

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

SISTEMA DE GENERACIÓN AUTOMATIZADA DE PERFILES DE CONOCIMIENTO MEDIANTE TÉCNICAS DE MINERÍA DE TEXTO

TESIS

PRESENTADO POR:

ING. ROGELIO VALDEZ ALMADA

COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

DIRECTOR DE TESIS: DR. OSCAR MARIO RODRÍGUEZ ELÍAS

Hermosillo, Sonora, México

Agosto de 2017





"Año del Centenario de la Promulgación de la Constitución Política de los Estados Unidos Mexicanos"

SECCIÓN: DIV. EST. POS. E INV.
No. OFICIO: DEPI/230/17.
ASUNTO: AUTORIZACIÓN DE IMPRESIÓN
DE TESIS.

09 de Agosto de 2017

**C. ROGELIO VALDÉZ ALMADA,
PRESENTE.**

Por este conducto, y en virtud de haber concluido la revisión del trabajo de tesis que lleva por nombre "Sistema de generación automatizada de perfiles de conocimiento mediante técnicas de minería de texto" que presenta para el examen de grado de la MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN, y habiéndola encontrado satisfactoria, nos permitimos comunicarle que se autoriza la impresión del mismo a efecto de que proceda el trámite de obtención de grado.

Deseándole éxito en su vida profesional, quedo de usted.

ATENTAMENTE

DR. OSCAR MARIO RODRÍGUEZ ELÍAS
DIRECTOR

M.C. CÉSAR ENRIQUE ROSE GÓMEZ
SECRETARIO

M.S.I. MARÍA DE JESÚS VELÁZQUEZ MENDOZA
VOCAL

M.C.O. ROSA IRENE SÁNCHEZ FERMÍN
JEFA DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN



S.E.P.

RISF/momv*

INSTITUTO TECNOLÓGICO
DE HERMOSILLO
DIVISIÓN DE ESTUDIOS
DE POSGRADO



Abstract

Nowadays, there is a tendency of increase in computers, mobile devices and internet usage. All these tools are used for all different kinds of tasks. This marks an increase in the demand for people who can design, implement, maintain and update the software of these tools. In Mexico, 70% of graduates declared to face difficulty to enter the labor market. Some software engineering graduates face difficulty finding a job according to their profile. Nevertheless, software engineering companies have issues hiring adequate personnel. The problem presents itself when it is needed to filter those candidates that are really compatible from hundreds or thousands of possibilities, making this process really costly.

The present work proposes a solution to this problem using natural language processing (NLP) and text mining (TM) techniques to identify the more adequate candidates for software engineering positions. All the former, with the automatized knowledge profile creator system that has been called KP GENERATOR. This system was specifically designed and implemented for this issue. For the analysis and design of the system, a careful analysis of the needs was made, requirements were gathered and a system architecture was created. For the implementation, several tools were used, which allowed us to accomplish the objectives of this project in a simple, yet effective, way. KP GENERATOR consists of four main elements which are the following: Preprocessing (curriculum extraction and preparation for processing), Processing (natural language processing, text mining techniques), Refinement (small adjustments that allow to improve the results of the processing), Presentation (graphic user interface in charge of handling the interaction with the user and results visualization).

Resumen

En la actualidad, hay una tendencia de incremento en el uso de computadoras y dispositivos móviles así como del internet. Estas herramientas son utilizadas para todo tipo de tareas cada día. Lo anterior marca una tendencia de aumento en la demanda de personas que puedan diseñar, implementar, mantener y actualizar el software de estas herramientas. En México, 70% de los egresados dicen tener dificultad para ingresar al mercado laboral. Algunos egresados en las áreas de ingeniería de software se les dificulta conseguir un trabajo acorde a su perfil. Sin embargo, las empresas desarrolladoras de software tienen problemas para conseguir el personal adecuado. El problema se presenta cuando se requiere filtrar entre cientos o incluso miles de posibilidades, aquellos candidatos que pudieran ser realmente compatibles con un puesto haciendo que este proceso sea muy costoso.

El presente trabajo propone una solución a esta problemática utilizando procesamiento de lenguaje natural y técnicas de minería de texto para identificar a los candidatos más adecuados para puestos en el área de ingeniería de software. Todo esto, con el sistema de generación automatizada de perfiles de conocimiento, denominado KP GENERATOR. Se diseñó e implementó este sistema específicamente para esta problemática. Para el análisis y diseño, se realizó un análisis de las necesidades, un levantamiento de los requerimientos y se creó una arquitectura para el sistema. Para la implementación se utilizaron herramientas que nos permitieron realizar la misma de manera sencilla pero efectiva, solucionando los objetivos del presente trabajo. KP GENERATOR consta de cuatro elementos principales los cuales son: Preprocesamiento (extracción y preparación del curriculum para el procesamiento), Procesamiento (procesamiento de lenguaje natural, técnicas de minería de texto), Refinamiento (pequeños ajustes que permiten mejorar los resultados del procesamiento) y Presentación (interfaz gráfica encargada de la interacción con el usuario y de la visualización de resultados).

Agradecimientos

Agradezco a mi director de tesis el **Dr. Oscar Mario Rodríguez Elías** por su apoyo incondicional, por todo el tiempo invertido en las revisiones de los capítulos de este trabajo, así como las publicaciones, cursos, entrevistas y viajes realizados; de igual manera, por el conocimiento, las enseñanzas y experiencias compartidas, ya que todo esto fue fundamental para la culminación del presente trabajo.

Agradezco a mi co-director **Dr. Samuel González López** y a mis asesores **M.C. César Enrique Rose Gómez** y **M.S.I. María de Jesús Velázquez Mendoza** por la disposición de revisar el presente trabajo, compartir sus observaciones y correcciones para mejorar el mismo; así como las pláticas, enseñanzas y experiencias compartidas.

Agradezco a **mis compañeros de maestría**, cada uno de los cuales aportó a mi crecimiento personal, como del presente trabajo. Agradezco el tiempo, paciencia y consejos que me dedicaron. También les agradezco todas las experiencias, chistes, pláticas, tramites, vivencias y por supuesto su amistad la cual permitió crear el mejor ambiente para realizar nuestros estudios de maestría.

Agradezco a la **Dra. Trinidad Serna Encinas**, por el apoyo desde el inicio de nuestros estudios de posgrado, cada clase, consejo, plática así como la búsqueda constante de apoyar a cada uno de los estudiantes es algo admirable y que me motiva a dar lo mejor en cada trabajo realizado.

Agradezco al **Consejo Nacional de Ciencia y Tecnología (CONACYT)** por el apoyo con la beca No. 708292/585476 otorgada durante veinticuatro meses durante mi estudio de maestría. Que siga apoyando a cada vez más estudiantes mexicanos en sus estudios de posgrado.

Dedicatoria

Primero que nada quisiera dedicar este trabajo a Dios, el creador de todo lo que conocemos, por prestarme vida y salud para permitirme realizar todo lo que he hecho, incluyendo el presente trabajo.

A mis padres, por su apoyo y amor incondicional durante esta etapa de estudios de posgrado y en todo momento. A mi madre por estar siempre pendiente de mí, brindándome su cariño, consejos y abrazos. A mi padre por siempre brindarme su consejo, enseñarme el valor del trabajo con excelencia y transmitirme su sabiduría. A mis hermanas por todo el apoyo que me brindan, por las pláticas, consejos, chistes y todo lo que compartimos. A toda mi familia ya que son una parte de mí que siempre llevo conmigo.

Por último y especialmente a mi persona favorita en todo el mundo, el amor de mi vida y la persona con la que comparto mi vida entera, a ti, por ser la persona que me inspira a ser mejor, que me reta a superarme para darte lo mejor de mí. Cada año que compartimos es el mejor porque estamos juntos, siempre estaré agradecido por tu apoyo y amor incondicional. Es un honor ser parte de tu vida y un privilegio compartir mi vida contigo.

Índice General

Abstract	i
Resumen	ii
Agradecimientos	iii
Dedicatoria	iv
Índice General	v
Índice de Figuras	vii
Índice de Tablas	vii
Capítulo 1: Introducción	1
1.1 Antecedentes	2
1.2 Planteamiento del problema	5
1.3 Preguntas de Investigación	5
1.4 Objetivos	6
1.5 Justificación	7
1.6 Alcances y delimitaciones	8
1.7 Metodología	8
1.8 Organización de la tesis	9
Capítulo 2: Estado del Arte	11
2.1 Recursos humanos	12
2.1.1 ¿Qué es recursos humanos?	12
2.1.2 Áreas de selección	13
2.1.3 Herramientas actuales de apoyo a RH	13
2.2 Gestión del conocimiento	17
2.2.1 Capital intelectual y capital humano	17
2.2.2 Gestión de talento.....	22
2.2.3 Perfil de conocimiento	23
2.2.4 Sistema de Valoración Difusa de Perfiles de Conocimiento (SVDPC).....	24
2.3 Minería de texto	25
2.4 Técnicas de extracción de información.	28
2.5 Procesamiento de lenguaje natural.	28
2.6 Sistemas dedicadas a promover las capacidades de las personas.	29
Capítulo 3: Análisis y Diseño del Sistema	33
3.1 Requerimientos del sistema	33
3.2 Arquitectura del sistema	42
Capítulo 4: Implementación del Sistema	49
4.1 Modelos y estilos del prototipo	49
4.2 Procesamiento de lenguaje natural	51
4.3 Definición y clasificación del conocimiento técnico	55
4.4 Funcionamiento del prototipo	57

Capítulo 5: Análisis de Resultados.....	64
5.1 Entorno de desarrollo de pruebas	64
5.2 Objetos de prueba	65
5.3 Resultado de pruebas	66
5.4 Análisis de resultados.....	70
Capítulo 6: Conclusiones	72
Referencias	75

Índice de Figuras

Figura 1.1 Usuarios de internet en México.....	3
Figura 2.1 El proceso de contratación para incorporar a otras personas.	13
Figura 2.2 Hogan Assessments analiza cuatro áreas en sus evaluaciones.	16
Figura 2.3 Cadena de valor de una empresa.....	18
Figura 2.4 Representación del capital intelectual.....	20
Figura 2.5 Etapas del procesamiento de lenguaje natural.....	29
Figura 3.1 Diagrama de contexto de KP GENERATOR.....	39
Figura 3.2 Diagrama de proceso de negocio.....	40
Figura 3.3 Diagrama de casos de uso.....	41
Figura 3.4 Arquitectura genérica de un sistema de minería de texto.....	43
Figura 3.5 Diagrama de despliegue.....	44
Figura 3.6 Diagrama de Componentes de KP GENERATOR.....	45
Figura 3.7 Esquema General de KP GENERATOR.....	47
Figura 4.1 Página principal de KP GENERATOR.....	50
Figura 4.2 CoreNLP Server.....	53
Figura 4.3 Archivo de propiedades.....	54
Figura 4.4 Archivo de reglas.....	54
Figura 4.5 Lista de conocimiento técnicos.....	57
Figura 4.6 Vista de categorías.....	59
Figura 4.7 Vista de conocimientos técnicos.....	59
Figura 4.8 Vista demo.....	60
Figura 4.9 Vista de confirmación.....	61
Figura 4.10 Vista de proceso satisfactorio.....	62
Figura 4.11 Vista de los perfiles.....	63
Figura 5.1 Cuerpo de curriculums para pruebas.....	66
Figura 5.2 Gráfica de resultados de pruebas.....	67
Figura 5.3 Relación de frecuencia vs recursos.....	70

Índice de Tablas

Tabla 2.1 Comparación de plataformas de búsqueda de empleo.....	31
Tabla 3.1 Características del sistema.....	38
Tabla 3.2 Descripción de casos de uso.....	41
Tabla 5.1 Tabla comparativa del tamaño de solicitudes.....	65
Tabla 5.2 Resultados del procesamiento del cuerpo de curriculums.....	69
Tabla 5.3 Totales de procesamiento de la prueba.....	70

Capítulo 1: Introducción

En México, 70% de los egresados dicen tener dificultad para ingresar al mercado laboral según un estudio del Instituto de Investigaciones para el Desarrollo de la Educación (INIDE, 2013). Algunos egresados en las áreas de ingeniería de software se les dificulta conseguir un trabajo acorde a su perfil. Sin embargo, las empresas desarrolladoras de software tienen problemas para conseguir el personal adecuado, por el cual estarían dispuestos a pagar salarios altos. Sólo realizar la búsqueda para llenar una posición, puede ser muy costosa en tiempo y dinero, todo esto sin garantía alguna de encontrar al candidato indicado para ese puesto.

En búsqueda de soluciones a la problemática anterior, en este trabajo se presenta el diseño y desarrollo de un Sistema de Generación Automática de Perfiles de Conocimiento conocido como KP GENERATOR (anteriormente SGAPC). La manera en que se busca resolver el tipo de problemas descrito, es mediante la generación de perfiles de conocimiento de diversas fuentes, resultando en un ahorro de recursos para las empresas. Si una empresa desarrolladora de software gasta menos en contratar personal, ésta puede utilizar esos recursos para mejorar las condiciones de sus empleados.

El propósito de KP GENERATOR es conectar a las empresas con los empleados que tengan el perfil deseado por las mismas, para ponerlos en un trabajo donde su conocimiento y habilidades son aprovechadas por las empresas para grandes proyectos. Así mismo, traer satisfacción a los empleados al tener un mejor empleo y un mayor desarrollo. Todo esto mediante la utilización de técnicas de minería de datos para crear perfiles de conocimiento de los empleados y facilitar la gestión de los mismos.

En el presente capítulo de esta tesis, se abordan los antecedentes de los principales temas involucrados en la problemática. Se define el problema así como las preguntas de investigación que nos llevaron al planteamiento del mismo. También se presentan los objetivos, seguidos de la justificación para el desarrollo del sistema propuesto. Posteriormente se describen los alcances, las delimitaciones y la metodología a seguir. Por último se enlista el contenido del resto de este documento.

1.1 Antecedentes

La minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados (Witten, Frank, & Hall, 2011). Los tipos de datos de la misma se dividen en datos estructurados, otros tipos de datos y datos no estructurados. Los datos no estructurados son los que provienen de la web y de otros tipos de repositorios de conocimiento.

La World Wide Web es el repositorio de información más grande y diverso de los existentes en la actualidad (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004). Por ello, hay gran cantidad de datos en la web de los que se puede extraer conocimiento. En la minería web muchos de los datos son no estructurados o semi-estructurados, lo que la vuelve un poco elaborada ya que es necesario encontrar métodos adecuados para la gestión de los mismos.

La minería web se organiza en torno a tres categorías, minería de estructura, minería del uso y minería de contenido. La minería de contenido consiste en encontrar patrones de los datos de las páginas web. Como menciona (Hernández Orallo et al., 2004), la diversidad de tipos de datos que hay en la web, ha llevado al desarrollo de varias técnicas para la minería de los mismo que, dependiendo del formato en el que se encuentran, se define el nombre y método del mismo. Minería de marcado (Markup Mining), minería multimedia (Multimedia Mining), minería de hipertexto (Hypertext Mining) y minería de texto (Text Mining). Esta última se utiliza

para el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado.

Según (Cuautle Cardenas, 2013), citando a (Indexmundi, 2011) el cual a su vez cita a (CIA World Factbook, 2015), el uso del internet en México va en aumento (Véase la Figura 1.1) y esta última fuente estima que en el año 2015 existían 69.915 millones de personas con acceso a internet en México. Así el uso del internet va en aumento en todo el mundo y cada vez que se utiliza se deja información en el mismo. Una encuesta de Universia y el portal Trabajando.com reveló que 82% se los jóvenes iberoamericanos se sienten muy influenciados por el internet y el 60% de los mismos lo consideran fundamental en sus vidas (ENDI, 2015). Por lo mismo, se sigue encontrando más información en la world wide web y ésta tiene mayor diversidad. Esto cambia la manera en que accedemos a la información.

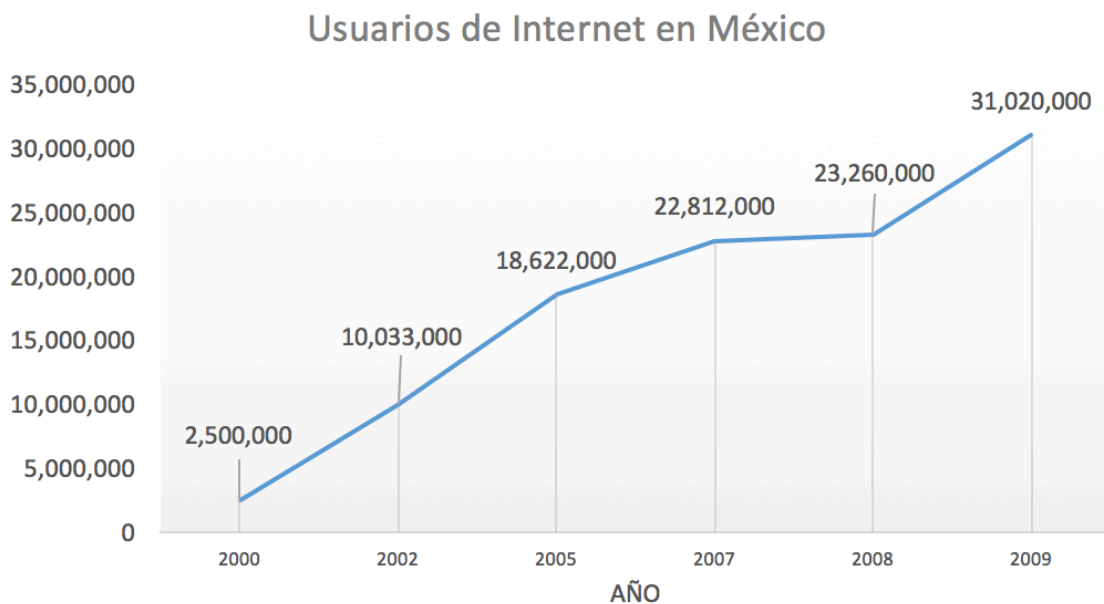


Figura 1.1 Usuarios de internet en México.
Extraído de (Indexmundi, 2011)

Encontramos todo tipo de páginas web: desde redes sociales y todo tipo de entretenimiento, hasta redes de trabajo, portales de empleo, foros de algún tema

especializado como programación. Todo esto requiere de almacenes de datos donde se tiene guardada esta información, la cual es muy valiosa, ya que puede dar información acerca de las personas para conocer sus aptitudes, sus capacidades y su experiencia. Lo anterior sirve como apoyo a la selección de personal para algún puesto. Esta información se encuentra de manera abstracta en estos almacenes, y se podría obtener mediante la minería de datos para crear perfiles de conocimiento de los posibles candidatos. Estos perfiles servirían para la gestión de talento (talent management) en una empresa.

La gestión del talento, según (Tejada Zabaleta, 2003), se basa en la legitimización de que el potencial humano se puede relacionar con los procesos productivos eficientes, eficaces y efectivos; además de sus funciones clásicas busca adquirir conocimiento que le permita detectar, apoyar, impulsar, poner a prueba e incentivar el talento de sus colaboradores y el propio (de la empresa). Por lo tanto nos permite tener un idea más concreta de las habilidades y competencias que existen dentro de la empresa, así como las que pueden llegar a incorporarse, al contratar nuevo talento.

La tecnología ha permitido la creación de herramientas que sirven para gestionar el talento. Estas herramientas tienen como objetivo servir de herramienta personal (gerentes, empleados, personal de recursos humanos), para el manejo de varias tareas. Sirven para el manejo de herramientas de reportes para ayudar a recursos humanos (RH) y al equipo ejecutivo, a comunicar estrategias, tendencias, identificar brechas y monitorear procesos.

Por otra parte, (Rosas Daniel, 2015) desarrolló un sistema de valoración difusa de perfiles de conocimiento (SVDPC), que sirve para clasificar a los candidatos según sus capacidades y aptitudes. El usuario ingresa perfiles de conocimiento de los candidatos y éste asigna un valor a la capacidad del mismo para el puesto. El problema es que el usuario del sistema necesita ingresar de manera manual los perfiles de

conocimiento, lo cual puede volver la gestión de talento un proceso tedioso, ya que conlleva tiempo y esfuerzo que puede ser utilizado en otras tareas. Un sistema de apoyo para la creación de los perfiles de conocimiento podría facilitar el ingreso de perfiles al SVDPC; por ejemplo, mediante la creación de dichos perfiles usando técnicas de minería de datos y minería de texto. La información extraída mediante las técnicas mencionadas, podría ser analizada para crear perfiles de manera automática, facilitando la tarea de ingresar los perfiles de conocimiento.

1.2 Planteamiento del problema

Las empresas constantemente requieren contratar personal, especialmente en el área de desarrollo de software. Los procesos de reclutamiento pueden llegar a consumir mucho esfuerzo en tiempo y dinero. El problema se presenta cuando se requiere filtrar entre cientos o incluso miles de posibilidades, a aquellos candidatos que pudieran ser realmente compatibles con un puesto.

Tomando en cuenta lo anterior, se observa la necesidad de desarrollar un sistema que pueda reducir la cantidad de recursos necesarios, para conseguir candidatos aptos y adecuados para un puesto, mediante el uso de herramientas que permitan automatizar al menos parte de este proceso.

1.3 Preguntas de Investigación

Derivado del planteamiento del problema, mencionado anteriormente, se obtiene la pregunta de investigación principal. De la pregunta de investigación principal surgen las preguntas específicas, las cuales nos permiten analizar el problema desde diferentes perspectivas, enfoques y tareas. Así mismo, son necesarias para asegurar un análisis y una solución completa. Las preguntas de investigación se describen a continuación.

Preguntas de investigación principal:

- ¿Cómo desarrollar un sistema inteligente, que pueda generar un perfil de conocimiento, con datos no estructurados, para ayudar a identificar a los candidatos más aptos para un puesto?

Preguntas de específicas:

- ¿Cómo identificar los perfiles de candidatos aptos para un puesto, mediante el análisis automático de datos no estructurados?
- ¿Cómo extraer conocimiento y habilidades de los candidatos para generar un perfil automáticamente?
- ¿Qué beneficios puede traer una aplicación de esta naturaleza?

1.4 Objetivos

Objetivo general:

Implementar una herramienta de software para apoyar en la identificación de candidatos aptos para puestos en empresas de desarrollo de software, mediante la identificación automática de perfiles de conocimiento de personal y candidatos a puestos en fuentes diversas, como LinkedIn, portales de empleo y curriculum vitae (CV), en formatos libres.

Objetivos específicos:

- Estudiar mecanismos de minería de texto e identificación de patrones o conceptos.
- Analizar los tipos de fuentes de datos curriculares y elegir uno o dos de ellos.
- Definir un mecanismo de extracción de información de dichas fuentes.

- Diseñar e implementar una herramienta que genere el perfil de conocimiento de los candidatos usando la información extraída.
- Analizar los beneficios que podría aportar la herramienta que se busca desarrollar.

1.5 Justificación

Actualmente, muchas empresas tienen problemas para encontrar candidatos aptos para cubrir algunos de sus puestos vacantes. Al segundo trimestre del año 2015, la Población Económicamente Activa (PEA) del país (52.6 millones de personas), registró una tasa de desempleo del 59.8%, aun cuando el 40.4%; es decir, un millón veintiséis mil de esos desempleados, cuenta con estudios profesionales (Cantillo, 2014). Sin embargo, 43% de las empresas mexicanas tienen dificultades para contratar personal (El Informador, 2010). Un estudio de Manpower Group muestra que esto afecta la capacidad del 42% de las empresas, de satisfacer las necesidades del cliente y reduce la competitividad y productividad. Así mismo existe un aumento en la rotación de empleados en el 30% de las empresas, y una menor motivación y compromiso en el 26% de los empleados (Manpower Group, 2015).

El sistema propuesto puede contribuir a ayudar a los empleados a encontrar un trabajo donde sus aptitudes coincidan con las requeridas por el puesto, aumentando su satisfacción con el empleo y por consecuencia, su motivación y compromiso. De igual manera, ayudar a los empleadores a disminuir su rotación de personal, aumentar su competitividad y aumentar su capacidad de satisfacer al cliente. Un cliente satisfecho se traduce en más ganancias para el empleador; más ganancias para el empleador significan estar más dispuesto a dar mejores sueldos a los empleados. Mejores sueldos para los empleados, significa estar más contento con su trabajo e incluso, tener un sentido de pertenencia con la empresa.

1.6 Alcances y delimitaciones

Para extraer los conocimientos técnicos del curriculum se aplicaron técnicas de minería de texto y procesamiento de lenguaje natural. Se utilizaron curriculums obtenidos en la web, que son fuentes de información de texto en formato no estructurado. El área de aplicación para esta tesis fue empresas de desarrollo de software.

Para la implementación del prototipo, se hizo uso del modelo de evaluación de perfiles de conocimiento de (Rosas Daniel, 2015). La experimentación del prototipo se llevó a cabo en el Laboratorio de Ingeniería de Software de la Maestría en Ciencias de la Computación, perteneciente a la División de Estudios de Posgrado e Investigación del Instituto Tecnológico de Hermosillo.

1.7 Metodología

La metodología que se siguió para el desarrollo del sistema propuesto consta de cuatro etapas. Estas etapas cubren a detalle todo el proceso de desarrollo, desde el análisis del estado del arte, hasta la última etapa, que son las pruebas de validación del mismo. Las etapas se describen a continuación.

En la primera etapa se realizó un análisis del estado del arte, en el cual se estudiaron los conceptos y aspectos teóricos necesarios para desarrollar el sistema. Es importante conocer las bases para poder comprender el análisis, diseño e implementación de la presente tesis. Los temas principales del estado del arte son: minería de texto, técnicas de extracción de información, sistemas de apoyo dedicados a promover las capacidades de las personas y sistemas de apoyo para contratación de recursos humanos.

La segunda etapa consistió en un análisis de las herramientas necesarias para el desarrollo del sistema de generación de perfiles de conocimiento como ayuda para la selección de personal. Así mismo, el desarrollo del diseño y la arquitectura del sistema. Para el diseño y modelado del sistema se utilizó el Lenguaje Unificado de Modelado (UML, por sus siglas en inglés).

En la tercera etapa se llevó a cabo la implementación del sistema y los módulos contenidos en la arquitectura del mismo. Se realizaron dos pruebas de funcionalidad del sistema, para verificar que los módulos implementados funcionan según lo esperado. La primera prueba, fue crear el perfil de conocimiento con base en un portal de datos curriculares. La segunda prueba utilizó como fuente de información, texto en formato no estructurado.

Finalmente en una cuarta etapa, se realizaron pruebas de validación, donde se estudiaron la percepción de utilidad y facilidad de uso del sistema propuesto con potenciales usuarios, que incluye empleados o futuros empleados, así como empleadores en el área del desarrollo de software.

1.8 Organización de la tesis

La presente tesis consta de seis capítulos y está organizada de la siguiente manera:

En el presente capítulo (Capítulo 1) se definió la problemática, los objetivos a cumplir, la importancia del presente trabajo, los alcances y delimitaciones, la metodología a seguir y la organización de la presente tesis.

El capítulo 2 presenta el estado del arte, en el cual se tratan los temas principales utilizados para fundamentar la solución del proyecto de tesis. Entre los temas principales se encuentran la minería de texto, técnicas de extracción de

información, sistemas de apoyo dedicados a promover las capacidades de las personas y sistemas de apoyo para contratación de recursos humanos.

El capítulo 3 es el análisis y diseño del sistema. En la primera parte se presenta el análisis, para el cual se realizó el levantamiento de necesidades, objetivos, requerimientos y restricciones del sistema. En la segunda parte se presenta el diseño para el cual se define la estructura y el funcionamiento general de KP GENERATOR.

El capítulo 4 es la implementación del sistema, es decir de KP GENERATOR. En la primera parte se explica los modelos y estilos utilizados en el prototipo. Seguido de la descripción de las herramientas utilizadas para su desarrollo, cerrando con la explicación del funcionamiento del prototipo.

El capítulo 5 es el Análisis de Resultados donde se estudian los resultados de la implementación del sistema propuesto en la presente tesis. El capítulo inicia con la descripción del entorno donde se desarrollaron las pruebas. En segunda instancia se explican los objetos analizados en las pruebas (curriculums). En consiguiente se analizan los resultados de las pruebas para cerrar con el análisis de los resultados.

Finalmente, el capítulo 6 son las conclusiones. En esta sección se describen las conclusiones obtenidas del presente trabajo y tesis, las aportaciones, así como el trabajo futuro.

Capítulo 2: Estado del Arte

Como se observó en el capítulo anterior, se busca desarrollar un sistema de generación automatizada de perfiles de conocimiento que sirva como apoyo para selección de personal. En este capítulo se abordan los temas más relevantes para entender la problemática y la solución que se está planteando, los cuales se encuentran organizados de la siguiente manera: primeramente se aborda el tema de Recursos Humanos (RH), por ser el área encargada de gestionar al personal dentro de una empresa (capital humano). Así mismo es necesario entender el proceso de selección de RH, lo que nos lleva a la gestión de talento, que se puede clasificar en cinco elementos: reclutamiento, la selección, la formación, las remuneraciones y la evaluación del desempeño. Otro tema importante es el capital intelectual, refiriéndonos al conocimiento tácito (conocimiento implícito) que existe dentro de la empresa por ser el fundamento y lo que lleva al desarrollo de talento dentro de la misma. De la misma manera, se estudian definiciones de gestión de conocimiento debido a que es necesario comprender la manera en que se administra dicho recurso para poder entender el concepto de perfil de conocimiento, lo cual es indispensable para lograr la creación del mismo.

En cuanto a la solución de la problemática, estudiaremos el Sistema de Valoración Difusa de Perfiles de Conocimiento (SVDPC) de Rosas Daniel (2015) que se divide en tres partes: perfilación, evaluación y valoración, las cuales se describirán más adelante. Habiendo estudiado los temas correspondientes a la problemática seguiremos con las herramientas y técnicas necesarias para el desarrollo de KP GENERATOR los cuales son: minería de texto, técnicas de extracción de información y procesamiento de lenguaje natural.

2.1 Recursos humanos

Recursos humanos o RH es un área indispensable en toda empresa. RH tiene la responsabilidad de ser el vínculo entre todas las partes de una empresa. A lo largo del tiempo, RH ha sido un área que ha recibido muchos cambios por lo que nos podemos encontrar que el término administración de recursos humanos (ARH) es intercambiado por gestión del talento humano, administración de asociados, administración de colaboradores, administración de capital humano, administración de capital intelectual entre otros; donde estos términos se relacionan con un mismo concepto. A continuación, veremos más a detalle que es RH y lo que comprende este departamento.

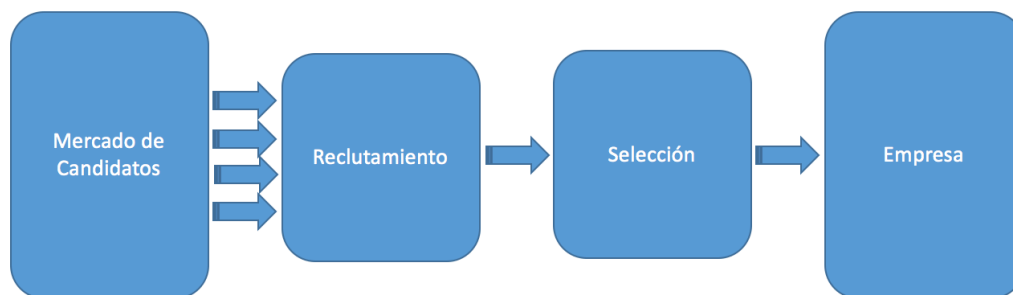
2.1.1 ¿Qué es recursos humanos?

El término RH puede tener tres significados: el primero es RH como departamento, el segundo RH como prácticas a seguir y el tercero RH como profesión (Chiavenato, 2009). RH debe lograr el bien común de todas las partes involucradas para que todos se vean beneficiados. (Chiavenato, 2009, p. 9), citando a (Dessler, 1997, p. 2), menciona que “la administración de recursos humanos (ARH) es el conjunto de políticas y prácticas necesarias para dirigir los aspectos administrativos en cuanto a las ‘personas’ o los recursos humanos, como el reclutamiento, la selección, la formación, las remuneraciones y la evaluación del desempeño”. El proceso empieza desde que se empieza a reclutar personal para trabajar en la empresa y de ahí en adelante nunca deja de estar en contacto con los trabajadores. RH tiene la misión de alinear los objetivos de la organización con los objetivos individuales de los trabajadores para que se cree una simbiosis en la que se buscan soluciones del tipo ganar-ganar. Por lo que RH es encargado del “conjunto integral de actividades de especialistas y administradores —como integrar, organizar, recompensar, desarrollar, retener y evaluar a las personas— que tiene por objeto proporcionar habilidades y competitividad a la organización.” (Chiavenato, 2009). RH es el área encargada de

administrar al personal dentro de una organización y podemos definirlo como el enlace entre una institución y sus individuos.

2.1.2 Áreas de selección

Las áreas de selección de RH representan el filtro que discierne quienes forman (o formarán) parte de la empresa, así mismo RH se encarga del proceso de contratación de personal (Véase la Figura 2.1). La primera etapa de este proceso es buscar dentro del mercado de candidatos las opciones disponibles a través de las herramientas o medios que consideren adecuados. La segunda etapa es el reclutamiento de candidatos donde se empieza a recibir curriculums de candidatos e incluso tener las primeras entrevistas que suelen ser breves. Al terminar la etapa de reclutamiento inicia la selección, etapa en la cual tienen que decidir qué candidato(s) quieren que formen parte de la empresa y se ven los últimos detalles para la contratación, para que entonces formen parte de la empresa de manera oficial. El área de selección distingue, entre los candidatos, a las personas que están mejor calificadas y que son más adecuadas para el puesto, para formar parte de la organización.



*Figura 2.1 El proceso de contratación para incorporar a otras personas.
Tomado de (Chiavenato, 2009, p. 136).*

2.1.3 Herramientas actuales de apoyo a RH

Existe una gran cantidad de herramientas que son utilizadas en RH para la selección de personal. Estas herramientas permiten conocer diferentes aspectos del

candidato; según el enfoque que tenga cada herramienta, puede resaltar ciertas aptitudes, virtudes, deficiencias, dificultades, entre otras características de los candidatos.

Cada empresa decide los métodos que mejor se ajustan a sus necesidades, al tiempo disponible y a su presupuesto. Dependiendo de la naturaleza del puesto o la empresa, se pueden aplicar etapas de manera menos rigurosas (o no aplicarse del todo). El proceso de selección común según (Ivancevich, 2005, pp. 232–245) consta de cinco etapas. Los métodos y herramientas que suelen ser utilizados por RH dependen de la etapa de la cual se trata. Las etapas y algunos tipos de herramientas utilizadas son los siguientes:

Etapas 1 - Reconocimiento preliminar: solicitud de empleo y entrevista.

Herramienta(s): Solicitud impresa. Usualmente, la primera etapa en el proceso de selección en las empresas es la solicitud de empleo (después de la publicación de la vacante). El candidato llena una solicitud de empleo con sus datos, experiencia y habilidades para que el empleador tenga conocimiento de las mismas.

Etapas 2 - La entrevista de empleo.

Herramienta(s): La entrevistas de trabajo. La entrevista se utiliza para tener un encuentro entre el entrevistador y el candidato para poder hacer preguntas referente al puesto, experiencia, requisitos, responsabilidades, entre otros, con el fin de resolver las inquietudes de ambas partes. Aunque usualmente suele ser en persona, puede llegar a ser a través de internet o por llamada telefónica. Los tipos de entrevista pueden variar según la estructura (con o sin estructura), la dimensión (antecedentes o situaciones hipotéticas).

Etapas 3 - Exámenes para el empleo.

Herramienta(s): Prueba de desempeño, prueba de facultades intelectuales, test psicológico. Las pruebas de desempeño en una muestra de trabajo se realiza en un

ambiente controlado para probar al candidato en una parte del funcionamiento. La prueba de facultades mentales busca evaluar ciertos conocimientos o habilidades del candidato, un ejemplo común es las pruebas que analizan las habilidades matemáticas y verbales. El examen psicológico no busca obtener una calificación numérica, lo que busca obtener es un análisis de la forma de pensar del candidato, en la que se pueda conocer si es adecuado para el puesto.

Etapa 4 - Verificación de referencias y antecedentes.

Herramienta(s): Carta de recomendación. En esta etapa se busca pedir datos de contacto como referencias tanto laborales o personales herramienta el empleador solicita cartas de recomendación a favor del candidato. Pueden ser del último empleador o personales.

Etapa 5 – Examen médico.

Herramienta(s): Certificados médicos, pruebas de antidoping: Los exámenes médicos son utilizados para descartar candidatos que no son aptos para la vacante. El certificado médico es una carta o constancia expedida por un médico que avale la salud y condición para trabajar del candidato.

Un ejemplo de una herramienta de apoyo es (Hogan Assessments, 2016) que obtiene cuatro diferentes mediciones que sirven para entender distintos aspectos de las personas.

- **Hogan Personality Inventory (HPI): The Bright Side of Personality.** Esta medida es el lado bueno o brillante de las personas. No desde el punto de bondad, sino del punto de vista cuando se encuentra relajado, sin mucha presión, su forma normal de relacionarse con las demás personas cuando se está en su mejor momento.

- **Hogan Development Survey (HDS): The Dark Side of Personality.** Describe el lado oscuro de la personalidad, es decir las cualidades que salen a flote en tiempos de mayor tensión y estrés; puede romper relaciones, dañarlas o alejar a las personas del éxito.
- **Motives, Values, Preferences Inventory (MVPI): The Inside of Personality.** Nos muestra el interior de la personalidad, las metas fundamentales, valores, motivaciones e intereses que determinan lo que deseamos por lo que nos esforzamos por obtener.
- **Hogan Business Reasoning Inventory (HBRI): How You Think.** Describe el tipo de razonamiento que tenemos, si razonamos tácticamente o estratégicamente. Esto afecta el tipo de decisiones que tomamos, nuestra habilidad para evaluar situaciones, cómo resolvemos los problemas y evitamos repetir los mismos errores.

Al analizar estas cuatro áreas (Véase la Figura 2.2) Hogan Assessments permite obtener datos para poder tomar una decisión más informada, ya que permite conocer hasta cierto punto, el carácter de los candidatos (bueno y malo), sus motivaciones y su tipo de razonamiento. Así es posible ver qué persona puede ser más adecuada según el objetivo de la empresa y el perfil para el puesto.

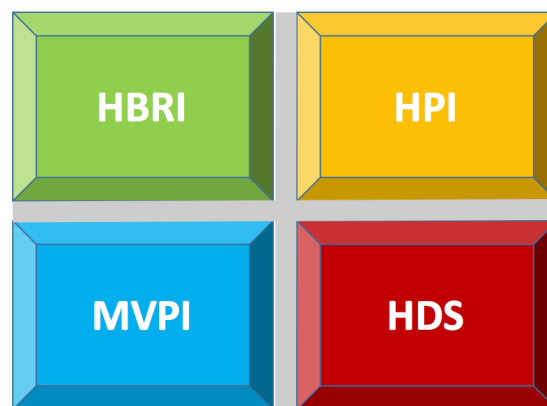


Figura 2.2 Hogan Assessments analiza cuatro áreas en sus evaluaciones. Tomado de (Hogan Assessments, 2016).

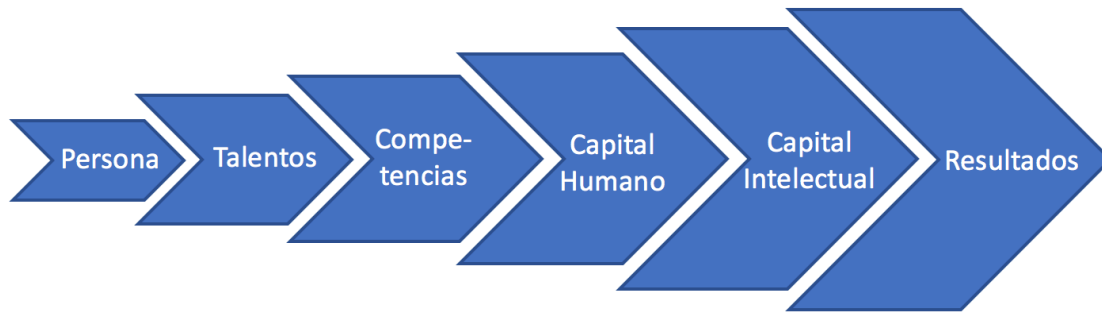
Leader Versatility Index (LVI) 360° es una herramienta de (Kaiser Leadership Solutions, 2016). Esta herramienta utiliza un cuestionario que toma como base la evaluación de los compañeros, superiores y subordinados en el área de trabajo para obtener los resultados del mismo. La evaluación pretende obtener una visión completa que es plasmada en una gráfica que se divide en doce secciones, en la cual existen variables que se encuentran en lados opuestos de la gráfica en las que usualmente, al tener mucho de una de ellas, se tiene menos de la opuesta; si se es muy directivo (forceful) se tiende a ser menos participativo (enabling) y si se es muy estratégico (strategic) se tiende a ser menos operacional (operational).

2.2 Gestión del conocimiento

La gestión del conocimiento es un área que ha ido en crecimiento a través de los años ya que se le ha dado un mayor enfoque a la administración del conocimiento y capital intelectual dentro de las organizaciones con el fin lograr una mayor eficiencia y control del mismo.

2.2.1 Capital intelectual y capital humano

Las personas son el cimiento de toda empresa. Las personas aportan el talento que a su vez crea competencias. Estas no llevan a formar el capital humano de la empresa. El capital humano es el fundamento del capital intelectual y este a su vez nos lleva a la obtención de resultados. Observe la Figura 2.3 donde se muestra el desarrollo de la cadena de valor en una empresa, iniciando con las personas como individuos hasta formar el capital intelectual dentro de la empresa. Antes de definir este último, se estudiarán algunas definiciones de capital humano, ya que precede al capital intelectual.



*Figura 2.3 Cadena de valor de una empresa.
Extraído de (Chiavenato, 2009, p. 56).*

Desde hace algunos años (Marshall, 1947) expresó la importancia del capital humano desde un punto de vista económico al describirlo como: “El capital más valioso de todos es el que se ha invertido en seres humanos”. La inversión en el capital humano es necesaria para el desarrollo del mismo. El capital humano es la base del capital intelectual, para poder impulsar este último es imprescindible tener un capital humano competente. Si no se valora la importancia (y por lo tanto invierte) en el capital humano, no se puede esperar un capital intelectual sólido.

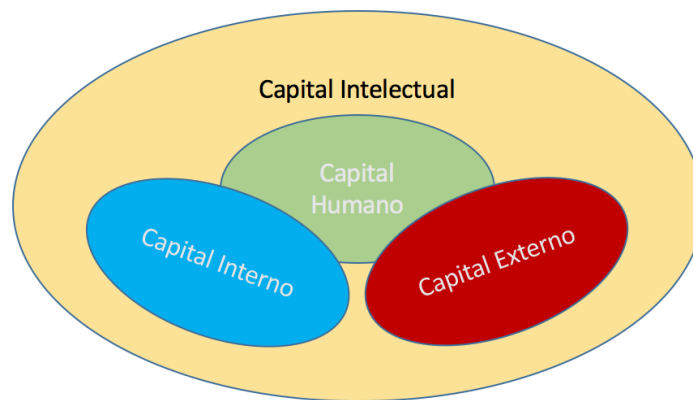
El capital humano según (Madrigal Torres, 2009, p. 69) “es el conocimiento que posee, desarrolla y acumula cada persona en su trayectoria universitaria o de formación, así como la laboral y organizacional”. Así mismo concluye que la mano de obra y el capital humano se distinguen en que la primera es de naturaleza física mientras que la segunda es de naturaleza mental, lo cual se relaciona con la acumulación de conocimiento, habilidades y competencias. Por lo tanto, se puede definir la gestión del capital humano como la administración del conocimiento de las personas desde el punto de vista mencionado anteriormente. La gestión del capital humano plantea que las personas son un recurso tangible en la organización y su valor va a depender del conocimiento y las habilidades que éste posea (Madrigal Torres, 2009).

El capital intelectual es la fuente de conocimiento de la empresa, por lo que no se trata únicamente de tener personal contratado, sino de que tenga el conocimiento clave para llevar a cabo las tareas clave definidas según el giro de la empresa. Chiavenato lo expresa diciendo: "Las personas son proveedoras de conocimientos, habilidades, competencias ... en este sentido, las personas constituyen parte integral del capital intelectual de la organización" (Chiavenato, 2009, p. 10). El capital intelectual no es tangible y es difícil de administrar ya que se trata de algo abstracto; sin embargo, de no hacerlo, la empresa se expone a tener fugas de esta parte fundamental, y no detectarlas. El capital intelectual es tan necesario como el inventario de los productos para un almacén. Un almacén que tiene mucho producto y desconoce cuánto tiene, por lo tanto no puede saber cuánto puede vender y cuánto requiere comprar. Este almacén se encontraría en grandes problemas haciendo ventas y comprar a ciegas. De la misma importancia es la necesidad de conocer el capital intelectual con el cual se cuenta en una empresa.

En la era industrial el centro de enfoque eran los factores tradicionales como la tierra, el capital de la organización y una gran fuerza de trabajo. En cambio, con la llegada de la era del conocimiento se ha iniciado un cambio de enfoque, el cual tiene la prioridad de concentrarse en mejorar los procesos de la organización, gestionar el capital humano e intelectual de las organizaciones, así como gestionar el conocimiento dentro de las mismas. Goldsmith & Carter lo describieron de la siguiente manera: "Los activos de una empresa pueden dividirse en dos partes: su gente, y todo lo demás" (Goldsmith & Carter, 2010). Esta gestión es la que permite a las compañías adaptarse a la revolución tecnológica que toma lugar hoy en día y evolucionar junto con el entorno. El capital intelectual consta de tres componentes (Véase la Figura 2.4): el capital humano, el capital interno y el capital externo (Chiavenato, 2009) .

- El **capital humano** es representado por los talentos; forman parte de la organización en la que tiene un estilo de administración participativo que invita a los mismos a involucrarse.

- El **capital interno** representa los conceptos, modelos, procesos y sistemas de información y administrativos. Son creados por las personas y utilizados por la organización.
- El **capital externo** representa la manera en que la organización ofrece soluciones a los problemas de los clientes, cómo se percibe su imagen, su reputación. Es decir todo lo ajeno a la misma organización incluyendo, más no limitándose a: relaciones con los clientes, proveedores, marcas, imagen y reputación.



*Figura 2.4 Representación del capital intelectual.
Tomado de (Chiavenato, 2009, p. 56)*

Chiavenato observa que la ARH (administración de recursos humanos), en su enfoque moderno, se centra más que nunca en el capital humano y la manera en que afecta al capital intelectual de la organización. Por lo tanto, la influencia de ARH se extiende a los trabajadores dentro de la empresa, a la organización y a los clientes.

En cuanto al conocimiento y capital intelectual (Chiavenato, 2009, p. 403), citando a (Nonaka & Takeuchi, 1997), dice que “el conocimiento es la base del capital intelectual de las organizaciones (empresas)”, donde este capital representa todos los conocimientos generados por la empresa y se divide en las siguientes áreas:

- Tecnología – patentes, procesos, productos.
- Información – clientes, proveedores, entorno.
- Habilidades desarrolladas por colaboradores.
- Solución de problemas en equipo, comunicación, administración de conflictos, desarrollo de inteligencia.

La administración del conocimiento busca identificar lo que se sabe o lo que se debe saber dentro de la organización. Significa crear, captar, archivar, organizar, sistematizar, difundir y aplicar el conocimiento, la información y la experiencia entre los colaboradores que tienen intereses y necesidades similares a efecto de crear innovaciones y generar rendimientos. En resumen, la administración del conocimiento está ligada a la decisión de cuál es el conocimiento relevante para la organización y que sirve para sus propósitos y proporciona resultados.

El conocimiento, así como el capital intelectual es un activo intangible por la que anteriormente, en la era industrial no se le daba mucha importancia. Las empresas buscaban tener el control de bienes tangibles y el conocimiento no formaba parte de ellos. Con el tiempo se dieron cuenta de la importancia de tener el conocimiento dentro de sus bienes. El buscar tener conocimiento dentro de la empresa, siendo este un activo intangible, requiere de especial atención y de ahí surge la gestión del conocimiento.

Los trabajadores del conocimiento son definidos por (Davenport, 2005) como los trabajadores con un alto grado de “expertise”, educación o experiencia, y que el principal propósito de su trabajo involucra la creación, distribución o aplicación de conocimiento.

El trabajo Intensivo en Conocimiento (Knowledge Intensive Work en inglés) es definido por (Wiig, 2004) como una función de diversos factores y la intensidad del conocimiento es una función de cuanto conocimiento y entendimiento una persona tiene que procesar y aplicar cuando se requiere realizar trabajo competente y estar preparado para lidiar con la incertidumbre y sorpresas. Los cuatro factores que intervienen son:

- Nivel y complejidad de conocimiento y entendimiento requerido para lograr el trabajo regular – La cantidad de conocimiento necesario para entregar trabajo competente bajo condiciones normales es parte de la determinación del grado de intensidad de conocimiento.
- Nivel de “expertise” requerido para manejo competente de variabilidad en el trabajo – El grado de intensidad de conocimiento es influenciado por la variedad de retos que una persona debe de estar preparado para manejar competentemente.
- Severidad de consecuencias de errores de trabajo potenciales – La intensidad del conocimiento aumenta cuando las consecuencias de un error aumentan.
- Rapidez de acción – La intensidad aumenta cuando se requieren acciones rápidas como un piloto de combate.

2.2.2 Gestión de talento

La gestión del talento es un área fundamental del nuevo enfoque de la ARH en el cual se le pone atención a áreas más allá que el producto y el cliente, si no también se enfoca la empresa en el estado interno propio. Podemos hacer una analogía con el cuerpo humano, cuando nuestro cuerpo está saludable y todo funciona bien internamente entonces nos vemos y sentimos bien; en cambio, cuando algún órgano no funciona o hay alguna enfermedad nuestro estado de ánimo, las fuerzas y el

rendimiento se ve truncado en gran manera. Así mismo la gestión del talento nos permite asegurarnos que el capital humano dentro de la empresa sea gestionado de forma que cada una de las personas use sus capacidades, pueda aprender nuevas capacidades, ser promovido y ser compensado según su rendimiento.

Este enfoque en la gestión del talento ha dado lugar a clasificar las áreas de la misma. La gestión del talento según (Docherty & Wasdin, 2007) se divide en cinco áreas:

- Reclutamiento y adquisición de talento
- Manejo de aprendizaje
- Manejo de rendimiento de empleados
- Planeación de fuerza de trabajo y sucesión
- Manejo de compensaciones

Al gestionar estas cinco áreas una empresa puede asegurar que el talento que tiene sea retenido e incluso atraer nuevo talento debido a que se vuelve una compañía más atractiva para formar parte de ella. Según (Chiavenato, 2009, p. 19) la administración moderna de recursos humanos consta de seis procesos: Integrar personas, organizar a las personas, recompensar a las personas, desarrollar a las personas, retener a las personas y auditar a las personas. Al aplicar la gestión del talento significa que la empresa busca darle un lugar a los colaboradores donde puedan aprender, crecer, ser evaluados y ser recompensados.

2.2.3 Perfil de conocimiento

Como se ha mencionado anteriormente, el Sistema de Generación Automatizada de “Perfiles de Conocimiento” creará los perfiles que son la salida del sistema, es decir el producto generado por el algoritmo de minería de texto, por lo que es importante elegir el enfoque con el cual definiremos lo que es un perfil de conocimiento.

No existe una definición estandarizada de lo que es un perfil de conocimiento, por lo que se utilizará la definición de (Velazquez Mendoza, 2013) la cual define perfil como “conjunto de rasgos descriptivos que caracterizan la personalidad de un individuo” y perfil de conocimiento como “conjunto de rasgos estructurados que describen el conocimiento requerido, asociado a los recursos y capacidades, que permitan habilitar la generación dinámica de competencias laborales, competencias de los procesos clave y competencias distintivas que generen valor a la organización”.

2.2.4 Sistema de Valoración Difusa de Perfiles de Conocimiento (SVDPC)

El SVDPC busca definir el valor de cumplimiento de individuos con respecto a un rol y observar el nivel de cumplimiento del perfil de manera general y específica de cada uno de los conocimientos analizados en el sistema. Este proceso consta de tres etapas las cuales se describen a continuación (Rosas Daniel, 2015).

- Definición del perfil de conocimiento (perfilación).
- Medición de los niveles de conocimiento de un candidato (evaluación).
- Comparación del perfil de conocimiento del rol con el perfil de conocimiento del candidato obtenido por medio de la evaluación (valoración)

El SVDPC define un valor de cumplimiento del individuo con respecto a un rol utilizando lógica difusa. Este nivel de cumplimiento es acorde al que es definido por el

usuario que evalúa acorde al nivel que requiere por parte de los candidatos. Las calificaciones posibles son: muy sub calificado, sub calificado, calificado, sobre calificado, muy sobre calificado. Así mismo permite observar el nivel con el cual se cumple con el perfil de manera general y particular a cada conocimiento.

2.3 Minería de texto

Antes de profundizar en la minería de texto se requiere definir la minería de datos ya que la primera se deriva de la segunda. La minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados (Witten et al., 2011). Los datos ya existen, usualmente dentro de un almacén de datos o data warehouse, pero la información no ha sido descubierta ya que se requiere de análisis de patrones y diversas técnicas para descubrirla.

En la minería de datos encontramos distintos tipos de datos (Hernández Orallo et al., 2004, p. 9) los cuales pueden clasificarse de la siguiente manera: Datos estructurados, conformados por las bases de datos relacionales. Otros tipos de datos estructurados, se componen de bases de datos, espaciales, temporales, textuales y de multimedia. Datos no estructurados, que son las que provienen de la web y de otros tipos de repositorios de conocimiento.

“La World Wide Web es el repositorio de información más grande y diverso de los existentes en la actualidad. Por ello, hay gran cantidad de datos en la web de los que se puede extraer conocimiento” (Hernández Orallo et al., 2004, pp. 11–12).

Según (Hernández Orallo et al., 2004) la minería web se organiza en torno a tres categorías: Minería de contenido, para encontrar patrones de los datos de las páginas web; minería de estructura, entendido por estructura de URLs e hipervínculos; y minería del uso que hace el usuario de las páginas web (navegación). Así mismo

este autor habla de dos tipos de modelos en la minería de datos: los descriptivos, que son aquellos que buscan obtener el conocimiento ya existente, y los predictivos, que buscan obtener información futura.

La minería de texto es definida por (Feldman & Sanger, 2007) como “un proceso intensivo en conocimiento, en el que un usuario interactúa con una colección de documentos usando una serie de herramientas de análisis”. Las siguientes definiciones del autor anterior nos permiten darnos una idea más precisa de los elementos que existen dentro de la minería de texto.

- **Caracteres:** En el nivel componente es el elemento individual, incluyendo letras, números, espacios y caracteres especiales que construyen los elementos de mayor nivel como, palabras, términos y conceptos.
- **Palabras:** Una palabra específica seleccionada de un documento “nativo” descrito como el nivel básico de riqueza semántica.
- **Término:** Son compuestas por frases de una o varias palabras seleccionadas directamente del cuerpo de un documento nativo por medio de metodologías de extracción de términos.
- **Conceptos:** Son características generadas para un documento por medio de metodologías de categorización, las cuales pueden ser manuales, estadísticas, basadas en reglas, o híbridas.

Dominio y conocimiento del dominio

Según (Feldman & Sanger, 2007) el dominio es un área especializada en la cual pueden ser desarrolladas: ontologías, lexicones, taxonomías de información dedicadas pueden ser desarrolladas. El dominio puede ser amplio como “desarrollo de software”, o específico como “lenguajes de programación de back end”. Así mismo

definen conocimiento de dominio, también conocido en la literatura como background knowledge o conocimiento de trasfondo, puede ser usado en la minería de texto para el preprocesamiento, para mejorar la extracción de conceptos y actividades de validación; ya que “pueden jugar un rol importante en el desarrollo de jerarquías de conceptos más significativas, consistentes y normalizadas”.

La minería de texto sigue el modelo general de la minería de datos el cual se compone de cuatro elementos: tareas de preprocesamiento, operaciones centrales de minería, componentes de la capa de presentación y funcionalidad de navegación, por último técnicas de refinamiento.

- **Tareas de preprocesamiento:** Todas las técnicas, métodos, procesos, enfoques relacionados a la preparación del texto para las operaciones centrales de minería de texto. Incluyendo convertir los datos originales en un formato aceptado antes de aplicar métodos de extracción de *features* para crear una nueva colección de documentos.
- **Operaciones centrales de minería:** Las funciones principales de un sistema de minería de texto incluyendo descubrimiento de patrones, análisis de tendencias y algoritmos de descubrimiento de conocimiento incrementales.
- **Componentes de presentación:** La interfaz gráfica del usuario (GUI, por sus siglas en inglés) y la funcionalidad de navegación de patrones así como el lenguaje para hacer solicitudes. Aquí también se incluyen las herramientas para crear o modificar los criterios de agrupamiento.
- **Técnicas de refinamiento:** Son los métodos que filtran información redundante y agrupan la información que está relacionada. Puede incluir aplicaciones más complejas como aquellas que se utilizan para el descubrimiento de optimizaciones, por ejemplo *supression*, *ordering*, *prunning*, *generalization*, y *clustering*.

2.4 Técnicas de extracción de información.

La extracción de información es descrita por (Meadow, Boyce, Kraft, & Barry, 2007) como un proceso que involucra encontrar información en un depósito de información o una base de datos. Así mismo especifica que la recuperación de información no es lo mismo que extracción de información, a menos que haya existido un proceso de búsqueda y selección.

Los sistemas de extracción de información pueden clasificarse en:

- **Búsqueda de tablas** – Se busca un campo de la tabla y al encontrarlo se extrae la información de los demás campos.
- **Búsqueda parcial en el valor de un campo** – Se busca solo parcialmente y nos permite evitar problemas con el orden en que se encuentran las palabras escritas.
- **Buscar el contenido de un registro por un empate total o parcial** de cualquier grupo de símbolos, usando algebra booleana para especificar los criterios de búsqueda – Esto es lo que se utiliza comúnmente al realizar búsquedas normales o en la world wide web (Meadow et al., 2007).

2.5 Procesamiento de lenguaje natural.

Para el Procesamiento de Lenguaje Natural (PLN) tenemos que conocer los niveles del Lenguaje Natural (LN) los cuales, según (Cortez Vásquez, Vega Huerta, & Pariona Quispe, 2009) son los siguientes.

- **Nivel Fonológico:** las palabras y los sonidos que representan.

- **Nivel Morfológico:** cómo las palabras se construyen a partir de unas unidades de significado más pequeñas llamadas morfemas.
- **Nivel Sintáctico:** las palabras pueden unirse para formar oraciones, fijando el papel estructural que cada palabra juega en la oración.
- **Nivel Semántico:** el significado de las palabras, significado de una oración y al significado independiente del contexto.
- **Nivel Pragmático:** cómo el significado de una oración se ve afectado por las oraciones inmediatamente anteriores.

El PLN es dividido por etapas (Herbrich & Graepel, 2010), las cuales se muestran en la Figura 2.5. Las etapas inician con lo superficial del texto, la primera empieza a procesar el lenguaje para que la computadora (el sistema) pueda procesarlo más no “comprenderlo”. Las últimas etapas involucran la definición del lenguaje, lo que quiere decir el que escribió algún texto, es decir el significado original del autor del texto analizado.

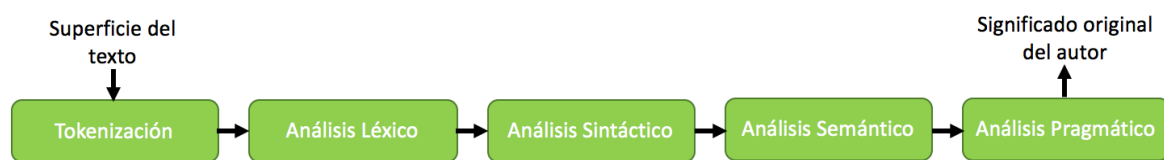


Figura 2.5 Etapas del procesamiento de lenguaje natural.

2.6 Sistemas dedicadas a promover las capacidades de las personas.

Existen diversos sistemas de apoyo para promover a las personas en cuanto a sus capacidades y aptitudes con el objetivo de ofrecer oportunidades laborales. En

este caso nos enfocaremos en los sistemas como portales de empleo en internet. Algunos ejemplos de estos portales son OccMundial, LinkedIn, Indeed, Stack Overflow, GitHub, incluso las bolsas de trabajo en las universidades. A pesar de que cada una de estas plataformas puede tener un objetivo y modelo de negocio diferente (Véase la Tabla 2.1), apoyan a las personas a encontrar empresas donde trabajar y viceversa. Podemos clasificar estos portales o plataformas según su objetivo principal en:

- **Portal de empleo:** Su objetivo principal es conectar a empresas que buscan cubrir una vacante con trabajadores aptos para esa vacante. Este es el caso de OccMundial y Indeed que buscan ofrecer oportunidades de trabajo a sus usuarios.
- **Red social de profesionistas:** Se enfocan en tener una red de contactos donde el perfil de los usuarios tiene la información laboral, por ejemplo: empleo actual, empleos anteriores, educación, habilidades, etc. y por lo tanto sirve para encontrar desde proveedores, clientes e incluso vacantes. Un claro ejemplo de este tipo de red es LinkedIn.
- **Foros:** El énfasis en este tipo de plataforma es servir para debatir acerca de un tema o hacer preguntas sobre un tema en específico. Stack Overflow es un ejemplo de este tipo de plataforma y también ayuda en la búsqueda de empleo a sus usuarios.
- **Repositorios:** Este tipo de plataforma permite guardar el trabajo que se hace en la nube e incluso trabajar en equipo simultáneamente. Un ejemplo de repositorios es GitHub que permite guardar proyectos y compartirlos con las personas.

Sin importar en cuál de las plataformas se trata, al inscribirse en una de estas plataformas, estas piden capturar información en formularios, lo que ocasiona que si el usuario utiliza varias plataformas diferentes tenga que actualizar cada una por separado. Algunas plataformas permiten guardar el curriculum en su base de datos como un documento adjunto pero solo para mostrárselo al empleador, mas no para procesarlo.

Como se mencionó anteriormente, dependiendo del enfoque de la plataforma son las herramientas que proporcionan a los usuarios, sean trabajadores o empleadores. Algunas herramientas buscan ofrecer parte de su funcionalidad gratuitamente, mientras que obtener la funcionalidad completa tiene un costo; otras ofrecen funcionalidad completa de manera gratuita.

Tabla 2.1 Comparación de plataformas de búsqueda de empleo.

Plataforma	Tipo de Plataforma	Tipo de Servicio/Costo	Captura de Datos en Formulario	CV como Archivo Adjunto
GitHub	Repositorio	Gratuito y Paga	Si	No
Indeed	Portal Empleo	Gratuito	Si	Si
LinkedIn	Red Social	Gratuito y Paga	Si	No
OccMundial	Portal Empleo	Gratuito y Paga	Si	Si
Stack Overflow	Foro	Gratuito	Si	No

Restricción: Prohibición de “Web Scrapping” en LinkedIn

Una restricción en cuanto al obtener los curriculums de una plataforma en línea es la siguiente. El “web scrapping” en LinkedIn se considera ilegal ya que viola las condiciones de dicha plataforma. LinkedIn declara que buscan proteger los datos de sus miembros, por lo que hacer uso de este tipo de herramientas puede causar que se restrinja o cierre la cuenta (LinkedIn, 2016). Por lo tanto no es posible extraer los curriculums de los usuarios de esta plataforma sin violar las condiciones de uso de la misma, por lo que solo permiten acceder a dicha información a través de las herramientas propias.

En el presente capítulo se analizaron los temas relevantes a la problemática del sistema así como de la solución de la misma. Donde se incluyó algunas estrategias y herramientas utilizados en RH para la contratación de personal. En el siguiente capítulo se describirá el análisis y diseño del sistema.

Capítulo 3: Análisis y Diseño del Sistema

En este capítulo se presenta el proceso para el diseño del sistema, el cual consta de dos etapas principales. La primera etapa es el análisis, que consiste en levantamiento de necesidades, objetivos, requerimientos y restricciones del sistema. La segunda es el diseño del sistema, que consiste en definir la estructura, el funcionamiento y la comunicación del sistema, es decir su arquitectura.

3.1 Requerimientos del sistema

Los requerimientos del sistema es una base fundamental del análisis de diseño del software, estos nos permiten conocer parámetros que necesitamos cumplir para poder asegurar satisfacer las necesidades o restricciones del sistema a diseñar.

Requerimientos funcionales:

- El sistema debe utilizar técnicas de minería de texto para la generación de los perfiles de conocimiento.
- El sistema debe poder generar el perfil de conocimiento a partir de datos no estructurados.

No funcionales:

- El sistema debe generar el perfil de conocimiento en el mismo tiempo (o menor) que tardaría un trabajador de RH. Este tiempo incluye la lectura del curriculum así como la capturar de sus datos y conocimientos para la creación del mismo.

- Debe de definir correctamente los perfiles de conocimiento el 95% de las veces.

Restricciones:

- El sistema debe ofrecer una salida compatible al sistema de valoración difusa de perfiles de conocimiento (SVDPC) para la generación de perfiles.

Atributos de calidad externo / interno:

- **Usabilidad / Operabilidad:** Debe tener la capacidad de generar 80 perfiles en una hora.
- **Eficiencia / Tiempo límite:** El tiempo límite en la generación de un perfil de conocimiento debe ser el mismo tiempo que tardaría un trabajador de RH en analizar el curriculum y capturar los datos.
- **Usabilidad / Protección contra errores de usuario:** Al aumentar la cantidad de perfiles de conocimiento a generar, se debe mantener la operabilidad y eficiencia.
- **Integración de datos:** El sistema debe proporcionar los resultados de los perfiles de manera que sean compatibles con el SVDPC.
- **Funcionalidad / Correctitud:** El sistema debe acertar en un 90% de los casos en la generación de los perfiles de conocimiento. Realizar pruebas para ver que los resultados siguen siendo consistentes.

Stakeholders:

- Los usuarios del sistema: gente que trabaja en RH y buscadores de talento.

- Las empresas beneficiadas del uso del sistema, gerentes de RH y encargados de las áreas de las cuales se va a contratar.
- Los candidatos a puestos los cuales el sistema usará su curriculum como fuente.
- Los portales de empleo y fuentes de curriculum.
- El sistema de valoración difusa al cual se le mandará la información.
- Director de tesis y asesores

Escenarios (ISO/IEC 25010: 2011):

- **Usabilidad / Operabilidad:** El usuario inicia la generación de perfiles durante la operación normal de sistema, el cual genera los perfiles 30% más rápido que el usuario promedio.
- **Eficiencia / Tiempo límite:** El desarrollador realiza pruebas de los perfiles de conocimiento y el sistema tarda en generar el perfil de conocimiento por lo que debe truncar la generación de dicho perfil y mostrar una notificación del problema.
- **Usabilidad / Protección contra errores de usuario:** El usuario ingresa una dirección invalida para el archivo que es una fuente de texto (CV) y el sistema responde mostrando una notificación al usuario pidiendo revisar el archivo.
- **Compatibilidad / Interoperabilidad:** El sistema termina de generar los perfiles y la salida de los resultados están disponibles para ser utilizados por el sistema de valoración difusa.

Objetivos del negocio:

- ON1 – Crear perfiles de CV automáticamente: El sistema debe generar los perfiles de conocimiento a partir de los CV sin intervención del usuario durante el proceso del sistema.
- ON2 – Visualizar los perfiles por conocimientos: El sistema debe permitir visualizar los resultados en relación a los conocimientos encontrados.
- ON3 – Visualizar los resultados específicos de cada perfil para ver con cuáles conocimientos cumple.

Objetivos específicos:

- OS1 – Se buscará cada conocimiento de la base de conocimiento (ingeniería de software) al analizar el CV utilizado para crear el perfil.
- OS2 – Para la creación del perfil deseado el sistema proporcionará la base de conocimiento del sistema para seleccionar los conocimientos deseados.

Necesidades del cliente o del mercado:

- Se requiere no modificar los archivos originales de los CV solo leerlos.
- Se requiere proporcionar acceso a los resultados para que pueda ser utilizado por otros sistemas.

Riesgos del negocio

- Falta de claridad en la descripción del CV por lo que el sistema excluya conocimientos debido a que no se encuentran en el mismo.

Características del sistema

Las características del sistema nos permite definir comportamiento o funcionamiento que es necesario o deseable en el sistema. Dentro de las características de KP GENERATOR (Véase la Tabla 3.1) se encuentra que el acceso sea a través de un navegador web, que la base de conocimiento sea proporcionada por el mismo sistema, que exista visualización general y específica de los resultados, así como la posibilidad de proporcionar información al SVDPC u otros sistemas de terceros. La columna Prioridad muestra letras donde “A” representa una prioridad mayor y “D” la prioridad menor en cuanto a las características del sistema. Observe que la prioridad mayor la tienen las características que involucran el funcionamiento fundamental, en segundo lugar la visualización, seguido de la visualización específica e interacción con sistemas de terceros.

Diagrama de contexto

El diagrama de contexto es una herramienta que nos permite visualizar las entradas y salidas del sistema, incluyendo otros sistemas con los cuales pudiese interactuar el sistema principal. KP GENERATOR interactúa con un usuario, que, al entregar los parámetros requeridos, se analiza el CV y se consulta la base de conocimiento para crear el perfil de conocimiento basándose en el CV (Véase la Figura 3.1).

Modelado del proceso del negocio

El diagrama de proceso del negocio nos permite visualizar el funcionamiento del sistema al mostrar el flujo de trabajo del mismo. Este diagrama muestra carriles que representan diferentes sistemas (o componentes) así como actores. Ponga atención a los diferentes carriles como lo son el usuario, el CV e incluso KP GENERATOR (Véase Figura 3.2).

*Tabla 3.1 Características del sistema.
Prioridad "A" representa mayor prioridad y prioridad "D" menor prioridad.*

ID	Descripción	Prioridad	Objetivo de Negocio Asociado
CAS1	El sistema debe permitir ingresar el CV desde un navegador web.	A	ON1
CAS2	El sistema debe de permitir procesar el CV en formato de texto libre.	A	ON1
CAS3	El sistema proporcionará la base de conocimiento.	A	ON2
CAS4	El sistema buscara cada conocimiento (ingeniería de software) dentro de la base de conocimiento.	B	ON1
CAS5	El sistema permitirá seleccionar los conocimientos deseados.	D	ON2
CAS6	El sistema debe de permitir visualizar los resultados del procesamiento del CV, es decir el perfil de conocimiento.	A	ON2
CAS7	El sistema debe de permitir guardar el perfil de conocimiento.	B	ON2
CAS8	El sistema debe de permitir acceso a los perfiles guardados.	B	ON2
CAS9	El sistema debe de permitir visualizar los resultados específicos del perfil.	C	ON3
CAS10	El sistema debe mostrar una comparación entre el/los perfiles generados y deseado.	D	ON3
CAS11	El sistema debe de proporcionar acceso al SVDPC y a sistemas de terceros.	D	ON2

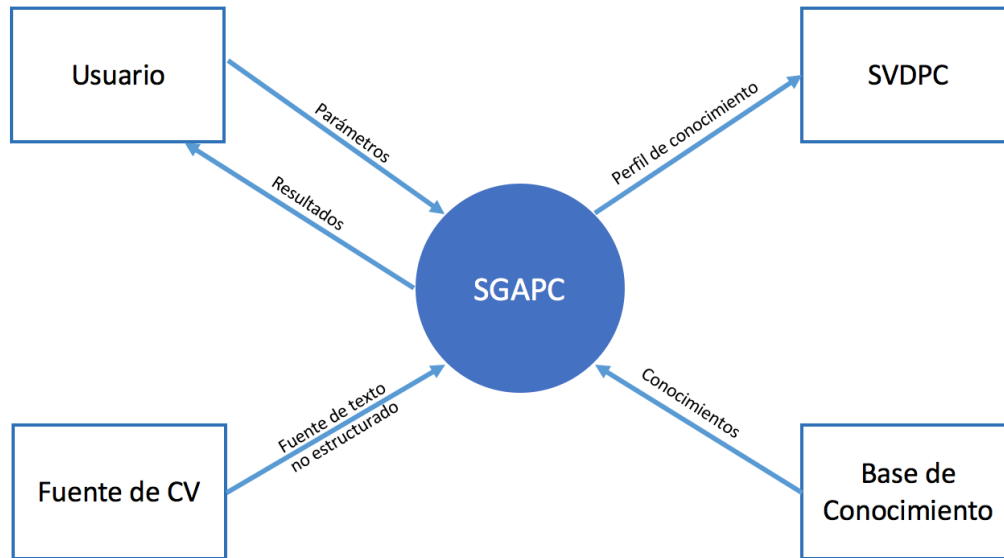


Figura 3.1 Diagrama de contexto de KP GENERATOR.

Diagrama de casos de uso (KP GENERATOR)

El diagrama de casos de uso nos permite visualizar a los actores del sistema y sus funciones. Los actores que intervienen (Véase Figura 3.3) son el usuario del sistema, el cual puede seleccionar el documento (o CV), procesar el mismo, visualizar los resultados y ver los detalles de esos resultados. El segundo actor es el SVPC el cual puede obtener los resultados de KP GENERATOR. La manera de proporcionar estos resultados es a través de un servicio web. La Tabla 3.2 muestra las descripciones de los casos de uso en el cual se puede observar tanto la descripción como las características asociadas al mismo.

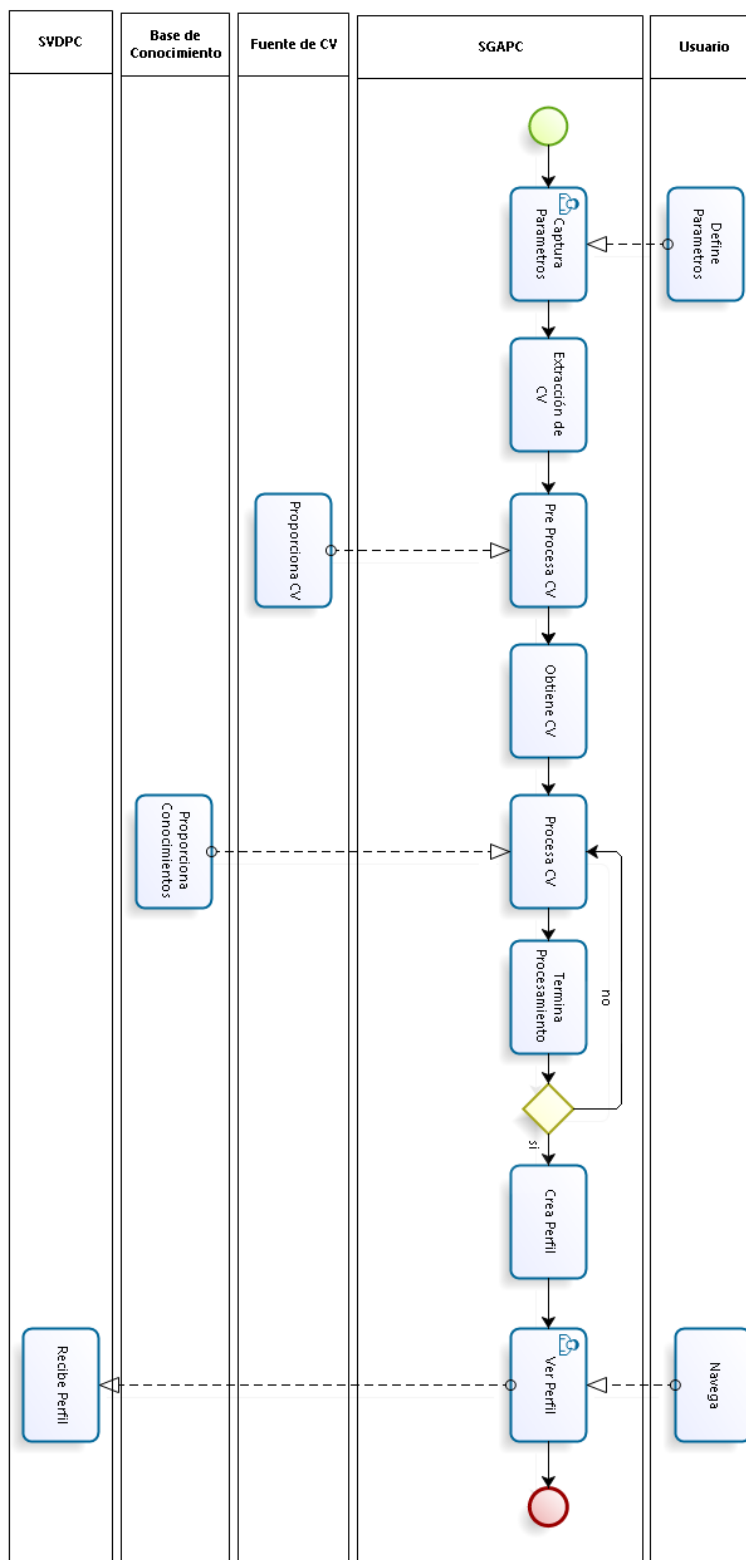


Figura 3.2 Diagrama de proceso de negocio.

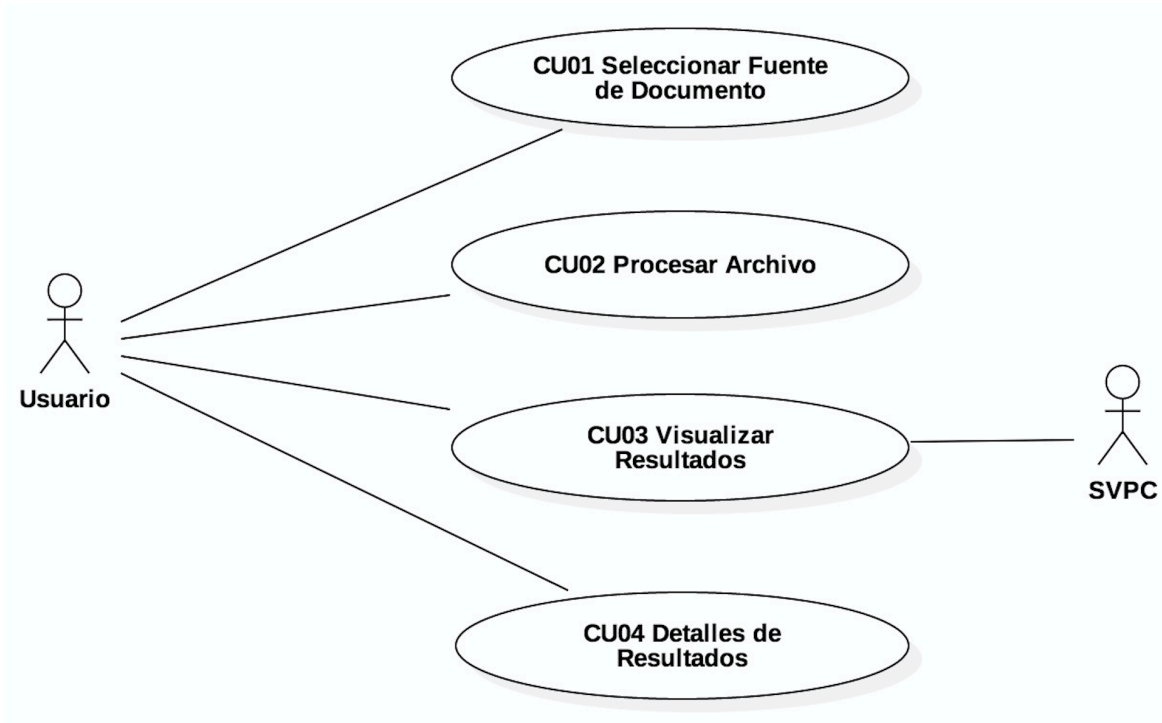


Figura 3.3 Diagrama de casos de uso.

Tabla 3.2 Descripción de casos de uso.

ID	Descripción	Características asociadas
CU01	Permite seleccionar la dirección del documento que fungirá como fuente de texto no estructurado en el cual se encuentra el CV.	CAS1, CAS2
CU02	Permite extraer el archivo de CV de la dirección especificada y procesarlo para crear el perfil de conocimiento.	CAS3, CAS4
CU03	Permite ver los resultados de la creación del perfil, los cuales pueden filtrarse por conocimientos para ver los que cumplen solo con ciertos conocimientos.	CAS5, CAS6, CAS7, CAS8, CAS10, CAS11
CU04	Permite visualizar a detalle el perfil de conocimiento de un candidato para observar en detalle el nivel de cumplimiento.	CAS5, CAS7, CAS8, CAS9

3.2 Arquitectura del sistema

En esta sección se describirá la arquitectura de KP GENERATOR. La arquitectura nos define la estructura del sistema con el propósito de organizar los componentes del sistema y facilitar su implementación. Como resultado, esto permite tener una base para la implementación del sistema y permitir la escalabilidad del mismo. KP GENERATOR busca servir como apoyo a la selección de personal en las empresas desarrolladoras de software, por lo que su arquitectura busca facilitar este propósito así como permitir la escalabilidad del sistema.

Para explicar la arquitectura general de un sistema de minería de texto podemos iniciar explicando el proceso de un sistema de minería de texto. Iniciamos con la fuente de texto no estructurado al cual se le hace un pre procesamiento de datos para obtener una colección de documentos. La colección de documentos es alimentada al algoritmo de minería de texto para analizar los documentos utilizando una base de conocimiento. Una vez que es procesada esta colección de documentos, se utilizan técnicas de refinamiento para mejorar la precisión del algoritmo y por lo tanto de estos resultados. Por último, tenemos la navegación, que es la que le permite al usuario visualizar los resultados obtenidos (Véase Figura 3.4).

Distribución general de la arquitectura

La distribución de la arquitectura nos determina la forma en que los componentes de un sistema se reparten en el entorno. KP GENERATOR tiene una arquitectura web distribuida, la cual se separa en Cliente, Servidor Web y Servidor de base de datos. Esta arquitectura fue seleccionada con el objetivo de facilitar su distribución y uso por los clientes (usuarios) alrededor del mundo. El diagrama de despliegue nos muestra dicha distribución de KP GENERATOR (Véase Figura 3.5).

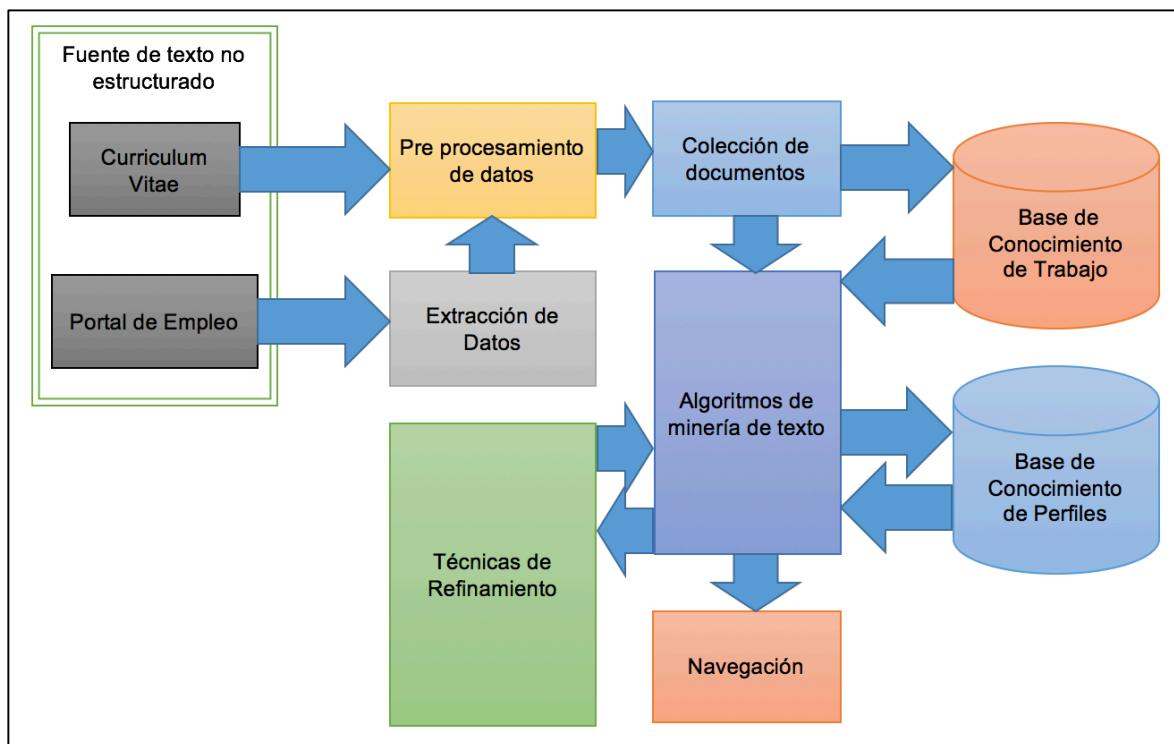


Figura 3.4 Arquitectura genérica de un sistema de minería de texto.
Adaptado de (Feldman & Sanger, 2007, p. 17).

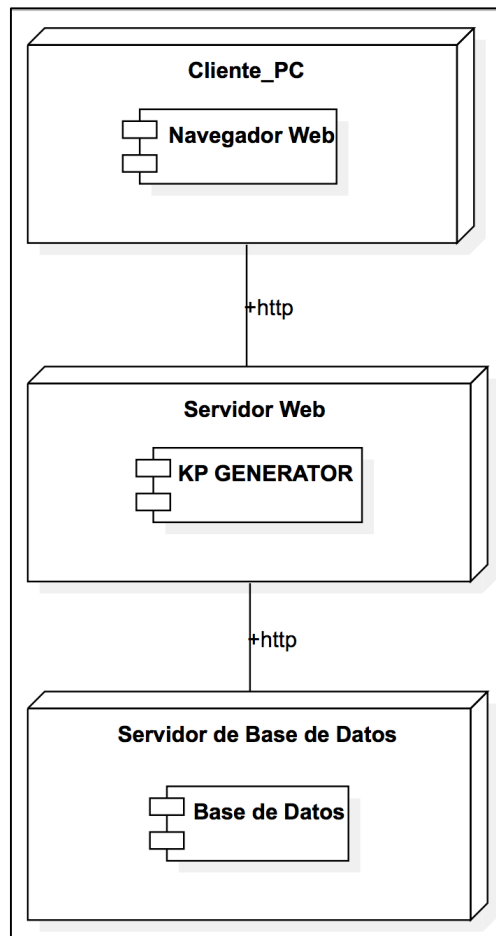


Figura 3.5 Diagrama de despliegue.

Componentes del sistema

KP GENERATOR contiene los cuatro componentes generales de un sistema de minería de texto, los cuales se pueden observar en el diagrama de componentes (Véase Figura 3.6). Los componentes son: preprocesamiento, minería de texto, refinamiento, y presentación.

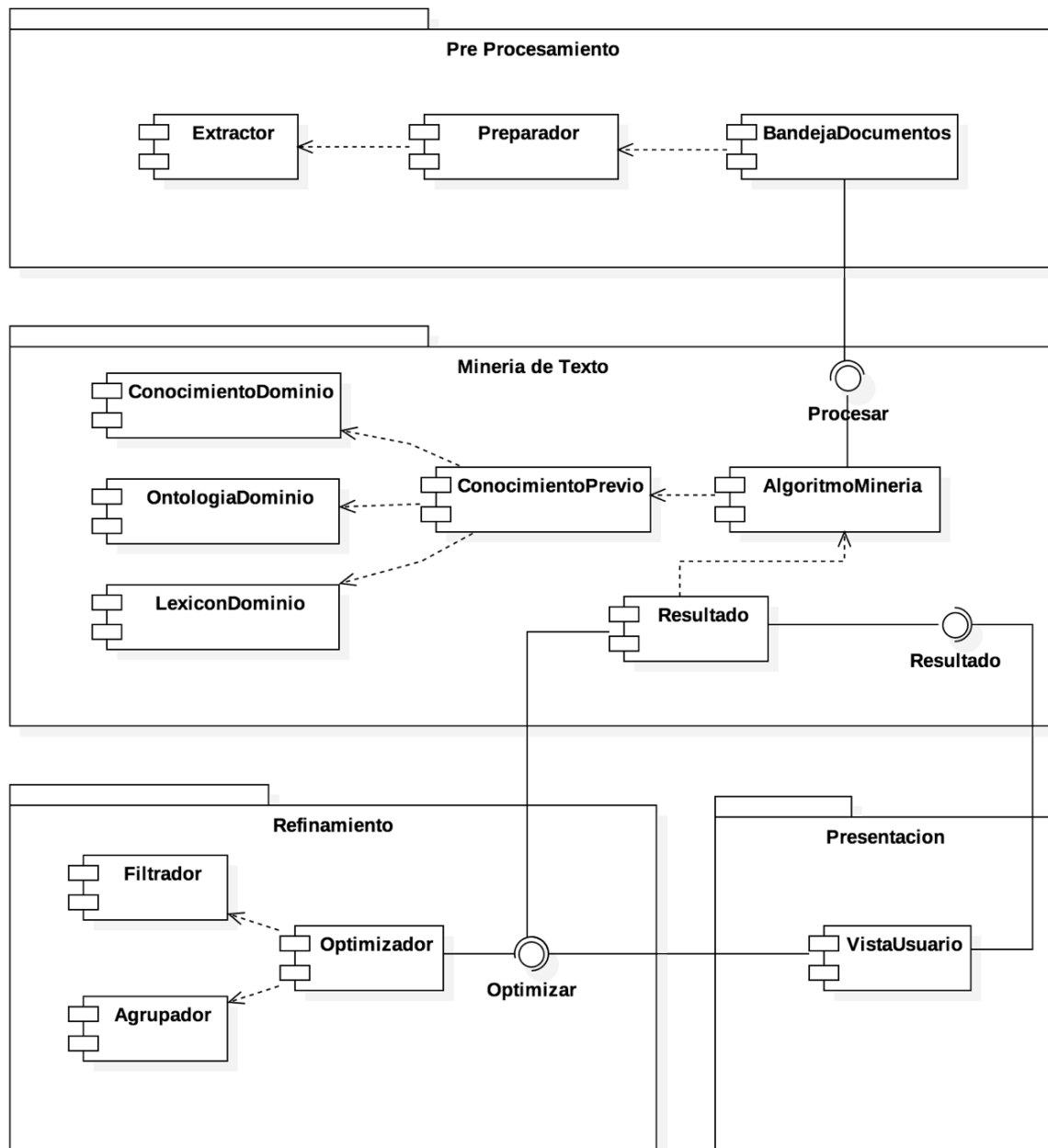


Figura 3.6 Diagrama de Componentes de KP GENERATOR.

- **Preprocesamiento:** Involucra todas las tareas previas de preparación del CV del candidato antes de iniciar a generar el perfil de conocimiento a partir del texto extraído del documento.
- **Minería de texto:** Es todo lo relacionado en el procesamiento del texto con el objetivo de crear el perfil de conocimiento incluyendo la base de conocimiento y el procesamiento de lenguaje natural.
- **Refinamiento:** Se encarga de los pequeños ajustes que nos permiten mejorar los resultados del procesamiento de los CV al obtener el perfil de conocimiento.
- **Presentación:** Es la interfaz con el usuario donde este interactúa con el sistema y puede visualizar los resultados.

Primero se prepara el texto a una estructura canónica para el procesamiento de lenguaje natural y la minería. Luego el procesamiento inicia cuando KP GENERATOR busca el conocimiento técnico en el texto y comienza a crear el perfil de conocimiento del candidato. Los resultados de este proceso se muestran en la interfaz gráfica cuando el usuario interactúa con el sistema. Por último el postprocesamiento es donde el usuario puede filtrar o agrupar la información de los perfiles de conocimiento para encontrar los resultados deseados según las necesidades del puesto.

La técnica de minería de texto utilizada por KP GENERATOR para crear los perfiles de conocimiento es categorización de texto (conocida en inglés como *text categorization* o *text classification*). La categorización de texto es definida por (Feldman & Sanger, 2007) como la tarea de clasificar una instancia de datos a categorías especificadas anteriormente. Para esta técnica de categorización se utilizó un enfoque de ingeniería de conocimiento (knowledge engineering) en el cual se crean reglas de clasificación manualmente para lograr el objetivo. Las reglas son definidas por un experto en el dominio, las cuales tienen condiciones suficientes para que un documento sea etiquetado con una categoría. Los datos de entrada al sistema son los

curriculums, las categorías pre especificadas son los conocimientos técnicos, el dominio es ingeniería de software y el proceso es encontrar los conocimientos técnicos correctos de cada curriculum.

Resumiendo todo lo anterior, KP GENERATOR contiene los cuatro componentes (y por lo tanto procesos) de un sistema de minería de texto. Cada uno de los cuales son fundamentales para la creación del perfil de conocimiento. El proceso general de KP GENERATOR es el siguiente (Véase la **Error! Reference source not found.**):

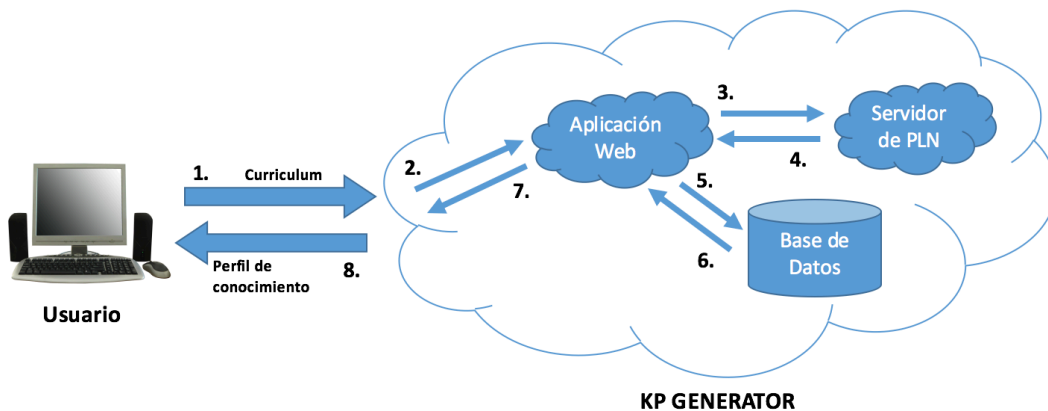


Figura 3.7 Esquema General de KP GENERATOR.

1. El usuario registra el curriculum en el sistema.
2. El usuario carga el curriculum a la aplicación web. El curriculum es preprocesado y preparado para ser enviado al Servidor de PLN.
3. El servidor recibe la solicitud e inicia el procesamiento del curriculum.
4. Una vez terminado el procesamiento del curriculum, el servidor regresa los conocimientos encontrados a la aplicación web.

5. La aplicación recibe los conocimientos encontrados y crea el perfil de conocimiento del curriculum; el curriculum y el perfil de conocimiento son almacenados en la base de datos.
6. Al navegar para visualizar los resultados, los perfiles de conocimientos se extraen de la base de datos.
7. La aplicación web muestra los resultados obtenidos.
8. El usuario obtiene el producto final, es decir, el perfil de conocimiento.

En el presente capítulo se describió el análisis y diseño del sistema de generación automatizada de perfiles de conocimiento conocido como KP GENERATOR, incluyendo su arquitectura. Se enlistaron las necesidades, los requerimientos y las restricciones del sistema, así como la estructura y funcionamiento general del mismo. En el siguiente capítulo se describirá la implementación del prototipo de KP GENERATOR.

Capítulo 4: Implementación del Sistema

En este capítulo se presenta la implementación de KP GENERATOR. Así mismo, se describirán las herramientas utilizadas para el desarrollo del mismo. El propósito de la implementación de KP GENERATOR es demostrar la factibilidad del sistema como apoyo a la selección de personal dentro de las empresas. Algunos de los diagramas, tablas, etiquetas, texto en las páginas o código puede estar en el idioma inglés debido a que se optó por este idioma por ser internacional, tener más alcance en cuanto al área de ingeniería de software, todo esto, con el propósito de maximizar el impacto del sistema.

4.1 Modelos y estilos del prototipo

Para el desarrollo del sistema se utilizó el modelo en cascada el cual nos permite desarrollar por etapas las cuales son: análisis de requerimientos, diseño del sistema (ambas descritas en el capítulo anterior) implementación (en el presente capítulo), verificación y mantenimiento. Esto nos permite trabajar de manera metodológica y ordenada para asegurar que se cumplan los objetivos del sistema. En caso de encontrar fallas en alguna etapa, es posible regresar a la etapa o etapas anteriores para hacer los ajustes necesarios e ir haciendo esas mejoras y llevando los cambios a las etapas posteriores.

KP GENERATOR fue concebido como una aplicación web para facilitar su distribución, esto nos permite hacer actualizaciones inmediatas sin necesidad de actualizar cada ordenador que utilice el sistema como es el caso de las aplicaciones que requirieren instalarse. Para la implementación del prototipo se utilizó el patrón arquitectónico Modelo-Vista-Controlador (MVC). Este patrón, de manera burda,

consiste en separar el acceso a datos (modelo), la lógica del negocio (controlador) y la interacción con el usuario (vista). Todo esto, con el objetivo de obtener componentes desacoplados que facilitan la implementación, actualización y mantenimiento del sistema. Para la implementación de este patrón se utilizó el Yii2, el cual es un framework en PHP. Yii2 se basa en este patrón arquitectónico (MVC) separando estos tres componentes y al mismo tiempo ofrecer herramientas que ayudan a hacer la implementación más fácil. La Figura 4.1 muestra la página principal de KP GENERATOR en la cual se pueden observar las pestañas de navegación las cuales nos permiten probar el prototipo (lo cual describiremos más adelante). Observe en la parte inferior derecha el logo de Yii que es el framework utilizado para el desarrollo de KP GENERATOR.

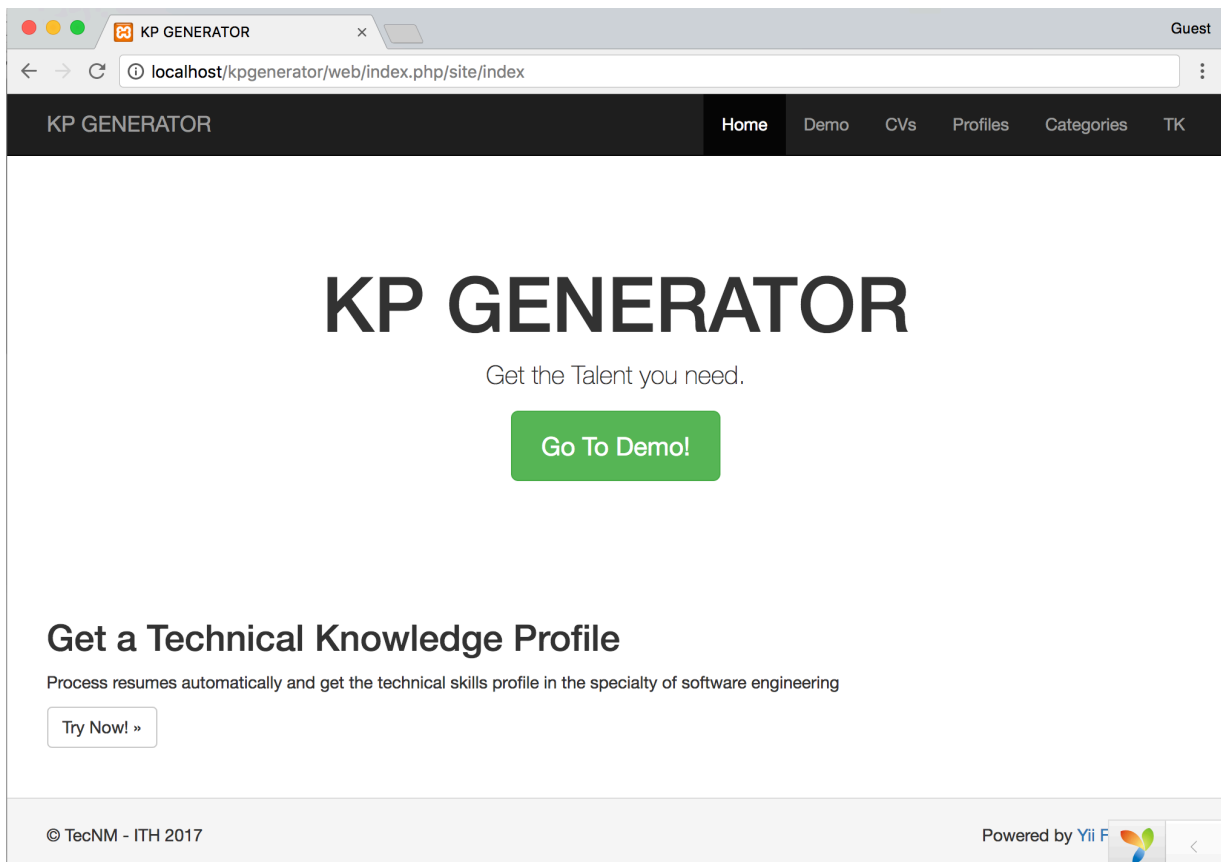


Figura 4.1 Página principal de KP GENERATOR.

KP GENERATOR está escrito en PHP. La razón por la cual se optó por este lenguaje es mantener la homogeneidad entre el SVDPC y KP GENERATOR ya que aunque usan diferentes frameworks y estilos de programación, estar escritos en el mismo lenguaje de programación permite facilitar la lectura de los sistemas, así como el mantenimiento y las actualizaciones. Otra razón para usar PHP es que es multiplataforma, por lo que se puede desarrollar e implementar desde cualquier sistema operativo y usar desde (prácticamente) cualquier navegador web.

4.2 Procesamiento de lenguaje natural

La herramienta utilizada para el Procesamiento de Lenguaje Natural (PLN) en la implementación de KP GENERATOR es la librería Stanford CoreNLP (Manning et al., 2014). Esta es una librería de procesamiento de lenguaje natural creada por el Grupo de Procesamiento de Lenguaje Natural de Stanford (Stanford Natural Language Processing Group), el cual es parte del Laboratorio de Inteligencia Artificial de Stanford (Stanford AI Lab). CoreNLP fue seleccionado por ser una de las librerías más completa, con buen soporte y documentación, así como por ser gratuito y tener licencia de código abierto.

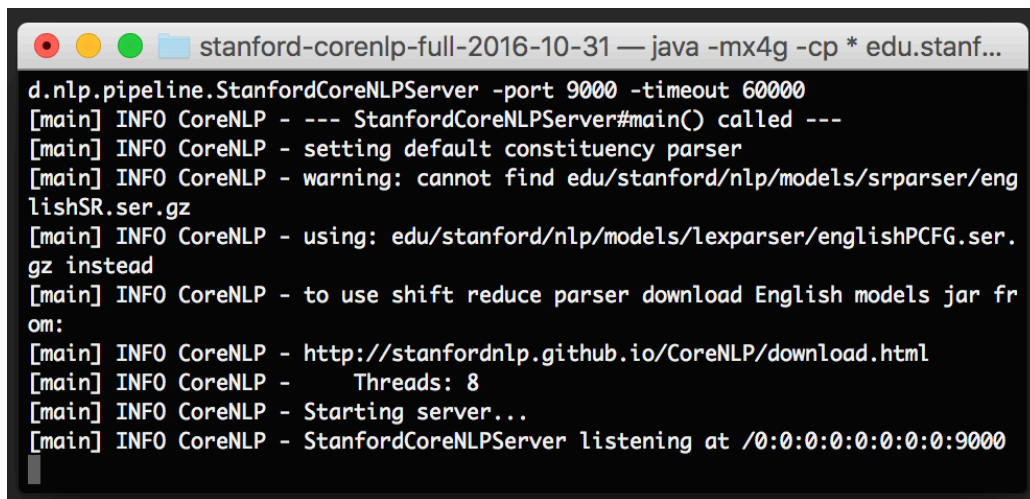
CoreNLP provee herramientas que permite analizar el lenguaje humano en diferentes idiomas, puede encontrar la forma base de las palabras, su categoría gramatical, los nombres propios y entidades, normalizar fechas, tiempo y cantidades numéricas, marcar las dependencias sintácticas y la estructura de las frases, entre otras características. CoreNLP procesa el texto usando anotadores dentro de su flujo de trabajo. Los anotadores (annotators en inglés) son los diferentes tipos de marcas o anotaciones que CoreNLP pone en las palabras como parte del procesamiento del texto, en este caso, los curriculums. Esta es una de las grandes ventajas de utilizar CoreNLP, en vez de utilizar un enfoque al nivel de caracteres, se trabaja con estas anotaciones, y como consecuencia mejorando las posibilidades y riqueza del resultado del procesamiento. Todo esto por trabajar en este nivel de abstracción más alto

utilizando PLN y las anotaciones. Existen diferentes tipos de anotaciones o anotadores en CoreNLP, los principales utilizados son los siguientes:

- tokenize: tokeniza el texto, guarda la posición de los caracteres, y ayuda a manejar texto razonablemente ruidoso o web. La tokenización puede ser considerada el primer paso de todo el procesamiento de lenguaje natural.
- ssplit: separa el texto tokenizado por oraciones.
- pos: Etiqueta los tokens según su categoría gramatical (part of speech en inglés).
- lemma: Genera la forma base o lema de todos los tokens.
- ner: Proviene de Reconocedor de Nombre de Entidades o Name Entity Recognizer (NER) en inglés. Este anotador ayuda en la identificación de entidades como personas, lugares, otros así como entidades numéricas como dinero, fechas y horarios.
- regexNer: NER de expresiones regulares permite al usuario incorporar anotadores personalizados usando reglas sencillas que pueden no estar presentes en el cuerpo de reglas de CoreNLP pero que son fáciles de reconocer.

Estos anotadores son solo algunos con los que cuenta CoreNLP y son los utilizados en KP GENERATOR. CoreNLP consiste en varios paquetes de software entre ellos StanfordTokensRegex (Chang & Manning, 2014). El cual es un framework para definir patrones sobre los tokens así como sus atributos y etiquetas. A partir de la versión 1.3.2 se agregó la funcionalidad para agregar anotadores de TokensRegex personalizados, los cuales son utilizados para crear las reglas de KP GENERATOR.

A partir de la versión 3.6.0, publicada a finales del año 2015, se agregó la funcionalidad para utilizar CoreNLP como un servidor (Véase Figura 4.2). Esto nos permite tener CoreNLP como un módulo independiente del resto de la aplicación web con la posibilidad de montar un servidor dedicado para el PLN con su propio puerto para conexiones y recibir solicitudes. Montar CoreNLP server como un servidor dedicado tiene la ventaja de que la velocidad del procesador, memoria y todos los demás recursos del mismo sean exclusivamente para el PLN y no para otras tareas o aplicaciones.

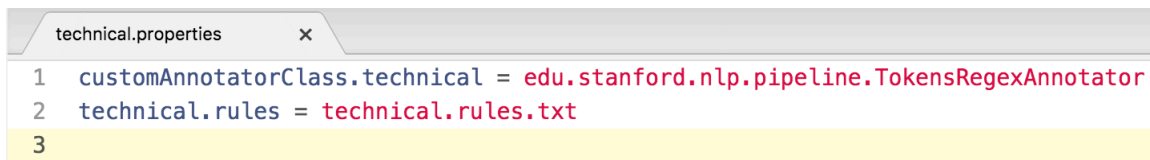


```
d.nlp.pipeline.StanfordCoreNLPServer -port 9000 -timeout 60000
[main] INFO CoreNLP - --- StanfordCoreNLPServer#main() called ---
[main] INFO CoreNLP - setting default constituency parser
[main] INFO CoreNLP - warning: cannot find edu/stanford/nlp/models/srparser/englishSR.ser.gz
[main] INFO CoreNLP - using: edu/stanford/nlp/models/lexparser/englishPCFG.ser.gz instead
[main] INFO CoreNLP - to use shift reduce parser download English models jar from:
[main] INFO CoreNLP - http://stanfordnlp.github.io/CoreNLP/download.html
[main] INFO CoreNLP - Threads: 8
[main] INFO CoreNLP - Starting server...
[main] INFO CoreNLP - StanfordCoreNLPServer listening at /0:0:0:0:0:0:0:9000
```

*Figura 4.2 CoreNLP Server.
Servidor listo para recibir solicitudes*

CoreNLP está escrito en el lenguaje de programación Java. Así mismo CoreNLP server proporciona un cliente java llamado “StanfordCoreNLPCClient” el cual se utiliza cuando la aplicación que hace solicitudes al servidor está programado en este mismo lenguaje. En caso contrario se requiere de un adaptador (interfaz) para conectarse al servidor. CoreNLP tiene una lista de adaptadores según el lenguaje de programación. En el caso del lenguaje de programación PHP, el cual es el lenguaje de programación de KP GENERATOR, se utilizó el adaptador “php-stanford-corenlp-adapter” creado por (Swart, 2016). Este adaptador el cual llamaremos “adaptador original” solo soporta funciones básicas de CoreNLP, por lo que se modificó para crear

lo que llamaremos el “nuevo adaptador” para poder utilizar los archivos personalizados con el anotador TokensRegex (Véase Figura 4.3), que fueron creados para encontrar el conocimiento técnico a través de reglas. Como se había mencionado anteriormente, esto es parte del enfoque de Knowledge Engineering para la categorización del texto basándose en el uso de reglas. Un ejemplo de estas reglas se muestra en la Figura 4.4 donde se muestra un fragmento de las mismas.



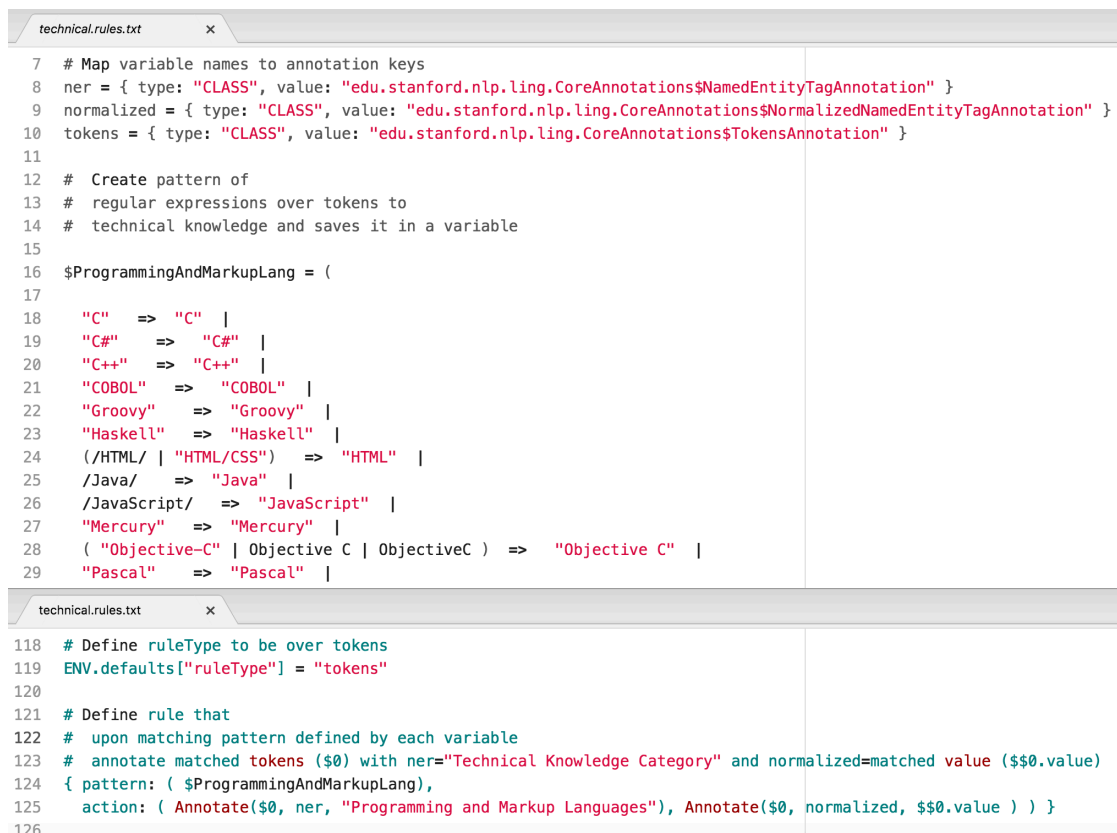
```

1 customAnnotatorClass.technical = edu.stanford.nlp.pipeline.TokensRegexAnnotator
2 technical.rules = technical.rules.txt
3

```

Figura 4.3 Archivo de propiedades.

Se agrega un anotador personalizado llamado “technical” al anotador TokensRegex.



```

7 # Map variable names to annotation keys
8 ner = { type: "CLASS", value: "edu.stanford.nlp.ling.CoreAnnotations$NamedEntityTagAnnotation" }
9 normalized = { type: "CLASS", value: "edu.stanford.nlp.ling.CoreAnnotations$NormalizedNamedEntityTagAnnotation" }
10 tokens = { type: "CLASS", value: "edu.stanford.nlp.ling.CoreAnnotations$TokensAnnotation" }
11
12 # Create pattern of
13 # regular expressions over tokens to
14 # technical knowledge and saves it in a variable
15
16 $ProgrammingAndMarkupLang = (
17
18 "C" => "C" |
19 "C#" => "C#" |
20 "C++" => "C++" |
21 "COBOL" => "COBOL" |
22 "Groovy" => "Groovy" |
23 "Haskell" => "Haskell" |
24 (/HTML/ | "HTML/CSS") => "HTML" |
25 /Java/ => "Java" |
26 /JavaScript/ => "JavaScript" |
27 "Mercury" => "Mercury" |
28 ( "Objective-C" | Objective C | ObjectiveC ) => "Objective C" |
29 "Pascal" => "Pascal" |

```

```

118 # Define ruleType to be over tokens
119 ENV.defaults["ruleType"] = "tokens"
120
121 # Define rule that
122 # upon matching pattern defined by each variable
123 # annotate matched tokens ($0) with ner="Technical Knowledge Category" and normalized=matched value ($0.value)
124 { pattern: ( $ProgrammingAndMarkupLang),
125   action: ( Annotate($0, ner, "Programming and Markup Languages"), Annotate($0, normalized, $0.value ) ) }
126

```

Figura 4.4 Archivo de reglas.

“Technical” es nombre del anotador personalizado.

El archivo de reglas de KP GENERATOR y del anotador personalizado es “technical” (nombrado así por *technical knowledge* o conocimiento técnico). Este anotador es el que nos permite extraer el conocimiento técnico del curriculum para poder crear el perfil de conocimiento. La manera de lograr lo anterior es que al cumplirse ciertas reglas en el texto, las palabras reciben una etiqueta conforme a las mismas. Lo que permite conocer las palabra de interés, y al finalizar el análisis del texto, se obtiene un conjunto de las mismas que nos permiten obtener los conocimientos encontrados en el curriculum y crear el perfil de conocimiento.

4.3 Definición y clasificación del conocimiento técnico

Para poder crear el perfil de conocimiento de los candidatos es necesario definir el área o tipo de conocimiento técnicos que se buscarán en el curriculum. El tipo de conocimiento técnico a extraer (definido en el primer capítulo del presente documento) es del área de ingeniería de software. Esta área tiene una gran cantidad de conocimientos técnicos que sigue creciendo y evolucionando constantemente, por lo que incluir todos estos conocimientos no está dentro de los objetivos de la implementación del prototipo. Por lo anterior, se realizó una selección de algunos conocimientos técnicos para ser utilizados por el prototipo en la búsqueda dentro de los curriculum. Esta lista no pretende estar completa, si no ser una lista no exhaustiva de algunos conocimientos técnicos (Véase la Figura 4.5) en el área de ingeniería de software.

Para facilitar la gestión de los conocimientos técnicos se clasificaron los conocimientos técnicos en ocho categorías. Estas categorías clasifican los conocimientos técnicos según el origen del mismo. Las categorías abarcan desde lenguajes de programación, bases de datos e incluso frameworks para realizar pruebas. Las categorías de clasificación del conocimiento técnico son las siguientes:

- **Programming and Markup Languages / Lenguajes de programación y de marcado:** Los conocimientos técnicos que son lenguajes utilizados para escribir aplicaciones de software.
- **Integrated Development Environments / Entorno de Desarrollo Integrado (IDEs):** Los conocimientos relacionados con el uso de estos entornos para desarrollar un tipo específico de aplicaciones como Android Studio para desarrollar aplicaciones móviles de los celulares que tienen el sistema operativo Android, o en el caso de Xcode, para el desarrollo de aplicaciones para los diferentes dispositivos Apple como iOS o macOS por mencionar algunos.
- **Back End Frameworks / Frameworks de Back End:** Conocimientos técnicos sobre frameworks que se relacionan a la lógica de la aplicación y los procesos que crean la funcionalidad de la misma.
- **Front End Frameworks / Frameworks de Front End:** Conocimientos técnicos sobre frameworks relacionados a la parte visual de las aplicaciones así como la interacción con el usuario.
- **Unit Testing Frameworks / Frameworks de Pruebas Unitarias:** Conocimiento técnico relacionado al uso de frameworks para realizar pruebas en aplicaciones.
- **Relational Database Management Systems / Sistemas Manejadores de Bases de Datos:** Conocimiento técnico relacionado al uso de manejadores de bases de datos relacionales (tradicionales).
- **Non-Relational Databases / Bases de Datos no Relacionales:** Conocimiento técnico relacionado al uso de bases de datos con enfoques diferente a el tradicional (relacional), como bases de datos de documentos.

- **Data Management Frameworks / Frameworks de Manejo de Datos:** Conocimiento técnico relacionado al uso de frameworks que ayudan en el acceso a las bases de datos o en el manejo de los mismos.

Programming and Markup Languages		Integrated Development Environments	Back End Frameworks	Front End Frameworks	Unit Testing Frameworks	Relational Database Management Systems	Non-Relational Databases	Data Management Frameworks
C	Objective C	Eclipse	.NET	AngularJS	Arquillian	Apache Hive	Apache Cassandra	Active Record
C#	Pascal	IntelliJ	Akka	Backbone	JUnit	BigQuery	BigTable	Entity Framework
C++	Perl	NetBeans	CodeIgniter	Boilerplate	KarmaJS	DB2	CouchDB	LINQ
COBOL	PHP	Visual Studio	Django	Bootstrap	Jasmine	MariaDB	MongoDB	Medoo
Groovy	Python	Xcode	Laravel	Ember	Mockito	Microsoft Access	Redis	SQLAlchemy
Haskell	Ruby		NodeJS	Foundation	NUnit	Microsoft SQL Server		
HTML	SmallTalk		Play	jQuery	PHPUnit	MySQL		
Java	Swift		Ruby on Rails	Kendo	PyUnit	Oracle		
JavaScript	Visual Basic		Yii	ReactJS	Selenium	PostgreSQL		
Mercury	XML		Zend	Semantic UI	VSUT Framework	SQLite		

*Figura 4.5 Lista de conocimiento técnicos.
Lista no exhaustiva de TK en ingeniería de software.*

4.4 Funcionamiento del prototipo

El prototipo de KP GENERATOR se desarrolló como una aplicación web. El flujo de trabajo de esta aplicación web inicia con su vista principal mostrada anteriormente (Véase la Figura 4.1). Su barra de navegación en la parte superior contiene los elementos Home, Demo, CVs, Profiles, Categories y TK, los cuales se describirán a continuación:

- **Home:** Muestra la página principal de la aplicación y la información general del sistema.

- **Demo:** Es donde el usuario puede probar el sistema introduciendo el texto de un curriculum para iniciar el proceso de crear el perfil de conocimiento a partir del mismo.
- **CVs:** Es la vista que muestra todos los curriculums guardados en la base de datos con la opción de poder visualizarlos a detalle.
- **Profiles:** Es la página donde se puede visualizar la relación entre los conocimientos técnicos (TKs) y los curriculums (CVs), es decir los perfiles de conocimiento.
- **Categories:** Muestra todas las categorías existentes en la base de datos para clasificar el conocimiento técnico con la opción de agregar, modificar o eliminar las mismas.
- **TK:** Aquí se muestran cada uno de los conocimientos técnicos que existen en la base de datos y que son los que se utilizan para relacionarse con los curriculums de los candidatos.

Las vistas de Categories y TK nos permite visualizar las categorías y conocimientos técnicos que son utilizados por KP GENERATOR para crear los perfiles de conocimiento (Véase la Figura 4.6 y Figura 4.7). Estos dos elementos son los catálogos que nos permiten categorizar el perfil de conocimiento al ser creado. Por lo tanto, estas dos vistas nos permiten hacer las cuatro operaciones básicas de bases de datos tanto para las categorías, como para los conocimientos técnicos. Estas operaciones son Create (Crear), Read (Leer), Update (Actualizar) y Delete (Borrar) o CRUD. En el caso de categorías podemos realizar estas solo haciendo cambios desde la vista. Por otro lado, en el caso del conocimiento técnico habría que hacer modificaciones tanto en esta vista como en las reglas utilizadas en el servidor de procesamiento de lenguaje natural ya que si se agrega un nuevo conocimiento en la vista (base de datos) pero no se modifican las reglas existiría un nuevo conocimiento

en la base de datos que nunca sería encontrado ya que no existen las reglas de extracción del mismo.

Home / Categories

Categories

Create Categories

Showing 1-8 of 8 items.

#	Id Category	Category Name	Category Description	
1	1	Front End Frameworks	Client side programming. The visual part of the applications.	
2	2	Back End Frameworks	The server side programming. It involves the business logic of the application.	
3	3	Relational DB Management Systems	The data structures. The model of the application.	
4	4	Programming and Markup Languages	Programming Language is a language used to code a software application.	
5	5	Integrated Development Environments	Integrated Development Environment or IDE is a software suite that consolidates the basic tools developers need to write and test software.	
6	6	Unit Testing Frameworks	Unit Testing is a software testing method.	
7	7	Non-Relational Databases	Non-Relational Database is any database that does not follow the relational model provided by traditional relational database management systems.	

Figura 4.6 Vista de categorías.

Home / Knowledges

Knowledges

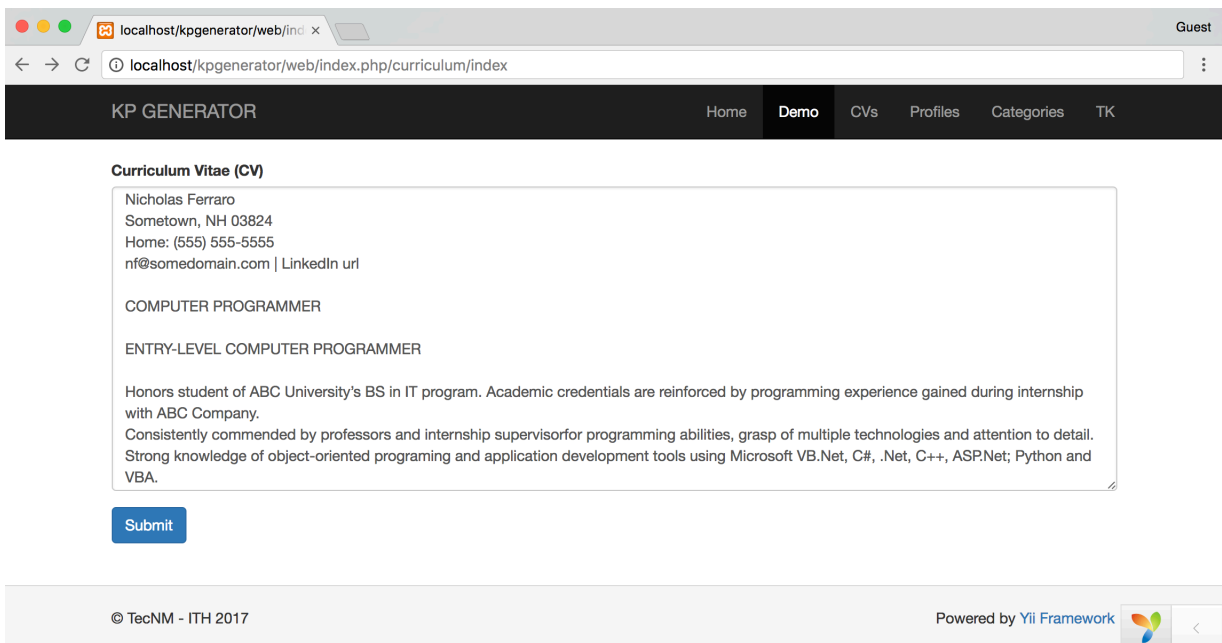
Create Knowledges

Showing 1-20 of 75 items.

#	Id Knowledge	Name	Description	Id Category	
1	1	C	C Programming Language is described here...	4	
2	2	C#	C# Programming Language is described here...	4	
3	3	C++	C++ Programming Language is described here...	4	
4	4	COBOL	COBOL Programming Language is described here...	4	
5	5	Groovy	Groovy Programming Language is described here...	4	
6	6	Haskell	Haskell Programming Language is described here...	4	
7	7	HTML	HTML Programming Language is described here...	4	
8	8	Java	Java Programming Language is described here...	4	
9	9	JavaScript	JavaScript Programming Language is described here...	4	
10	10	Mercury	Mercury Programming Language is described here...	4	
11	11	Objective C	Objective C Programming Language is described here...	4	

Figura 4.7 Vista de conocimientos técnicos.

Para iniciar el proceso de analizar el curriculum y creación del perfil de conocimiento debemos de ir a la pestaña Demo donde aparece una vista con un cuadro de texto donde podemos insertar el texto del curriculum (Véase la Figura 4.8). Una vez insertado el texto se somete el curriculum e inicia el procesamiento del mismo. En esta etapa del prototipo se optó por permitir insertar el texto directamente en la aplicación web en vez de insertar un archivo (en formato PDF, Word, TXT, o dirección web) para facilitar la demostración del sistema y permitir el análisis de los resultados durante cada etapa del procesamiento en el desarrollo del prototipo de KP GENERATOR.



*Figura 4.8 Vista demo.
Inserción del texto del curriculum para procesarlo.*

Una vez que el texto es ingresado al sistema, inicia el preprocesamiento del texto, en esta etapa, el sistema se encarga de preparar el mismo para enviar las solicitudes al servidor de PLN. En esta etapa se incluyen procesos como el filtrado de caracteres especiales y todo lo relacionado con normalizar el texto para cumplir con los requerimientos de solicitudes por parte del servidor de PLN. Una vez terminado el preprocesamiento el adaptador recibe el texto y envía la solicitud o solicitudes (según

la cantidad de texto) al servidor. Al recibir las solicitudes, CoreNLP inicia inmediatamente el procesamiento del texto para obtener los conocimientos técnicos del curriculum. Una vez analizado el texto se obtienen los resultados en una vista de confirmación (Véase la Figura 4.9) que nos muestra todo el texto recibido (después del preprocesamiento) así como el perfil de conocimiento, es decir, los conocimientos técnicos encontrados. Esta vista de confirmación existe en el prototipo con el propósito de mostrar los resultados de ese curriculum y la funcionalidad del sistema, ya que se podría guardar los resultados directamente para mostrar el resultado del análisis de una bandeja completa de curriculums.

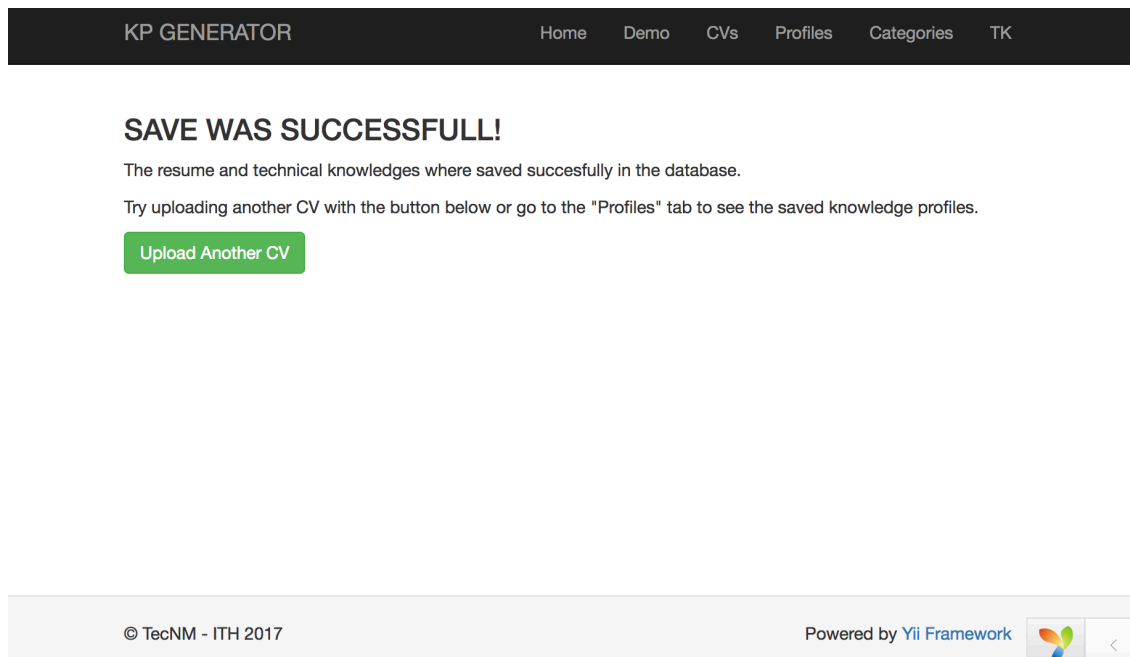
The screenshot shows a web application interface for 'KP GENERATOR'. The navigation bar includes 'Home', 'Demo' (highlighted), 'CVs', 'Profiles', 'Categories', and 'TK'. The main content area is titled 'Technical Knowledge Found:' and displays a grid of 12 yellow boxes, each containing a technical skill: Visual Basic, C, C#, C++, HTML, Java, MySQL, .NET, Oracle, PHP, Python, and Microsoft SQL Server. Below the grid is a green 'Save' button. Underneath, there is a section titled 'Raw Text in CV:' followed by a block of text representing the original curriculum content.

*Figura 4.9 Vista de confirmación.
Muestra el resultado del procesamiento de un curriculum.*

El siguiente paso después de observar los resultados del procesamiento es guardarlos en la base de datos. Al presionar el botón “Save” (Guardar) se procede a almacenar el curriculum y el perfil de conocimiento en la base de datos. Una vez que la información ha sido almacenada se muestra una pantalla de proceso exitoso (Véase la Figura 4.10) donde se muestra un mensaje confirmando que el procedimiento fue

completado exitosamente. Los curriculums guardados se pueden observar en la pestaña CVs.

Al completarse el procesamiento y almacenamiento del curriculum podemos observar los resultados navegando a la pestaña Profiles para ver los perfiles de conocimiento (Véase la Figura 4.11). Esta vista se obtiene de la relación de los conocimiento técnico y los curriculums. Por lo tanto, la vista Profiles nos permite ver cada uno de los conocimientos técnicos que se encontraron en el curriculum. Así mismo, se pueden observar todas las relaciones entre curriculums y conocimientos o se puede filtrar para encontrar los conocimientos de un curriculum específico.



*Figura 4.10 Vista de proceso satisfactorio.
Permite guardar el curriculum y el conocimiento técnico encontrado.*

Home / Profiles

Profiles

Create Profiles

Showing 1-3 of 3 items.

#	Id Cv	CV Description	Knowledge Name
	<input type="text" value="14"/>		
1	14	AMANDA COOPER WEB DEVELOPER RESUME Career Summary An energetic and imaginative young web developer who is able to work alongside other talented IT professionals in creating websites to the very h	HTML
2	14	AMANDA COOPER WEB DEVELOPER RESUME Career Summary An energetic and imaginative young web developer who is able to work alongside other talented IT professionals in creating websites to the very h	JavaScript
3	14	AMANDA COOPER WEB DEVELOPER RESUME Career Summary An energetic and imaginative young web developer who is able to work alongside other talented IT professionals in creating websites to the very h	Perl

© TecNM - ITH 2017 Powered by Yii Framework

*Figura 4.11 Vista de los perfiles.
Se observan cada uno de los conocimientos relacionados con el curriculum con id 14.*

En este capítulo se describió los componentes y explicó la funcionalidad del sistema. Esta funcionalidad incluye la descripción de la vista principal de la aplicación web, el proceso de ingresar el curriculum para ser analizado así como una explicación del procesamiento que se realiza para crear el perfil de conocimiento. Del mismo modo, se procedió a explicar la navegación en la aplicación, por ejemplo: como acceder a los curriculums y perfiles de conocimiento e incluso a las categorías y conocimientos técnicos. El siguiente capítulo explica los resultados obtenidos al utilizar el prototipo de KP GENERATOR.

Capítulo 5: Análisis de Resultados

En esta sección de resultados se describirá el proceso para realizar las pruebas de validación de KP GENERATOR, el cual propone crear perfiles de conocimiento a partir de curriculums en formato libre o no estructurado. Así mismo se describirá el entorno en el cual se realizaron las pruebas, el origen y contenido del texto a analizar para las pruebas, se mostrarán los resultados de una manera clara explicando las variables o puntos más relevantes de las pruebas. Todo esto, con el propósito de comprobar la factibilidad del objetivo de KP GENERATOR mencionado anteriormente.

5.1 Entorno de desarrollo de pruebas

Las pruebas fueron desarrolladas utilizando como servidor web XAMPP versión 5.6.30, el cual es una distribución de Apache que incluye: Apache 2.4.25, MariaDB 10.1.21 y PHP 5.6.30. Se utilizó una computadora MacBook Pro con sistema operativo macOS Sierra versión 10.12, con un procesador Intel Core i7 de 2.5 GHz y 16 Gb de RAM. Así mismo, al servidor de procesamiento de lenguaje natural (CoreNLP server) se le asignaron 4Gb de RAM (de los 16Gb totales de la computadora). Ninguno de los servidores eran dedicados, si no que compartían recursos y se montaron como servidores locales en el ordenador mencionado anteriormente. Por lo tanto los recursos eran compartidos entre los dos servidores y los recursos de programas que utilizaba la computadora. El navegador utilizado para las pruebas fue Google Chrome Versión 58.0.3029.110 (64-bit).

Como se ha mencionado anteriormente, KP GENERATOR hace solicitudes al servidor de PLN. Estas solicitudes incluyen bloques de texto el cual es enviado para ser procesado. En las etapas de preprocesamiento KP GENERATOR divide todo el texto ingresado, es decir, el texto del curriculum y lo divide en bloques de cierta

cantidad de palabras que llamaremos “tamaño de solicitudes”. El tamaño de solicitudes puede ser ajustado para obtener diferentes resultados en el procesamiento de los curriculums. Un tamaño de solicitudes pequeño puede ocasionar muchas solicitudes lo que significa perder tiempo por la gran cantidad de conexiones, pero al mismo tiempo procesamiento rápido de cada solicitud por ser poca cantidad de texto. Por otro lado, tener un tamaño de solicitud grande significa menos conexiones al servidor de PLN por lo que ahorraría tiempo en ese aspecto, aunque significaría tiempo de procesamiento lento por la gran cantidad de texto e incluso saturación del servidor, deteniendo el procesamiento del curriculum.

Tabla 5.1 Tabla comparativa del tamaño de solicitudes.

TAMAÑO DE SOLICITUD	VENTAJAS	DESVENTAJAS
SOLICITUDES PEQUEÑAS	Procesamiento rápido por ser poco texto.	Se pierde tiempo por ser más conexiones al servidor de PLN.
SOLICITUDES GRANDES	Se ahorra tiempo por ser menos conexiones al servidor de PLN.	Se vuelve más lento el procesamiento por ser más complejo debido a la gran cantidad de texto.

5.2 Objetos de prueba

Los objetos de pruebas para crear los perfiles de conocimiento fueron curriculums en formato libre, es decir, texto no estructurado. Estos curriculums fueron tomados de la web, buscando y seleccionando aquellos relacionados con el área de ingeniería de software. Se seleccionaron los primeros cuarenta curriculums encontrados, los cuales forman el cuerpo de curriculums utilizado para las pruebas (Véase la Figura 5.1). Dentro de los objetos de prueba se incluyeron plantillas de

ejemplo de curriculums, los cuales son capaces de simular un curriculum real de un candidato y de demostrar el comportamiento del prototipo de KP GENERATOR.

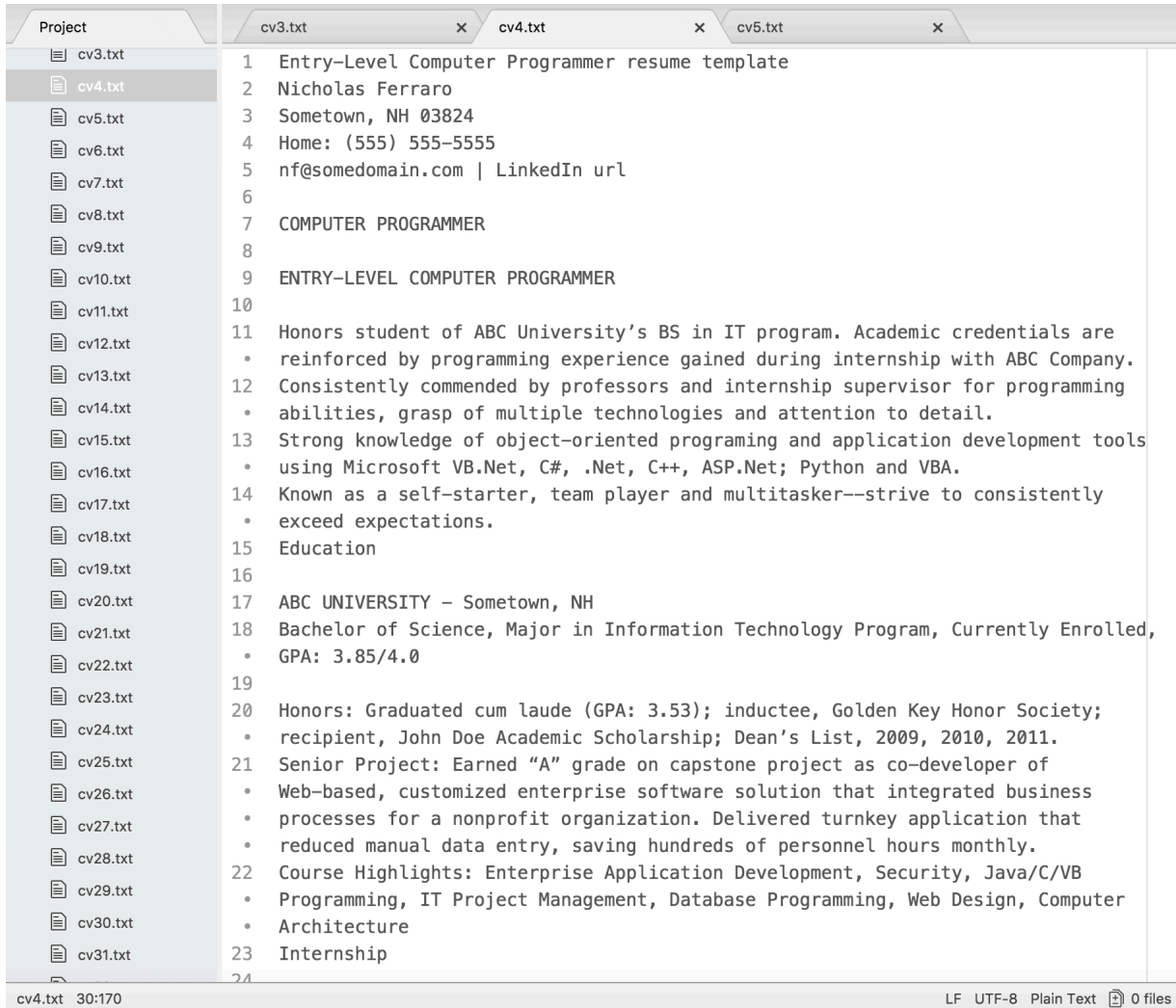


Figura 5.1 Cuerpo de curriculums para pruebas.

5.3 Resultado de pruebas

Los resultados de las pruebas realizadas se obtuvieron del procesamiento de los cuarenta curriculums mencionados anteriormente. Este procesamiento incluye las cuatro etapas del sistema mencionadas en el tercer capítulo del presente documento. Una vez procesados, se guardaron en la base de datos del sistema y se procedió a

analizar los resultados obtenidos. La Figura 5.2 muestra los resultados generales de KP GENERATOR acerca de los conocimientos técnicos encontrados en todo el cuerpo de curriculums. Estos resultados permiten observar la cantidad de veces que los conocimientos técnicos se encontraron. Por lo tanto se puede saber cuáles fueron los conocimientos con mayor frecuencia (que aparecieron en más curriculums) y los de menor frecuencia (aparecieron en una menor cantidad de curriculums) e incluso los de frecuencia cero (no aparecieron en ningún curriculum). Se asume que la razón de esta frecuencia cero se debe a que el presente conocimiento técnico no estaba presente en ninguno de los curriculums.

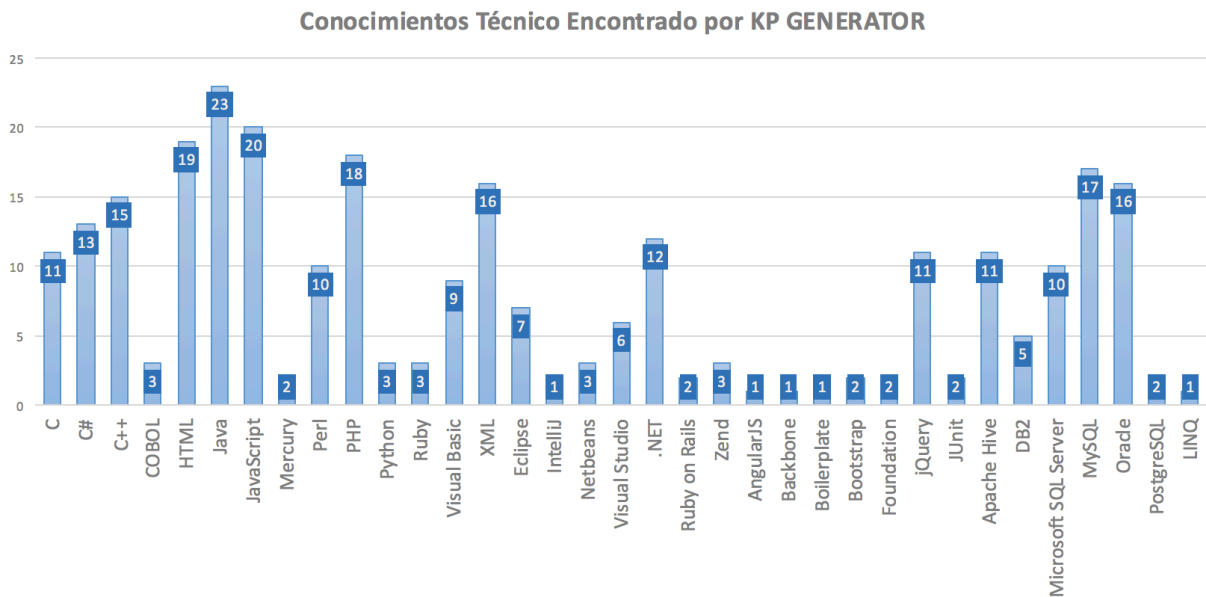


Figura 5.2 Gráfica de resultados de pruebas.

Se puede observar en la Figura 5.2 que el conocimiento técnico con mayor frecuencia fue Java, encontrándose en 23 de los curriculums analizados, seguido de JavaScript con una frecuencia de 20 y HTML con una frecuencia de 19. Así mismo, se observa que algunos de los conocimientos técnicos con menor frecuencia incluyen IntelliJ, AngularJS, Backbone, entre otros. Por último, los conocimientos técnicos con

frecuencia cero son aquellos que se encuentran en la lista de conocimientos técnicos del sistema pero que no aparecen en la figura mencionada anteriormente.

Al analizar el cuerpo de curriculums se registraron el tiempos de procesamiento, el número de solicitudes al servidor, la cantidad de palabras procesadas y las palabras por segundo procesadas. La descripción de estas variables es la siguiente:

- **Tiempo de procesamiento:** Nos permite saber cuánto tiempo duró el procesamiento de ese curriculum en segundos.
- **Número de solicitudes al servidor:** Nos dice cuántas solicitudes se hicieron al servidor de PLN para completar el análisis de ese curriculum.
- **Cantidad de palabras procesadas:** Nos muestra el número de palabras que existen en el curriculum y por ende la cantidad de palabras que se procesaron para crear el perfil de conocimiento.
- **Palabras por segundo procesadas:** Nos muestra la velocidad de procesamiento para ese curriculum calculando cuantas palabras se analizan en un segundo.

La Tabla 5.2 nos muestra las variables mencionadas anteriormente para cada uno de los curriculums. Así mismo, la Tabla 5.3 nos muestra los totales del procesamiento de los curriculums durante la prueba. Cada una de estas variable nos permite tener una referencia del rendimiento y capacidad del prototipo. Esta referencia puede servir para realizar una comparación en el tiempo en analizar los curriculums y capturar el conocimiento técnico del mismo sin utilizar la herramienta contra el tiempo usando KP GENERATOR. Así mismo esta referencia sirve para asegurarse de mejorar el rendimiento del prototipo al hacer modificaciones en el mismo.

Tabla 5.2 Resultados del procesamiento del cuerpo de curriculums.

Número de Curriculum	Tiempo de Procesamiento (segundos)	Número de Solicitudes al Servidor	Cantidad de Palabras Procesadas	Palabras por Segundo Procesadas
1	45.27	9	636	14.05
2	45.41	9	523	11.52
3	25.15	5	369	14.67
4	25.16	5	330	13.12
5	55.32	11	730	13.20
6	95.59	11	806	8.43
7	40.24	8	622	15.46
8	65.37	13	912	13.95
9	75.58	15	1073	14.20
10	116.60	23	1579	13.54
11	70.35	14	879	12.49
12	10.06	2	136	13.52
13	60.37	12	851	14.10
14	40.31	8	500	12.41
15	45.22	9	568	12.56
16	25.52	5	277	10.85
17	20.10	4	252	12.54
18	39.61	5	272	6.87
19	30.14	6	338	11.21
20	27.31	5	273	10.00
21	30.15	6	341	11.31
22	30.65	5	259	8.45
23	37.71	7	350	9.28
24	43.59	5	269	6.17
25	20.11	4	251	12.48
26	25.12	5	253	10.07
27	26.52	5	283	10.67
28	25.11	5	267	10.63
29	60.29	12	754	12.51
30	36.44	6	302	8.29
31	35.25	4	204	5.79
32	48.90	8	584	11.94
33	80.62	15	966	11.98
34	160.79	32	1792	11.15
35	80.57	16	1212	15.04
36	112.48	20	1350	12.00
37	39.39	4	255	6.47
38	40.20	8	562	13.98
39	40.20	8	549	13.66
40	80.42	16	1117	13.89
Promedio	50.33	9.25	596.15	11.61

Tabla 5.3 Totales de procesamiento de la prueba.

Tiempo Total de Procesamiento (segundos)	2013.16
Total de Palabras Procesadas	23846
Promedio de Palabras / Segundo	11.85
Total de Solicitudes al Servidor (65 palabras)	370
Tiempo Promedio Por Solicitud (65 palabras / segundos)	5.44

5.4 Análisis de resultados

Observe como al analizar los resultados globales de los curriculums y obtener los conocimientos técnicos de los candidatos, nos podemos dar una idea del mercado de candidatos y conocimientos que existen para la vacante de una empresa. Se puede saber qué conocimiento tiene una frecuencia baja, y qué conocimiento es común entre los candidatos. Incluso puede darle a la empresa una idea de qué tan competitiva es la vacante para la que está solicitando por su frecuencia (Véase la Figura 5.3). Es decir, si la vacante incluye conocimientos con frecuencias altas, será más fácil cubrir la vacante y si las frecuencias de los conocimientos para la vacante son bajas, se puede esperar tomar más tiempo para encontrar al candidato adecuado para el puesto por estar en búsqueda de conocimientos que son menos comunes.

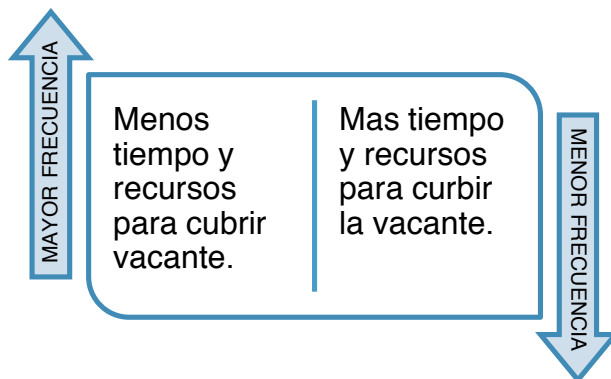


Figura 5.3 Relación de frecuencia vs recursos.

Como se mencionó anteriormente, la velocidad de procesamiento es afectada por diversos factores (como el tamaño de solicitudes al servidor). Otro factor que afecta la velocidad es el tamaño de las oraciones. Entre más grandes son las oraciones, se requiere de más memoria y tiempo para procesarlas. Por lo tanto, la velocidad para procesar un curriculum depende de muchas variantes como es de esperarse en el lenguaje natural. El tiempo promedio de procesamiento de los curriculums durante la prueba fue de 50.33 segundos (Véase la Tabla 5.2). Podemos concluir que en la mayoría de los casos se puede procesar el curriculum en formato libre así como obtener y capturar los conocimientos técnicos del candidato en menos de un minuto. Por lo tanto, se puede concluir que es posible crear perfiles de conocimiento a partir de curriculums en formato no estructurado usando técnicas de minería de texto y procesamiento de lenguaje natural.

En este capítulo se explicó el entorno de las pruebas realizadas, se describió los objetos de pruebas, se mostraron los resultados obtenidos por KP GENERATOR al analizar satisfactoriamente cuarenta curriculums de texto no estructurado y se discutieron los resultados obtenidos. En la siguiente sección se explican los beneficios obtenidos con el presente trabajo, las aportaciones y el trabajo futuro.

Capítulo 6: Conclusiones

Con base en el presente trabajo podemos concluir que es posible crear un sistema que utilizando técnicas de procesamiento de lenguaje natural y técnicas de minería de texto, sirva de herramienta de apoyo a RH dentro de las empresa. Lo anterior, permitiendo la identificación de candidatos aptos para puestos dentro de las empresas.

Los objetivos del trabajo fueron cumplidos satisfactoriamente. Se logró cumplir el objetivo general al implementar KP GENERATOR ya que es una herramienta de software que apoya en la identificación de candidatos aptos para un puesto en empresas de desarrollo de software, mediante la identificación automática de perfiles de conocimiento de los candidatos a puestos. Así mismo se cumplieron cuatro de los cinco objetivos específicos: 1) al estudiar mecanismos de minería de texto e identificación de patrones o conceptos, 2) analizar los tipos de fuentes de datos curriculares 3) diseñar e implementar una herramienta que genere el perfil de conocimiento de los candidatos utilizando la información extraída y 4) analizar los beneficios que podría aportar una herramienta de esta naturaleza. El objetivo específico relacionado a la extracción de curriculum de un portal de empleo (LinkedIn) utilizando alguna técnica como web scrapping no fue cumplido debido a restricciones legales enumeradas al final del segundo capítulo del presente trabajo. No obstante, por los resultados obtenidos al analizar los currículos, y considerando que es el texto disponible el que se analiza, consideramos que la herramienta podría ser también de utilidad para analizar la información disponible en este tipo de portales, siempre que describan habilidades o conocimientos de una persona.

Aportaciones

El presente trabajo aportó una herramienta para apoyar la solución de la problemática, la cual es la contratación de personal dentro de las empresas de desarrollo de software identificando a los posibles candidatos más aptos para puestos dentro de las mismas. Así mismo aportó tres artículos:

- Publicación de artículo en el 6to Simposio Sobre Avances de Investigación en Ingeniería en el Estado de Sonora (AVIIES 2016) titulado “Propuesta de una Arquitectura para un Sistema Generador de Perfiles de Conocimiento”
- Un artículo aceptado en el 5th International Conference in Software Engineering Research and Innovation (CONISOFT 2017), titulado “*Natural Language Processing and Text Mining to Identify Knowledge Profiles for Software Engineering Positions*”.
- Se sometió el artículo para el 7mo Simposio Sobre Avances de Investigación en Ingeniería en el Estado de Sonora (AVIIES en su edición 2017) titulado “Generación Automatizada de Perfiles de Conocimiento Utilizando Procesamiento de Lenguaje Natural y Técnicas de Minería de Texto”.

El presente trabajo también aportó una estancia en Samara, Rusia para el curso de verano aeroespacial de nanosatélites 2016 en la Universidad de Samara de Nanosatélites, donde se tuvo la oportunidad de conocer tanto la historia, como los últimos avances tecnológicos en este rubro. Se cursó un taller de Altium Designer, un software para el diseño de circuitos impresos en PCB. Así mismo, se trabajó en el concepto de una misión aeroespacial con nanosatélites en un equipo con integrantes de diferentes países. Todo esto, permitiendo el aprendizaje y crecimiento en el área aeroespacial.

Aportó un curso intensivo de WEKA, un software de minería de datos, en el Instituto Tecnológico de Nogales. Este curso sirvió para aprender el uso de esta herramienta, así como fortalecer los conocimientos generales de minería de datos.

Trabajo futuro

Existen diversos caminos que pueden tomarse en cuanto al trabajo futuro, los cuales representan áreas de oportunidad para mejorar. Hacer más eficiente el sistema y buscar la mayor velocidad de procesamiento ajustando el tamaño de las solicitudes al servidor de procesamiento de lenguaje natural. Implementar el sistema dentro de una empresa para encontrar los beneficios que se obtienen al utilizar la herramienta. Hacer la comparación entre el trabajador promedio de RH utilizando las herramientas usuales y compararlas con un trabajador de RH utilizando KP GENERATOR dentro de las mismas. Por último, otro trabajo futuro es enfocarse en analizar las disposiciones legales del web scrapping y las plataformas de empleo para: a) encontrar las plataformas que permiten esta técnica en sus páginas b) desarrollar un sistema (subsistema o módulo) que permita extraer el curriculum de esta plataforma para trabajar en conjunto con KP GENERATOR.

Como conclusión final, a partir del presente trabajo se obtiene un sistema que ayuda en la problemática de contratación de personal que, si bien se probó en el área de ingeniería de software, podría ser aplicada prácticamente en cualquier área dentro de las empresas; permitiéndonos identificar candidatos aptos para las vacantes dentro de dichas áreas donde el conocimiento es un factor clave a tomar en cuenta en la contratación. De igual manera, se obtienen diversas publicaciones que registran los avances que se hicieron con el presente trabajo, la oportunidad de continuar con este trabajo para aumentar su impacto, y la experiencia que nos permite seguir innovando y desarrollando sistemas que sirvan para el bien común.

Referencias

- Cantillo, P. (2014). Dinero en Imagen: El 40.4% de desempleados cuenta con estudios profesionales. Recuperado el 7 de octubre de 2015, a partir de <http://www.dineroenimagen.com/2014-08-25/42344>
- Chang, A. X., & Manning, C. D. (2014). *TokensRegex: Defining cascaded regular expressions over tokens*.
- Chiavenato, I. (2009). *Gestión del Talento Humano*. México: Mc Graw Hill.
- CIA World Factbook. (2015). The World Factbook. Recuperado el 6 de marzo de 2017, a partir de <https://www.cia.gov/library/publications/the-world-factbook/geos/mx.html>
- Cortez Vásquez, A., Vega Huerta, H., & Pariona Quispe, J. (2009). Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática*, 6(2), 45–54. Recuperado a partir de <http://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923>
- Cuautle Cardenas, M. de L. (2013). *El Uso del Internet Como un Recurso Didáctico, En Las Licenciaturas del Área de Ingenierías de la Universidad del Valle de México, Campus San Rafael*. Instituto Politécnico Nacional.
- Davenport, T. H. (2005). *Thinking for a Living: how to get better performance and results from knowledge workers*. Harvard Business School Publishing.
- Dessler, G. (1997). *Human Resource Management*. NJ.
- Docherty, B., & Wasdin, M. (2007). *Talent Management For Dummies*. Wiley Publishing, Inc.
- El Informador. (2010). Dificultades para contratar personal. Recuperado el 30 de septiembre de 2015, a partir de <http://www.informador.com.mx/economia/2010/215974/6/dificultades-para-contratar-personal.htm>
- ENDI. (2015). Internet influencia 82 % de jóvenes iberoamericanos. Recuperado el 4 de noviembre de 2015, a partir de

- <https://www.elnuevodia.com/tecnologia/tecnologia/nota/internetinfluencia82dejovenesiberoamericanos-2121662>
- Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook. Imagine*. New York, United States: Cambridge University Press.
- Goldsmith, M., & Carter, L. (2010). *Best practices in talent management: how the world's leading corporations manage, develop, and retain top talent*.
- Herbrich, R., & Graepel, T. (2010). *Handbook of Natural Language Processing*. (N. Indurkha & F. J. Damerau, Eds.) (2nd ed.). Boca Raton: Chapman & Hall/CRC.
- Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la Minería de Datos*. Madrid: PEARSON EDUCACIÓN S.A.
- Hogan Assessments. (2016). Predict Performance. Recuperado el 1 de enero de 2016, a partir de <http://www.performanceprograms.com/hogan-assessments/hogan-assessments-2/>
- Indexmundi. (2011). México Número de usuarios de Internet. Recuperado el 6 de marzo de 2017, a partir de http://www.indexmundi.com/es/mexico/numero_de_usuarios_de_internet.html
- INIDE. (2013). Tiene dificultad 70% de egresados para encontrar trabajo. Recuperado el 20 de septiembre de 2016, a partir de <http://www.inidedelauia.org/2013/09/tiene-dificultad-70-de-egresados-para.html>
- Ivancevich, J. M. (2005). *Administración de Recursos Humanos* (9a ed.). Ciudad de México: McGraw-Hill Interamericana.
- Kaiser Leadership Solutions. (2016). Modern Tools for Modern Leader. Recuperado el 1 de enero de 2016, a partir de <http://kaiserleadership.com>
- LinkedIn. (2016). Prohibition of Scraping Software. Recuperado a partir de <https://www.linkedin.com/help/linkedin/answer/56347/prohibition-of-scraping-software?lang=en>
- Madrigal Torres, B. E. (2009). Capital humano e intelectual : su evaluación. *Observatorio Laboral Revista Venezolana*, 2, 65–81.
- Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of*

- 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. <http://doi.org/10.3115/v1/P14-5010>
- Manpower Group. (2015). Encuesta de escasez de talento 2015.
- Marshall, A. (1947). *Principios de Economía*.
- Meadow, C. T., Boyce, R. R., Kraft, D. H., & Barry, C. (2007). *Text Information Retrieval Systems* (3rd Editio). Academic Press.
- Nonaka, I., & Takeuchi, H. (1997). *Creation with Knowledge in the Company*. Rio de Janeiro.
- Rosas Daniel, J. A. (2015). *Construcción de un modelo de lógica difusa para la validación de perfiles de conocimiento de personal*. Instituto Tecnológico de Hermosillo.
- Swart, D. (2016). PHP adapter for Stanford CoreNLP. Recuperado a partir de <https://github.com/DennisDeSwart/php-stanford-corenlp-adapter>
- Tejada Zabaleta, A. (2003). Los modelos actuales de gestión en las organizaciones. Gestión del talento, gestión del conocimiento y gestión por competencias. *Psicología desde el Caribe, Universidad del Norte No. 12*, 115–133.
- Velazquez Mendoza, M. de J. (2013). *Construcción de un modelo para el diseño de perfiles de conocimiento*. Instituto Tecnológico de Hermosillo.
- Wiig, K. (2004). *People-Focused Knowledge Management : how effective decision making leads to corporate success*. Elsevier Inc.
- Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining: Practical Machine Learning Tools and Techniques. Complementary literature None*.