frases clave, por la orientación a la que están dirigidos. Sin embargo, se dio el caso en que un extractor de términos tuvo mejores resultados que uno de frases clave.

En conclusión, la comparación de resultados nos ayudó a conocer el nivel que presentan los sistemas de extracción de frases clave en la actualidad con los ya evaluados sobre el conjunto de datos SemEval2010.

CAPÍTULO 6.

Conclusiones y Trabajo Futuro

En este capítulo, se presentan las conclusiones de este trabajo y se mencionan líneas futuras de investigación a partir de esta tesis.

6.1 Conclusiones

En esta tesis, se realizó la evaluación de sistemas que extraen una lista de frases clave de un texto con sistemas de licencia libre y comercial que se encuentran disponibles en internet para su uso o descarga, con el fin de conocer el desempeño que tiene sobre un conjunto de artículos científicos y encontrar las frases clave que fueron asignadas para cada artículo por un ser humano, mismo objetivo que se buscó en la tarea compartida del SemEval-2010 "*Tarea 5: Extracción automática de frases clave de artículos científicos*".

Recordando el planteamiento del problema: *¿Cuál es el desempeño de los sistemas instalables y en línea disponibles en 2016, para la extracción automática de frases clave en comparación con las asignadas por un ser humano, sobre un conjunto de artículos científicos de la colección SemEval-2010?*

La respuesta a la pregunta de investigación fue comparar los sistemas de extracción automática de frases clave disponibles e implementar la metodología que se propuso para evaluar a los sistemas, con lo que se pudo clasificarlos de acuerdo a su desempeño sobre la colección SemEval2010. Con lo mencionado anteriormente, se logró contestar la pregunta de investigación que se planteó, y comprobar que la hipótesis fue correcta.

Como conclusiones particulares se menciona lo siguiente:

- Dentro de las tres asignaciones que contienen las frases clave estándar. El sistema Extractor obtuvo el primer lugar en las frases clave asignadas por el autor con 14.31%, mientras que el sistema Alchemy obtuvo el mayor resultado en las frases clave que asignó el lector con 19.31% y de igual forma en las frases clave combinado con 21.37% (ver capítulo 5).
- El sistema en línea que presentó el mejor desempeño fue Alchemy 21.37% en las frases combinado, mientras que de los sistemas instalables lo fue Extractor con 14.31% para las frases clave del autor.
- El sistema que obtuvo el primer lugar en dos de las características (lector y combinado) Alchemy. El servicio que ofrece en línea como versión gratuita, tiene un uso fácil para el usuario, además de que contiene varias características en el análisis del texto.
- Los sistemas de extracción de frase clave instalables como el caso de Extractor, KEA y Wordstat presentaron buenos resultados pero la instalación y uso para un usuario que tiene pocos conocimientos en informática representan un inconveniente al momento de utilizarlo.
- Se conoció el nivel que presentan los sistemas disponibles en 2016 con los ya evaluados en la tarea 5 del SemEval-2010.

En esta tesis, se cumplieron con los siguientes objetivos:

- Se mostraron sistemas disponibles que realizan la extracción automática de frases claves de textos.
- Se describieron los pasos para realizar la extracción de frases clave en cada uno de los sistemas evaluados.
- Se extrajo una lista de frases clave para cada artículo del SemEval-2010 con cada uno de los sistemas de extracción.
- Se implementó la metodología propuesta en este trabajo para evaluar a los sistemas.
- Se utilizó como en la tarea 5 del SemEval-2010, el programa que calcula las métricas de evaluación para clasificar a los sistemas de acuerdo a su resultado.
- Se presentaron los resultados de los sistemas en las tres asignaciones de las frases clave: autor, lector y combinado.

6.2 Trabajo futuro

En el área de extracción automática de frases clave, el estado del arte presenta trabajo con técnicas diferentes para cumplir con la extracción de frases clave. El trabajo futuro con base en esta investigación es proponer un método para la extracción de frases clave, utilizando características de sistemas del estado del arte y probando con n-gramas sintácticas [Sidorov 13], [Sidorov 13a] y secuencias frecuentes maximales [García 06], [Ledeneva 08], [Ledeneva 14]. Otra dirección es comparar el método de extracción de frases clave propuesto por [Hernández 16] y los sistemas de extracción automática de frases clave sobre otra colección de artículos con asignaciones de frases clave.

Con la comparación de sistemas libres y versiones gratuitas de sistemas comerciales en esta tesis surgen las siguientes preguntas:

¿El uso de un vocabulario en sistemas que implementan esta característica como el caso de KEA influiría en sus resultados?

¿Qué resultado se obtendrían al ingresar el texto completo de los artículos científicos en los sistemas en su versión completa como Extractor y Skyttle, ya que su versión gratuita tiene un límite de caracteres?

¿Cómo sería el desempeño de los sistemas sobre un conjunto de datos diferente?

¿Qué resultados se obtendrían al evaluar sistemas de extracción automática de frases clave comerciales, (es decir cuando se debe de pagar una licencia)?

Referencias

[AlchemyAPI 15] AlchemyAPI Keyword Extraction API;
Url: http://www.alchemyapi.com/products/demo/alchemylanguage; Consultado 23/12/15.

[Baeza 99] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.

[Beliga 14] Beliga, S. (2014). Keyword extraction: a review of methods and approaches.*University of Rijeka, Department of Informatics, Rijeka*.

[Berdend 10] Berend, G., & Farkas, R. (2010). SZTERGAK: Feature engineering for keyphrase extraction. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 186-189). Association for Computational Linguistics.

[Bordea 10] Bordea, G., & Buitelaar, P. (2010). DERIUNLP: A context based approach to automatic keyphrase extraction. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 146-149). Association for Computational Linguistics.

[Bulgarov 15] Bulgarov, F., & Caragea, C. (2015). A Comparison of Supervised Keyphrase Extraction Models. In *Proceedings of the 24th International Conference on World Wide Web Companion* (pp. 13-14). International World Wide Web Conferences Steering Committee.

[Decong Li 10] Li, D., Li, S., Li, W., Wang, W., & Qu, W. (2010). A semi-supervised key phrase extraction approach: learning from title phrases through a document semantic network. In *Proceedings of the ACL 2010 conference short papers*(pp. 296-300). Association for Computational Linguistics.

[El-Beltagy 10] El-Beltagy, S. R., & Rafea, A. (2010). Kp-miner: Participation in semeval-2. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 190-193). Association for Computational Linguistics.

[Extractor] Extractor; Extractor Live Content Demonstration; Url:
http://www.dbitech.com/trials/dbi_TrialDownloads.aspx; Consultado 23/12/15.

[fivefilters 15] fivefilters Term Extraction;

Url: http://fivefilters.org/term-extraction/ Consultado 23/12/15.

[Frank 99] Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G. (1999). Domain-specific keyphrase extraction. In *IJCAI* (Vol. 99, pp. 668-673).

[Garcia 06] García-Hernández, R. A., Martínez-Trinidad, J. F., & Carrasco-Ochoa, J. A. (2006). A new algorithm for fast discovery of maximal sequential patterns in a document collection. In *Computational Linguistics and Intelligent Text Processing* (pp. 514-523). Springer Berlin Heidelberg.

[HaCohen 03] HaCohen-Kerner, Y. (2003). Automatic extraction of keywords from abstracts. In *Knowledge-Based Intelligent Information and Engineering Systems* (pp. 843-849). Springer Berlin Heidelberg.

[Hasan 14] Hasan, K. S., & Ng, V. (2014). Automatic Keyphrase Extraction: A Survey of the State of the Art. In *ACL (1)* (pp. 1262-1273).

[Hasan 10] Hasan, K. S., & Ng, V. (2010). Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 365-373). Association for Computational Linguistics.

[Hulth 03] Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 216-223). Association for Computational Linguistics.

[Jiménez 03] H. Jiménez, David Eduardo Pinto Avendaño; Recuperación de Información; Notas de Academia; 2003.

[KEA 15] Kea; Keyphrase extraction algorithm; Url: http://www.nzdl.org/Kea/ ; Consultado 23/12/15;

[Kim 10] Kim, S. N., Medelyan, O., Kan, M. Y., & Baldwin, T. (2010). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In*Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 21-26). Association for Computational Linguistics.

[Kim 13] Kim, S. N., Medelyan, O., Kan, M. Y., & Baldwin, T. (2013). Automatic keyphrase extraction from scientific articles. *Language resources and evaluation*, *47*(3), 723-742.

[Ledeneva 08] Ledeneva, Y., Gelbukh, A., & García-Hernández, R. A. (2008). Terms derived from frequent sequences for extractive text summarization. In *Computational Linguistics and Intelligent Text Processing* (pp. 593-604). Springer Berlin Heidelberg.

[Ledeneva 14] Ledeneva, Y., García-Hernández, R. A., & Gelbukh, A. (2014). Graph ranking on maximal frequent sequences for single extractive text summarization. In*Computational Linguistics and Intelligent Text Processing* (pp. 466-480). Springer Berlin Heidelberg.

[Lopez 10] Lopez, P., & Romary, L. (2010). HUMB: Automatic key term extraction from scientific articles in GROBID. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 248-251). Association for Computational Linguistics.

[Manning 09] C. Manning, Prabhakar Raghavan, Hinrich Schutze; An Introduction to Retrival Information; Cambridge University Press; Cambridge England; Online Edition; 2009.

[Mihalcea 04] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. Association for Computational Linguistics.

[Medelyan 06] Medelyan, O., & Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 296-297). ACM.

[Medelyan 09] Medelyan, O., Frank, E., & Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3* (pp. 1318-1327). Association for Computational Linguistics.

[Medelyan 09a] Medelyan, O. (2009). *Human-competitive automatic topic indexing* (Doctoral dissertation, The University of Waikato).

[Nguyen 07] Nguyen, T. D., & Kan, M. Y. (2007). Keyphrase extraction in scientific publications. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers* (pp. 317-326). Springer Berlin Heidelberg.

[Nguyen 10] Nguyen, T. D., & Luong, M. T. (2010). WINGNUS: Keyphrase extraction utilizing document logical structure. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 166-169). Association for Computational Linguistics.

[Park 10] Park, J., Lee, J. G., & Daille, B. (2010). UNPMC: Naive approach to extract keyphrases from scientific articles. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 178-181). Association for Computational Linguistics.

[Pianta 10] Pianta, E., & Tonelli, S. (2010). KX: A flexible system for keyphrase extraction. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 170-173). Association for Computational Linguistics.

[SemEval 10] SemEval-2; Evaluation Exercises on Semantic Evaluation - ACL SigLex event *5th workshop on semantic evaluation. ACL in Uppsala*; URL: http://semeval2.fbk.eu/semeval2.php; Consultado 20/11/15.

[SemEval 12] SemEval-2012: Semantic Evaluation Exercises; *First Joint Conference on Lexical and Computational Semantics; NAACL, Montreal, Canada*; URL: https://www.cs.york.ac.uk/semeval-2012/; Consultado 20/11/2015;

[SemEval 13] SemEval-2013: Semantic Evaluation Exercises; *International Workshop on Semantic Evaluation; NAACL North American Association of Computational Linguistics; Atlanta, Georgia USA*; URL: https://www.cs.york.ac.uk/semeval-2013/; Consultado 20/11/15.

[SemEval 14] SemEval-2014: Semantic Evaluation Exercises; *International Workshop on Semantic Evaluation; 25th International Conference on Computational Linguistics and *SEM 2014, Second Joint Conference on Lexical and Computational Semantics; Dublin, Ireland*; URL: http://alt.qcri.org/semeval2014/; Consultado 20/11/15.

[SemEval 15] SemEval-2015: Semantic Evaluation Exercises*; International Workshop on Semantic Evaluation; NAACL-HLT 2015, 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies and *SEM 2015, Third Joint Conference on Lexical and Computational Semantics, Denver USA;* URL: http://alt.qcri.org/semeval2015/; Consultado 20/11/15.

[SemEval 16] SemEval-2016: Semantic Evaluation Exercises; *International Workshop on Semantic Evaluation;* URL: http://alt.qcri.org/semeval2016/; Consultado 22/11/15.

[SemEval-Wikipedia 16] SemEval: Semantic Evaluation URL: https://en.wikipedia.org/wiki/SemEval; Consultado 23/11/15.

[Senseval 15] Senseval-2015*; Evaluation Exercises for the Semantic Analysis of Text Organized by ACL-SIGLEX;* University of North Texas; URL: http://www.senseval.org/; Consultado 20/11/2015.

[Siddiqui 15] Siddiqi, S., & Sharan, A. (2015). Keyword and Keyphrase Extraction Techniques: A Literature Review. *International Journal of Computer Applications*, *109*(2).

[Sidorov 13] Sidorov, G. (2013). Syntactic dependency based n-grams in rule based automatic English as second language grammar correction. *International Journal of Computational Linguistics and Applications*, *4*(2), 169-188.

[Sidorov 13a] Sidorov, G. (2013). N-gramas sintácticos no-continuos. *Polibits*, (48), 69-78.

[Skyttle 15] Skyttle; Skyttle.api; Url: http://www.skyttle.com/demoin; Consultado 23/12/15.

[Termine15] Termine; Termine web demonstration; URL:http://www.nactem.ac.uk/software/termine; Consultado 23/12/15.

[TexLexan 15] TexLexAn; TexLexAn Analyze, Classify and Summarize any text; Url: http://texlexan.sourceforge.net/; Consultado 23/12/15.

[Turney 99] Turney, P. (1999). Learning to extract keyphrases from text.

[Turney 2000] Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval*, *2*(4), 303-336.

[Turney 03] Turney, P. (2003). Coherent keyphrase extraction via web mining.

[Witten 99] Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA: Practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries* (pp. 254-255). ACM.

[Wordstat 15] Wordstat 7; Software de análisis de contenido y minería de texto; Url:http://provalisresearch.com/es/products/software-de-analisis-de-contenido/; Consultado 23/12/15.

[Hernández 16] Hernández C. Y(2016) Extracción de frases clave usando patrones léxicos en artículos científicos (Tesis de Maestria).UAEM Tianguistenco, Edo de México.

[Zahang 08] Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, *4*(3), 1169-1180.

Anexo 1. Resultados obtenidos en el Top 5

1.1 Resultados obtenidos en el Top 5

Resultados en las frases clave asignadas por el autor

En el top 5 para las frases clave asignadas por el autor, KEA se posiciona como el sistema con mayor resultado con Precisión 15.2%, Recuerdo 19.64% y F-medida de 17.14%. En la tabla 12 se muestran los resultados y se marca el valor más alto (ver tabla 12).

		Top 5		
Sistema	Rank	P	R	F
Kea	1	**15,2**	**19,64**	**17,14**
Extractor	2	14,8	19,12	16,68
Alchemy	3	14,6	18,86	16,46
Wordstat	4	14,4	18,6	16,23
Genia	5	14	18,09	15,78
Tree tagger	6	13,4	17,31	15,11
Skyttle	7	8,2	10,59	9,24
Fivefilters	8	6	7,75	6,76
Texlexan	9	5,8	7,49	6,54
Translatedlab	10	5,6	7,24	6,32

Tabla 12. Resultados de los sistemas en las frases clave asignadas por el autor en el top 5.

En la figura 18, se muestran los resultados de manera gráfica del top 5, para las frases clave del autor clasificados por F-medida y se señala el mayor resultado.

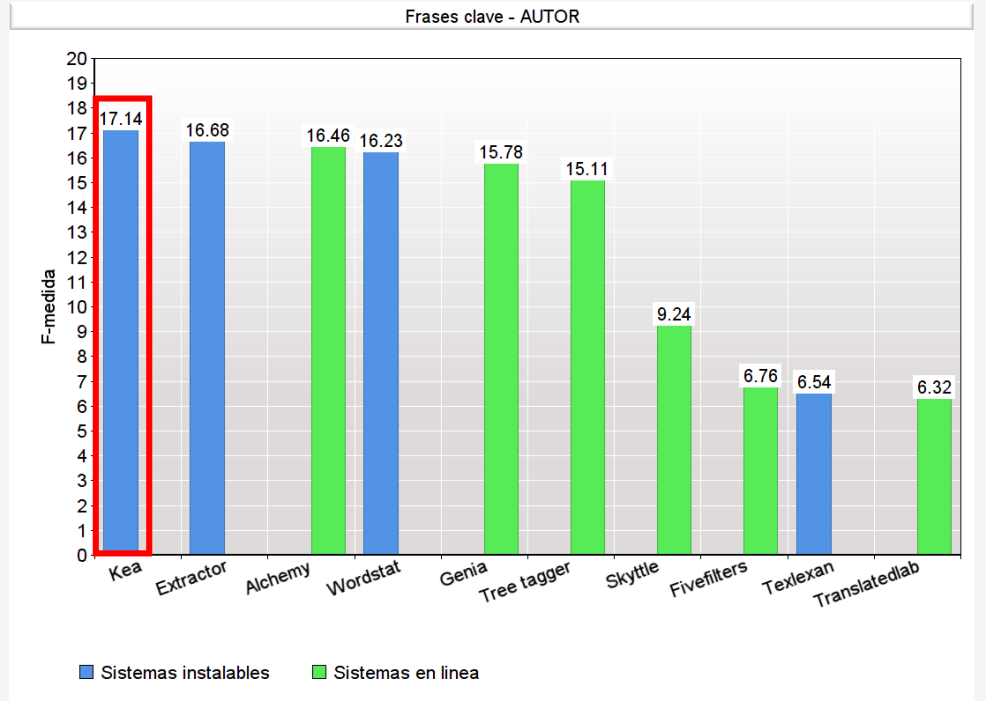


Figura 18. Desempeño de los sistemas sobre las frases clave asignadas por el autor en el top 5.

Resultados en las frases claves asignadas por el lector

En el top 5 para las frases clave asignadas por el lector, Wordstat se ubica como el sistema con mayor porcentaje con Precisión 26.2%, Recuerdo 10.88% y F-medida de 15.38% (ver tabla 13).

Top 5

Sistema	Rank	P	R	F
Wordstat	1	**26,2**	**10,88**	**15,38**
Alchemy	2	25,2	10,47	14,79
Tree tagger	3	25,2	10,47	14,79
Genia	4	24,4	10,13	14,32
Kea	5	20	8,31	11,74
Extractor	6	19	7,89	11,15
Fivefilters	7	13,2	5,48	7,74
Skyttle	8	12,4	5,15	7,28
Translatedlab	9	11,2	4,65	6,57
Texlexan	10	10	4,15	5,87

Tabla 13. Resultados de los sistemas en las frases clave asignadas por el lector en el top 5.

En la figura 19, se muestran los resultados de manera gráfica del top 5, para las frases clave del lector clasificados por F-medida y se señala el mayor resultado.

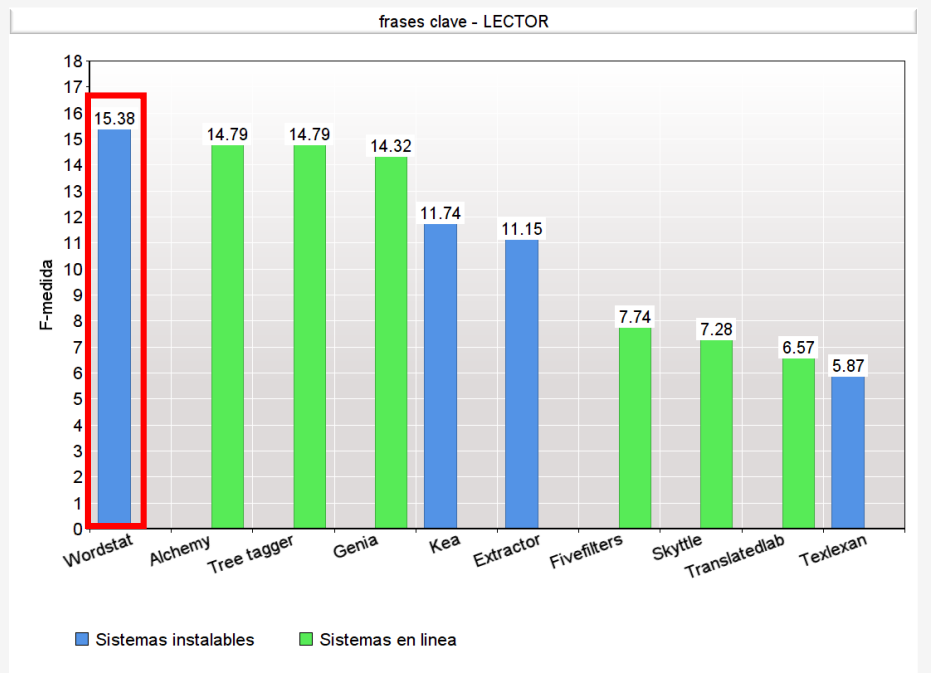


Figura 19. Desempeño de los sistemas sobre las frases clave asignadas por el lector en el top 5.

Resultados en las frases clave combinado

En el top 5 para las frases clave combinado, Wordstat se ubica como el sistema con mayor resultado con Precisión 32.2%, Recuerdo 10.98% y F-medida de 16.38% (ver tabla 14).

Top 5

Sistema	Rank	P	R	F
Wordstat	1	**32,2**	**10,98**	**16,38**
Alchemy	2	31,2	10,64	15,87
Tree tagger	3	30	10,23	15,26
Genia	4	29,6	10,1	15,06
Kea	5	27,8	9,48	14,14
Extractor	6	27	9,21	13,73
Fivefilters	7	16,4	5,59	8,34
Skyttle	8	16,2	5,53	8,25
Translatedlab	9	14	4,77	7,12
Texlexan	10	13,4	4,57	6,82

Tabla 14. Resultados de los sistemas en las frases clave combinado en el top5.

En la figura 20, se muestran los resultados de manera gráfica del top 5, para las frases clave combinado clasificados por F-medida y se señala el mayor resultado.

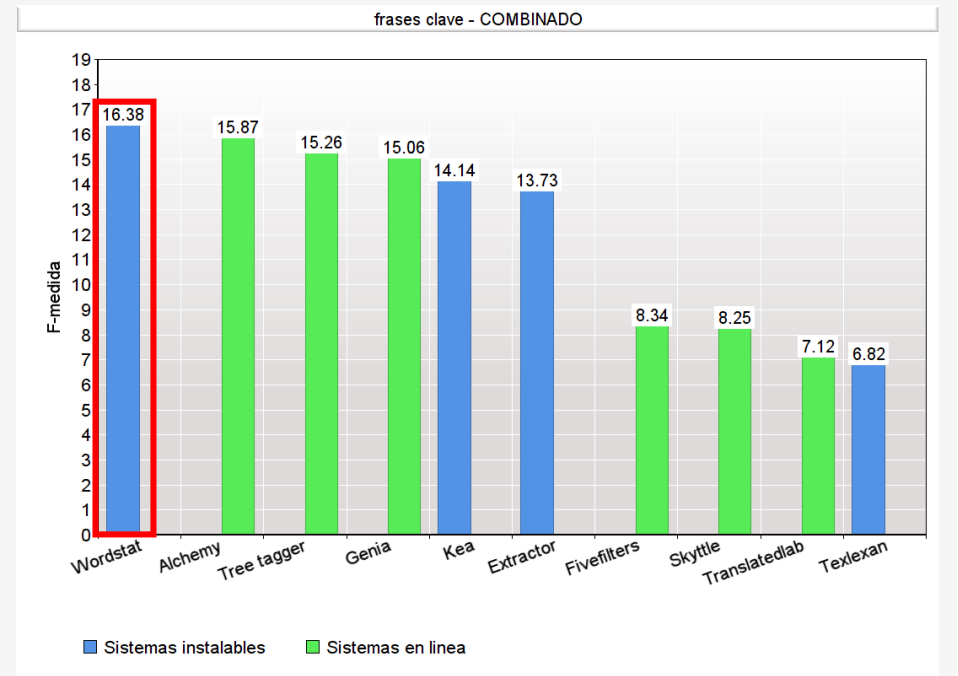


Figura 20. Desempeño de los sistemas sobre las frases clave combinado en el top 5.

Anexo 2. Resultados obtenidos en el Top 10

1.2 Resultados obtenidos en el Top 10

Resultados en las frases clave asignadas por el autor

En el top 10 para las frases clave asignadas por el autor, KEA obtiene el resultado más alto con Precisión 11.2%, Recuerdo 28.94% y F-medida 16.15% (ver tabla 15).

		Top 10		
Sistema	Rank	P	R	F
Kea	1	**11,2**	**28,94**	**16,15**
Extractor	2	10,4	26,87	15
Alchemy	3	10,2	26,36	14,71
Wordstat	4	10	25,84	14,42
Genia	5	9,5	24,55	13,7
Tree tagger	6	9,1	23,51	13,12
Skyttle	7	6,4	16,54	9,23
Fivefilters	8	5,5	14,21	7,93
Texlexan	9	4,8	12,4	6,92
Translatedlab	10	4	10,34	5,77

Tabla 15. Resultados de los sistemas en las frases clave asignadas por el autor en el top 10.

En la figura 21, se muestran los resultados de manera gráfica del top 10, para las frases clave del autor clasificados por F-medida y se señala el mayor resultado.

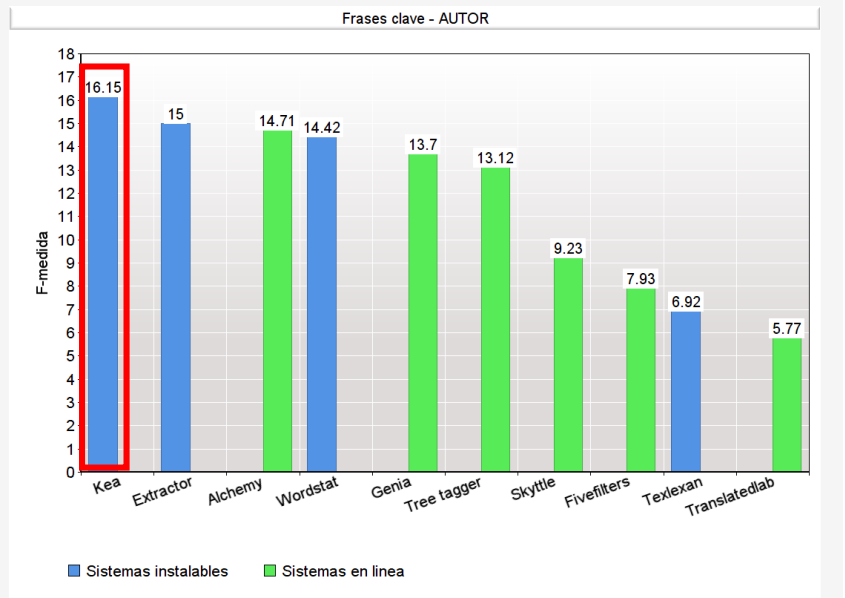


Figura 21. Desempeño de los sistemas sobre las frases clave asignadas por el autor en el top 10.

Resultados en las frases clave asignadas por lector

En el top 10 para las frases clave asignadas por el lector, Wordstat y Alchemy obtiene los valores más altos con Precisión 20.1%, Recuerdo 16.69% y F-medida de 18.24% (ver tabla 16).

Top10

Sistema	Rank	P	R	F
Alchemy	1	**20,1**	**16,69**	**18,24**
Wordstat	2	**20,1**	**16,69**	**18,24**
Genia	3	18,9	15,7	17,15
Tree tagger	4	18,5	15,37	16,79
Extractor	5	17,2	14,29	15,61
Kea	6	17	14,12	15,43
Fivefilters	7	11,7	9,72	10,62
Skyttle	8	10,8	8,97	9,8
Translatedlab	9	9,1	7,56	8,26
Texlexan	10	7,6	6,31	6,9

Tabla 16. Resultados de los sistemas en las frases clave asignadas por el lector en el top 10.

En la figura 22, se muestran los resultados de manera gráfica del top 10, para las frases clave del lector clasificados por F-medida y se señala el mayor resultado.

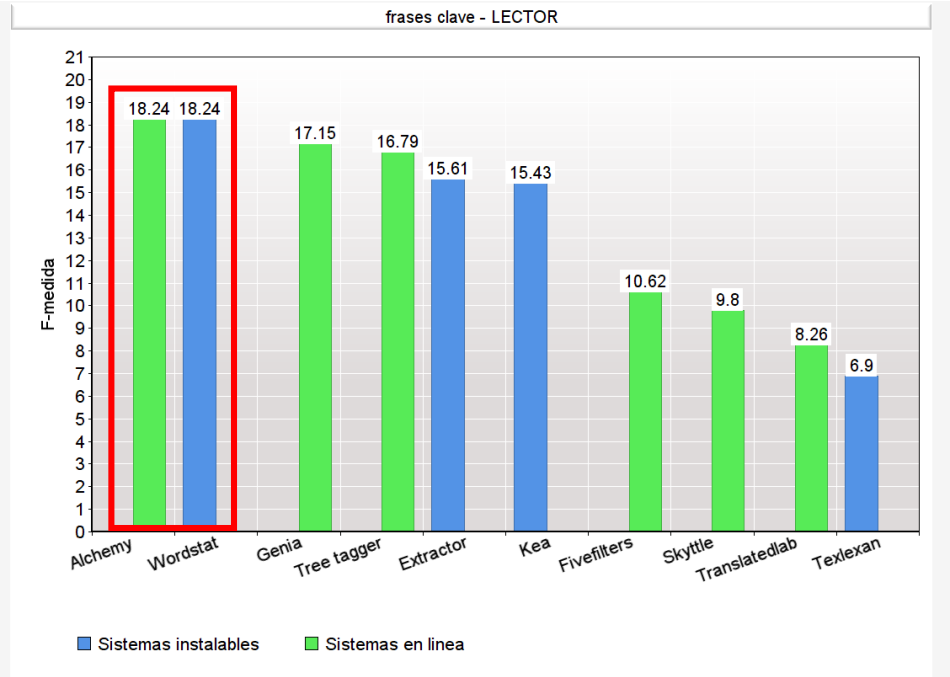


Figura 22. Desempeño de los sistemas sobre las frases clave asignadas por el lector en el top 10.

Resultados en las frases clave combinado

En el top 10 para las frases combinado. Wordstat se ubica en el primer lugar con Precisión 24.5%, Recuerdo 16.71 y F-medida de 19.87% (ver tabla 17).

Top10

Sistema	Rank	P	R	F
Wordstat	1	**24,5**	**16,71**	**19,87**
Alchemy	2	24,4	16,64	19,79
Kea	3	22,8	15,55	18,49
Genia	4	23	15,69	18,65
Tree tagger	5	22,3	15,21	18,08
Extractor	6	22,1	15,08	17,93
Fivefilters	7	14,4	9,82	11,68
Skyttle	8	13,9	9,48	11,27
Translatedlab	9	10,8	7,37	8,76
Texlexan	10	10,7	7,3	8,68

Tabla 17. Resultados de los sistemas en las frases clave combinado en el top 10.

En la figura 23, se muestran los resultados de manera gráfica del top 10, para las frases clave combinado clasificados por F-medida y se señala el mayor resultado.

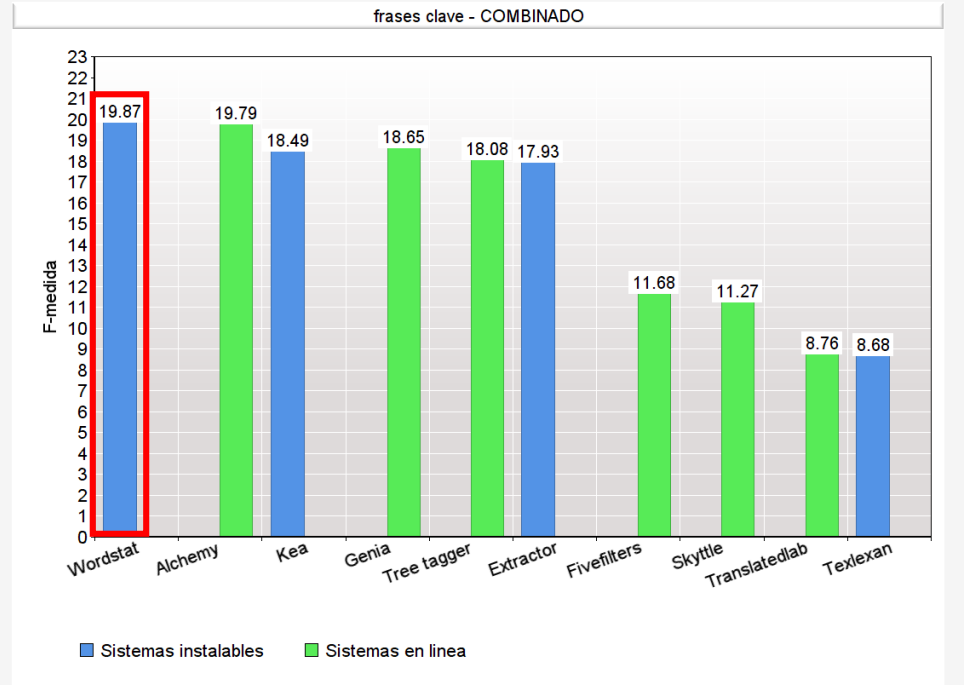


Figura 23. Desempeño de los sistemas sobre las frases clave combinado en el top 10.

Anexo 3. Organización del Corpus SemEval2010

La organización del corpus SemEval2010 se encuentran de la siguiente manera:

1.- Carpeta SemEval2010: Carpeta raíz la cual contiene a las carpetas.

- test
- test_answer
- train
- trial

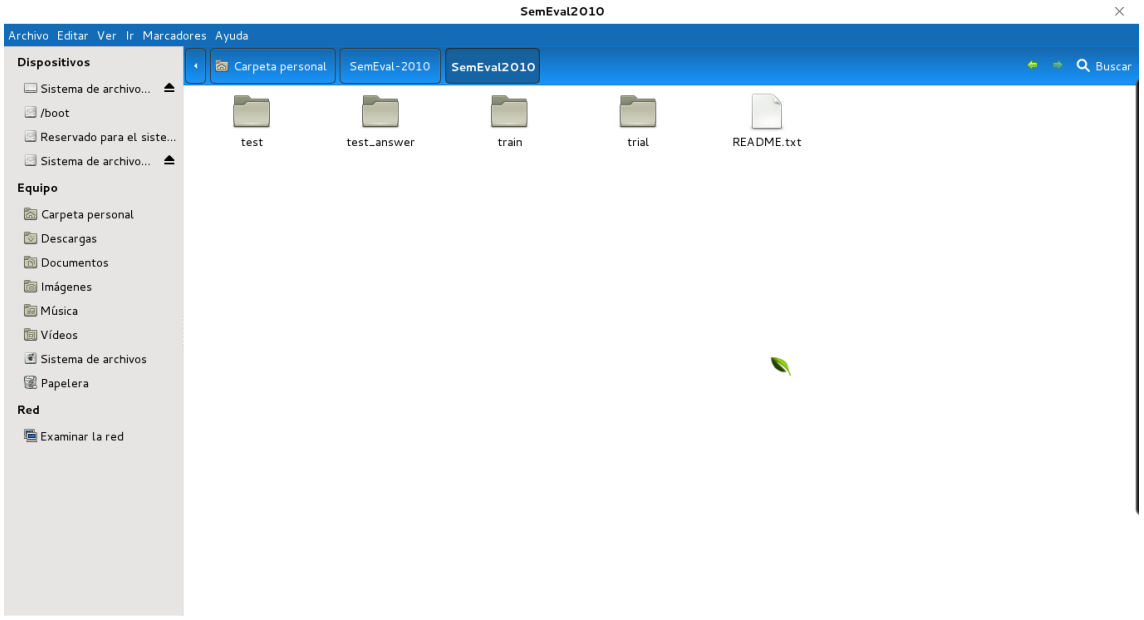


Figura 24. Archivos que integran a la carpeta raíz SemEval2010.

2.- test: carpeta que contiene 100 artículos científicos distribuidos entre cuatro áreas de investigación.

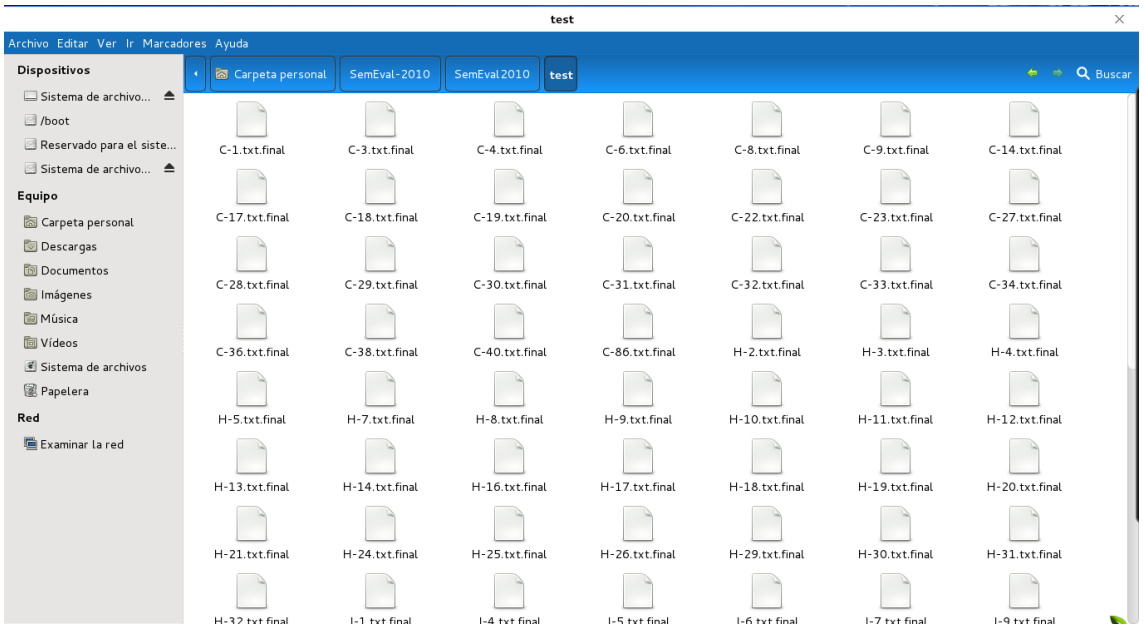


Figura 25. Contenido carpeta "test".

3.- test_answer: carpeta que contiene 3 archivos en donde se encuentran las frases clave propuestas por el autor, lector y combinadas.

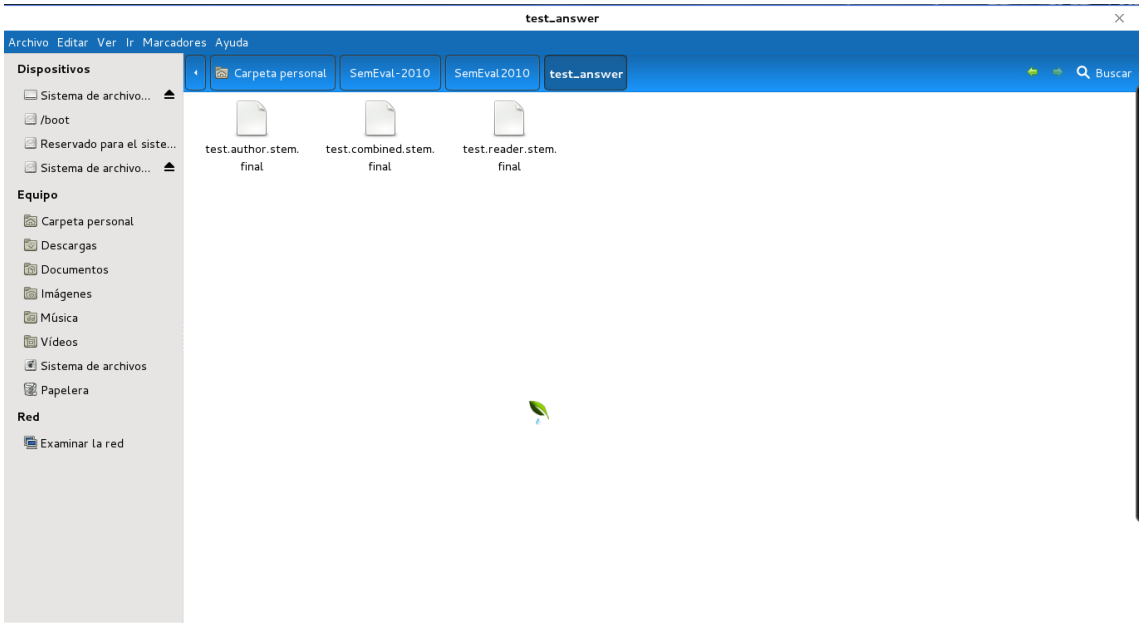


Figura 26. Contenido carpeta "test answer".

4.- train: Carpeta que contiene 144 archivos de entrenamiento.

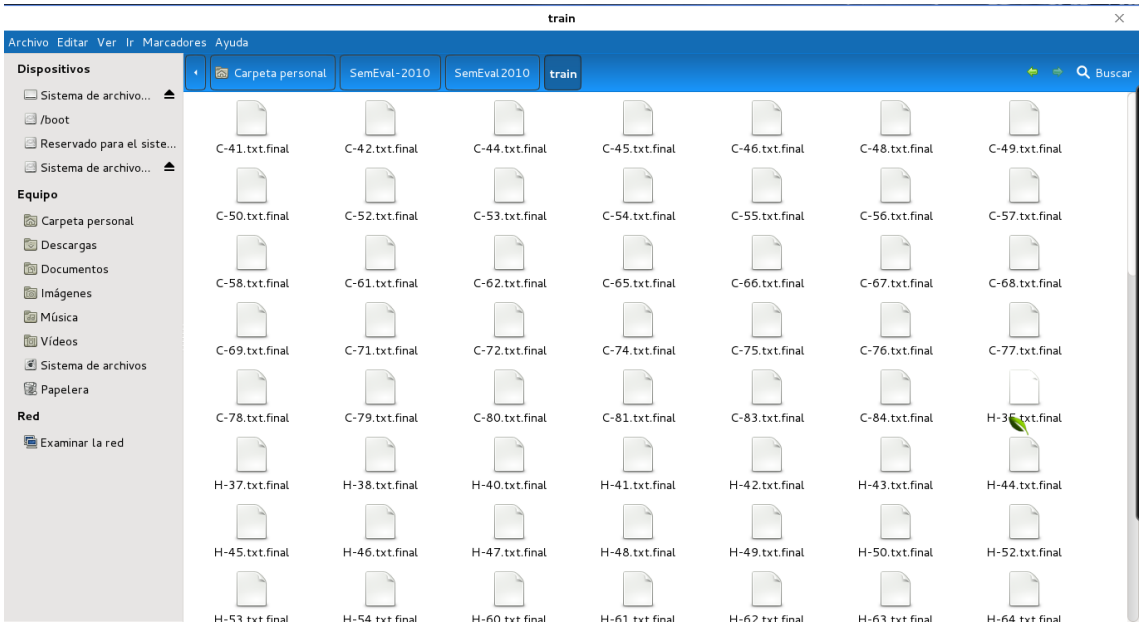


Figura 27. Contenido carpeta "train".

5.- trial: Carpeta la cual contiene 40 archivos de ensayo

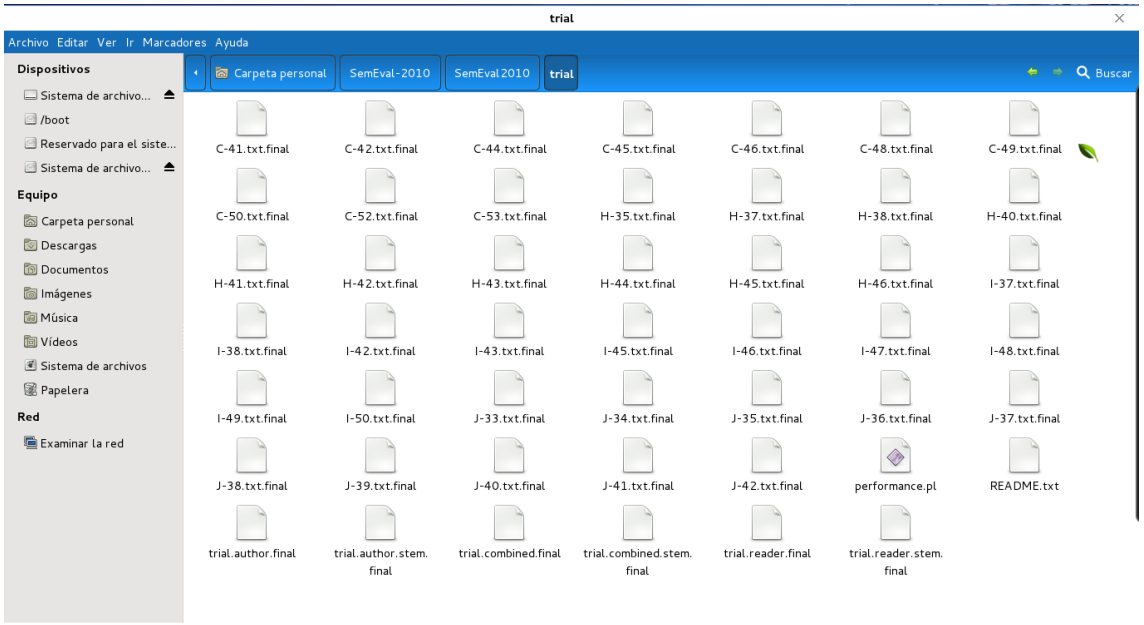


Figura 28. Contenido carpeta "trial".

Anexo 4. Proceso de extracción automática de frases clave con los sistemas a evaluar

Anexo 4.1 Sistemas de extracción automática de frases clave en línea

AlchemyIBM (Comercial)

AlchemyAPI ofrece 12 funciones de la API como parte de su servicio de análisis de texto, cada una de las cuales utiliza sofisticadas técnicas de procesamiento de lenguaje natural para analizar su contenido y añadir alto nivel de información semántica [AlchemyAPI 15].

1.- Abrir cualquier navegador de internet e ingresar al siguiente enlace

http://www.alchemyapi.com/products/demo/alchemylanguage

El enlace dirige directamente al Demo de AlchemyAPI.

2.-En la página que contiene el demo se debe realizar lo siguiente:

1.- Desplazarse hacia la parte derecha de la tabla y dar clic en la opción "*Enter you own text*"
2.- Pegar el texto que se desee analizar, en nuestro caso cada uno de los 100 artículos científicos que contiene la carpeta "test" del corpus SemEval-2010.

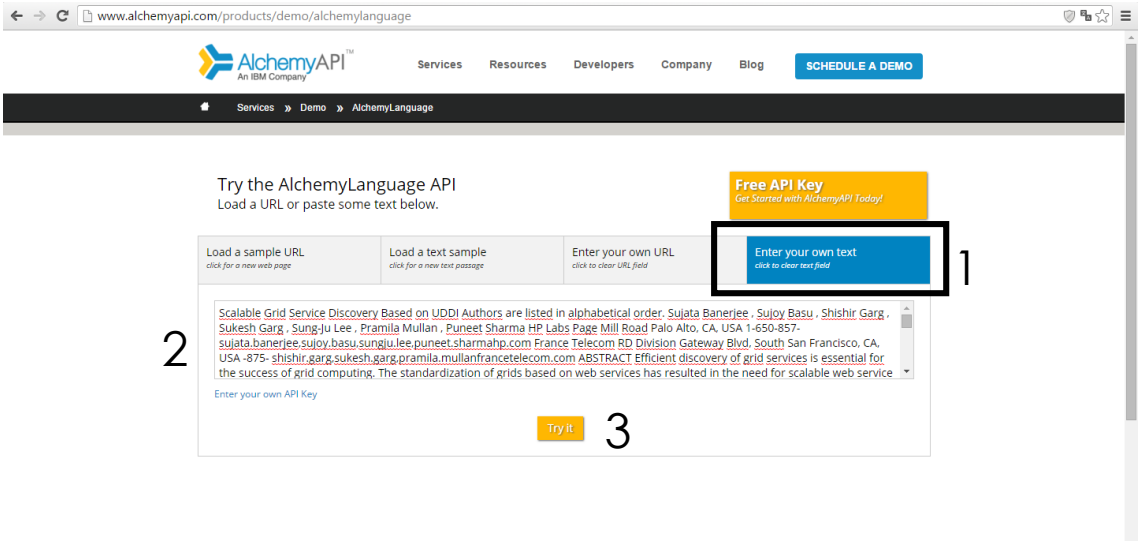


Figura 29. Demo AlchemyAPI.

4.- Al dar clic en el botón "Try" empezará la extracción de frases clave y de las otras características que ofrece AlchemyAPI. Para ver las frases clave dar clic en la pestaña de lado izquierdo "*Keywords*" (Ver figura 30).

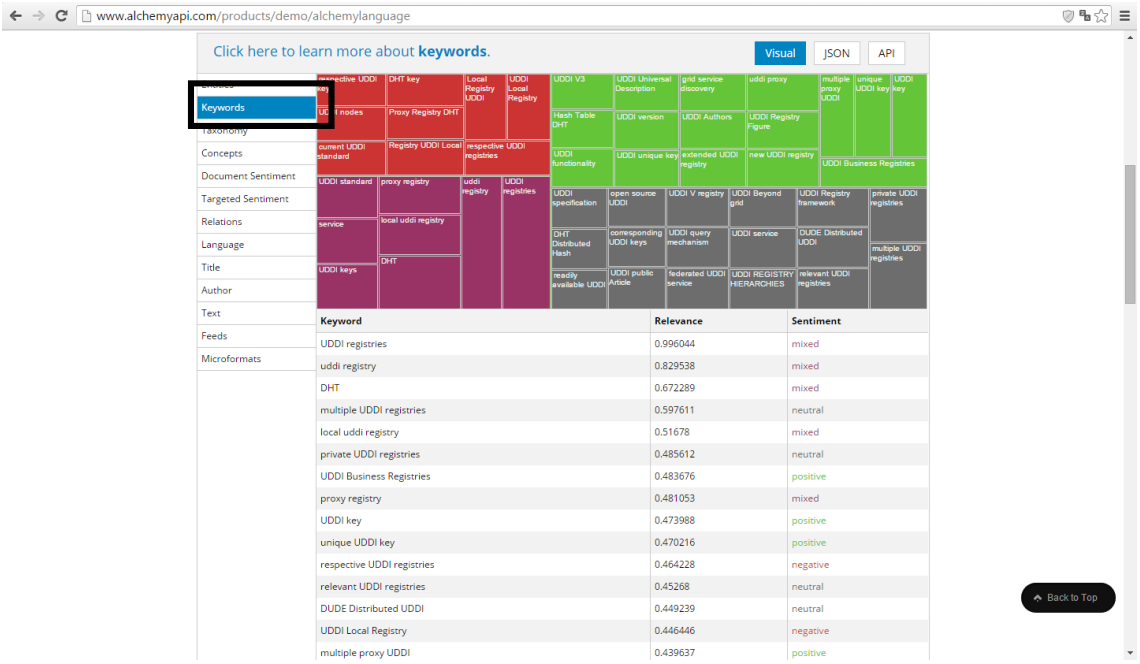


Figura 30. Salida de resultados AlchemyAPI.

fivefilters.org (Libre)

Extracción de términos de FiveFilters.org es un proyecto de software libre para ayudar a extraer términos (por ejemplo, para su uso como etiquetas) a través de un servicio web. Dado algún texto devolverá una lista de términos con el más relevante primero [fivefilters.org 15] (La diferencia entre frases clave y términos se describen el capítulo 2).

1.- Abrir cualquier navegador de internet e ingresar al siguiente enlace:

http://fivefilters.org/term-extraction/

El enlace dirige directamente al demo de fivefilters.

2.- Para realizar el proceso de extracción:

1.-Pegar el texto a analizar.

2.-Ajustar los parámetros de extracción.

Los parámetros que pide fivefilters para realizar la extracción de términos son los siguientes:
-Máximo de ítems: tamaño de la lista de resultados (por ejemplo una lista de 5 términos).
-Output: formato de salida en la que se mostrarán los resultados (HTML.JSON, XML, TEXT, PHP).
-Max words per term: número de palabras que puede contener un término (por ejemplo, 3 palabras por término).
-Terms in lowercase: devuelve los resultados en letra minúscula.

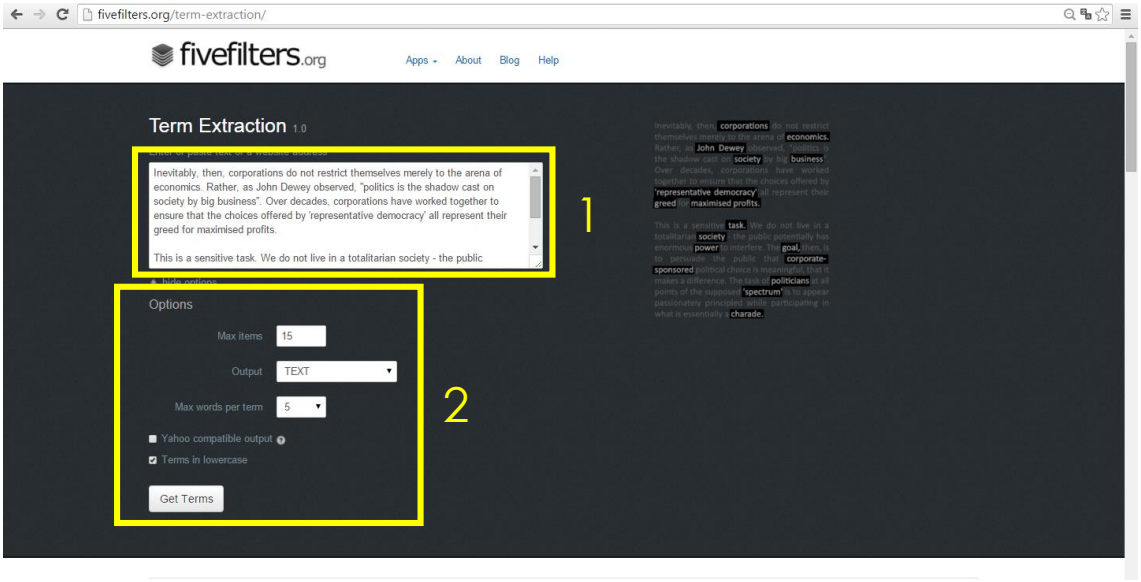


Figura 31. Demo fivefilters.

3.-Clic en el botón "*Get Terms*" y se mostrarán los resultados dependiendo del formato de salida que se haya indicado. En la figura 32, se muestran los resultados en formato texto plano.

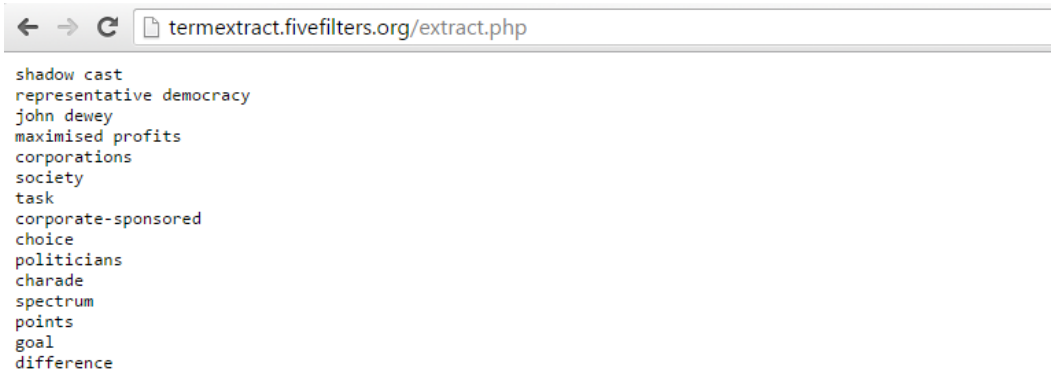


Figura 32. Salida de resultados fivefilters.

Skyttle (Comercial)

Skyttle es un (SaaS) sistema que proporciona análisis de texto de servicios para extraer patrones interesantes de texto y almacenarlos en un formato estructurado para el análisis de datos en profundidad [Skyttle 15].

1.-Abrir cualquier navegador de internet e ingresar al siguiente enlace:

http://www.skyttle.com/demoin

El enlace dirige directamente al Demo de Skyttle.

2.-Para realizar la extracción se hace lo siguiente:

1.-Pegar el texto en el área de texto.

2.-Dar clic en el botón "Submit".

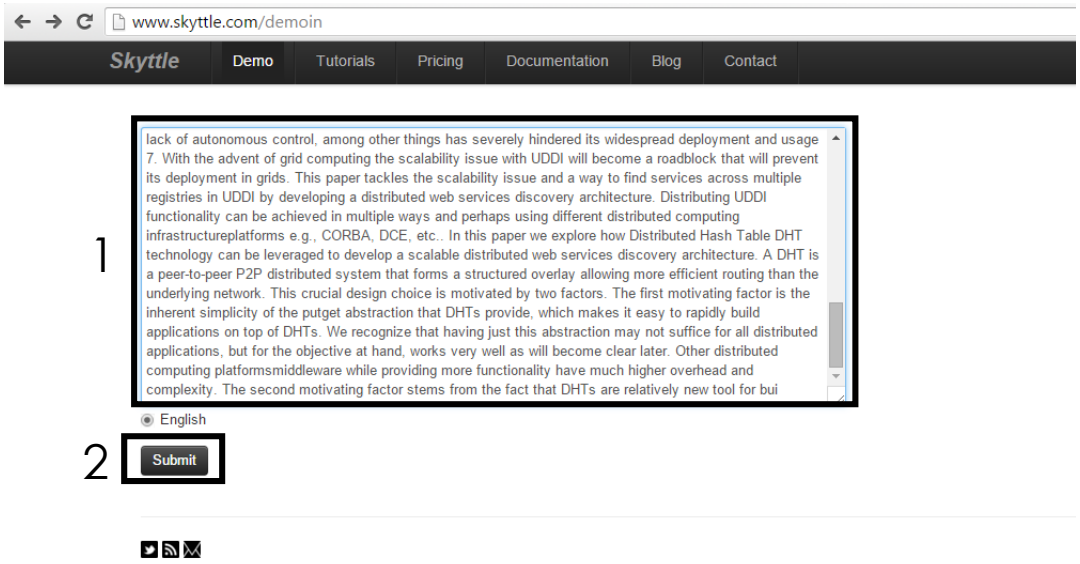


Figura 33. Demo Skyttle.

3.- Después de dar clic al botón "*Submit*" se muestra la página con los resultados propuestos por Skyttle, en la parte derecha de la página se muestra las frases clave que son propuestas.

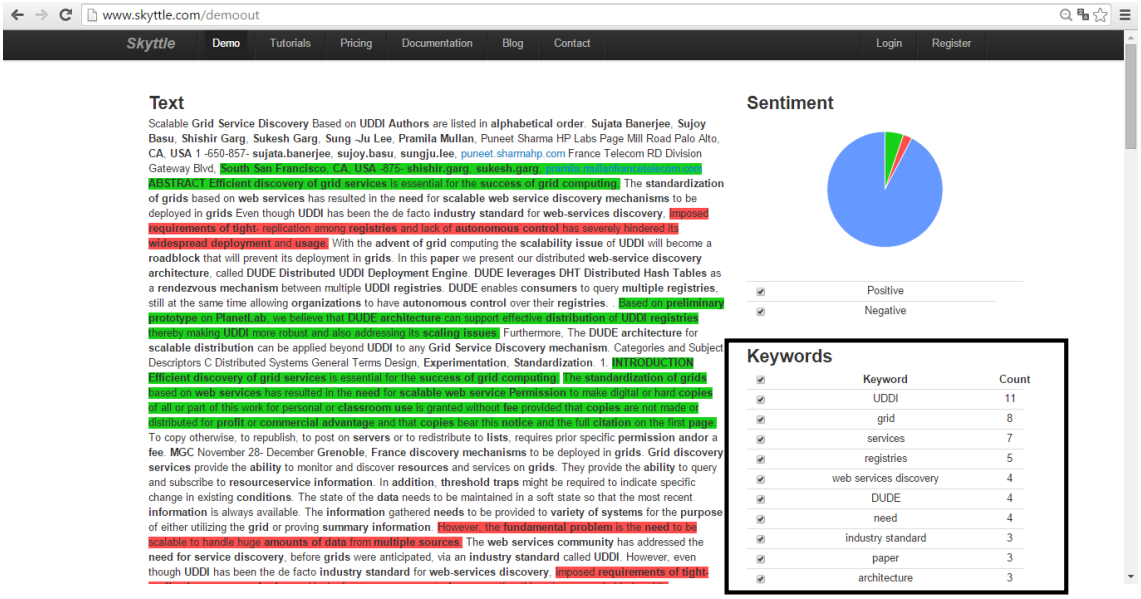


Figura 34. Salida de resultados Skyttle.

Translated Labs

Translated labs ofrece servicios lingüísticos, uno de ellos es la extracción de la terminología de un texto en la cual se trata de identificar términos que mejor describan a un documento dentro de un dominio.

1.- Abrir cualquier navegador de internet e ingresar al siguiente enlace:

http://labs.translated.net/terminology-extraction/

2.- Al dar clic en el enlace se muestra la página para la extracción de términos, para realizar la extracción se hace lo siguiente:

1.-Pegar el texto a extraer en el cuadro de texto.

2.-Eligir el idioma en el que se encuentra el texto.

3.-Clic en el botón "Terminlogy extraction".

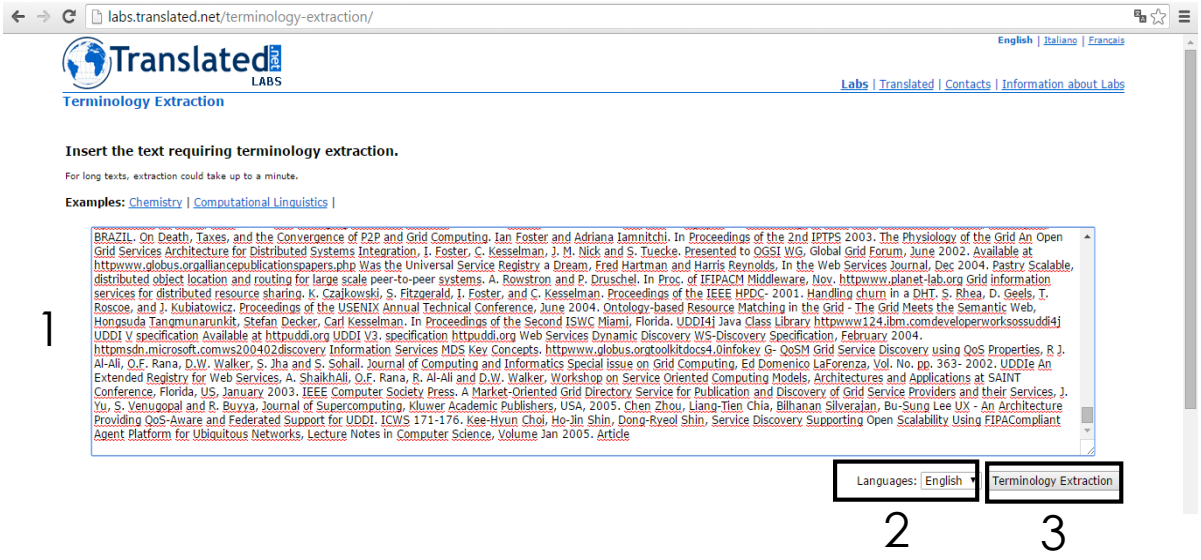


Figura 35. Demo Translated Labs.

3.- Después de dar clic en el botón "*Terminology Extraction*" se muestran los resultados que propone Translated labs.

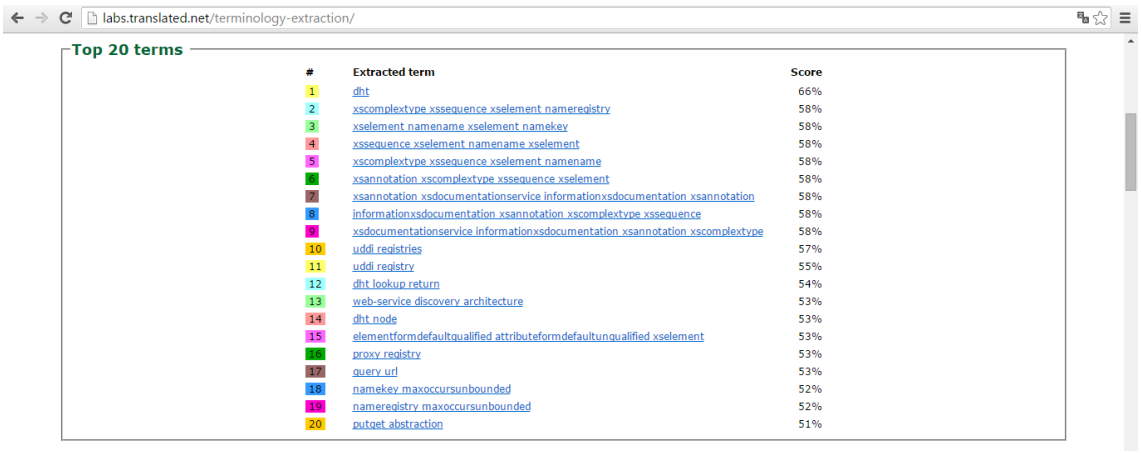


Figura 36. Salida de resultados Translated Labs.

TerMine (Genia & Tree tragger)

TerMine es un sistema de gestión de términos que identifica frases clave en el texto [Termine 15].

1.- Abrir cualquier navegador de internet e ingresar al siguiente enlace:

http://www.nactem.ac.uk/software/termine/

Al entrar al enlace se muestra la página de TerMine, la cual contiene a los dos extractores de términos que son Genia Tagger y Tree Tagger.

2.- Para realizar la extracción de términos se ubica la sección "*Web Demostration*" y se realiza lo siguiente:

1.-Pegar el texto en la área de texto (en nuestro caso los artículos de SemEval2010).

2.-Seleccionar el extractor de términos.

Genia tagger: textos de biomédica

Tree tagger: textos genéricos

(Para este ejemplo se usa Genia tagger)

3.-Dar clic al botón "Analyze".

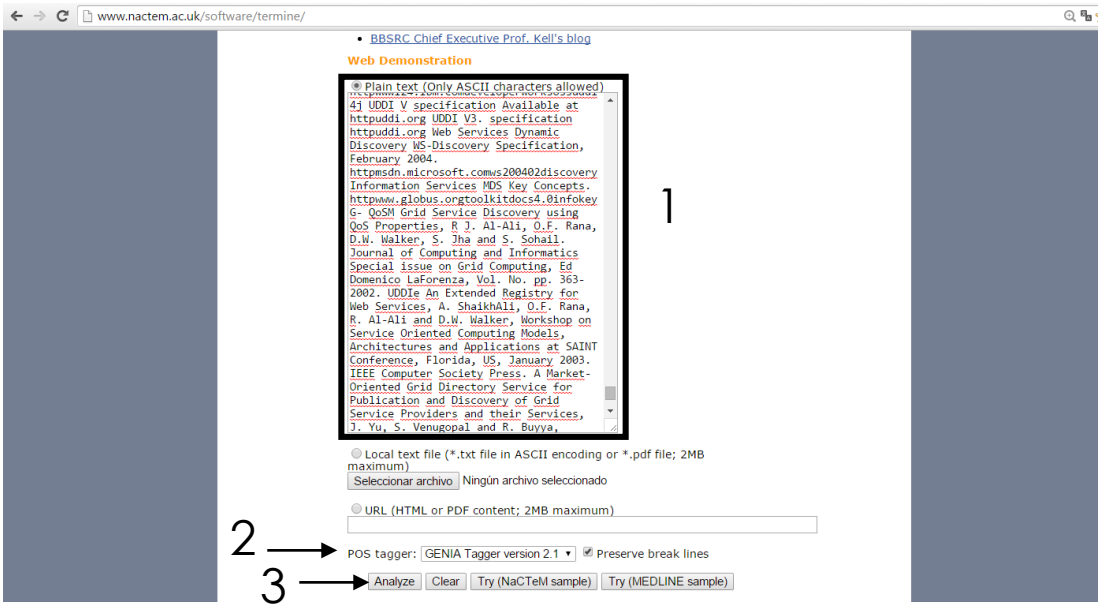
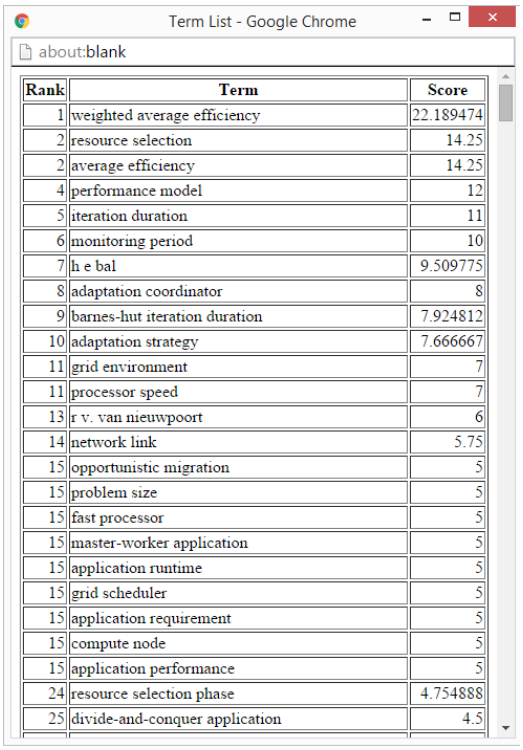


Figura 37. Demo TerMine.

3.-Despues de dar clic en el botón "*Analyze*", se muestra el texto y aparece resaltado las palabras candidatas, para ver los resultados en la parte superior izquierda se encuentran 2 opciones: "*table*" que muestra los resultados en una tabla y "*text*" que muestra los resultados en texto plano, al dar clic en alguno de los dos se muestra una ventana emergente con los resultados (en este ejemplo se elige la salida de resultados en tabla).



Rank	Term	Score
1	weighted average efficiency	22.189474
2	resource selection	14.25
2	average efficiency	14.25
4	performance model	12
5	iteration duration	11
6	monitoring period	10
7	h e bal	9.509775
8	adaptation coordinator	8
9	barnes-hut iteration duration	7.924812
10	adaptation strategy	7.666667
11	grid environment	7
11	processor speed	7
13	r v. van nieuwpoort	6
14	network link	5.75
15	opportunistic migration	5
15	problem size	5
15	fast processor	5
15	master-worker application	5
15	application runtime	5
15	grid scheduler	5
15	application requirement	5
15	compute node	5
15	application performance	5
24	resource selection phase	4.754888
25	divide-and-conquer application	4.5

Figura 38. Salida de resultados en Termine con Genia tagger.

Anexo 4.2 Sistemas de extracción automática de frases clave instalables

A continuación se describe la instalación de los sistemas libres y comerciales que pueden ser instalados en el equipo

Extractor (Comercial)

Extractor es una tecnología de resumen de contenido que automáticamente, sin intervención humana sesgada, analiza el contenido - noticias, información no estructurada, documentos, correo electrónico, páginas web, contextualmente preciso en palabras clave y frase clave resúmenes [Extractor 15].

1.-Abrir cualquier navegador de internet e ingresar al siguiente enlace:

http://www.extractor.com/

Al entrar al enlace se presenta la página principal, Extractor ofrece un demo en línea y uno demo instalable de su software (en este trabajo se descargó el instalable). Para poder descargarlo se da clic en "*Sample Application*" ubicado en la parte izquierda de la página.

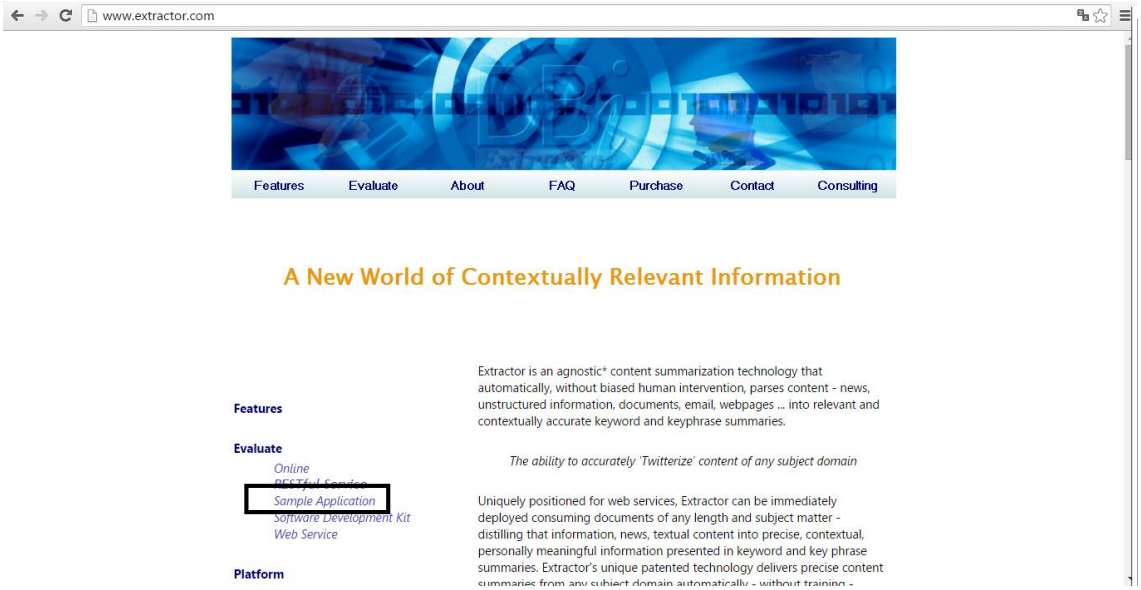


Figura 39. Página principal de Extractor.

2.-Al dar clic en "*Sample Application*" nos enviará a la página de descarga, para descargar el demo de Extractor se debe ubicar en la tabla de los demos disponibles y dar clic en el demo de Extractor 7.2 (ver figura 40).

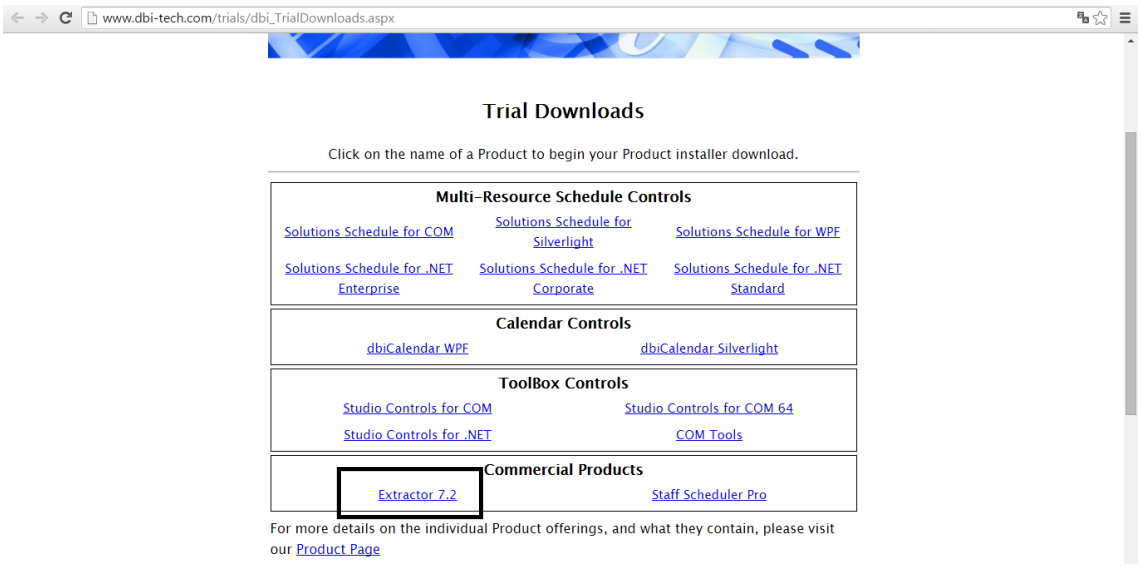


Figura 40. Página de Descarga de Extractor.

3.- Terminada la descarga lo último es la instalación (en este trabajo se ejecutó en la plataforma de Windows 8) para poder empezar a usar Extractor.

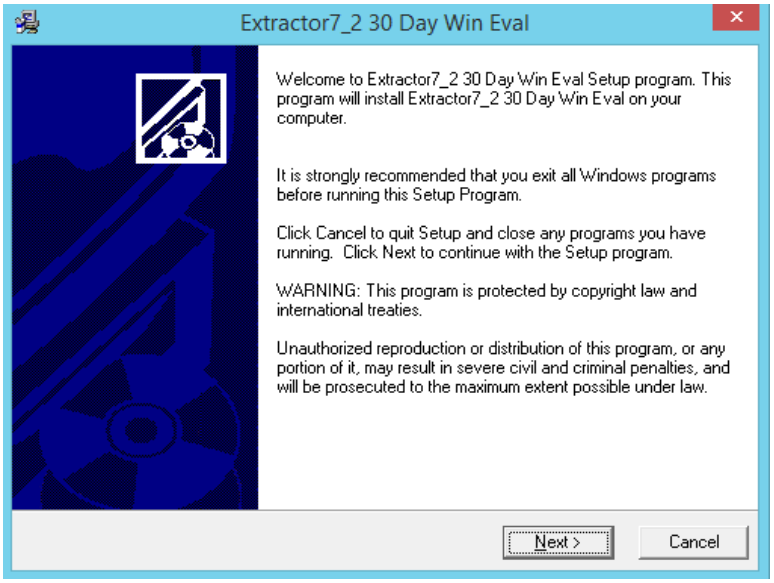


Figura 41. Setup Extractor.

4.- Después de descargar el demo y haber realizado la instalación para poder realizar la extracción de frases clave:

1.- Se ejecuta Extractor 7.2.

2.-Se pega el texto que se desee analizar (en nuestro caso los artículos científicos).

3.-Se indica el número de frases a extraer.

4.- Se ingresan las stopwords.

5.-Clic en el botón "Extract".

6.-Resultados de las frases y palabras clave propuestas por Extractor.

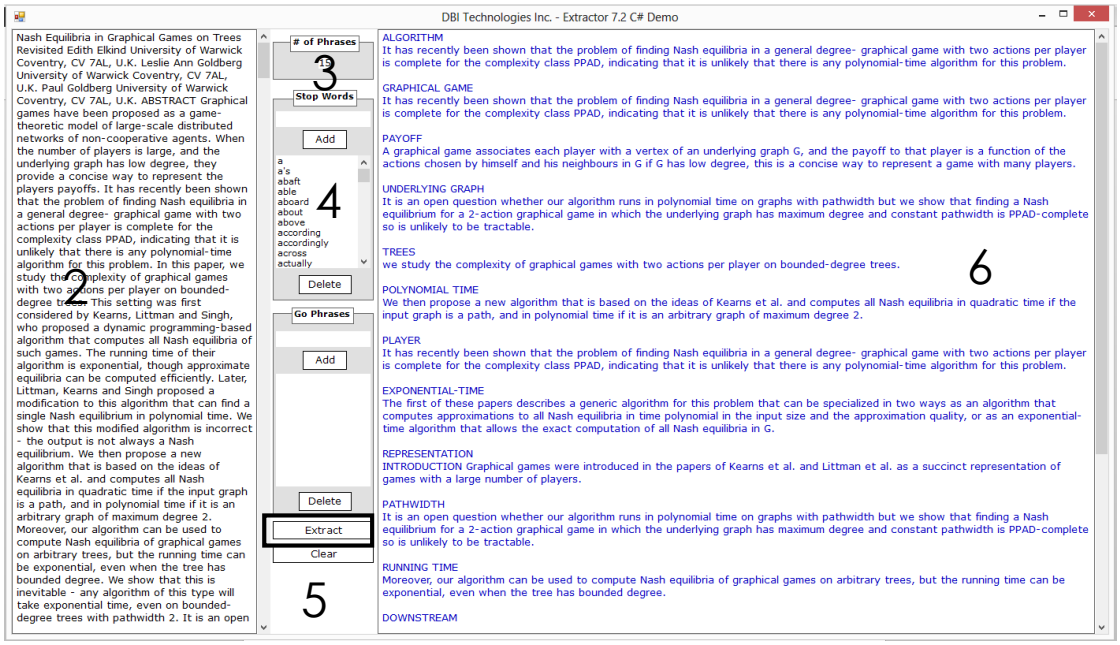


Figura 42. Interfaz de Usuario Extractor.

KEA (Libre)

KEA es un algoritmo para la extracción de palabras clave a partir de los documentos de texto. Puede ser utilizado tanto para la indexación libre o para la indexación con un vocabulario controlado. KEA está implementado en Java y es independiente de la plataforma. Se trata de un software de código abierto distribuido bajo la Licencia Pública General de GNU [KEA 15].

1.- Abrir cualquier navegador de internet e ingresar al siguiente enlace

http://www.nzdl.org/Kea/

2.-Al entrar al enlace aparece la página principal del sitio oficial KEA keyphrase extraction algorithm, se muestra la descripción del algoritmo y características referentes a su proceso de extracción. Para descargar el archivo que contiene al sistema, en la página principal en la parte superior se da clic en el hipervínculo "Download" y nos enviara a la página que contiene el archivo de descarga, para realizar la descarga se da clic en el enlace "Download Kea from its Google Code project page" como se marca en la figura 43.

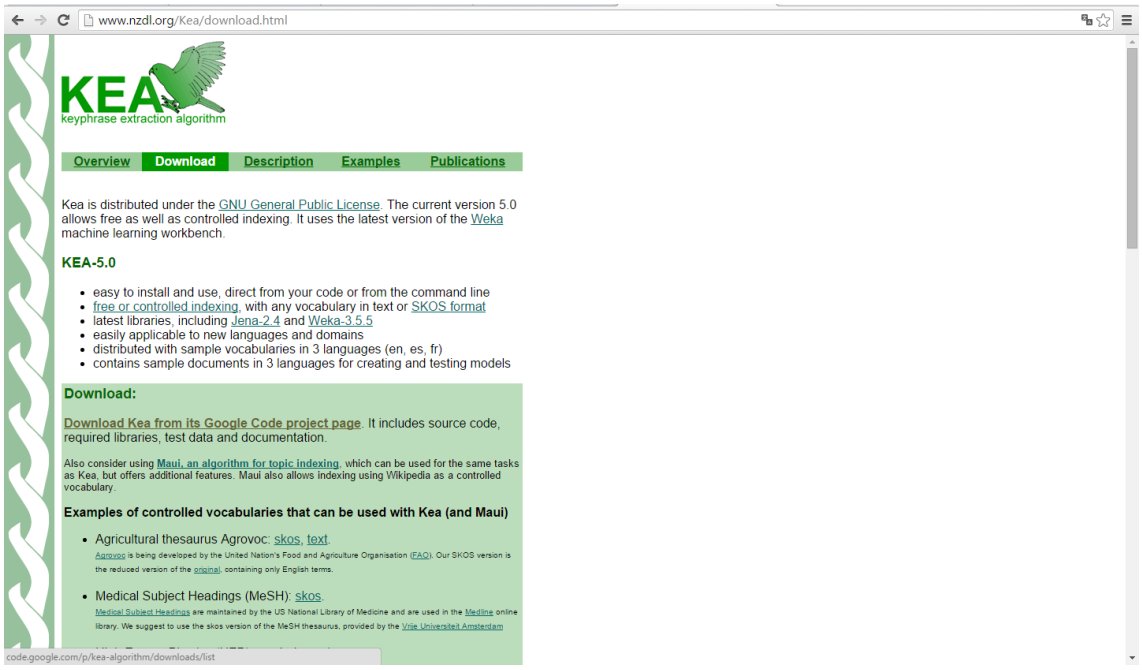


Figura 43. Sección de descarga del sitio oficial KEA

3.- Después de dar clic en el enlace se muestra la página que contiene diferentes versiones del algoritmo KEA, para descarga se da clic en la versión y la descarga comenzará
En este trabajo se descargó la versión 5

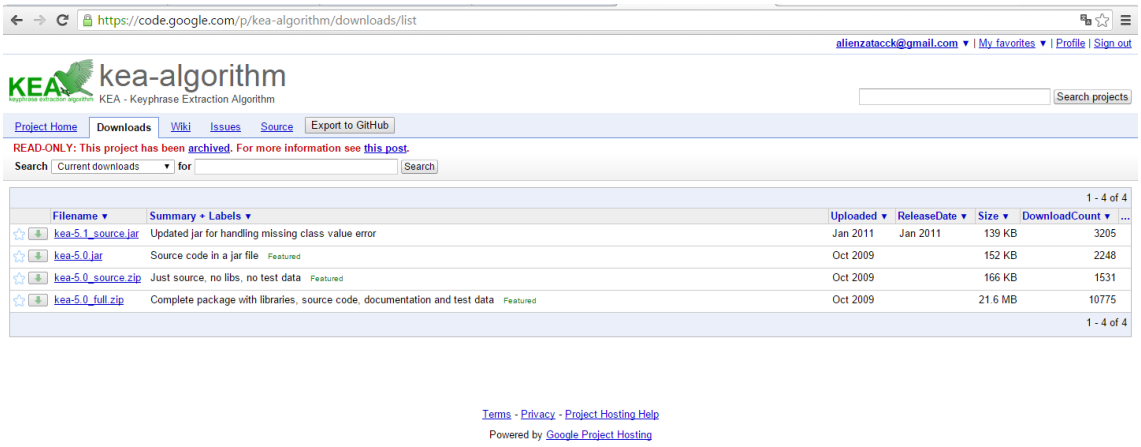


Figura 44. Página de descargas Kea.

4.-Al finalizar la descarga tendremos una carpeta con el nombre de la versión descargada y procederemos a instalar el sistema en el equipo, el sistema está implementado en Java por lo que es multiplataforma, (Como ejemplo se instala en la Distribución Debían 7 de Linux). Como primer paso para usar KEA hay que tener instalado el jdk de Java en el equipo si no se cuenta con él se puede descargar del sitio oficial de Oracle.

http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html

5.- Después de haber instalado el jdk de Java, dentro de la carpeta que contiene a KEA se encuentra un archivo nombrado "Kea-5.0 Readme" en el cual se indican las instrucciones para la instalación como los nombres para las variables de entorno con java, formato de los documentos de los cuales se va a extraer las frases clave, formato de los datos de entrenamiento, vocabularios etc.

6.-Teniendo declaradas las variables de entorno de java para ejecutar KEA en la carpeta se encuentra un archivo con el nombre TestKea.java desde el cual se puede ejecutar, este archivo contiene:
1.-Los parámetros con los cuales se crea el modelo de formación (en nuestro caso se usan los artículos de formación de la carpeta "train" del corpus SemEval-2010)

2.-La ruta de los documentos de los cuales se desean extraer frases clave
3.-Las características con las cuales se extraerán las frases clave (longitud máxima de frase clave, mínimo de ocurrencia, número de frases a extraer, vocabulario etc.)

7.-Estableciendo los parámetros correspondientes de extracción para ejecutar KEA se abre una terminal con la ruta de la carpeta raíz que contiene los recursos de KEA y se compila el archivo TestKea.java mediante el comando

Javac TestKea.java

8.-Después de compilar con éxito el archivo "TestKea.java", se ejecuta para realizar la extracción de frases clave mediante el comando.

Java TestKea

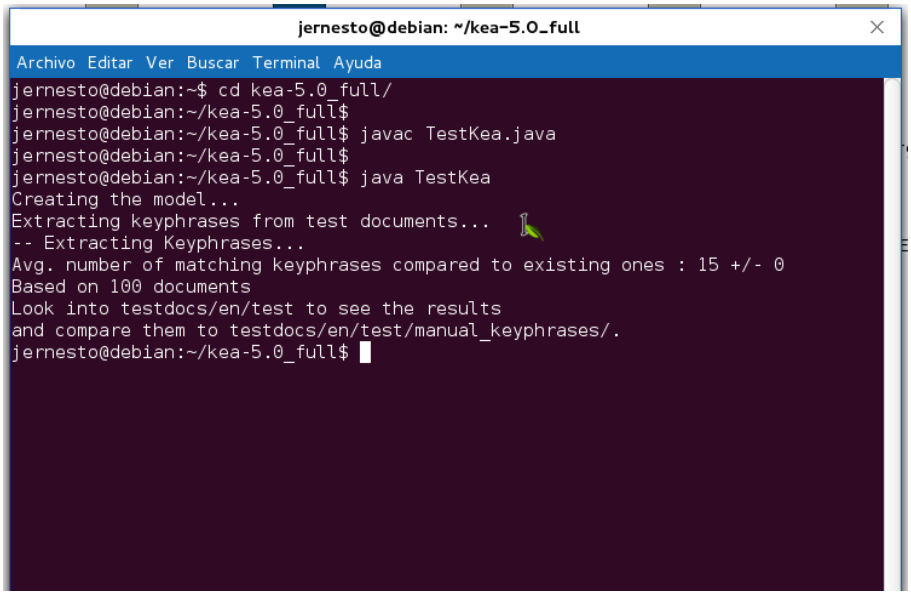


Figura 45. Ejecución de TestKea.java.

9.-Cuando finaliza la ejecución se debe dirigir a la carpeta en donde se colocaron los archivos a analizar *testdocs/en/test/* y aparecerán archivos con el nombre del articulo y extensión .key los cuales contiene las frases clave que propuso KEA para cada artículo en nuestro caso del corpus SemEval (ver figura 46).

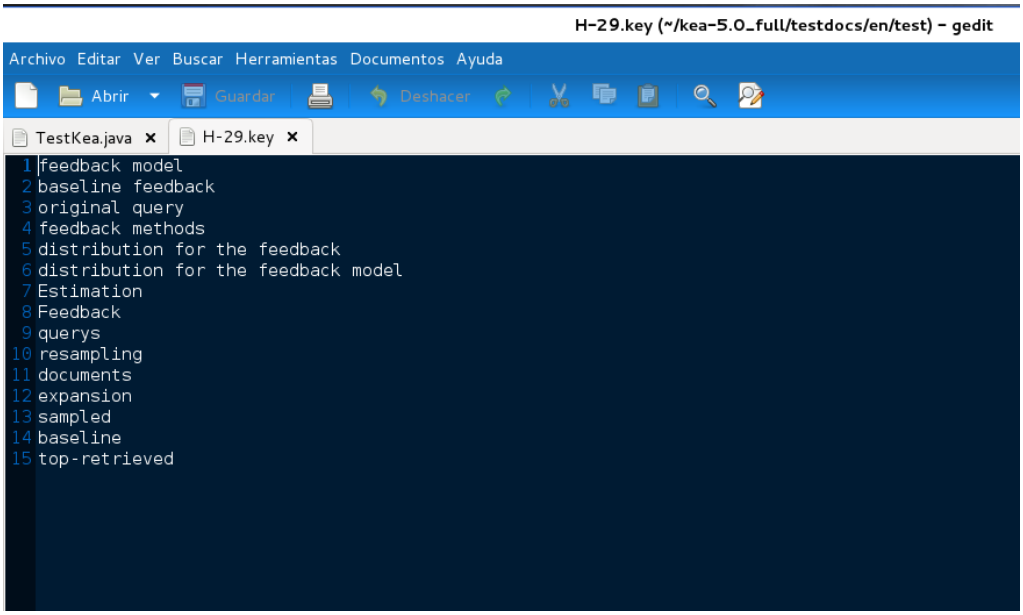


Figura 46. Formato de salida de las frases clave del archivo H-29.key.

TexLexan (Libre)

TexLexAn es el proyecto de un analizador de texto automático, clasificador y de resúmenes [TexLexan 15].

1.-Abrir cualquier navegador de internet e ingresar al siguiente enlace

http://texlexan.sourceforge.net/

Con lo que se muestra la página principal del proyecto TexLexAn.

Figura 47. Página principal TexLexan.

En la página principal se encuentra la información acerca del sistema, características, instalación, uso, mecanismo etc. Por ser un software libre se puede descargar, para descargarlo se debe ubicar en la página principal la sección "*Sources and Links*".

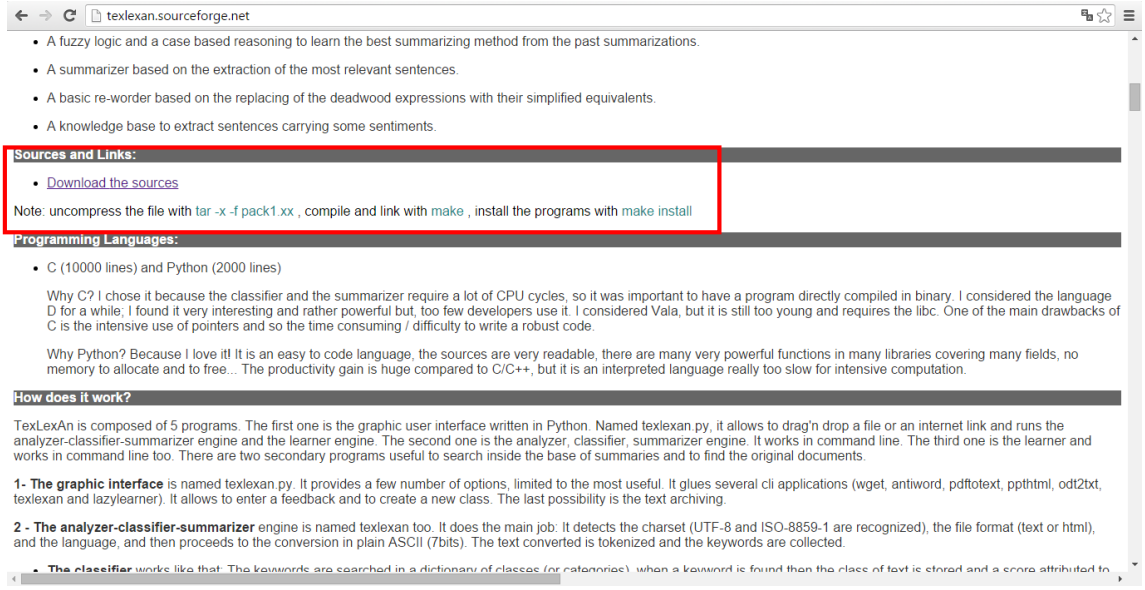


Figura 48. Enlace para Descarga Texlexan.

2.-Dar clic en el enlace y enviará a la página que contiene los archivos de descarga.

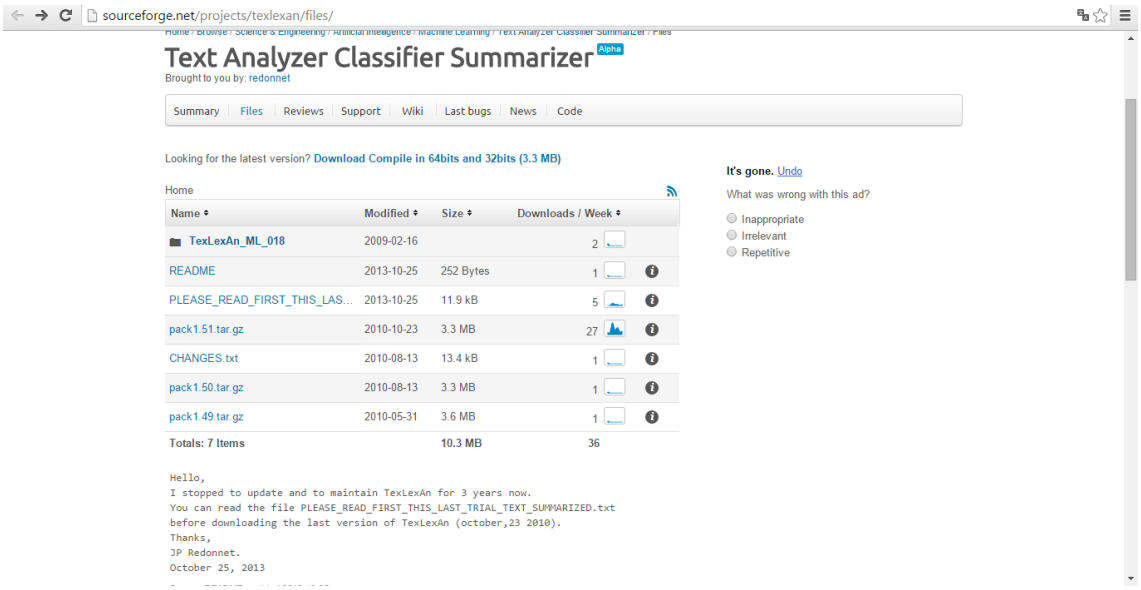


Figura 49. Página de descargas TexLexan.

3.-Al finalizar la descarga se debe realizar la instalación para poder ejecutarlo se debe tener instalado Python en el equipo, en este trabajo se ejecutó en Kali Linux ya que solo está disponible en la plataforma Linux, para poder instalarlo en la página principal se encuentran los comandos correspondientes. Después de haberlo instalado en la "Carpeta personal" deben encontrarse las siguientes carpetas:

- texlexan_archive
- texlexan _cfg
- texlexan_dico
- texlexan_doc
- texlexan_prog
- texlexan_result

4.-Para ejecutar el programa se debe de abrir una terminal y entrar a la carpeta texlexan_prog que es donde se encuentra el programa principal, para ejecutar TexLexAn se ingresa el siguiente comando :

Python texlexan.py

Después de haber ejecutado el comando anterior se ejecuta la interfaz de TexLexAn.

Para realizar la extracción de palabras clave en la interfaz de TexLexan:

1.- Marcamos la opción de "List key terms".

2.- Pegamos el texto (en nuestro caso los artículos de SemEval-2010).

3.-Clic en el botón "Aceptar".

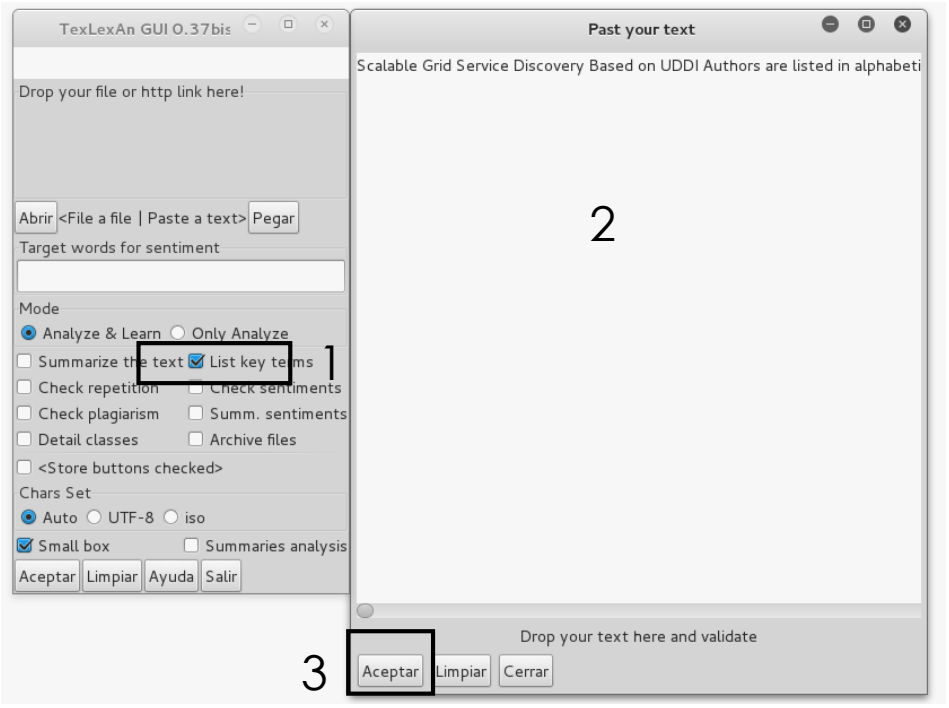


Figura 50. Interfaz de usuario TexLexan.

5.- Al dar clic en el botón aceptar se abre una ventana nueva la cual contiene los resultados de las opciones marcadas en nuestro caso "*List key terms*".

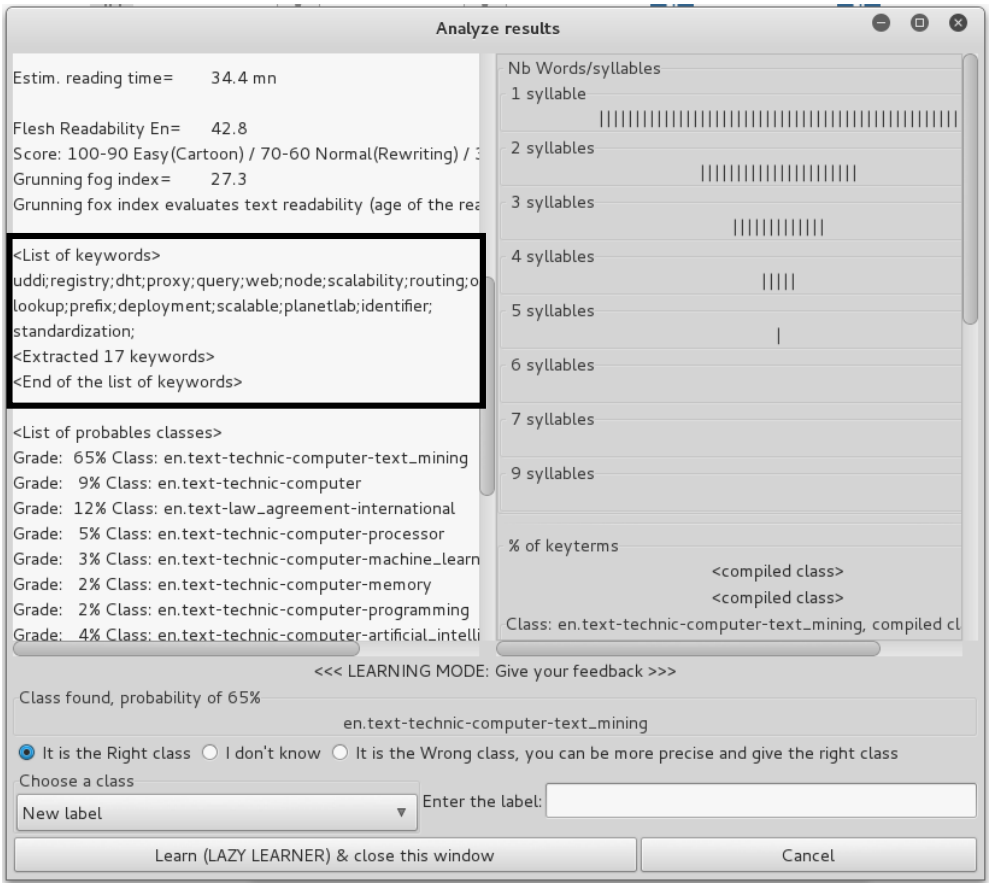


Figura 51. Salida de resultados TexLexAn.

Wordstat 7 (Comercial)

WordStat es un módulo de análisis de texto específicamente diseñado para estudiar la información textual, como respuestas a preguntas abiertas, entrevistas, títulos, artículos de revistas, discursos públicos, las comunicaciones electrónicas, etc. WordStat se puede utilizar para la clasificación automática de texto usando un enfoque de diccionario o varios métodos de minería de texto. WordStat es un módulo que se debe ejecutar desde cualquiera de los siguientes productos básicos: SimStat,QDA Miner [Wordstat 15].

1.- Abrir cualquier navegador de internet e ingresar al siguiente enlace

http://provalisresearch.com/es/descargar/trial-versions/

2.-Para este trabajo se descargó la última versión de Wordstat como es un módulo se descarga el programa principal que lo contiene en este caso QDA Miner.

Figura 52. Página de descargas Provalis research.

3.- Después de tener el Setup de QDA Miner se realiza la instalación.

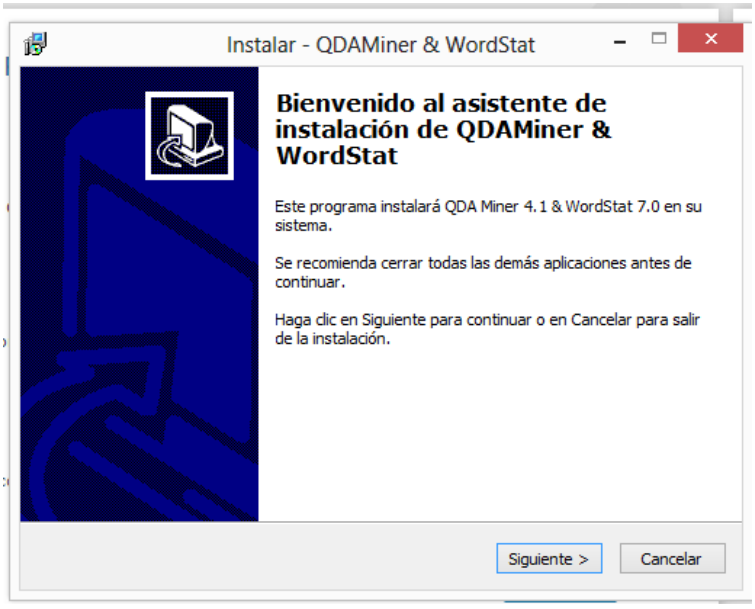


Figura 53. Instalación de QDA Miner.

4.-Para poder abrir cada uno de los 100 artículos de SemEval primero se convierten al formato de proyectos de QDA Miner, con el convertidor que se incluye después de la instalación "Document Conversion Wizard v2.0".

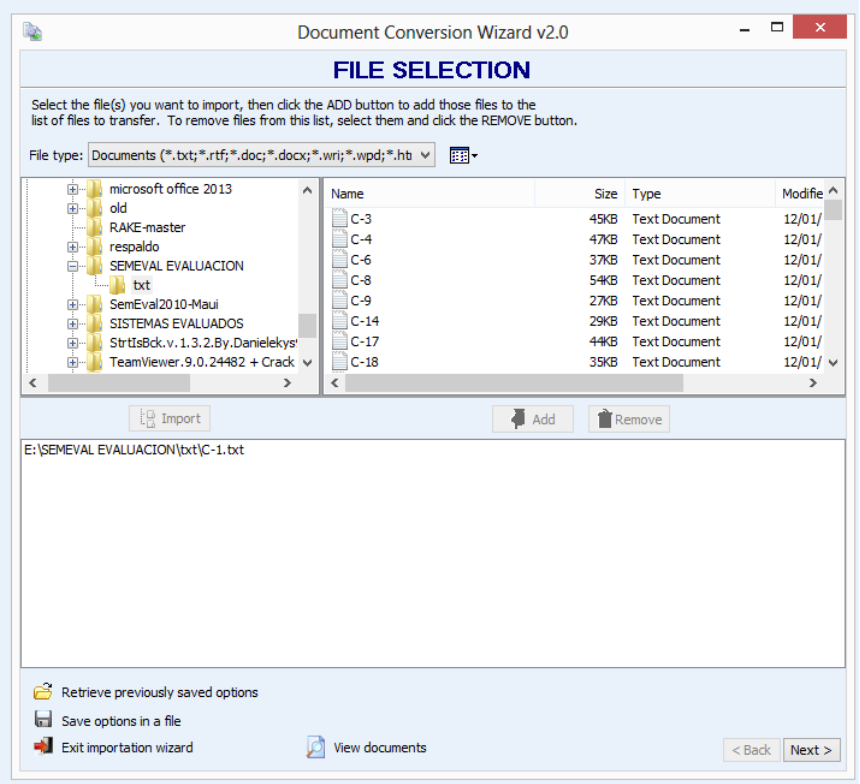


Figura 54. Convertidor de documentos Wizard v2.0.

Para convertir cada uno de los artículos primero se debe de elegir la ubicación de los mismos y posteriormente dar clic en el botón "Next" para que se pueda convertir en proyecto de QDA Miner y ser abierto desde el entorno gráfico.

5.-Después de convertir un artículo a proyecto de QDA Miner está listo para poder trabajar en el (Ver figura 55).

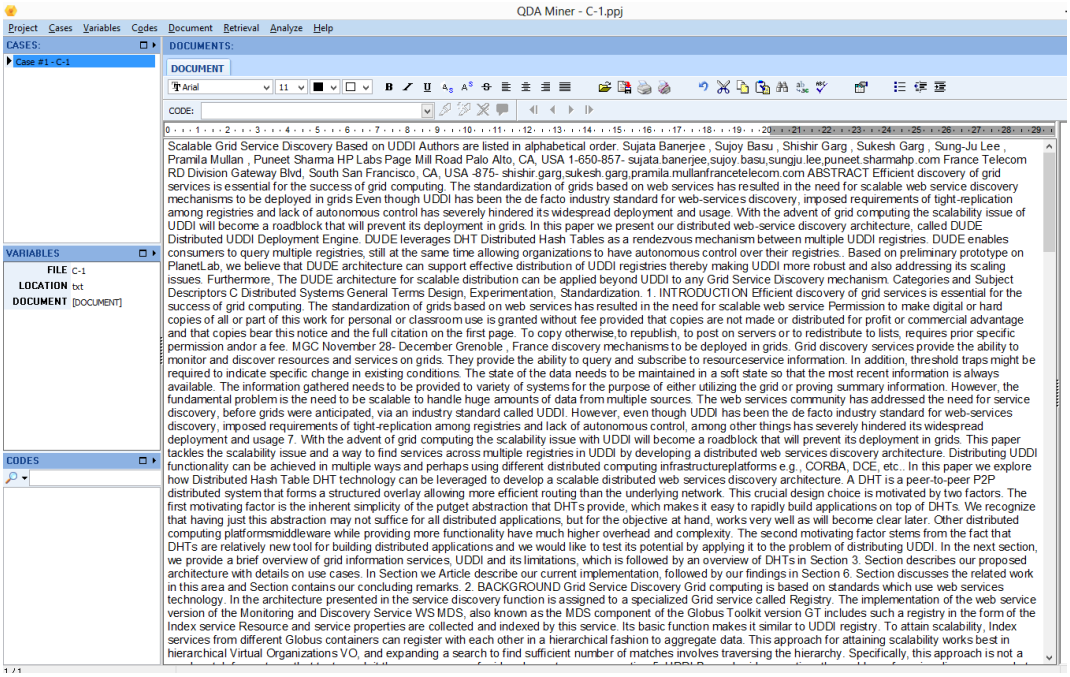


Figura 55. Entorno de trabajo de QDA Miner sobre un artículo del corpus SemEval2010.

Después de instalar QDA Miner en el equipo.

6.- Para comenzar con la extracción de frases clave utilizando Wordstat se da clic en el menú "*Analyze*" ubicado en la parte superior en la ventana del entorno de QDA Miner (Ver figura 56).

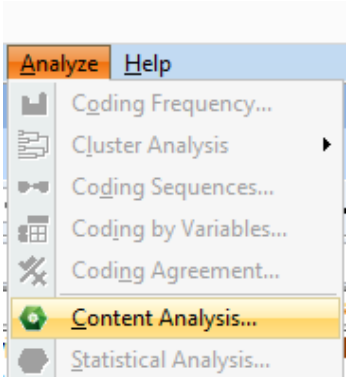


Figura 56. Menú Analyze - QDA Miner.

Con lo cual se ejecutara Wordstat 7.

7.-Al ejecutarse Wordstat, QDA Miner se bloquea ya que son programas independientes uno del otro, estando en el entorno de Wordstat:

1.-Clic en el menú "Extraction".

2.-Clic en la pestaña "Phrases" para extraer las frases del texto.

3.-Ajustar parámetros.

Para poder extraer las frases Wordstat 7 solicita parámetros de extracción:
Min words: número mínimo de palabras que puede tener una frase, por defecto 2.
Max words: número máximo de palabras que puede contener una frase.
Min Frequency: se refiere al número mínimo a partir del cual una palabra puede ser considerada como candidato a frase.
4.-Dar clic en el botón "*Search*" (ver figura 57).

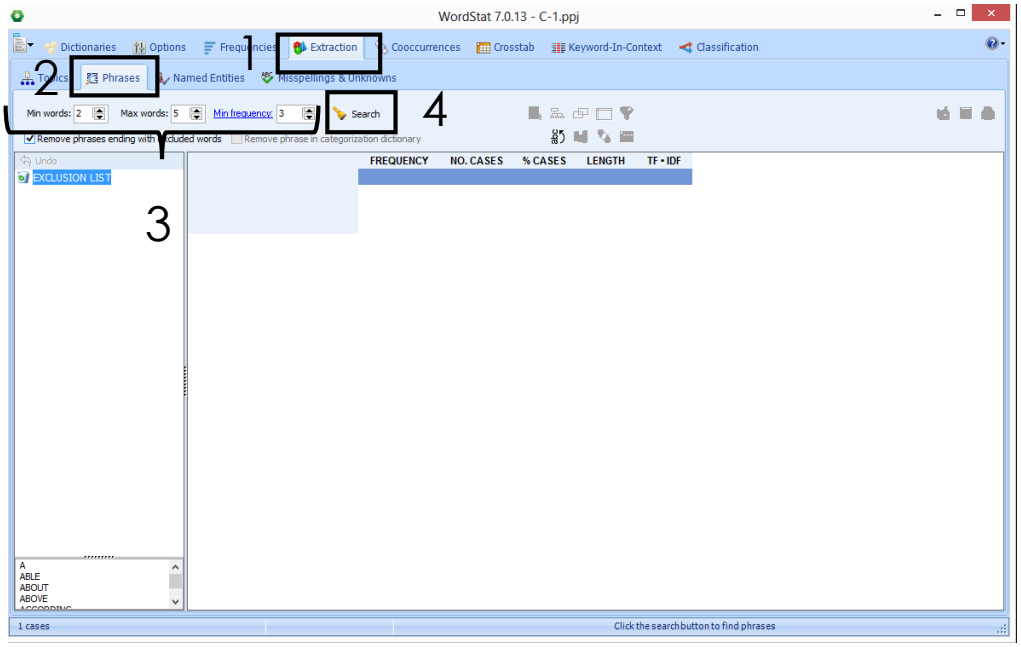


Figura 57. Menú extracción Wordstat 7.

8.- Después de dar clic en el botón "*Search*" se muestran los resultados, en la parte superior izquierda se encuentra un botón con el icono guardar para guardar las frases en formato .txt.

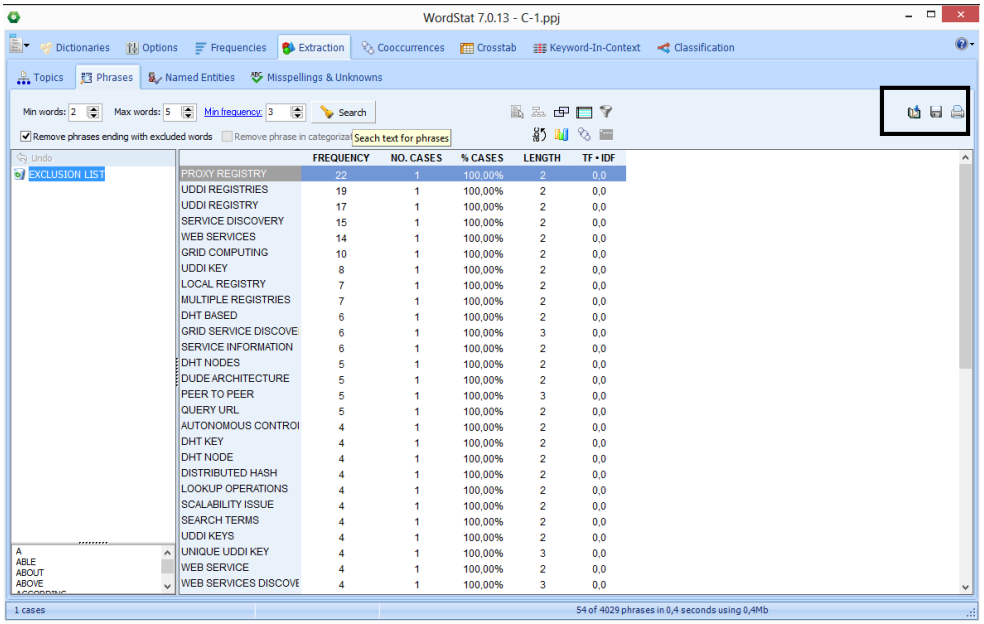


Figura 58. Salida de resultados Wordstat 7.