MODELO PARA LA GESTIÓN DE LECCIONES APRENDIDAS BASADO EN EL PROCESAMIENTO DEL LENGUAJE NATURAL Y APRENDIZAJE AUTOMÁTICO.

JORGE ESTEBAN RIVERA MUÑOZ

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA DE ECONOMÍA, ADMINISTRACIÓN Y NEGOCIOS

MEDELLÍN

2020

MODELO PARA LA GESTIÓN DE LECCIONES APRENDIDAS BASADO EN EL PROCESAMIENTO DEL LENGUAJE NATURAL Y APRENDIZAJE AUTOMÁTICO.

JORGE ESTEBAN RIVERA MUÑOZ

Trabajo de grado para optar al título de magister en gerencia de proyectos

Asesora

ALEJANDRA CUADROS MEJÍA
Ph.D Dirección de Proyectos

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA DE ECONOMÍA, ADMINISTRACIÓN Y NEGOCIOS

MAESTRÍA EN GERENCIA DE PROYECTOS

MEDELLÍN

2020

13 de abril de 2020

JORGE ESTEBAN RIVERA MUÑOZ

“Declaro que este trabajo de grado no ha sido presentado con anterioridad para optar a un título, ya sea en igual forma o con variaciones, en ésta o en cualquiera otra universidad”. Art. 92, parágrafo, Régimen Estudiantil de Formación Avanzada.

Firma del autor (es)

MODELO PARA LA GESTIÓN DE LECCIONES APRENDIDAS BASADO EN EL PROCESAMIENTO DEL LENGUAJE NATURAL Y APRENDIZAJE AUTOMÁTICO.

LESSONS LEARNED MANAGEMENT MODEL BASED ON THE NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING.

Jorge Esteban Rivera Muñoz 1[1]

Resumen

La gestión del conocimiento es el proceso de recopilación, desarrollo, intercambio y manejo eficiente de la información dentro de una organización o proyecto y su propósito es garantizar que la información correcta esté disponible para las personas adecuadas, el tratamiento de las lecciones aprendidas es uno de los principales insumos para lograr este propósito.

Como una propuesta a esta gestión, este articulo presenta el desarrollo de un modelo de analítica de textos (procesamiento, clasificación y predicción) que permite realizar el procesamiento del lenguaje natural y mediante el uso algoritmos de inteligencia artificial como de máquina de soporte vectorial y aprendizaje profundo, se plantean alternativas de opciones de clasificación automática y predictivas.

De esta manera, ante el surgimiento de una nota u observación en la documentación de un proyecto relacionado al de la base de datos modelada, se podría conocer su clasificación de manera automática, dar alertas tempranas o elementos para la toma de decisiones, basado en un modelo entrenado con lecciones aprendidas de proyectos previos.

Palabras Clave: gestión de conocimiento, lecciones aprendidas, analítica de textos, procesamiento del lenguaje natural, máquina de soporte vectorial, aprendizaje profundo.

Abstract

Knowledge management is the process of gathering, developing, sharing, and the efficient handling of information within an organization or project and its purpose is to ensure that the right information is available to the right people. The treatment of lessons learned is one of the main inputs to achieve this purpose.

As a proposal to this management, this paper exposure the development of a text analytics model (processing, classification and prediction) that allows natural language processing and through the use of artificial intelligence algorithms such as machine support vector machine (SVM) and deep learning (ANN), it pose alternatives of classification and prediction in automatic way.

In this way, given the input of a note or observation in the documentation of a project related to the modeled database, its classification could be known automatically and give early alerts or elements for decision-making, based on a model previous training with lessons learned from previous projects.
……..

Keywords: Knowledge Management, Lesson Learned, Text Analytics, Natural Language Processing, Support Vector Machine, Deep Learning.

[1] Estudiante de Maestría en Gestión de Proyectos, Universidad Pontificia Bolivariana, Medellín - Colombia.

1. Introducción

Una de las actualizaciones más significativas de la última guía de proyectos del PMI, el PMBOK V6.0, es la inclusión de la gestión del conocimiento dentro del grupo de procesos de ejecución y es allí donde se especifica que una de las entradas, son los documentos del proyecto, los cuales incluyen las lecciones aprendidas de proyectos previos (PMI, 2017), de manera consecuente se tiene que una de las herramientas y técnicas para la gestión de dichas entradas son la gestión de conocimiento, que incluye los métodos de codificación, registro de nuevas lecciones, sistemas de información entre otras; lo anterior trasciende en una actualización al plan de la dirección del proyecto mediante la respectiva solicitud de cambio y actualizaciones a los activos de los procesos de la organización (Ritchie, Paul | Jorgensen, 2007).

Es entonces uno de los retos de los gerentes de proyectos aprovechar el conocimiento existente para mejorar los resultados y que el conocimiento creado esté disponible para apoyar fases o proyectos futuros, pero ¿cómo gestionar la enorme cantidad de información disponible que pudieses estar estructurada o no?, ¿indexada o no? y que realmente una lección aprendida de un proyecto previo no pase desapercibida frente una potencial recomendación que hubiese podido brindar (Morris, 2002).

1.1 Inteligencia artificial IA y la gerencia de proyectos

De esta manera entonces se da entrada al concepto de inteligencia artificial IA como un conjunto de técnicas y herramientas que le van a permitir al gerente de proyectos enfocarse en las verdaderas acciones que agregan valor y centrase en la toma de decisiones, dejando de lado o trasladando el procesamiento de información a sistemas computacionales de aprendizaje automático, sería entonces (Hosley, 1987) quien hablaría inicialmente sobre la inteligencia artificial aplicada a la gerencia de proyectos en diferentes tópicos como:

- Preparación de una justificación comercial para el proyecto.
- Construcción de un equipo de proyecto efectivo.
- Selección de líderes de equipo de proyecto.
- Determinación de necesidades especiales de capacitación.
- Desarrollar una lista de inquietudes del proyecto.
- Selección de software de planificación de proyectos.
- Preparación de un cronograma y presupuesto óptimo del proyecto.
- Realización de análisis de valor y riesgo.
- Diseño de especificaciones para calidad, seguridad y confiabilidad.
- Diagnóstico y resolución de problemas técnicos y personales.
- Mejora del rendimiento del proyecto y la eficacia del liderazgo.
- Cálculo / adjudicación de bonos de rendimiento.

Consecuentemente el PMI en su publicación In-Depht Report asociada al PMI´s Pulse of the Profession 2019, confirma que está ocurriendo una disrupción de la inteligencia artificial, donde el 81 por ciento de los encuestados informa que su organización está siendo afectada por las tecnologías de IA, y el 37% de los encuestados dice que la adopción de estas tecnologías de IA es una alta prioridad para su organización, lo que provocó un cambio en los enfoques de gestión de proyectos. En los próximos tres años, los profesionales en proyectos esperan que la proporción de proyectos que administran utilizando IA aumente del 23% al 37% (PMI, 2019a).

Incluso las empresas del sector de la construcción y la ingeniería saben que no pueden ser indiferentes a la IA; se cita entonces que la empresa Bechtel, una empresa de construcción e ingeniería con sede en EE. UU., buscaba mejorar sus tasas de productividad y recurrió al Deep Learning informando que la compañía ahora usa una red neuronal 3D que permite a los equipos de proyectos probar diferentes secuencias virtuales que maximizan su productividad.

Sin embargo, mientras algunas organizaciones lideran el camino en IA, otras se están quedando atrás, un poco más de un tercio de los encuestados en dicho reporte dice que adoptar tecnologías de IA es una prioridad baja en su organización, donde alrededor del 31% informan que menos del 5% de sus proyectos han usado IA en los últimos tres años (PMI, 2019b).

1.2 Analítica de textos de lecciones aprendidas

El presente trabajo se centra en el desarrollo y modelamiento del procesamiento de textos extraídos de lecciones aprendidas de un contexto de proyectos específicos con la ayuda de inteligencia artificial.

El reto con respecto a analizar estos textos o datos cualitativos y tratar de extraer patrones significativos e ideas útiles que pudiesen ser beneficiosas para la organización o los proyectos, es que a pesar de que se cuenta con una gran cantidad de técnicas de aprendizaje automático y análisis de datos, la mayoría de ellas están sintonizadas para trabajar con datos numéricos, por lo tanto, se recurre a áreas como el procesamiento del lenguaje natural (PNL)(Sarkar, 2019).

Dado que la entrada de variables es cualitativa, mediante técnicas de "tokenización", "stemming", "lemmatization", "n-grams" se puede representar de forma vectorial y matricial los textos extraídos(Anandarajan, Hill, & Nolan, 2019) y así realizar modelos de clasificación y predicción basados en máquina de soporte vectorial o SVM por sus siglas en inglés, redes neuronales o "Deep Learning" y un caso particular de este último como lo son las redes neuronales convolucionadas, que han sabido ganarse un lugar importante en procesamiento de imágenes y de textos, dada la exactitud de sus predicciones con pequeños lotes de validación (Heaton, 2015).

A partir del procesamiento de una base de datos de lecciones aprendidas, es decir, de proyectos probablemente finalizados, que tiene tabulados diferentes registros y variables de control, se pueden predecir situaciones de riesgo potencial en la ejecución de proyectos futuros con la documentación que se va generando durante su ciclo de vida, dándole herramientas para la toma de decisiones al gerente de proyectos y de ahí determinar o proponer los controles de cambios correspondientes (Rozenes, 2006).

2. Marco teórico

2.1 Gestión del conocimiento en proyectos

Cada proyecto es diferente, nuevo y distinto de abordar, pero el director de proyectos trae consigo la experiencia de proyectos anteriores y así dispone de la información adquirida previamente y aprovecharla al máximo para alcanzar los objetivos del proyecto en curso (Morris, 2013).

En la guía PMBOK v6.0 el proceso de gestionar el conocimiento del proyecto pertenece al grupo de procesos de ejecución y su principal beneficio es aprovechar el conocimiento existente para mejorar los resultados del proyecto y que el conocimiento creado esté disponible para apoyar fases o proyectos futuros (PMI, 2017).

El conocimiento se divide usualmente en conocimiento explícito, que es el que se puede codificar a través de palabras números e imágenes y es precisamente el que podemos documentar y compartir; por otro lado, está el conocimiento tácito, que es el personal o inherente de cada individuo, son creencias, experiencias, percepciones y comúnmente llamado el "know how" (Rao, 2004).

2.1.1 Lecciones aprendidas

Las entradas del proceso de gestionar el conocimiento del proyecto son el plan para la dirección del proyecto como guía principal para la ejecución del proyecto, documentos del proyecto, que incluyen registro de lecciones aprendidas, asignaciones del equipo de proyecto, estructura

desglose de recursos y registro interesados (PMI, 2017).

Las herramientas y técnicas del proceso de gestionar el conocimiento del proyecto son:

- Juicio de expertos: que se debería acudir a la pericia de individuos o grupos expertos especialmente en gestión del conocimiento, gestión de la información aprendizaje organizacional herramientas, de gestión del conocimiento y la información e información relevante de otros proyectos.

- La gestión del conocimiento: las herramientas y técnicas de la gestión de conocimiento conectan a las personas desarrollando el trabajo en equipo de manera que compartan su conocimiento tácito (Todorović, Petrović, Mihić, Obradović, & Bushuyev, 2015).

  El director del proyectos debe usar toda su experiencia, ya que las herramientas y técnicas dependen de la naturaleza del proyecto su complejidad y la diversidad de los miembros del equipo, que pueden ser entre otras, la creación de relaciones de trabajo, incluidas la interacción social informal y las redes sociales, comunidades de práctica y grupos de interés especial, reuniones, incluidas las reuniones virtuales, aprendizaje por observación, foros de discusión, eventos de intercambio de conocimiento como seminarios, talleres y capacitación (Al-Zayyat, Al-Khaldi, Tadros, & al-Edwan, 2010). Todas las herramientas y técnicas pueden aplicarse presencial o virtualmente y una vez establecidas servirán para mantener las relaciones.

- Gestión de la información: son efectivas para compartir conocimiento explícito siempre inequívoco y codificado, incluye entre otras métodos para codificar el conocimiento explícito, registro de lecciones aprendidas, servicios de biblioteca, recopilación de información y sistemas de información para la dirección de proyectos; d)habilidades interpersonales y de equipo, las cuales incluyen entre otras la escucha activa, la facilitación, el liderazgo y la creación de relaciones de trabajo y además la conciencia política (Yeong & Lim, 2011).

Las salidas del proceso de gestionar el conocimiento del proyecto son:

-Registro lecciones aprendidas: este registro puede incluir la categoría y la descripción de la situación, el impacto, las respectivas recomendaciones, las acciones propuestas y relacionadas con la situación;

- Actualización al plan de la dirección del proyecto: cualquier cambio en el plan para la dirección del proyecto pasa por el proceso de control de cambios de la organización mediante una solicitud de cambio (PMI, 2017);

- Actualizaciones a los activos de los procesos de la organización: todos los proyectos generan un nuevo conocimiento y este puede codificar, separar, mejorar los procesos y procedimientos y dado que el conocimiento se encuentra en la mente de las personas y no se les puede forzar a compartir lo que saben o aprender el conocimiento de otros, el director del proyecto debe enfocar todo su esfuerzo en generar un clima de confianza para que todos los miembros del equipo estén motivados, el conocimiento pueda ser compartido y asegurar el éxito del proyecto (Gasik, 2011).

2.2 Analítica de textos

Se define la analítica de textos o minería de texto, como el descubrimiento automático de información nueva, previamente desconocida a partir de datos textuales no estructurados. Los términos análisis de texto y minería de texto son a menudo usado indistintamente. La minería de texto también se puede describir como el proceso de extraer información de alta calidad del texto (Anandarajan et al., 2019).

Este proceso implica tres tareas principales: información-recuperación, extracción de información y minería de datos. El análisis de texto ha sido influenciado por muchos campos y ha hecho contribuciones significativas a muchas disciplinas, las aplicaciones modernas de análisis de texto abarcan muchas disciplinas y objetivos, además de tener orígenes multidisciplinarios y campos de investigación (Sarkar, 2019), incluyendo:

• Biblioteca y ciencias de la información.
• Ciencias Sociales
• Ciencias de la Computación
• Bases de datos
• Minería de datos
• Estadísticas
• Inteligencia artificial
• Ligüística computacional

2.2.1 Procesamiento del lenguaje natural

Entendiendo en primera instancia que el lenguaje natural es como el ser humano expresa pensamientos o sentimientos a través de un lenguaje determinado, se puede entrar a definir el procesamiento del lenguaje natural como la capacidad de las tecnologías computacionales y o lingüística computacional, para procesar el lenguaje natural humano, es también un campo de la informática, inteligencia artificial y lingüística computacional relacionada con las interacciones entre computadoras y lenguajes humanos y que puede definirse como automático o semiautomático (Wilks, 1996), no obstante se podría mencionar que abarca otras áreas de conocimiento como la estadística, habilidades duras como se muestran en la Figura 1.

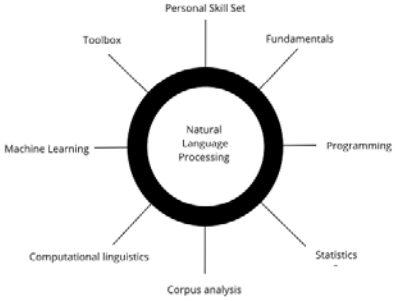

Figura 1. Conceptos sobre NLP (Thanaki, 2017).

Dentro de las técnicas básicas para el procesamiento de textos tenemos:

2.2.1.1 tokenización

El proceso de tokenización consiste en convertir los textos en tokens, los cuales son componentes textuales independientes y mínimos que tienen una sintaxis y semántica definida. Un párrafo o un documento tienen varios componentes, incluidos oraciones, que pueden desglosarse en cláusulas, frases y palabras. Las técnicas de tokenización más populares incluyen la tokenización de oraciones y palabras que se usan para dividir un documento de texto en oraciones y cada oración en palabras, por lo tanto, la tokenización se puede entender como el proceso de descomposición o división datos textuales en componentes más pequeños y significativos llamados tokens (Beysolow II, 2018).

2.2.1.2 Stemming y lemmatization

Entendiendo los morfemas como la unidad mínima de la lengua que posee significado léxico o gramatical. En este sentido, es la unidad mínima aislable de análisis gramatical, y, por ello mismo, no puede ser dividida en unidades menores (Coelho, n.d.), se entiende entonces que los morfemas están compuestos por una raíz y unos afijos, es entonces las técnicas y algoritmos asociados que nos llevan a devolver los morfemas a su raíz como el proceso de “stemming” tal y como se muestra en la Figura 2 con la palabra “present”, teniendo en cuenta que la palabra raíz pudiese estar o no contenida en un diccionario y/o tener un significado.

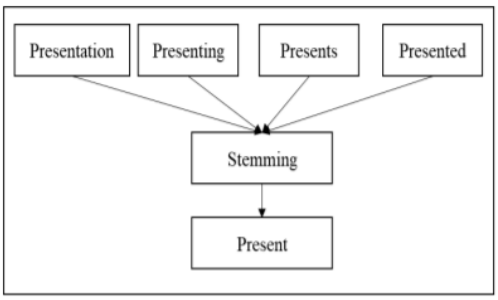

Figura 2. Ejemplo de un Stemming (Thanaki, 2017).

Lemmatization es un proceso similar al anterior, pero puede obtener la forma canónica de una palabra en función de su lema, es decir la palabra resultante tiene un significado concreto por consecuentemente se encuentra contenida en un diccionario. Por ejemplo: (beautiful) y (beautifully) son “lemmatized” a (beautiful) y (beautifully) respectivamante sin cambiar el sentido de las palabras, pero (good), (better) y (best) son “lemmatized” a (good), ya que todas tienen significado similar.
En la figura 3 se muestran las diferencias entre las técnicas de “stemming” y “lemmatization”.

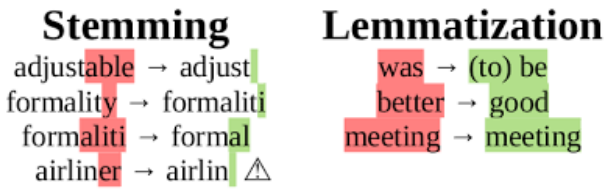

Figura 3. Diferencia entre Stemming y Lemmatization (Kushwah, 2019)

2.2.1.3 Parts of Speech (POS) Tagging o Etiquetas

Llamado también etiquetado de partes del discurso, es el proceso donde se le asignan etiquetas dado el comportamiento asociado, para un idioma en particular, en general se utilizan las etiquetas de sustantivos, pronombres, verbos, adjetivos, adverbios, etc,(Tobergte & Curtis, 2016).

2.2.1.4 Stop words o Palabras Parada

Uno de los pasos más importantes dentro del procesamiento de textos, es remover aquellas palabras ya convertidas en tokens que no aportan al entendimiento o modelamiento, por lo tanto, no tienen mucho peso en la interpretación y deben ser removidas, basadas en una colección prestablecida en una librería determinada, por ejemplo: the, end, and, for, of, etc.(Xie, Le, Zhou, & Raghavan, 2018).

2.2.1.5 Bag of Words o Bolsa de Palabras

La bolsa de palabras o “Bag of Words” es el modelo de representación de espacio vectorial más simple para textos no estructurados. Un modelo de espacio vectorial es un modelo matemático para representar un texto o cualquier otro dato como vectores numéricos, de modo que cada dimensión del vector sea una característica o atributo específico. El modelo Bolsa de palabras representa cada documento de texto como un vector numérico donde cada dimensión es una palabra específica y el valor podría ser su frecuencia en el documento, ocurrencia (denotado por 1 o 0), o incluso valores ponderados, este modelo recibe este nombre porque cada documento está representado literalmente como una bolsa de sus propias palabras, sin tener en cuenta su orden, las secuencias o la gramática (Mc Tear, Callejas, & Griol, 2014).

2.2.1.5 N-grams y Bag of N-Grams

Una palabra es solo un token, a menudo conocido como unigrama o 1- Grams y como se mencionó anteriormente el modelo Bolsa de palabras no considera el orden de las palabras, pero en ocasiones se requiere tener en cuenta las frases o la colección de palabras que aparecen en una secuencia, es decir, N- Grams que es básicamente una colección de tokens de palabras de un documento de texto.

El modelo Bag of N-Grams es solo una extensión del modelo Bag of Words que aprovecha las características basadas en N-Gram (Wang, McCallum, & Wei, 2007). Por ejemplo, si un texto contiene las palabras New York, requerimos que se forme la estructura –New York- o sea un 2-Grams y no –New- y –York- de forma separada como 1-Grams. Este tipo de reconocimientos generalmente se logran gracias a la técnica de Asignación Latente de Dirichlet o LDA por sus siglas en inglés.

El LDA es un modelo en grafo y generativo que permite que conjuntos de observaciones puedan

ser explicados por grupos no observados que explican por qué algunas partes de los datos son similares, para el particular caso de estudio, si las observaciones son palabras en documentos, presupone que cada documento es una mezcla de un pequeño número de categorías y la aparición de cada palabra en un documento se debe a una de las categorías a las que el documento pertenece. (Blei, Ng, & Jordan, 2003).

2.2.1.6 Word embedding

Si bien las técnicas de Bag of Words son métodos efectivos para extraer características del texto, debido a que la naturaleza inherente del modelo es solo una bolsa de palabras no estructuradas, se puede perder información importante como la semántica, estructura, secuencia y contexto en torno palabras cercanas en cada documento de texto. Esto forma una motivación suficiente para que explorar modelos más sofisticados que pueden capturar esta información y dar características que son representaciones vectoriales de palabras, conocidas popularmente como “embeddings” o embebidos.(Levy & Goldberg, 2014).

Con respecto a los sistemas de reconocimiento de voz o imagen, toda la información ya está presente en forma de vectores embebidos de características densas en conjuntos de datos de alta dimensión como espectrogramas de audio e intensidades de píxeles de imagen. Sin embargo, cuando se trata de datos de texto sin procesar, especialmente modelos basados en conteo como Bag of Words, se trata con palabras individuales que pueden tener sus propios identificadores y no capturan la relación semántica entre palabras, esto conduce a enormes vectores de palabras dispersas para datos de texto, por ejemplo, si no tenemos suficientes datos, podemos terminar obteniendo modelos pobres o incluso sobreajustar (overfitting) los datos debido al sesgo en la dimensionalidad.

Una de los modelos más reconocidos de Word embedding es Word2Vec, desarrollado por Google en 2013 y es un modelo de aprendizaje predictivo profundo para calcular y generar vectores densos continuos, distribuidos y de alta calidad, representaciones de palabras que capturan similitudes contextuales y semánticas. Esencialmente estos son modelos no supervisados que pueden incorporar cuerpos de texto masivos, crear un vocabulario de palabras posibles, y generar palabras embebidas densas para cada palabra en el espacio vectorial que representa ese vocabulario.(Sarkar, 2019).

2.2.2 *Modelamiento, clasificación y predicción*

2.2.2.1 Máquina de Soporte Vectorial (SVM)

Originalmente presentados por Vapnik en 1995 como “Support-Vector Networks”, son un algoritmo de aprendizaje automático para problemas de clasificación de dos grupos, donde el algoritmo de entrenamiento construye un modelo que predice si un nuevo dato cae en una categoría u otra.(Chen, Lu, Yang, & Li, 2010)

Esta técnica se basa en la idea de encontrar la mejor línea, plano o hiperplano acorde a la dimensión, que divida un conjunto de datos en dos clases y para ello utiliza los conceptos de hiperplano, el cual es un plano que separa y clasifica linealmente un conjunto de datos; los vectores de apoyo, como los puntos críticos de datos más cercanos a la línea del hiperplano y que al eliminarse alterarían la posición del hiperplano; el margen, definida por la distancia entre el hiperplano y el dato más cercano de cualquiera de los conjuntos; el núcleo o Kernel, que comprende el conjunto de funciones matemáticas utilizadas para transformar los datos de entrada en la forma deseada y por último la Dimensión VC (Vapnik-Chervonekis) como el conjunto de funciones disponibles y la dimensión, es la cantidad máxima de puntos que se pueden separar de todas las formas posibles mediante el conjunto VC. Cuanto mayor sea la dimensión, menor será el error del conjunto de entrenamiento y la confianza crecerá.(Numerentor.org, n.d.).

Esta técnica se desarrolló en primer lugar para resolver los problemas de clasificación, pero luego

se han extendido al dominio de los problemas de regresión, conservando todas las propiedades principales que caracterizan el algoritmo de margen máximo, como dualidad, dispersión, núcleo y convexidad. Una diferencia notable es que los SVM de regresión introducen una función de pérdida que ignora los errores que están dentro de una cierta distancia del valor verdadero.(Moraes, Valiati, & Gavião Neto, 2013).

2.2.2.2 Aprendizaje profundo (Deep Learning)

2.2.2.2.1 Redes Neuronales Perceptrón Multicapa

Las redes neuronales son modelos matemáticos que se componen de una serie de neuronas o nodos interconectados que emulan el aprendizaje humano; dichos nodos se encuentran alojados en las denominadas capas, las cuales son de entrada, ocultas y de salida, donde cada una de los cuales recibe un cierto número de entradas o "inputs" y suministra una salida o "output" como se muestra en la figura 4.

Cada uno de los nodos para otorgar un valor de salida desempeña cálculos de suma ponderada en los valores que reciben de entrada y luego generan un valor que utiliza funciones de transformación no lineal simples en estas sumas, dichas funciones pueden ser la sigmoide o tangente hiperbólica, entre otras.

Para la red neuronal seleccionada, la perceptrón multicapa, las correcciones a los pesos se realizán en respuesta a errores o pérdidas individuales que exhiben las redes en los nodos de salida, dichas correcciones generalmente utilizando la técnica de gradiente descendiente estocástico, llamado retropropagación.

Los principales factores que distinguen diferentes tipos de redes entre sí son cómo son los nodos que están conectados y el número de capas. Redes básicas en el que todos los nodos se pueden organizar en capas secuenciales, con cada nodo recibiendo entradas solo de nodos de capas anteriores, se conocen como feedfoward neural networks (FFNN) (Lecun, Bengio, & Hinton, 2015).

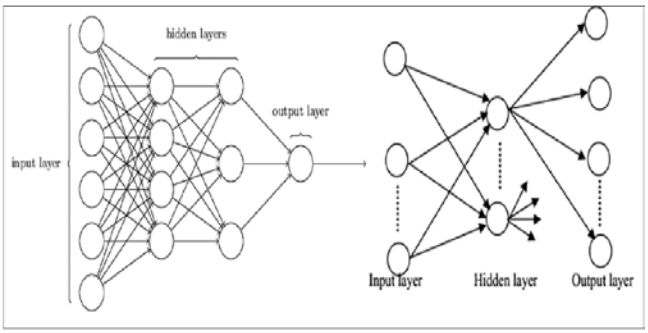

Figura 4. Ejemplo de redes neuronales FFNN.

2.2.2.2.2 Redes Neuronales Convolucionadas

Las redes neuronales convolucionadas (CNN) hacen parte del aprendizaje profundo o "Deep Learning" y están inspiradas en la corteza visual de los animales y por ende han ganado buen reconocimiento dada la adaptación para el reconocimiento de imágenes y el reconocimiento de la escritura a mano. Su estructura se basa en el muestreo de ventanas o partes de una imagen, detectando sus características y luego utilizando las características para construir una representación (Zhang & LeCun, 2015).

Las CNN son bastante similares a las redes neuronales convencionales, están formadas por neuronas con pesos que se pueden aprender de los datos, cada neurona recibe algunas entradas y realiza un producto de puntos, tienen una función de pérdida en la última capa totalmente conectada y pueden usar una función no lineal.

En general, hay tres capas principales en un CNN simple. Son la capa de convolución, la capa de agrupación y la capa totalmente conectada (Kim, 2017). En la figura 4 podemos observar los diferentes pasos que conforman el procesamiento de texto a través de una red neuronal convolucionada.

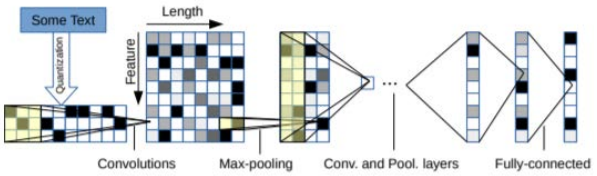

Figura 5. Esquema de una red neuronal convolucionada. (Zhang & LeCun, 2015).

3. Metodología

Mediante el uso del software MATLAB y su librería de funciones de "text analytics" se realizó el procesamiento, clasificación y modelamiento predictivo de la base de datos de lecciones aprendidas de programas y proyectos de la NASA, la cual es de consulta libre dentro de su programa "NASA Public Lessons Learned System" y disponible en formato HTML, por lo se convirtío a formato *.CSV para su análisis. La base de datos contaba con 1648 registros capturados hasta el año 2017.

3.1 Preparación de los datos

Consta de la lectura de la base de datos, la cual contiene las variables de N°, Fecha, Abstract, Category, Lesson Learned y Safety, posterior a dicha lectura se asignan las variables de entrada de texto y categóricas, se procede a realizar la tokenización de los textos de las variables Abstract y Lesson Learned, se remueven las palabras de parada (stop words), se remueven los signos de puntuación, se realiza el proceso de "Stemming" y "lemmatization", posteriormente se crea la bolsa de palabras o Bag of words. Con el fin de conocer el resultado del proceso previo se evidencian en la tabla 1 las 10 palabras más frecuentes:

Tabla 1: Palabras más frecuentes en la bolsa de palabras para los textos de lecciones aprendidas.

Word	Count
"system"	911
"design"	637
"project"	629
"test"	622
"flight"	500
"requirement"	495
"result"	442
"program"	433
"failure"	414
"process"	411

Se retiran las palabras no frecuentes, que para el caso del modelamiento serían menos de dos apariciones en la bolsa de palabras, finalmente mediante un gráfico de nubes de palabras procedemos a comparar el texto antes y después de procesamiento. En la figura 6 se evidencia las palabras más frecuentes antes y después del procesamiento de textos, mediante la técnica de nube de palabras.

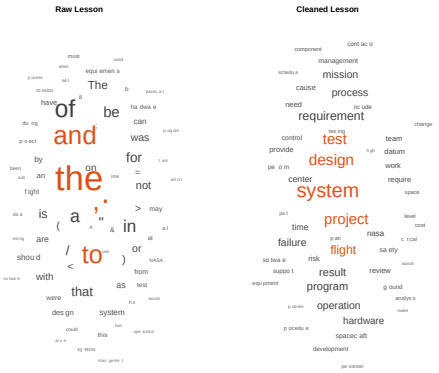

Figura 6. Comparativo de las palabras más frecuentes antes y después de procesamiento para 1-Grams.

Es importante considerar en el momento del modelamiento la influencia e importancia de los diferentes N-Grams, por lo que se evalúa 2-Gram y 4 tópicos mediante la técnica de LDA, por ejemplo, en la figura 7 se muestran las dos palabras consecutivas o cadenas de palabras más frecuentes (2-Grams) antes y después del procesamiento de textos, mediante la técnica de nube de palabras.

Figura 7. Tópicos 1 y 2 del procesamiento con LDA para 2-Grams.

Y ahora por ultimo se revisan las cadenas de tres palabras más repetida tanto en la variable Abstract como Lesson Learned, De esta manera la figura 8 nos muestra los "3-Grams".

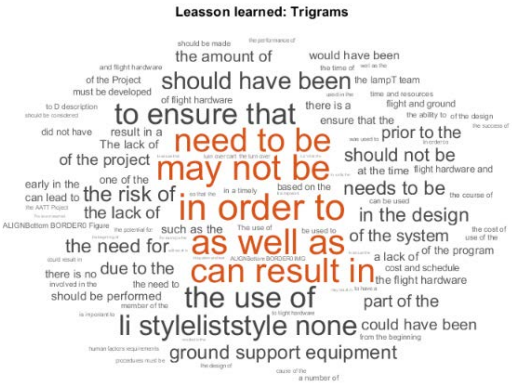

Figura 8. Nube de datos para 3-Grams para Lesson learned.

3.2 Clasificación y modelamiento con SVM

Dentro del ejercicio de modelamiento y predicción, se elaborarán dos modelos, el primero que prediga a través de la descripción de la lección aprendida contenido en la variable "Lesson Learned" en que categoría se encuentra, es decir, la asociación automática a la variable "category" y un segundo modelo que mediante la descripción del resumen del problema o situación ocurrida contenida en la variable "abstract" pueda identificar de manera automática si represento un riesgo para la seguridad, es decir, la asociación con la variable "safety".

Para efectos de entrenamiento y validación del modelo, se procede a particionar la base de datos en un 90% y 10% respectivamente, posteriormente se asigna como categóricas a las variables independientes, es decir, "category" y "safety", luego al tratarse de variables potencialmente desbalanceadas, se descartan las que aparezcan menos de 10 veces estableciendo así una normalización.

Se determina mediante un histograma tanto la probabilidad de ocurrencia de cada categoría, como de la seguridad comprometida. En la figura 9 se muestra el histograma de las diferentes clases de lecciones aprendidas y de la misma manera en la figura 10 se presentan las categorías variable seguridad o "safety", es decir, si se vio comprometida (TRUE) o no (FALSE).

Figura 9. Histograma de las categorías de lecciones aprendidas.

Figura 10. Histograma de la presencia de riesgo negativo en las lecciones aprendidas.

Con el 90% de la base de datos se realiza el respectivo procesamiento de textos, se construye una "Bag of words" o bolsa de palabras conformando una matriz , para finalmente junto con las variables categoricas extraidas, modelar los SVM bajo la función **mdl** y **fitcecoc,** que permiten modelar multiples categorias, ya que el modelo básico de SVM solo permite dos.

Se construye una matriz con los datos del 10% de validación mediante al función **encode** y se calcula la precisión o "accuracy" de los modelos, basado en los datos de validación tanto para la predicción de la categoria (Category) partiendo de la descripción de la lección aprendida (Lesson Learned), así como la situación de riesgo (Safety) partiendo de la descrición del problema (Abstract).

Finalmente, se toman 3 lecciones aprendidas y resumenes de los años 2018 y 2019 (no incluidos en la base de datos) con el fin de validar nuevamente el modelo.

3.2 Clasificación y modelamiento con Deep Learning

3.2.1 Clasificación y modelamiento con Redes Neuronales Perceptrón Multicapa

Para el modelamiento mediante aprendizaje profundo o "Deep Learning", se empleará uno de los tipos de red neuronal denominado perceptron multicapa (MLP), así que, de manera análoga a los pasos anteriores, se procede con el procesamiento de texto básico y nuevamente con el establecimiento de los dos modelos planteados de predicción de categorías frente al texto de las lecciones aprendidas y un segundo modelo que predice la presencia de riesgo ante la descripción del problema.

Se inicia con la partición de la base de datos en 70% 15% y 15%, con el fin de tener un conjunto de entrenamiento, validación y test respectivamente, luego se realiza el mismo algoritmo de preprocesamiento y normalización de variables que se utilizó para los SVM.

Se entrena un tipo de red neuronal LSTM (Long short-term memory), la cual requiere que los datos de entrenamiento estén convertidos en índices numéricos de secuencia y para esto se emplea la función **wordencoding**, se a los hiperparámetros de la red neuronal y se comienza con la construcción del modelo.

En la figura 11 se evidencian las diferentes épocas de entrenamiento del modelo de redes neuronales en Matlab, así como la tendencia porcentual de la precisión.

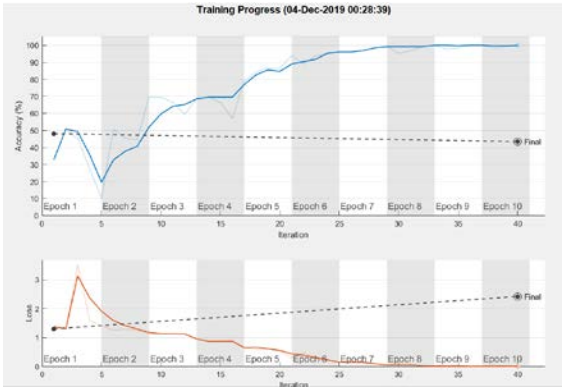

Figura 11. Entrenamiento de la red neuronal a través de las iteraciones. (Matlab)

Se calcula nuevamente la precisión de los modelos, basado en los datos de validación tanto para la perdición de la categoría (Category) partiendo de la descripción de la lección aprendida (Lesson Learned), así como la situación de riesgo (Safety) partiendo de la descripción del problema (Abstract), así como la prueba de las tres lecciones aprendidas que se probó en los SVM.

3.2.2 Clasificación y modelamiento con Redes Neuronales Convolucionadas

Con el fin de tratar de encontrar modelos de aprendizaje profundo que brinden más precisión, se acude a una de las variaciones denominadas redes neuronales convolucionadas, la cual para clasificar datos de texto, se deben convertir inicialmente en imágenes, para hacer esto, los registros deben tener una longitud constante S y convertir los documentos en secuencias de vectores de palabras de longitud C utilizando "word embedding", así se puede representar un documento como una imagen de 1 x S x C (una imagen con altura 1, ancho S y C canales).

Es importante anotar que al igual que con el modelo de perceptrón, se inicia con la partición de la base de datos en 70% 15% y 15%, con el fin de tener un conjunto de entrenamiento, validación y

test respectivamente, luego se realiza el mismo algoritmo de preprocesamiento y normalización.

En la tabla 2 se muestra como una descripción de lección aprendida tiene asociada una clase “TRUE” o “FALSE” y posterior a la conversión de esta descripción en la tabla 3 se presentan como predictores al haber sido convertido en imágenes en secuencias de vectores.

Tabla 2: muestra de 8 registros antes de ser convertidos en imágenes.

abstract	safety
Power system distribution at	'TRUE'
The sampling and analysis processes,	'FALSE'
An operator error involving	'TRUE'
The Restore Project obtained	'FALSE'
The JPL focus on primarily one	'FALSE'
Three employees were exposed	'FALSE'
Thermal-vacuum testing of a	'TRUE'
Following thermal-	'FALSE'

Tabla 3: muestra de registros de la Tabla 2 después de ser convertidos en imágenes.

predictors	responses
1×100×300 single	TRUE
1×100×300 single	FALSE
1×100×300 single	TRUE
1×100×300 single	FALSE
1×100×300 single	FALSE
1×100×300 single	FALSE
1×100×300 single	TRUE
1×100×300 single	FALSE

Se parametriza la red convolucionada en su capa de entrada, el primer bloque de 1-Grams, la capa de concatenación profunda, la capa totalmente conectada (fully conected layer), la softmax layer y la capa de clasificación. Para los N-Grams de dos en adelante, se crea un bloque de convolución, un bache de normalización, un ReLU (rectified linear unit), una capa de abandono (dropout layer) y una capa Max Polling. En la figura 12 se presenta la arquitectura diagramada mediante Matlab de la red convolucionada para las lecciones aprendidas, similar a lo cual se había presentado de manera esquemática en la figura 5 para este tipo de redes en general.

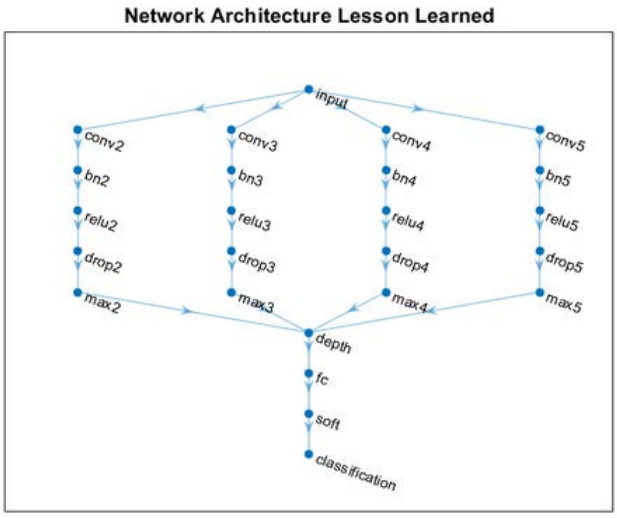

Figura 12. Arquitectura de la red neuronal convolucional para los textos de lecciones aprendidas.

Se realiza el entrenamiento con la arquitectura construida; en la figura 13 se evidencian las diferentes épocas de entrenamiento del modelo de redes neuronales convolucionadas en Matlab, así como la tendencia porcentual de la precisión.

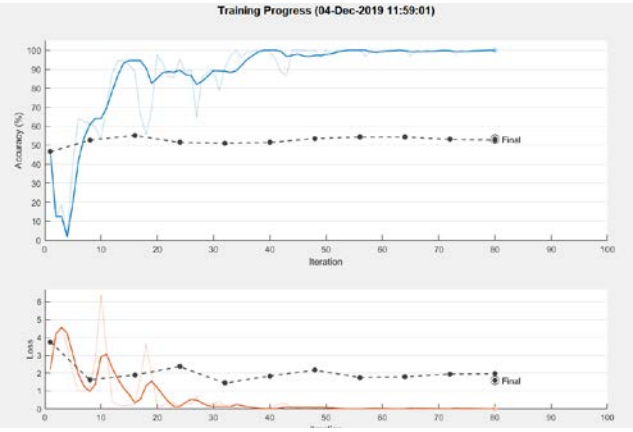

Figura 13. Entrenamiento de la red neuronal convolucional a través de las iteraciones.(Matlab)

Una vez mas se calcula la presición de los modelos, basado en los datos de validación tanto para la perdición de la categoría (Category) partiendo de la descripción de la lección aprendida

(Lesson Learned), así como la situación de riesgo (Safety) partiendo de la descripción del problema (Abstract).

4. Resultados y Discusión

Luego de realizar el modelamiento de clasificación y de predicción de las variables categóricas "category" y "safety" basados en el lenguaje natural almacenado en los campos "Lesson learned" y "Abstract" respectivamente, se presenta en la tabla 4 la precisión asociada a cada una de las técnicas empleadas.

Tabla 4: Precisión de las técnicas para los modelos

	Lesson Learned - Category	Abstract - Safety
SVM	52.38%	79.16%
Deep Learning	47.66%	69.53%
CNN	55.92%	74.69%

De los resultados obtenidos se puede inferir que, en el modelo entrenado con la base de datos disponible, la elección de la variación de redes neuronales convolucionadas obtuvo una precisión superior al modelo clásico de aprendizaje profundo de perceptrón multicapa, dando indicios del porque se viene consolidando como una técnica a considerar dentro del procesamiento del lenguaje natural y de la misma manera se observa como mediante los vectores soporte de máquina se obtiene una mejor precisión al tratarse de variables dicotómicas como "TRUE" y "FALSE", ya que se debe recordar que para el modelo de lecciones aprendidas y su categoría, se debió realizar un ajuste al algoritmo, proporcionado por una función de Matlab, debido a que originalmente está concebido para este tipo de datos.

En la tabla 5 se muestran las pruebas realizadas con los modelos de SVM y Deep Learning tras ingresar la descripción de tres lecciones aprendidas no incluidas para la construcción de los modelos y sus respectiva causa o "abstract", dando una clasificación apropiada. Dichas lecciones correspondían a registros de los años 2018 y 2019.

Tabla 5: Prueba de clasificación

	SVM	Deep Learning	Clasificación Original
Muestra de Lesson Learned	Aeronautic Research Aeronautic Research Exploration Systems	Aeronautic Research Aeronautic Research Exploration Systems	Aeronautic Research Aeronautic Research Exploration Systems
Muestra de Abstract	FALSE FALSE FALSE	FALSE FALSE FALSE	FALSE FALSE FALSE

Se videncia que tanto el modelo de vectores soporte de máquina como la red neuronal perceptrón multicapa perteneciente al aprendizaje profundo, clasificaron correctamente los 3 registros ingresados.

4. Conclusiones

La clasificación de textos y los modelos predictivos asociados a estos son de vital importancia para la gestión de proyectos, dada la cantidad de información que se genera y la que se trae de proyectos previos, dado que, ante la inexistencia de un sistema automático de alarmas tempranas, quedaría relegado la gestión de conocimiento de lecciones aprendidas al conocimiento tácito de sus integrantes y a su disposición de adoptarlas.

Los modelos construidos a través de la analítica de textos y los diferentes algoritmos de aprendizaje automático mostraron como mediante la descripción de un problema y su respectiva lección aprendida, se puede modelar y posteriormente predecir las categorías a la que pertenece y las consecuencias de seguridad o aparición de un riesgo, lo anterior puede ayudar a determinar la salud de un proyecto en línea toda vez que las documentación de las diferentes novedades del proyecto se comparen con los modelos entrenados o dar elementos para la toma de decisiones oportunas.

La precisión y convergencia de los modelos es una condición supeditada a muchas variables,

entre las más importantes esta la calidad de los datos de ingreso, en este caso, la forma de redacción de las lecciones aprendidas y las palabras usadas, otra variable es la cantidad de registros que nos permitan tener un modelo aceptable, esta condición solo es posible determinarla ante el ingreso de más registros y comparar. Para efectos de uso del modelo se deben tener contextos de proyectos semejantes dada la terminología usada, por ejemplo, la base de datos empleada pertenecía a proyectos del ámbito aeroespacial, si se ingresara una lección aprendida o descripción de un problema del sector de la construcción, probablemente, no brindaría pronósticos o clasificaciones acertadas.

Para los gerentes de proyectos comprender los usos potenciales de la tecnología es tan importante como comprender las estrategias de mejora de procesos y las mejores prácticas para lograr proyectos exitosos, con miras a adquirir e incrementar lo que se conoce como Technology Quotient TQ o cociente tecnológico, el cual es la capacidad de una persona para adaptar, administrar e integrar tecnología basada en las necesidades de la organización o el proyecto en cuestión y es precisamente la adopción de la inteligencia artificial una de las principales.(PMI, 2019b).

5. Trabajos Futuros

Este modelamiento tuvo la gran restricción de trabajar con una base de datos pública, la cual para efectos de de confidencialidad se publican solo apartes, por lo que un trabajo futuro será el modelamiento de lecciones aprendidas de un sector productivo especifico y bajo un lenguaje de código abierto como Python, tambien se puede materializar el código dentro de un aplicativo funcional así como la optimización del mismo.

Adicionalemte, dentro del Deep Learnig existen muchas más tipologias de redes que día a día muestran mejor desempeño y menores tiempo de procesamiento, uno de los talones de aquiles de estas técnicas.

5. Bibliografía

Al-Zayyat, A. N., Al-Khaldi, F., Tadros, I., & al-Edwan, G. (2010). The Effect of Knowledge Management Processes on project Management. *IBIMA Business Review Journal*, *2010*, 1–6. https://doi.org/10.5171/2010.826105

Anandarajan, M., Hill, C., & Nolan, T. (2019). *Practical Text Analytics* (Vol. 2). https://doi.org/10.1007/978-3-319-95663-3

Beysolow II, T. (2018). *Applied Natural Language Processing with Python*. *Applied Natural Language Processing with Python*. https://doi.org/10.1007/978-1-4842-3733-5

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

Chen, N., Lu, W., Yang, J., & Li, G. (2010). Support Vector Machine in Chemistry. In *Support Vector Machine in Chemistry* (p. 19). https://doi.org/10.1142/9789812794710

Coelho, F. (n.d.). “Morfema - Qué es, Definición y Ejemplos”. Retrieved December 4, 2019, from https://www.diccionariodedudas.com/morfema/

Gasik, S. (2011). A model of project knowledge management. *Project Management Journal*, *42*(3), 23–44. https://doi.org/10.1002/pmj.20239

Heaton, J. (2015). *AIFH, Volume 3: Deep Learning and Neural Networks*. *Journal of Chemical Information and Modeling* (Vol. 3). https://doi.org/10.1017/CBO9781107415324.004

Hosley, W. N. (1987). The application of artificial intelligence software to project management, 73–75. Retrieved from

https://www.pmi.org/learning/library/application-artificial-intelligence-software-pm-5234
Kim, Y. (2017). Convolutional neural networks for sentence classification. *2017 43rd Latin American Computer Conference, CLEI 2017*, *2017-Janua*, 1–5. https://doi.org/10.1109/CLEI.2017.8226381
Kushwah, D. (2019). What is difference between stemming and lemmatization? - Quora. Retrieved December 5, 2019, from https://www.quora.com/What-is-difference-between-stemming-and-lemmatization
Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539
Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, *3*(January), 2177–2185.
Mc Tear, M., Callejas, Z., & Griol, D. (2014). The Conversational Interface. In *Integrative and Eclectic Counselling and Psychotherapy* (p. 166). https://doi.org/10.4135/9781446280409.n3
Moraes, R., Valiati, J. F., & Gavião Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621–633. https://doi.org/10.1016/j.eswa.2012.07.059
Morris, P. W. G. (2002). Managing project management knowledge for organizational effectiveness. Retrieved from https://www.pmi.org/learning/library/managing-pm-knowledge-organizational-effectiveness-1945
Morris, P. W. G. (2013). Reconstructing project management reprised: a knowledge perspective. Project Management Journal, 6–23. Retrieved from https://www.pmi.org/learning/library/project-management-discipline-knowledge-3804
Numerentor.org. (n.d.). Máquina de Soporte Vectorial SVM. Retrieved December 4, 2019, from http://numerentur.org/svm/
PMI. (2017). *A Guide to the PROJECT MANAGEMENT of BODY OF KNOWLEDGE. Project Management Institute* (Vol. 53). https://doi.org/10.1002/pmj.21345
PMI. (2019a). AI Innovators. *PMI´s Pulse of the Profession. In-Depth Report*.
PMI. (2019b). the Future of Work - Leading the Way With Pmtq. *Pulse of the Profession*, 8. Retrieved from https://www.pmi.org/-/media/pmi/documents/public/pdf/learning/thought-leadership/pulse/pulse-of-the-profession-2019.pdf?sc_lang_temp=en
Rao, M. (2004). *Book Review: "Knowledge Management: Concepts and Best Practices." Journal of Information & Knowledge Management* (Vol. 03). https://doi.org/10.1142/s0219649204000717
Ritchie, Paul | Jorgensen, K. (2007). Project Management Knowledge. *Project Management Knowledge Management Moving from Standards to Leadership, Project ma*. Retrieved from https://www.pmi.org/learning/library/knowledge-management-standards-leadership-7399
Rozenes, S. | V. G. | S. S. (2006). Project control: literature review. Project Management Journal, 5–14. Retrieved from https://www.pmi.org/learning/library/project-control-gap-project-planning-implementation-2567
Sarkar, D. (2019). *Text Analytics with Python*. *Text Analytics with Python*. https://doi.org/10.1007/978-1-4842-4354-1
Thanaki, J. (2017). Python Natural Language Processing, Explore NLP with machine learning and deep learning techniques. In *Python Natural Language Processing, Explore NLP with machine learning and deep learning techniques* (p. 10). BIRMINGHAM - MUMBAI. Retrieved from http://nemertes.lis.upatras.gr/jspui/bitstream/10889/5243/1/Σταυλιώτης.pdf
Tobergte, D. R., & Curtis, S. (2016). *Mastering Natural Language Processing with Python*.

Packt Publishing (Vol. 53). https://doi.org/10.1017/CBO9781107415324.004

Todorović, M. L., Petrović, D. T., Mihić, M. M., Obradović, V. L., & Bushuyev, S. D. (2015). Project success analysis framework: A knowledge-based approach in project management. *International Journal of Project Management*, *33*(4), 772–783. https://doi.org/10.1016/j.ijproman.2014.10.009

Wang, X., McCallum, A., & Wei, X. (2007). Topical N-grams: Phrase and topic discovery, with an application to information retrieval. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 697–702. https://doi.org/10.1109/ICDM.2007.86

Wilks, Y. (1996). Natural Language Processing. *Communications of the ACM*, *39*(1), 60–62. https://doi.org/10.1145/234173.234180

Xie, Y., Le, L., Zhou, Y., & Raghavan, V. V. (2018). *Deep Learning for Natural Language Processing*. *Handbook of Statistics* (Vol. 38). https://doi.org/10.1016/bs.host.2018.05.001

Yeong, A., & Lim, T. T. (2011). Integrating knowledge management with project management for project success. *Journal of Project, Program & Portfolio Management*, *1*(2), 8. https://doi.org/10.5130/pppm.v1i2.1735

Zhang, X., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification, 1–9. Retrieved from http://arxiv.org/abs/1502.01710