UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

ÁREA TÉCNICA

TITULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Análisis de la información de foros en cursos MOOC mediante técnicas de procesamiento de Lenguaje Natural

TRABAJO DE TITULACIÓN

AUTOR: Peñarreta León, Santiago René

DIRECTOR: Riofrío Calderón, Guido Eduardo, ING

LOJA – ECUADOR

2016

Febrero, 2016

APROBACIÓN DEL DIRECTOR DEL TRABAJO DE TITULACIÓN

Ingeniero.

Guido Eduardo Riofrío Calderón

DOCENTE DE LA TITULACIÓN

De mi consideración:

El presente trabajo de titulación: Análisis de la información de foros en cursos MOOC mediante técnicas de procesamiento de Lenguaje Natural realizado por Santiago René Peñarreta León , ha sido orientado y revisado durante su ejecución, por cuanto se aprueba la presentación del mismo.

Loja, febrero de 2016

f). . . . . .

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

Yo Santiago René Peñarreta León declaro ser autor (a) del presente trabajo de titulación: Análisis de la información de foros en cursos MOOC mediante técnicas de procesamiento de Lenguaje Natural, de la Titulación Ingeniería en Sistemas Informáticos y Computación, siendo Ing. Guido Eduardo Riofrío Calderón director (a) del presente trabajo; y eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones legales. Además certifico que las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo investigativo, son de mi exclusiva responsabilidad.

Adicionalmente declaro conocer y aceptar la disposición del Art. 88 del Estatuto Orgánico de la Universidad Técnica Particular de Loja que en su parte pertinente textualmente dice: “Forman parte del patrimonio de la Universidad la propiedad intelectual de investigaciones, trabajos científicos o técnicos y tesis de grado o trabajos de titulación que se realicen con el apoyo financiero, académico o institucional (operativo) de la Universidad”

F.....

Peñarreta León Santiago René

1104331739

DEDICATORIA

A mis Padres Galo Francisco y María Esperanza

Quienes son mi principal ente de motivación y me han brindado todo el apoyo durante todo mi proceso educativo, ellos son coautores también del presente proyecto.

AGRADECIMIENTO

De la manera más cálida y afectuosa expreso un fuerte agradecimiento al Ing. Guido Riofrío, quien ha dirigido este proyecto de la manera más comprometida, por todo el aprendizaje ofrecido durante el transcurso , la confianza vertida en mi persona, y por todos los aportes ideológicos ofrecidos que de una u otra forma permitieron robustecer la calidad del presente proyecto..

INDICE DE CONTENIDOS

RESUMEN

El aprendizaje online es una de las características de este milenio, debido a la gran cantidad de información que se genera a diario y la accesibilidad a dispositivos para acceder a la web. A partir de este cambio en la perspectiva del aprendizaje varias instituciones a nivel mundial han concebido el nuevo paradigma de la enseñanza online como una oportunidad para globalizar sus programas. Así desde ya hace unos cuantos años, se dictan cursos, seminarios e incluso carreras universitarias por medio de la web, lo que ha generado una gran cantidad de interacción entre docentes y estudiantes ubicados en diferentes partes del mundo, interacción que crea masividad de datos que no han sido explorados, por lo que se pretende en este proyecto realizar tal exploración mediante técnicas de inteligencia artificial.

En el presente trabajo se realiza la extracción de los mensajes en foros que se generan a partir de interacción de estudiantes en los cursos abiertos masivos online (MOOC) de la plataforma Udacity (mediante técnicas de scraping), con el fin de rescatar patrones de texto mediante el uso de técnicas de procesamiento de lenguaje natural (como el reconocimiento de entidades y el análisis de n-gramas) y técnicas de aprendizaje automático (utilizando el algoritmo Naïve Bayes).

PALABRAS CLAVE: MOOC, Scrapy, Udacity, pln

ABSTRACT

The online learning is one of the characteristics of this millennium, due to the large amount of information generated daily and accessibility devices to access the web. From this change in the perspective of several learning institutions worldwide they have conceived the new paradigm of online education as an opportunity to globalize their programs. So from a few years ago ,offers courses, seminars and even university courses via the web, which has generated a lot of interaction between teachers and students located in different parts of the world, creating massive interaction data that have not has been explored, so it is intended to conduct such exploration project using artificial intelligence techniques.

This proyect makes the extraction and analysis of data generated from interaction of students in massive open online course (MOOC) of Udacity platform by scraping techniques, in order to rescue text patterns using techniques of processing natural language(like entity recognition and analysis of n-grams) and machine learning techniques (using Naïve Bayes algorithm).

KEYWORDS: MOOC, Scrapy, Udacity, pln

INTRODUCCIÓN

El análisis de datos es una subárea dentro de la ciencia de los datos donde se utilizan técnicas para identificar relaciones significativas, patrones y tendencias de un conjunto de datos que permitan descubrir información subyacente y oculta, lo que conlleva al descubrimiento de nuevas relaciones entre las variables de un conjunto de datos.

Clasificar texto, realizar predicciones a futuro, descubrir nuevas cadenas de ADN, analizar los mercados industriales, son algunas de las aplicaciones que se realizan en la actualidad mediante el análisis de datos, por lo que es un campo importante dentro del marco de la investigación académica como de la investigación empresarial.

En este proyecto se intenta identificar patrones de texto de estudiantes que se forman académicamente mediante plataformas MOOC, analizando las incertidumbres expuestas dentro de los foros de discusión, previo a esto se realizará la extracción de los datos aplicando técnicas de Web Scrapy sobre la plataforma Udacity para un curso en específico.

Con la llegada del aprendizaje en línea sobre los últimos años, la construcción de cursos con metodologías innovadoras ha sido un desafío para las instituciones educativas, por lo que esta investigación es importante dentro de este contexto pues realiza un análisis automatizado de la participación de los estudiantes en un curso MOOC para ayudar a entender la problemática dentro de este lo que podría en trabajos futuros ayudar a mejorar esas metodologías

El primer capítulo consta de una revisión teórica de los temas que se hacen referencia durante este documento, temas como el aprendizaje online, el modelo de aprendizaje de los MOOC's, las herramientas para realización extracción de datos de la web, las técnicas principales de procesamiento de lenguaje natural y también conceptos relacionados al aprendizaje automático son vistos de manera general en este capitulo

En el segundo capítulo se analiza las principales plataformas que ofrecen cursos MOOC a nivel mundial como son coursera, edx, udacity y miriadax, con el ánimo de realizar una comparación que permita elegir la plataforma correcta para la extracción de mensajes de sus foros, teniendo en cuenta algunas variables que se definieron importantes para el presente proyecto y que son descritas en ese mismo capítulo.

En el tercer capítulo se realiza la extracción de datos de los mensajes del curso "Introducción a la psicología" de la plataforma Udacity, por medio del framework Scrapy

para el lenguaje Python. También constan en este capítulo los modelos en los que se estructuró la información extraída de los foros.

En el cuarto capítulo se realiza un análisis exploratorio de los datos por medio de técnicas de procesamiento de lenguaje natural, extrayendo palabras que son más frecuentes en los mensajes previo proceso de segmentación y normalización, reconociendo entidades de las que más se enfatiza en el curso, y por último se analizó la polaridad de los mensajes utilizando el algoritmo de aprendizaje automático Naïve Bayes.

CAPITULO I

ESTADO DEL ARTE

1.1. El E-learning

El anglicismo e-learning significa en idioma castellano “aprendizaje electrónico”, y se define como el proceso educativo llevado por los medios que proporciona la web , como son el correo electrónico, las páginas web, los foros de discusión ,la mensajería instantánea, el video chat y las redes sociales, rompiendo el modelo clásico del encuentro físico entre estudiante e instructor.

De manera más general se puede definir al e-learning como “la instrucción entregada sobre un dispositivo digital, como un ordenador o un dispositivo móvil que tiene la intención de apoyar el aprendizaje” (Garrison, 2011).

Así el e-learning elimina barreras de espacio y de tiempo permitiendo al estudiante inmediatez de contenidos, actualización constante, gestión real del conocimiento y la reducción de costos para obtener contenidos de calidad.

Por otro lado cabe mencionar algunas desventajas que se podrían presentar en el e-learning, como la carencia de habilidades para el aprendizaje autónomo y colaborativo que los estudiantes posean, las pocas competencias tecnológicas por parte del profesor y estudiantes y la adaptabilidad a nuevos modelos donde prima el compromiso y la motivación.

El e-learning por lo tanto altera la naturaleza del aprendizaje y combate las deficiencias de la educación superior como las largas charlas magistrales y permite presentar nuevas soluciones pedagógicas para reavivar los objetivos de la educación. (Garrison & Anderson, 2010)

1.1.1. Surgimiento.

El término fue acuñado a mediados de los 90, con el desarrollo de la World Wide Web, donde los pocos beneficiados que se comunicaban por la red creaban espacios para grupos de discusión asíncronos independientes del tiempo y la localización. Estas comunidades tenían como finalidad la discusión y reflexión con el fin de construir el aprendizaje de forma mutua. (Garrison, 2011)

El aprendizaje se generaba a partir de la declaración de un problema, la discusión por parte de los participantes del foro y finalmente la resolución al problema, dando lugar al aprendizaje mutuo y colaborativo.

Por tanto no existe un ente creador del e-learning pues se trata de un fenómeno que surgió con el descubrimiento del internet y las tecnologías asociadas, se puede decir que nació por la necesidad mundial de compartir los conocimientos sin limitaciones tempo-espaciales.

1.1.2. Comunidades de Aprendizaje.

"Una comunidad de aprendizaje es una comunidad cohesionada que encarna una cultura de aprendizaje. Los miembros están involucrados en un esfuerzo colectivo de entendimiento" (McConnell,2006,p.19). Es por tanto responsabilidad de cada integrante el participar con experiencias, problemas y soluciones que enriquezcan a la comunidad y así mismo, de otra manera no se cumplen con los objetivos del aprendizaje colaborativo.

El e-learning debe mantener el espíritu con el que surgió, de manera que una comunidad de aprendizaje debe interactuar siempre con el objetivo de facilitar, construir y validar la comprensión de los integrantes, ese contexto ofrece resultados sociales positivos. (Garrison & Anderson, 2010)

Los tres elementos básicos para una comunidad de aprendizaje según (Garrison & Anderson, 2010) deben ser la presencia cognitiva que se refiere a los resultados educativos pretendidos y cognitivos, la presencia social que refiere a la capacidad de los participantes de proyectarse a sí mismos dentro de la comunidad y un tercer elemento también fundamental que es la presencia docente como equilibrador y coordinador de las actividades educativas y también como encargado de realizar la crítica constructiva hacia los integrantes. Por tanto el ambiente que generen estos tres factores es la base del aprendizaje e-learning teniendo en cuenta cuanto influye una comunidad equilibrada entre los participantes para sus correctas proyecciones.

En las comunidades de aprendizaje el conocimiento es construido colectivamente, sin embargo puede ser relativo pues existen múltiples verdades e interpretación y el aprendizaje es basado en un problema que los mismos integrantes exponen, a diferencia de los modelos clásicos donde los problemas los coloca el profesor y el estudiante los aprende, limitando la interpretación y creatividad. (McConnell, 2006)

1.1.3. El e-learning en el siglo XXI.

La adopción masiva de la comunicación electrónica, nos ha movido a una etapa donde el e-learning de desarrolla de manera global, la educación y aprendizaje por medios electrónicos en el siglo XXI es ya un hecho en el que todos los involucrados en el proceso educativo quieren sumergirse.

El rasgo esencial del e-learning no se limita solo a la facilidad de acceder a la información sino en su potencial comunicativo e interactivo, es decir en como el conocimiento se puede formar dinámica e intelectualmente y de manera estimulante mediante mejores vías para procesar información, darle sentido y recrear esa información. (Garrison & Anderson, 2010).

Son varias instituciones de renombre que ofrecen en la actualidad diferentes tipos de programas educativos por medio de la red, el costo-beneficio de cambiar el método clásico de aprendizaje por un método recursivo online es enorme tomando en cuenta los costos por parte de las organizaciones educativas en infraestructura, servicios, materiales, cantidad de profesores y también el beneficio de los estudiantes en cuanto a costos de transporte, materiales educativos, tiempo empleado en clases presenciales, pago a las instituciones, todos estos costos prácticamente se eliminan.

La utilización de la red para ofrecer programas educativos ha sido dispareja pudiendo distinguirse 3 grupos a nivel mundial. (Martínez, 2012)

El primer grupo conformado por países como EEUU, Australia, Reino Unido y Canadá que llevan varios años formando programas maduros a través de la red. El segundo grupo lo conforman países de la unión europea e hispanos donde el e-learning se encuentra en desarrollo. El número de universidades y estudiantes que se forman por la web, está en crecimiento y las metodologías de aprendizaje online aún se encuentran en evaluación. El tercer grupo se encuentra conforma por países de África y Asia donde apenas existe, debido al escaso poder tecnológico o por haberse incorporado tarde al desarrollo de las nuevas tecnologías. (Martínez, 2012)

Como centro del e-learning según (Garrison & Anderson, 2010) se halla una transacción constructiva que requiere cooperación, el e-learning dicen “es estimulante porque da valor tanto al contexto como a los contenidos. El reto se encuentra en diseñar un contexto con niveles suficientes de presencia social, que sea coherente con los contenidos y refuercen los objetivos educativos.”

1.1.4. Características frente a la educación clásica.

En la tabla 1 tomada de (Zavando, 2011) se realiza una comparativa del e-learning y la educación tradicional tomando ventajas y desventajas de ambas.

Tabla 1 Educación Clásica Vs E-learning

Categorías	e-Learning	Enseñanza Clásica
Flexibilidad	Puede ser seguida al propio ritmo del estudiante, sin horarios fijos ni predefinidos o bien con horarios programados con anticipación	Las sesiones tienen que ajustarse a factores de la organización, del profesorado y de los estudiantes.
Cobertura	Se puede acceder desde cualquier parte del mundo	Los estudiantes y el profesor deben estar presentes en el mismo lugar geográfico
Acceso	El estudiante requiere tener acceso durante un tiempo suficiente a la infraestructura tecnológica que le permita realizar su aprendizaje.	Se requiere una sala para que se produzca la interacción entre instructor y estudiantes
Costos de operación	No existen costos de transporte y estadía para estudiantes ni instructores, no existe costos de salas y equipos, pero se paga por servidores para dictar el curso. No hay costos de materiales impresos	Se incurre en costos por cada sesión realizada en horas del instructor, transporte y estadía para instructor y estudiantes, uso de salas y equipos
Costos de inversión	Dependiendo del tipo de infraestructura que se utilice podría ser alta , sin embargo actualmente existen plataformas gratuitas	El costo de preparación del curso puede ser bastante bajo
Estilos de aprendizaje	El estudiante elige el método más apropiado para sus habilidades, existe total autonomía en el estilo.	La enseñanza se focaliza a un estilo de aprendizaje promedio , por lo que los estudiantes con diferentes ritmos de aprendizaje tienen menores tasas de retención

Contenidos	Perfecto para capacitar en conceptos y habilidades, pero no para desarrollar habilidades personales o cambiar actitudes.	Ciertos temas necesariamente requieren la interacción física entre instructor y alumno, como la danza o el teatro.

Fuente: (Zavando, 2011)

1.2. MOOC (Massive Open Online Course).

MOOC es el acrónimo en inglés de Massive Online Open Courses (o Cursos online masivos y abiertos), su objetivo es pasar de entornos cerrados a un entorno totalmente abierto sin limitaciones de ninguna clase, con la posibilidad de que miles de personas de todo el mundo se unan a diferentes iniciativas educativas.

Este fenómeno educativo es relativamente reciente, siendo el primero creado en el año 2008, el mismo que generó considerable atención y relevancia en instituciones educativas de alto prestigio y deslumbro una nueva oportunidad de negocio a ser explotada. (Yuan & Powell, 2013). El termino MOOC fue introducido por Dave Cormirer y Bryan Alexander para designar un curso realizado por George Siemens y Stephen Downes en el año 2008.El curso se titulaba “Connectivism and connective knowledge”, fue seguido por 2300 alumnos por medio de Internet. (Vázquez, López, & Sarasola, 2013)

Pensar estudiar en las instituciones del mayor prestigio a nivel mundial desde nuestro hogar era casi impensable hace una década, los MOOC’s son el principio de este sueño, actualmente se imparten cursos de manera masiva, abierta y online con el fin de globalizar la educación queriendo así eliminar las barreras que los recursos económicos, la distancia y tiempo que no permitían acceder a una educación llamada de las élites. Ahora podemos acceder a un curso de la Universidad de Harvard, Stanford, Cambridge sin movernos de la silla.

Este nuevo tipo de formación supone un reto para todos pues se debe reorientar las metodologías actuales para adentrarse en formas de diseñar materiales interactivos, colaborativos y ubicuos. (Vázquez et al., 2013)

Los MOOC están teniendo repercusiones en aspectos socioculturales educativos y tecnológicos debido a los siguientes factores de cambio a nivel mundial según (Vázquez et al., 2013).

- La globalización y el fuerte proceso de internalización
- La creciente demanda de acceso a la Educación Superior
- La necesidad constante de información a lo largo de la vida
- El acceso cada vez mayor a tecnologías de la nueva era
- El cambio de modelo de negocio en el ámbito educativo.

1.2.1. Características.

Las siglas MOOC se forman debido a las 4 principales características que debe cumplir un curso de este tipo:

Masivos: Deben estar orientados a asimilar a cientos de miles de alumnos registrados. Por lo tanto su infraestructura tecnológica (hardware y software) deben ser de altas prestaciones.

Abiertos: Desde el enfoque de gratuidad cualquier persona en el mundo debe poder registrase sin pagar anticipadamente para acceder a este. Los costos sin embargo podrán ser por certificados de aprobación con lo que se podrá costear parte de las iniciativas.

Online: Todos los cursos deben ser totalmente accesibles por medio de internet, ninguna de las actividades como exámenes deberá ser de manera clásica.

Curso: Debe constar de una estructura y metodología que permita que el contenido se oriente al aprendizaje de los participantes. Así un curso debe tener actividades como tareas, evaluaciones, proyectos, y otras actividades que refuercen los contenidos, de tal forma que se pueda parametrizar si el estudiante ha logrado los objetivos de enseñanza del curso. Deberá existir también una calendarización que permita una línea base del curso para el seguimiento de los tutores a sus estudiantes en sus actividades.

Los MOOC han acaparado un interés mundial debido a su gran potencial para ofrecer una formación gratuita, de calidad y accesible a cualquier persona independientemente de su país de procedencia, su formación previa y sin la necesidad de pagar por su matrícula. (Liyanagunawardena, 2013)

"Desde comienzos del año 2010, la incursión de estos cursos empezó a ser vista desde una perspectiva más academicista cuando diferentes universidades de reconocido prestigio iniciaron sus actividades masivas, entre otras, Stanford, Yale, Harvard, MIT, Universidad de Pennsylvania y la Universidad de Toronto" (Vázquez & López, 2014).

Actualmente estos cursos no solo son ofrecidos por universidades sino también por sociedades bastante conocidas a nivel mundial como National Geographic Society , The Museum of Modern Art o The World Bank con el propósito de que su conocimiento sea libre y gratuito para todo el mundo.

1.2.2. Surgimiento

Fathom fue un portal de aprendizaje online creado en el año 2000 por la Universidad de Columbia. Se ofrecía cursos de Columbia y de otras instituciones académicas por un valor determinado. Llegó a tener 65.000 usuarios en más de 50 países. Sin embargo las tasas de matrícula que se cobraban a los estudiantes no llegaban para cubrir los costos de realización de los materiales y terminó cerrándose en el 2003. (Vinader & Abuín, 2013)

AllLearn fue otra plataforma creada por las universidades de Stanford, Yale y Oxford en 2001. En su inicio ofrecía cursos solo para alumnos de estas universidades, pero posteriormente permitió el acceso a usuarios de otras instituciones con el objetivo de masificar el número de estudiantes. En este caso los cursos tenían un valor de 200 dólares, pero el modelo de negocio también fracasó y en 2006 la plataforma cerró sus puertas virtuales (Vinader & Abuín, 2013)

En agosto de 2007, David Wiley, profesor de la Universidad Estatal de Utah experimento con un curso abierto donde se unieron 50 alumnos de 8 países lo que demostraba la aceptabilidad que podrían tener lo cursos dándoles mayor empuje económico. George Siemens y Stephen Downes, ambos profesores de la Universidad de Manitoba en Canadá, un año más tarde crean el curso "*Connectivism and Connective Knowledge"* atrayendo a 2500 personas de diferentes partes del mundo a inscribirse.

Pero el impulso definitivo a este tipo de cursos se produce en los Estados Unidos, principalmente en la Universidad de Stanford, con Sebastian Thrun y Peter Norvig por un lado, Andrew Ng y Daphne Koller, por el otro. Los primeros crean, a finales del año 2011, un MOOC denominado "*Introduction to Artificial Intelligence"* con más de 160.000 alumnos inscritos en más de 200 países, este evento marca un hito en la historia de los MOOC's. Como resultado de este experimento, en enero de 2012, se da paso a la primera iniciativa privada para la distribución de MOOC con el nombre de Udacity. (Gutiérrez & Nava, 2014)

En Octubre de 2011 Ng y Koller, también profesores en Stanford, crean el curso "*Introduction to Databases"* con más de 100.000 estudiantes online. Dos meses después del nacimiento de Udacity, Koller y Ng crearán su principal competidor, la plataforma Coursera" (Gutiérrez & Nava, 2014) con un capital de inicio procedente de la inversión privada por 16 millones de dólares y teniendo a inicios de 2012 la participación ya de tres universidades más.

El Massachusetts Institute of Technology(MIT) y la Universidad de Harvard lanzan en mayo del 2012 otra iniciativa, la plataforma EDX, a la que se ha unido la Universidad de Berkeley y la Universidad de Texas. Con una inversión de 60 millones de dólares para soporte de infraestructura, licencias y recursos humanos; cuenta actualmente con más de 1 millón de estudiantes registrados en los cursos a través del mundo. (Harvard, 2012)

En Diciembre de 2012 por iniciativa de Universia (red de universidades de habla hispana y portuguesa) y Telefónica Learning Services –compañía especializada en ofrecer soluciones integrales de aprendizaje online para la Educación, con el fin de fomentar la difusión del conocimiento en abierto en el espacio iberoamericano de Educación Superior, ponen a disposición de las 1.232 universidades iberoamericanas que integran la red Universia, una plataforma para crear e impartir MOOC's sin costo alguno bajo el nombre de MIriadaX y bajo el auspicio del Banco Santander y Telefónica. (miriadaX, 2014)

1.2.3. Los MOOC como modelo de negocio

La aparición del modelo educativo basado en MOOC's encaja en el patrón evolutivo de la teoría de Christensen's llamada innovación disruptiva. (Mazoue, 2013)

Cinco características en particular, definen un núcleo extensible para una educación de precisión según Mazoe.

1. Su metodología basada en la investigación produce arquitecturas de cursos de aprendizaje-optimizado.
2. Es efectivo al máximo porque hace el aprendizaje individualizado.
3. Es eficiente ya que se basa en competencias específicas.
4. Es escalable.
5. Es rentable.

"Al cerrar el diferencial de calidad entre los planes de estudio basados en MOOC y la instrucción tradicional, los cursos fabricados con precisión eliminarían gradualmente la distinción entre la "gama alta " y "gama baja" de la educación. Sólo habría un tipo de aprendizaje - optimizado para cada individuo. Efectivamente diseñado y organizado en un

plan de estudios coherente, los Mooc's tendrían el potencial de marcar el comienzo de una nueva modelo de negocio de educación a alto nivel" (Mazoue, 2013)

Sin embargo no todo es tan perfecto como parecería para los MOOC's, pues una de las características como la rentabilidad aún no se ha resuelto del todo y es un punto en contra de este modelo, debido a los altos costos en infraestructura inicial, mantenimiento, docentes y marketing es prácticamente imposible mantenerlos si no absorben ingresos de alguna parte en contradicción con su filosofía de ser totalmente abiertos.

Algunas iniciativas han optado por implantar el cobro de una cantidad simbólica en concepto de matrícula, esto puede mejorar notablemente la rentabilidad ejemplificando que si se cobraría un dólar por alumno se podría recaudar miles de dólares dependiendo de la cantidad de alumnos. (Vinader & Abuín, 2013).

Otras iniciativas como Coursera, Udacity y Edx siguen ofreciendo los cursos de manera gratuita, pero colectan una tarifa económica en caso de que el estudiante requiera el certificado que acredite los conocimientos que se ha impartido en el curso.

1.2.4. El modelo pedagógico de los MOOC

Una filosofía acorde a como se distribuye al usuario el MOOC, resulta una característica que promoverá al alumnado hacia la adquisición de competencias y que evitará el abandono de los cursos. Existe una bifurcación en cuanto a metodologías de los MOOC's, los cMOOCS se basan en el aprendizaje distribuido en red, el contenido es escaso y la búsqueda la debe realizar el estudiante como ente autónomo del aprendizaje pero compartirla con su red pública. Y Los xMOOCs basan su metodología en cambio en pruebas estandarizadas y objetivas que permiten la acreditación del estudiante, su proceso de aprendizaje consta de video simulación, tareas semanales, autoevaluaciones e incluso proyectos de curso. (Vázquez et al., 2013)

1.2.5. Plataformas que Ofrecen MOOC's

1.2.5.1. Udacity

Según la página oficial de Udacity, esta nació como un experimento de la Universidad de Stanford en el que Sebastián Thrun y Peter Norvig, ofrecieron el curso online "Introducción a la Inteligencia Artificial" abierto para cualquier persona y de forma gratuita. Más de 160.000 estudiantes en más de 190 países se inscribieron y no mucho después, Udacity tuvo su nacimiento.

Para octubre del 2013 ya contaba con 1.6 millones de estudiantes registrados, y una inversión de 15 millones; pero Thrun se sentía decepcionado pues menos del 10% de alumnos inscritos en los cursos terminaban sus clases, así declaro que no estaba educando a la gente como los demás deseaban o como él deseaba, literalmente dijo que Udacity era un producto malísimo. (Chafkin, 2013)

Luego Thurun se dedicó a realizar cursos con metodologías lo más dinámicas posibles, por lo que lanzo un curso de Estadística con clases totalmente fáciles de entender, pero esta iniciativa no resolvía el problema de abandono, por lo que se entusiasmó por otro tipo de iniciativas para incentivar a los estudiantes a acabar los cursos ofreciendo crédito por lo que se vendía cursos por 150 dólares cada uno. Los resultados fueron desastrosos ,entre los alumnos que tomaron el curso llamado "remedial math" sólo el 25 % aprobó y cuando se comparó con las clases tomadas presencialmente fueron aún mas desalentadoras pues una persona tenía 52 % más probabilidades de pasar en modalidad presencial que tomándolo como una clase de Udacity. (Chafkin, 2013)

Udacity no ha brindado estadísticas para el año 2014, aunque se asoció con el Instituto George Tech para una maestría online para sus estudiantes con un costo de 6600 dólares, bajo el auspicio de AT&T con lo que se pretende recaudar fondos para mantener la iniciativa de cursos MOOC libre de costos.

A la fecha Udacity ofrece 45 cursos en 6 categorías: Data Science, Web Development, Software Engineering, Android, Georgia Tech Masters in Cs (En asociación con Georgia Tech y AT&T ofrece certificaciones por parte de Georgia Tech con un costo de 6600 dólares por 3 semestres) y una última categoría de cursos no asociados a la tecnología "Non-Tech Classes"

1.2.5.2. EDX

En mayo de 2012 los presidentes de la Universidad de Harvard y el Instituto Tecnológico de Massachusetts (MIT) anuncian la inversión de sesenta millones de dólares para la creación de edX con el ánimo de mejorar la educación en el aula, no suplantarla y actualmente está respaldada por un consorcio formado por veinte universidades de todo el mundo. (Kolowich, 2013)

Se considera una plataforma sin ánimo de lucro sin embargo algunos de los cursos cobran una tarifa por certificación. El dinero recaudado se reinvierte, según esta iniciativa, en la mejora de la plataforma. Según se explica en su propia página web, éstas pueden ser de tres tipos:

Honor Code: acredita sin coste alguno la superación del curso pero no refleja la identidad del alumno.

ID Verified Certificate: permite certificar que el alumno ha completado el curso, identificando al estudiante con su nombre y su foto. Este tipo de diploma tiene un coste asociado y sólo está disponible en una selección limitada de cursos.

Por último, el XSeries Certificate para aquellos estudiantes que haya superado una serie de cursos vinculados a una materia determinada.

Edx tiene disponibles 277 cursos divididos en 29 categorías principales: Arquitectura, Arte y Cultura, Biología y Ciencias de la vida, Negocios y Administración, Química, Comunicación, Computación Científica, Economía y Finanzas, Educación, Electrónica, Energía y Ciencias de la Tierra, Ingeniería, Estudios Medioambientales, Comida y Nutrición, Salud y Seguridad, Historia, Humanidades, Derecho, Literatura, Matemáticas, Medicina, Música , Filantropía, Filosofía y Ética, Física , Ciencia, Ciencias Sociales y Estadísticas y Análisis de Datos.

1.2.5.3. Coursera

Fue fundada en el otoño de 2011 por Daphne Koller y Andrew Ng en la Universidad de Stanford, como una empresa con fines de lucro con una inversión inicial de 22 millones y las universidades de Stanford, la universidad de Princeton y las Universidades de Michigan y Pennsylvania como socias. (Yuan & Powell, 2013)

En términos cuantitativos, Coursera quizá sea la plataforma más importante. Cuenta actualmente con 447 cursos distribuidos en 23 categorías diferentes y, aunque la mayor parte son en idioma inglés, también pueden cursarse en francés (12), español (12) o chino (10).

Según sus propias fuentes, en septiembre del 2013 Coursera contaba con 17 millones de alumnos inscritos procedentes de 190 países en más de 440 cursos. Teniendo el curso más popular 240.000 estudiantes. Ha conseguido generar 251.9 millones de minutos de clases y 590.000 hilos de discusión (Vinader & Abuín, 2013)

Una característica única de Coursera fue llamada Signature Track, ofreciendo un certificado del curso pero que requiere la verificación de identidad de los estudiantes. La verificación de identidad del estudiante involucrado consiste en el análisis del rostro mediante cámara web durante los exámenes.

Ofrece a las empresas una manera de colaborar en la formación continua de sus trabajadores, de este modo, las empresas que así lo deseen pueden poner a disposición de

sus recursos humanos una amplia variedad de cursos que, a cambio de una cantidad económica, la plataforma se encargará de validar y certificar convenientemente. Se trata, por tanto, de un modelo de negocio alternativo que ayuda al mantenimiento de la iniciativa gratuita. (Vinader & Abuín, 2013)

1.3.5.4. MIriadaX

Miríada X es un proyecto de formación en línea que tiene su origen a principios del año 2013 por el Banco Santander y Telefónica, a través de la Red Universia y Telefónica Learning Services y basado en la plataforma de software libre WEMOOC. (miriadaX, 2014)

Según datos oficiales de su página web en su primer aniversario, el 21 de noviembre de 2013, Miríada X contó con la participación de 28 universidades de seis países iberoamericanos: Argentina, Colombia, España, Perú, Puerto Rico y República Dominicana; 730 profesores y 96 cursos impartidos. (miriadaX, 2014)

Su éxito la ha llevado a convertirse en una plataforma de formación online de referencia no solo a nivel español sino también europeo, en el que más de un 35% de los Mooc's provienen de universidades españolas según Open Educación Europa, siendo MiríadaX un factor clave en la evolución educativa española según muestra el informe de la Sociedad de la información en España en el año 2013. Al día 01 de septiembre de 2014, Miríada X alcanzó 868.000 usuarios inscritos en sus cursos y esta puesta a disposición de 1262 universidades iberoamericanas.

Muchas universidades de habla hispana dictan sus cursos actualmente mediante Miriadax, como la Universidad Politécnica de Valencia(España), la Universidad Rey Juan Carlos(España), la Universidad Nacional de Quilmes(Argentina), la Universidad Pompeu Fabra(España), la Universidad de Málaga(España), la Universidad de Ibagué (Colombia), la Universidad compútense de Madrid(España),la Universidad Carlos III de Madrid(España), La universidad Católica Santo Toribio de Mogrovejo(Perú), la universidad de Celaya (México) entre muchas otras universidades.

1.3. Exploración y Extracción de Datos en la Web (Web Scraping).

1.3.1. Web Scrapy.

Es una técnica utilizada para la extracción de información proveniente de páginas web no estructuradas, mediante la cual un robot programado accede y guarda el contenido de las páginas simulando la navegación de un ser humano automatizando y acelerando la extracción de datos.

Los datos extraídos se transforman en datos estructurados que pueden ser almacenados y analizados en una base de datos local o en cualquier otra forma que permita su análisis o utilización como archivos csv, json, xml u otros formatos de texto.

Para realizar web scraping se necesita llevar a cabo varias tareas secuenciales empezando por realizar la conexión al sitio, realizar los métodos de autentificación si se requiere, gestión de cookies, y la navegación de distintos enlaces en el mismo sitio y precedente a todas estas tareas se realiza un análisis de la estructura de sus páginas para conocer sus elementos html que identifiquen estructuralmente la página.

Sin embargo no todas las páginas son accesibles a estas técnicas debido a las medidas de seguridad como usos de sistemas de verificación, servicios comerciales antibots y antiscraping, uso de javascript, ajax, bloqueo de direcciones ip, monitorización del exceso de tráfico proveniente de una ip, uso de programación en capas y certificados de seguridad.

Existen 3 niveles de web scraping según “The Global Scraping Intelligence Plataform”. El nivel amateur que no satura el servidor y utiliza pocos números de direcciones ip, el nivel profesional que intenta ocultar sus objetivos usando scripts, programas y servicios en varias ips. Y un último nivel avanzado, que distribuye sus búsquedas usando un rango extenso de direcciones ips y cambia su comportamiento rápidamente para no ser descubierto.

1.3.2. Técnicas y Herramientas.

A partir del año 2010 el porcentaje de crecimiento de scraping a sitios web incremento paulatinamente pasando del 17 % en el 2010 al 23% en el 2014 según la empresa ScrapySentry Inc(empresa a nivel mundial dedicada a la lucha contra el scraping), por lo que se han incrementado también los métodos que permiten extracción de datos en diferentes plataformas y con diferentes tecnologías , por lo que se realizó una investigación de estas tecnologías para elegir la más asertiva a cumplir con los objetivos.

Algunas de las técnicas que se pueden utilizar para capturar la información de los sitios web son las peticiones mediante sockets con lo que se realiza peticiones http a un servidor remoto para obtener el contenido de varias páginas, el uso de expresiones regulares como medio para recuperar textos, parsers que se utilizan para recuperar información de documentos html , visión computarizada en la que se identifica y se extrae información como un humano lo haría y por último también se utilizan plugins para exploradores web que permiten ir capturando la información que se requiera y luego automatizar este proceso.

1.3.2.1. Web Scraper

Es una extensión únicamente para el explorador Chrome y se lo puede instalar desde las aplicaciones de google Chrome. Esta extensión permite seleccionar los elementos a los que se realizará la extracción de datos e ir configurando una araña web de manera muy fácil

Su uso resulta de mucha utilidad pues no se necesita programar ninguna línea para realizar la extracción de datos de cualquier página, solo se necesita seleccionar los elementos padre y los hijos y los datos que se extraerá de cada iteración.

Esta extensión permite crear mapas de sitios para irlos navegando e ir extrayendo los datos de diferente tipo como texto, links, imágenes, tablas y otros. Permite exportar los datos que hayan sido escarbados a un formato csv en su versión libre y a una base de datos en su versión pagada.

Sus características principales son descritas a continuación:

- Escarbado de múltiples páginas de inicio
- Escarbado mediante la opción de mapas de sitio.
- Múltiple tipo de selectores de datos
- Extracción de datos de páginas dinámicas
- Exportación de datos a CSV
- Importación y Exportación de Mapas de Sitio.
- Dependiente solo del Explorador Chrome

1.3.2.2. Curl

Es una librería de programación que permite conectarse a servidores por medio de los siguientes protocolos FTP, FTPS, HTTP, HTTPS, TFTP, SCP, SFTP, Telnet, DICT, FILE y LDAP según las especificaciones oficiales en su página web http://curl.haxx.se/ (CURL, 2014).

Es bastante flexible debido a que permite simular las acciones de un navegador web y extraer la información de manera automatizada permitiendo la utilización de cookies, conexión proxy, autentificación mediante usuario y contraseña. Esta biblioteca es de código abierto por lo que se la puede implementar en más de 30 lenguajes de programación distintos como php, java, c#, c++ , python etc.

Según su página oficial CURL fue creada a finales de 1996 por Daniel Stenberg cuando se le ocurrió que podía hacer cálculos para realizar cambio de moneda, pero todos los datos necesarios se encontraban en la web, por lo que necesitaba automatizar la recuperación;

Daniel adoptó la herramienta de código abierto en línea de comandos existentes HTTPGET que se había lanzado recientemente, le introdujo nuevas características y lanzó una propia versión más tarde. (CURL, 2014)

1.3.2.3. Import.io

Es un navegador web específico para la extracción de datos de manera visual y simplificada creado en el año 2012 por David White, Mateo Pintor y Andrew Fogg con el ánimo de hacer que los datos estén disponibles para todos

Los datos que se recogen de la extracción se almacenan en los servidores de import.io para que puedan ser descargados en formato csv,xls, google sheets o json. Para usuarios de mayor conocimiento import.io ofrece la recuperación de datos mediante la creación de un api y así poder integrar los datos con páginas web

Entre otras de sus características permite combinar un conjunto de datos de hasta 100 fuentes en un solo conjunto, permite convertir un cuadro de búsqueda de un sitio web en una api consultable y una de las características más impresionantes es el alto rendimiento que tiene pues realiza la adquisición de datos de forma paralelizada distribuyendo automáticamente el escarbado a su arquitectura en la nube.

1.3.2.4. Scrapy Python

Es un marco de colaboración para extracción de datos de la web que puede ser usado en diferentes aspectos, como la minería de datos, el procesamiento de información o de archivos históricos, la extracción de imágenes y video para la generación de bibliotecas, la búsqueda automatizada de información y otros fines tanto comerciales como no comerciales.

Según su página oficial (Scrapy, 2015), Scrapy nació en el año 2008 y está escrito en Python en Código Abierto , cuenta con características como extracción de datos en fuentes HTML y XML, soporte integrado para limpieza de datos mediante una colección de filtros reutilizables, soporte integrado para generar exportaciones en formatos JSON,CSV y XML, extensiones incorporadas para el manejo de Cookies y sesión, compresión HTTP, autentificación HTTP, cache HTTP, suplantación de agente de usuario, restricción de profundidad de rastreo y archivos robots.txt.

Sus altas prestaciones y al ser parte de un lenguaje como Python permiten que sea una opción para el escarbado de datos con infinidad de soluciones como soportes de rastreo según URLs, resoluciones de cache DNS, depuración mediante consola de Telnet, seguimiento y control de robots mediante servicio web y algunas otras.

1.3.2.5. Helium Scraper

Es un software escarbador que utiliza un algoritmo de búsqueda donde asocia los elementos a ser extraídos con sus propiedades HTML, este método permite la extracción y manipulación con ayuda de Javascript y scripts SQL.

Helium Scraper es un software de pago con un costo desde $99 dólares el paquete básico con una sola licencia y que permite realizar labores de escarbado multihilo hasta $699 por 10 licencias de usuario.

Cuenta con plantillas prediseñadas para la exportación de datos, y formatos como CSV, XML, bases de datos MS Access y archivos de comando para MySql. Con 2 características únicas como el manejo de la información con Java Script se convierte en una herramienta potente que permite desde la búsqueda por árboles hasta la generación de datos nuevos a partir de los extraídos.

1.3.2.6. Apache Nutch

Según la información compartida en la wiki de apache, Nutch es una herramienta de rastreo web menos atractiva para principiantes pues no cuenta con una interfaz gráfica pero es ampliamente utilizada por su extensibilidad, escalabilidad y por ser de Código Abierto.

Apache Nutch esta codificado en lenguaje Java, su robot fue escrito desde 0 especialmente para este proyecto por lo que cuenta con una arquitectura altamente modular , esto permite a desarrolladores escribir plugins muy completos para la extracción y guardado.

Por ser una herramienta de una potencialidad enorme que permite ser ejecutado en un grupo de hasta 100 máquinas, varios buscadores han sido construidos gracias a este proyecto entre ellos Wikia Search,Krugle, DiscoverEd y Creative Commons

Nutch es un proyecto con propósitos para rastreo de datos a gran escala en la web compatible con bases de datos como Apache Casandra, bases de datos distribuidas como Apache HBase, sistemas de socialización de datos como Apache Avro y sistemas de datos distribuidos como Haddop.

1.3.3. Cuestiones Legales del Scrapy

Los precedentes acerca de la legalidad de realizar web scraping son escasos por lo que se debe se debe acudir a las condiciones de uso y derechos de autor de los documentos a acceder o sitios para conocer sobre la legalidad o ilegalidad del uso de sus datos. (Mitchell, 2013).

La legalidad del web scraping está asociada a los derechos de autor, por lo que implica el uso que se le dé a la información que se extraiga el motivo de legalidad al que debe atarse los problemas legales. Mientras que muchos sitios en los que la información es la razón de comercio, como la venta de artículos o libros, el acto de scraping hacia esos documentos y la publicación de estos mismos infringirían en un robo de documentos.

Otro aspecto interesante es que buscadores como google, yahoo o bing utilizan robots para surfear entre las páginas de sitios web para indexarlas a sus índices sin ningún consentimiento de los propietarios de los sitios, esto podría resultar ventajoso o indeseado para los dueños de las páginas web pero no ha sido denunciado como un acto indebido.

1.3.4. Araña Web

También conocida como spider, consiste en un programa que ingresa a páginas web para conocer sus elementos e información, el método que utiliza es visitar una url, identificar los links que existen en dicha url y los empieza a recorrerlos de manera metódica, hasta los niveles que sean necesarios.

Muchos de los buscadores de la web como google y bing utilizan estas técnicas para mejorar el servicio de búsqueda, indexando aquellas páginas con mayor número de referencias al principio de las búsquedas del usuario.

1.4. Procesamiento de Lenguaje Natural (PLN)

El PLN se encarga del estudio de métodos usados para la comprensión de la comunicación entre humanos por parte de las máquinas, lo que conlleva a otras áreas como la lingüística, la inteligencia artificial y la programación para poder llevar a cabo con su objetivo principal.

Entre las más ambiciosas investigaciones del PLN se encuentra el diseño de lenguajes de entrada y salida que permitan que las máquinas usen el lenguaje fluidamente y flexiblemente así como lo realizamos los humanos. (Dale, Moisi, & Somers, 2000).

Por tanto se puede definir como los métodos para el análisis y entendimiento de cualquier lenguaje por medio de la tecnología. Pueden ser procesos tan simplistas como el análisis de

frecuencias de palabras para comparar diferentes estilos o tan complejos como el entendimiento de lenguajes poco formales. (Bird, Klein, & Loper, 2009).

1.4.1. Origen

El desarrollo de los lenguajes matemáticos formales en los años 30 y 40 tuvieron un profundo efecto en el campo computacional en años posteriores tanto así que se generaron nuevas áreas de investigación que trataban de simular funciones cognitivas humanas por medio de la representación de lenguajes formales en el computador. (Dale et al, 2000).

En los años 50, los modelos de Turing de computación algorítmica intentaban hacer un modelo simplificado de neuronas descrito en términos de lógica proposicional, en esta misma época también aportaron los trabajos de Kleene sobre automatización y expresiones regulares. Shannon aplicó modelos probabilísticos de modelos ocultos de Markov para automatizar el procesamiento del lenguaje. Chomsky considero los estados finitos de las máquinas para caracterizar la gramática y definir un lenguaje de estado finito como un lenguaje. Estos modelos dejaron asentada la teoría formal de los lenguajes con el uso de algebra y símbolos. (Kumar, 2011).

1.4.2. Técnicas en el Procesamiento de Lenguaje Natural

1.4.2.1. Tokenización y Segmentación de Oraciones

Es el proceso mediante el cual se segmenta un texto mediante la separación de palabras a estas palabras se las conoce como tokens. El método de programación que usualmente se utiliza es separarlos por espacios o por los signos de puntuación para separar oraciones o segmentos de texto que deberían ser tratados en diferente contexto. El dilema que se presenta es cuando no se utiliza correctamente la puntuación y el significado cambia sin que sea intención del texto.

1.4.2.2. Análisis Morfológico

Se intenta mediante la morfología (rama de la lingüística que estudia la estructura interna de las palabras) la correcta definición y clasificación de las palabras dentro del contexto de un discurso mediante uso de raíces, diccionarios, unidades léxicas compuestas y otros.

1.4.2.3. Análisis Sintáctico

Realiza el análisis de un conjunto de palabras de acuerdo a las reglas gramaticales del lenguaje, haciendo hincapié en la separación de los sujetos, verbos y predicados en una

oración. Este método permite deducir cual es el sentido de un discurso según como se ordenan las palabras y como se conjugan dentro de este.

1.4.2.4. Análisis Semántico

El análisis semántico se encarga de realizar un análisis estructurado de una declaración en un texto para intentar determinar su significado. Se intenta realizar un análisis real del texto haciendo uso de la morfología, la sintaxis de las palabras dentro de la oración y su interpretación.

1.4.2.5. Análisis Pragmático

Dentro de un análisis pragmático se intenta resolver el significado de referentes (Russell & Norvig, 2004), considerando que en el lenguaje muchas de las veces se evade palabras por saberse anticipadamente de que se está hablando, para una máquina es mucho más complicado inducir esto, por lo que este tipo de análisis intenta resolver las referencias que no se escriben como por ejemplo al escribir <<viajaremos a quito hoy>> debemos resolver el sujeto que no se escribe directamente y el tiempo que es relativo al escribir la palabra hoy, en vez de la fecha exacta.

1.4.3. Aplicaciones del PLN

1.4.3.1. Categorización de Textos

Es una de las aplicaciones del Procesamiento de Lenguaje Natural que consiste en catalogar documentos en una o varias clases. Es una subárea importante debido a la cantidad enorme de documentos que se generan y que sin etiquetado resultaría imposible realizar búsquedas.

La categorización se realiza en dos fases: la primera consiste en una fase de entrenamiento donde se obtiene una generalización inductiva de un conjunto de documentos y la segunda fase del test que se encarga de evaluar la efectividad del mismo por lo que se necesita de un conjunto de textos clasificados manualmente (Sierra, 2006).

Categorizar texto implica varias técnicas como la limpieza de los datos, que conlleva a realizar una eliminación de palabras sin contenido semántico (stop words en inglés). En segunda instancia se procede a utilizar tokenizadores, lematizadores, etiquetadores morfosintácticos y analizadores sintácticos que permitan identificar de manera adecuada de lo que el texto trata y se elimine cualquier tipo de ambigüedad. (Benítez, Escudero, Kanaan, & David, 2013)

1.4.3.2. Traducción Automática

Es un área que se especializa en la investigación y utilización de técnicas para traducir texto o habla de un lenguaje a otro. Se utilizan reglas lingüísticas en niveles sintácticos y semánticos de los dos idiomas por lo que es necesaria la utilización de grandes diccionarios e incluso métodos estadísticos basados en corpus ya traducidos.

Traductores automáticos como los que utilizan el explorador Chrome hacen uso de estas técnicas de procesamiento de lenguaje para traducir una página web a varios idiomas. También sitios como youtube permiten realizar incluso las traducciones de un idioma a otro desde el audio de una conversación, sin que sea del todo confiable debido a tener que depender de otro campo del pln como el reconocimiento del habla.

Las traducciones complejas incluyen procesos de alineación de palabras según las reglas morfológicas que cada lenguaje utiliza, métodos estadísticos que permiten decidir según la posición de la palabra en la oración darle el significado más apropiado, otros métodos de traducción se realizan por aprendizaje automático donde el computador aprende de cuerpos de texto ya traducidos para entender como traducir un nuevo cuerpo de texto a otro idioma.

1.4.3.3. Análisis de Discurso

Es una disciplina que usa el discurso para descubrir patrones de lenguaje socio-psicológicos de la persona que relata el discurso. Se intenta descubrir al discurso como un hecho de comunicación en un contexto que engloba varios aspectos socio cultural. Las máquinas que realizan este proceso deben simular una comprensión crítica aproximada de lo que el discurso transmite.

1.4.3.4. Generación automática de Resúmenes

Estas aplicaciones buscan realizar una abstracción de lo que el cuerpo de un texto pretende comunicar, por lo que realizan un análisis de su conjunto sintáctico para transformarlo a un conjunto semántico y poder sintetizarlo en una plantilla mucho más compacta que el documento original. También se utilizan métodos más sencillos como el conteo de frecuencias de palabras para realizar un resumen mediante algoritmos.

1.4.3.5. Reconocimiento de Entidades

El reconocimiento de entidades mediante procesamiento de lenguaje natural permite extraer nombres de personas, lugares y organizaciones para clasificarlos según la categoría a la que corresponden. Es una tarea de bastante utilidad cuando se busca en la web en fuentes

infinitas de contenidos, por lo que varios sitios web utilizan esta tarea para categorizar sus documentos o subpáginas.

1.4.3.6. Análisis de Sentimientos

El análisis de sentimientos o minería de opiniones busca analizar texto para determinar la polaridad del locutor indicando si es positivo, negativo, neutro o si refleja alguna emoción en especial. Se emplean diccionarios donde se clasifica las palabras por emociones y así poder clasificar el texto, lo que conlleva también a tener que aplicar al texto reglas de desambiguación, sinonimia y reglas propias para cada lenguaje

1.4.3.7. Reconocimiento y Predicción del Habla

El reconocimiento y predicción del habla es una de las tareas más complicadas en el campo del procesamiento del lenguaje natural que pretende que sea el computador quien reconozca información contenida en una señal de voz emitida por un ser humano.

Su complejidad se deduce a que interfieren incertidumbres de carácter fonético, acústico, fonológico, sintáctico y pragmático; que no permiten se logre una interpretación correcta del mensaje analizado.

En la actualidad varios son los usos que se le da al reconocimiento y predicción del habla como en los controles por comando, los sistemas diseñados para discapacitados, los sistemas de emergencia por voz e incluso sistemas de domótica. Sin embargo debido a las dificultades nombradas la confiabilidad de estas aplicaciones aún no está en un rango aceptable.

1.5. Aprendizaje Automático

Es el campo de estudio de la computación que pretende darles a los ordenadores habilidades para aprender sobre algo para lo que no han sido explícitamente programados por medio de diferentes tipos de algoritmos de aprendizaje. Estos algoritmos tienen el objetivo de optimizar un criterio de desempeño a partir de un conjunto de datos o experiencias pasadas. (Alpaydin, 2004)

El aprendizaje automático es necesario cuando no es posible abstraer conclusiones de manera manual sobre un conjunto de datos ya sea por su tamaño, por existir demasiadas variables cambiantes o simplemente porque el problema cambia en el tiempo o depende de un medio ambiente particular con lo que se necesita programas que sean adaptables a estos cambios continuos. (Alpaydin, 2004)

Un agente de aprendizaje puede ser diseñado con un elemento de acción que es el elemento que decide como va actuar a futuro en base a un elemento de aprendizaje que le ayuda a tomar mejores decisiones. Este elemento de aprendizaje se ve afectado por tres aspectos: ¿qué componentes del elemento de acción tienen que aprenderse?, ¿qué realimentación está disponible? y ¿qué tipo de representación se usa para los componentes?'. (Russell & Norvig, 2004)

1.5.1. Tipos de aprendizaje

1.5.1.1. Aprendizaje Supervisado

Consiste en aprender una función a partir de ejemplos de sus entradas y sus salidas. Las entradas deben ser ejemplos positivos y negativos. Por ejemplo una empresa desea realizar una clasificación sobre el tipo de carros que las familias prefieren. Las muestras positivas serán las que una familia se encuentre en el carro, y las negativas serán las de los otros carros. El algoritmo debe encontrar cuales son las expectativas de un carro familiar por tanto debería aprender cuáles carros serán preferidos por una familia y cuáles no. Después de algunas discusiones con expertos sobre el tema, las conclusiones son que las características que separan a una familia de un carro a otro son el precio y el motor. Estos dos atributos se convierten en las entradas para reconocer las clases de salida. (Alpaydin, 2004).

Otro ejemplo muy utilizado para su comprensión es de a partir de un conjunto de pacientes con tumores en el organismo (malignos y benignos), se tiene información sobre el tamaño del tumor y sobre la edad de los pacientes. A partir de estos datos se puede utilizar un algoritmo para el aprendizaje supervisado que tendría como resultado un clasificador de si un tumor será benigno o maligno a partir de ingresar la edad del paciente y el tamaño del tumor.

En conclusión el aprendizaje supervisado parte de un conjunto de objetos descritos por un vector de características y la clase a la que pertenecen cada uno de ellos (conjunto de entrenamiento) (Sierra, 2006).

Este tipo de aprendizaje has sido utilizado en numerosos problemas de distinta índole tales como el diagnóstico de enfermedades , la aprobación de créditos en la banca, la predicción de quiebra en empresas, la detección de anomalías en cromosomas y otros (Sierra, 2006).

1.5.1.2. Aprendizaje No Supervisado

“Consiste en aprender a partir de patrones de entradas para los que no se especifican los valores de sus salidas” (Russell & Norvig, 2004).

Se utiliza este tipo de aprendizaje en análisis de redes sociales identificando grupos cohesivos de amigos , organización de clústeres para saber cuáles deberían trabajar más cerca , en segmentaciones de mercado para ofrecer los productos de empresas a las personas correctas, también en análisis de datos astronómicos en la clasificación de galaxias.

Clasificar objetos en diferentes grupos es necesario para poner orden en el mundo, los métodos llevados a cabo para realizar un aprendizaje no supervisado requiere de métodos estadísticos que permitan agrupar casos sobre los cuales se miden diferentes variables (Sierra, 2006).

1.5.1.3. Aprendizaje por Refuerzo

El aprendizaje por refuerzo cubre métodos en los cuáles un agente necesita saber que algo bueno a ocurrido, es decir una recompensa o refuerzo, como en un juego de ping-pong cuando se realiza un punto, la recompensa es un subir al marcador, por tanto el agente en este tipo de aprendizaje debe saber aprender de las acciones que le llevan a una recompensa y también de las que no lo llevan (Russell & Norvig, 2004).

Se pueden identificar 4 elementos claves en el aprendizaje por refuerzo: una política que define el comportamiento en un momento dado , una función de recompensa que define una meta a corto plazo, una función de valor que define el estado a largo plazo y un modelo de ambiente que define comportamientos cambiantes del ambiente.

Se puede dividir aprendizaje por refuerzo en pasivo y en activo. En el aprendizaje por refuerzo pasivo la política del agente está fijada en el estado s, su meta es aprender la bondad de la política, en cambio en el aprendizaje por refuerzo activo debe decidir qué acciones tomar a través de la experimentación.

Uno de los dilemas que aparece en el aprendizaje por refuerzo activo es el de la exploración o explotación debido a que el agente tiene que tomar decisiones en cuanto a la exploración que ha hecho, pero también puede seguir explorando para conseguir mejores resultados a futuro. (Cetina, 2012). Un agente en este caso debe tener un compromiso entre la explotación de manera que maximice su recompensa y la exploración para maximizar su comportamiento a largo plazo.

CAPITULO II.

ELECCIÓN DE LA PLATAFORMA

2.1. Elección de la Plataforma

En el presente proyecto se realizó el análisis de 4 plataformas para cursos MOOC, explorando las estructuras de sus cursos, la codificación html utilizada en sus páginas y las características que las constituyen .Se realizó el análisis a las siguientes plataformas: Coursera, EdX, Udacity, y MiríadaX por ser las de mayor prestigio y cubrir el modelo de Masividad, Apertura, En línea y ser de tipo Curso.

2.1.1. Parámetros a evaluar

Debido a que el objetivo del presente proyecto es extraer la información que se genera en los foros de discusión de los MOOC's, se evaluó los siguientes parámetros para realizar un análisis comparativo de la plataforma con mayor factibilidad para ejecutar la extracción de mensajes de sus foros.

2.1.1.1. El número de estudiantes inscritos en los cursos.

El número de estudiantes debe reflejar el concepto de masividad teniendo que ser superior al número de estudiantes que podrían tomar un curso de manera presencial. Cursos con números de estudiantes superiores a 1000 debería ser la métrica mínima para que el curso sea elegible para la presente investigación aunque este número no es un indicador suficiente debido a la alta tasa de abandono en los cursos de todas las plataformas investigadas.

2.1.1.2. La interactividad entre estudiantes mediante mensajes en foros

Los foros de discusión en un MOOC reflejan la cantidad de estudiantes que siguen presentes en el curso, por tanto son superfluos cursos con grandes cantidades de estudiantes inscritos, sino se refleja mediante el uso de los foros.

Los foros son la parte viva de un MOOC pues al no existir la interacción física entre estudiantes es el único medio junto al chat que permiten la sociabilización de ideas, criterios, dudas, sugerencias y reclamaciones. Se preferirá por lo tanto plataformas donde el nivel de interacción entre estudiantes por medio de los foros sea alta considerando que deben existir al menos 100 hilos de discusión por curso.

2.1.1.3. La factibilidad de acceder a la página mediante técnicas "web scraping".

Como se describió en el capitulo anterior existen métodos por parte de las páginas web que no permiten que sus páginas sean extraídas como el uso autentificaciones, el uso de

captchas (comprobación electrónica para saber si el usuario es o no humano), el usos de javascript y ajax, denegación de permisos, el uso de programación dinámica en capas y uso de programas antiscraping.

Este parámetro es de vital interés para nuestro objetivo pues aunque se cumple con los demás parámetros, sino existe el método que permita la extracción de datos no se lo podrá llevar a cabo la extracción de los datos. Por lo tanto la plataforma a elegir no deberá utilizar métodos antiscraping citados anteriormente que interrumpan las técnicas de extracción de datos a emplear.

2.1.1.4. Los metadatos asociados a una entrada en un foro

Algunos metadatos como el número de respuestas, los likes, el número de veces que ha sido leído el mensaje, la fecha de posteo, las etiquetas asociadas, los tipos de filtro para las entradas en un foro serán también tomados en cuenta para la elección de la plataforma, pues indican la relevancia de las entradas en los foros y son indicadores válidos para realizar análisis de datos.

2.1.1.5. La temática de discusión entre estudiantes.

Las temáticas de discusión deberán ser de carácter específico a la materia, por tanto las plataformas deberán tener categorizadores que permitan filtrar otras discusiones ajenas a la materia.

Se preferirá aquellos cursos con temáticas de carácter socio humanístico que no involucren problemas de carácter numérico que son propios de materias como matemáticas, física o química, esto debido a que se necesita información que pueda ser analizada con técnicas de procesamiento de lenguaje natural y los datos números tienen poca probabilidad de ser analizados correctamente

2.1.1.6. Idioma en que se dictan los cursos

Dando prioridad a las plataformas que dicten cursos en idioma español, aunque el peso dentro de nuestro análisis es menor a las otras características teniendo en cuenta que las plataformas con más aceptación dictan sus cursos en idioma Inglés, y solamente Miriadax ofrece cursos totalmente en idioma español

2.1.1.7. Disponibilidad de cursos al culminar

Debido a que existe variabilidad entre el tiempo que dura un curso y algunas instituciones cierran definitivamente sus cursos una vez terminadas las actividades con los estudiantes, se dará prioridad a las plataformas que permitan el acceso a los datos generados de los cursos una vez que culmine, debido a que el tiempo necesario para el desarrollo de scripts puede superar al tiempo de disponibilidad de datos del curso

2.1.2. Análisis comparativo de las plataformas candidatas

La plataforma Udacity como ya se mencionó en el capítulo anterior es una plataforma de alta aceptación con más de 2 millones de estudiantes entre todos sus cursos, tiene una media de 64741 estudiantes por curso (se puede ver detalle en el anexo 6), un nivel alto de interactividad registrando una mínimo de 2000 entradas por curso , la entrada a los foros es de manera libre (no requiere autentificaciones) y sus elementos de programación de la página no son complejos, sus 40 cursos son dictados en idioma Inglés , como metadatos existentes a cada entrada de un foro están la cantidad de vistos de la entrada (para mayor detalle en el anexo 1) , el número de respuestas, los votos positivos o negativos a la entrada(ver anexo 2) , la ultima vez por quien fue posteado, la fecha exacta en que se ingreso el mensaje, el contenido al que se refiere la entrada y las temáticas a las que se refieren los participantes si son descriptivas(anexo 3 y anexo 4) a pesar que en ocasiones se refieren a temas propios de la plataforma o sobre las calificaciones. Sus características son perfectas para nuestro objetivo de extracción de datos.

Coursera es la plataforma con mayor popularidad entre los MOOC's, cuenta con una media de estudiantes registrados por curso de 80.000, su interactividad entre estudiantes es alto registrando más de 10000 discusiones por foro. (Collazos, 2014) Existen metadatos asociados a una discusión como los puntos asignados, las veces que se ha ingresado, los likes, las fechas de publicación y el tag asignado al mensaje, existen cursos en varios idiomas incluyendo el Inglés, el Español, Chino, Francés y otras, el contenido que se encuentra en los foros es de varios tipo como reclamos, preguntas ajenas a la materia y problemas técnicos.

El inconveniente la plataforma Coursera en nuestro análisis es la autentificación que requiere para dirigirse a las discusiones cuenta con elementos programados en javascript y ajax, lo que dificulta el acceso de las arañas web, y el tiempo de respuesta de los servidores es en ocasiones demorado por lo que aunque siendo la plataforma de mayor relevancia no se realizo la extracción de datos de esta plataforma.

EdX por su parte ha tenido una buena acogida por los estudiantes en sus dos años de funcionamiento, tiene un promedio estimado de 700000 estudiantes por curso, la interactividad entre estudiantes tiene una media de 5000 respuestas por curso pero su modelo de foros es menos intuitivo por lo que algunos cursos se redirigen a otras plataformas, los metadatos asociados a un mensaje en el foro son el número de votos a favor, el tiempo trascurrido desde que se publicó y el número de respuestas. Existen cursos en idioma Inglés y Chino. El contenido de discusión permite el filtrado de categorías lo que favorecería nuestra extracción de datos, sin embargo lo realiza mediante tecnologías dinámicas que no nos permitirán acceder a todas sus categorías por scraping. Si requiere autentificación para el acceso a las discusiones, y su programación contiene elementos ajax y javascript, por lo que no es conveniente para el objetivo del presente proyecto.

MIríadax cuenta con un menor número de estudiantes que las plataformas antes mencionadas con un promedio estimado de 292 por curso (se puede ver más detalle en el Anexo 7), la interactividad entre estudiantes genera un aproximado medio de 1100 mensajes(Anexo 7) por curso, lo que significa una interactividad media tomando en cuenta el número promedio de inscritos, los metadatos asociados a una entrada son el número de accesos, la fecha exacta de posteo, la categoría a la que pertenece, la ultima respuesta por quien fue hecha, y la valoración positiva o negativa del mensaje. En esta plataforma encontramos en su mayoría cursos impartidos en idioma Español. Cuenta con categorización de los mensajes, pero se requiere autentificarse para acceder a los foros y el formulario de autentificación está programado en javascript lo que dificulta el acceso de nuestra araña web. El poco número de estudiantes inscritos frente a sus competidoras también es un punto en contra en nuestro análisis.

Tabla 2. Comparación de Plataformas de MOOC's

Parámetro	Coursera	EdX	Miríadax	Udacity
Número de Estudiantes Inscritos por curso (MEDIA)	Estimada media de 80000 estudiantes	Estimada media de 70000 estudiantes	Media de 292 estudiantes (Anexo 7)	Media de 64741 estudiantes (Anexo 6)
Interactividad entre Estudiantes	Alta	Alta	Media	Alta(Anexo 2 y Anexo 5)

Factibilidad de Acceso para aplicación de Scraping	Codificada en ajax y java script que no permiten el fácil acceso.	Tiene elementos ajax y java script que no permiten el fácil acceso de una araña web.	Diseño de elementos HTML cambiante de curso a curso	Elementos HTML tienen comportamiento predecible
Requiere Autentificación	Si	Si	Si	No
Idioma de Cursos	Inglés, Español, Chino, Francés y Otros	Inglés, Chino, Español, Francés	Español	Inglés
Metadatos Existentes en una entrada	Más de 5	Menos de 5	Más de 5	Más de 5 (Anexo 4)
Temáticas de Discusión	Todos los asuntos del curso, problemas técnicos y reclamos	Todos los asuntos del curso por categorías pero son páginas que utilizan javascript o ajax.	Existe categorías para filtrar las temáticas propias de la materia y las del curso en general	Varían las temáticas aunque existe filtrado de mensajes.(En el Anexo 3 se muestra una gráfica de los mensajes categorizados)
Disponibilidad	Pocos cursos	En su mayoría	Algunos cursos	En su mayoría

2.1.3. Elección de Plataforma para Extracción de Información

Una vez descritas todas las plataformas plataforma se decidió realizar la extracción de las discusiones de los foros de la plataforma Udacity por ser una plataforma con una alta interactividad entre estudiantes, su diseño de elementos de elementos HTML es de fácil acceso, cada entrada de discusión cuenta con más de 5 metadatos, su modelo de foros no requiere autentificación para la lectura de mensajes y no contiene elementos antiscraping Una de las desventajas para el presente proyecto es que sus cursos son netamente dictados en inglés, por lo que el análisis de sus datos tendrá que ajustarse a este idioma.

En la tabla 3 se muestra los datos informativos del curso al que se realizó la extracción de las discusiones en sus foros, siendo un curso de un área socio humanístico que tiene un considerable número de alumnos inscritos (46630) y un número de hilos de discusión mayor a 100.

Tabla 3. Datos de Curso MOOC “Introducción a la Psicología”

Curso	Introducción a la Psicología
Resumen	Introducción a la Psicología es un viaje a través de todos los principales conceptos y principios psicológicos. El conocimiento obtenido de este curso permitirá a los estudiantes para evaluar críticamente la investigación psicológica y tener un conocimiento más profundo del pensamiento y el comportamiento humano
Duración	4 meses , con un trabajo de 6 horas semanales
Nivel:	Principiante
Número de Estudiantes:	46630
Iniciado por:	San Jose State University
Requisitos y Requerimientos:	No hay prerrequisitos, pero los estudiantes deben ser curiosos por la naturaleza humana
Contenidos del Curso:	Lección 1: Introducción a la Psicología Lección 2: Métodos de Investigación en Psicología Lección 3: La Biología del Comportamiento Lección 4: Sensación y percepción Lección 5: Desarrollo Humano Lección 6: Conciencia

	Lección 7: Aprender Lección 8: Memoria Lección 9: Lenguaje y Pensamiento Lección 10: Inteligencia Lección 11: Motivación y Emoción Lección 12: Estrés y Salud Lección 13: Personalidad Lección 14: Comportamiento Social Lección 15: Trastornos Psicológicos Lección 16: Tratamientos de los Trastornos Psicológicos
Instructores	Susan Snycerski- Instructor- Greg Feist-Instructor Lauren Castellano- Course Developer

CAPITULO III

DESARROLLO DE SCRIPTS PARA LA OBTENCIÓN DE DATOS Y ESTRUCTURACIÓN DE LA INFORMACIÓN

3.1. Herramienta para Extracción de Datos

Para la extracción de datos se eligió el framework de desarrollo Scrapy para Python por ser de altas prestaciones, ser de código abierto, estar programado en un lenguaje muy flexible como Python lo que permite realizar la limpieza de datos de forma eficiente mediante programación, tener la posibilidad de guardar los datos en diferentes estructuras y por ser un lenguaje utilizado de manera usual en proyectos de investigación dentro de la comunidad científica.

3.1.1. Arquitectura Scrapy Python

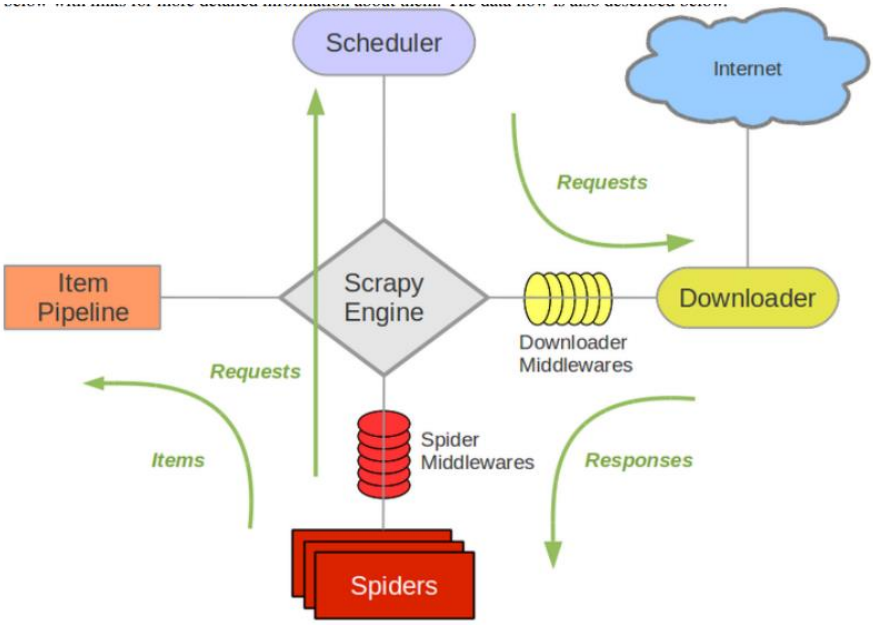


Figura 1 Arquitectura Scrapy
Fuente: (Scrapy, 2015)

3.1.1.1. Componentes de la Arquitectura de Scrapy

Según la documentación oficial de Scrapy Python en http://doc.scrapy.org (Scrapy, 2015) el diseño de funcionamiento del framework consta de 7 componentes que evidencian su funcionamiento en la figura 1, y que se describen a continuación cada elemento.

3.1.1.1.1. Scrapy Engine(Motor)

El motor de Scrapy controla el flujo de datos por los demás componentes desde el Dowloader pasando hacia las Spiders para su raspado y luego hacia ItemPipelines, este

proceso se repite hasta que Scheduler no tenga más solicitudes, en ese momento el motor cierra el dominio.

3.1.1.1.2. Scheduler(Programador)

Se encarga de recibir solicitudes del motor y devolverlas nuevamente a este de manera que no colapsen cuando el motor las solicite nuevamente.

3.1.1.1.3. Spiders(Arañas):

Son clases escritas de manera personalizada para extraer los datos necesarios de las respuestas, se incluyen en estas clases las reglas necesarias de comportamiento de la araña (rastreo) (Scrapy, 2015)

3.1.1.1.4. Dowloader (Descargador):

Se encarga de buscar páginas web y pasar el contenido hacia del motor para que este las transmite a las arañas.

3.1.1.1.5. Item Pipeline(Artículos):

Una vez que los datos han sido extraídos en forma de artículos es necesario estructurarlos en algún modelo de datos, de estas tareas se encarga ítem pipeline al realizar la limpieza, validación y persistencia de los datos.

3.2. Estructuración de la Información

3.2.1. Modelo de Objetos JSON De Foros Udacity

3.2.1.1. Descripción de JSON

JSON (Javascript Object Notation) según su página web oficial (Json, 2015) es un formato ligero para el intercambio de datos que es fácil de leer para los humanos y las máquinas . Es construido sobre dos estructuras: una colección de pares nombre/valor que se conocen en diferentes lenguajes como objetos, diccionarios, tablas hash, listas codificadas o array asociados, y la segunda estructura es una lista de valores como un array, un vector, una lista o una secuencia. Debido a que estas estructuras son universales, Json es independiente de cualquier lenguaje de programación pero muy útil para su intercambio.

En la estructuras de un archivo Json un objeto comienza con “{“ (llave de apertura) y termine con “}” (llave de cierre). Cada nombre es seguido por: (dos puntos) y los pares nombre/valor están separados por “,” (coma). Un formato bastante sencillo entendible para máquinas y humanos.

3.2.1.2. Diseño de archivos JSON

3.2.1.2.1. Archivo JSON para cursos

Este archivo es el más sencilla de todos por lo que se compone de 1 objeto con 3 valores informativos para cada curso, su identificador, el resumen sobre lo que trata el curso, y el título del curso.

```
{
    "idCourse": "",
    “summary": "",
     "title": "",
}
```

El segundo formato de salida es el archivo más complejo del modelo debido a que cada página que se extrae consta de discusiones, respuestas, y comentarios, y se las debe guardar en una misma estructura json.

De esta tabla se deriva un identificador para cada hilo de discusión (idPost), el texto principal de la discusión (text), la relación que tiene con los contenidos del MOOC (relatedTo), la fecha en que fue ingresada la discusión (date), las veces que ha sido visto por los participantes del curso (seen), el número de votos obtenidos (votes) y el número de respuestas que acumula este hilo de discusión (numberRespuests)

Puesto que cada discusión tiene comentarios se añade en esta estructura un arreglo con los identificadores de los comentarios(idComment), un arreglo para los identificadores de los usuarios que postearon comentarios(idUserComment), las fechas en que fueron escritos los comentarios(dateComment), y el texto de estos(comments). Las posiciones de los arreglos están directamente relacionadas entre sí.

También a cada discusión están asociadas respuestas por lo que se agregan los identificadores de las respuestas en un arreglo (idRespuest), el texto de la respuesta (respuest) ubicado en un arreglo, la fechas en que se posteo tales respuestas (respuestDate) ,los identificadores de los usuarios a los que pertenece cada respuesta(user) y el número de votos que acumula esta respuesta(voteRespuest). Las posiciones de estos arreglos pertenecientes se corresponden entre sí en el número de posición.

```
{
        "idPost": "",
        "text": “",
        "relatedTo": "",
        "date": "",
        "seen": "",
        "votes”: "",
        “numberRespuests”: "",
        “idComment”: [ ],
        "idUserComment": [ ],
        "dateComment": [ ],
         "comments": [ ],
        "idRespuest": [ ],
        "respuest": [ ],
        "respuestDate": [ ],
        "user": [ ],
        “voteRespuest”: [ ]
}
```


El ultimo archivo json de usuarios corresponde a la información relativa a cada usuario componiéndose de un ítem para el identificador del usuario (user), su biografía (bio) siendo un dato opcional, el nombre del usuario en udacity(name), desde cuando es miembro de udacity(memberSince), la edad (age) siendo opcional , la última vez visto en el curso (lastSeen) , el karma que es el número que significa cuantos puntos ha acumulado por realizar discusiones y mensajes con mayor aceptación, y la localización que se refiere a de donde proviene este participante, siendo un dato no obligatorio debido a que pocos estudiantes proveen estos datos a la plataforma.

{

"user": " ",

"bio": " ",

"name": " ",

"memberSince": " ",

"age": "",

"lastSeen": "",

"karma": " ",

"location": " "

}

3.2.2. Modelo Relacional de Base de Datos de Foros de Udacity

3.2.2.1. Descripción del modelo

En un principio del proyecto solo se proyecto realizar la estructuración de los datos extraídos a archivos Json, pero por pedido del director del presente proyecto se previno la necesidad de realizar también la estructuración en una base de datos libre como MySql, con la finalidad de comprender mejor las relaciones de los hilos de discusión y estructurar de manera relacional los datos.

El modelo lógico que se diseño (Figura 2) para almacenar la información se compone de 5 tablas descritas en idioma inglés para evitar confusiones de traducción entre la plataforma elegida Udacity y la base de datos.

La primera tabla (Course) es una tabla informativa de los cursos que provee la plataforma a los estudiantes, en esta se guardara el id del Curso (idCourse), un titulo del curso (titleCourse), y un resumen acerca de lo que se estudiará en cada curso (summaryCourse).

La tabla principal (post) se refiere a cada entrada en el foro propuesta por un usuario , esta tabla contiene campos como el título del mensaje en el foro(titlePost), el cuerpo del texto(textPost) , el identificador del curso(idCourse), la fecha en la que se posteo(datePost), el número de veces que ha sido visto por otros usuario(seen), el identificador del

post(idPost), el número de respuestas(numberAnswers), y los votos o likes que tiene cada hilo de discusión(votes).

Una tercera tabla (Answer) está diseñada para las respuestas a cada entrada en la tabla principal Post, esta tabla se compone de un identificador de Respuesta(id_Answer) el texto de la respuesta(textAnswer), la fecha en que se hizo cada respuesta(dateAnswer), el identificador del usuario que realizo la respuesta(idUser), el número de votos(likes) que tiene cada respuesta (votesAnswer) y el identificador del Post (idPost) al que responde .

La cuarta tabla corresponde a cada usuario que ha posteado dentro de los foros de Udacity en el curso que se está escarbando, en esta tabla es importante conocer el nombre del usuario, desde cuando es miembro, la edad, la última vez que visito la página, el karma que se refiere a cuántos puntos ha ganado de popularidad, su biografía, y de donde proviene.

La quinta tabla es una tabla que re refiere a cada comentario (Comment) dentro de los post generados, pudiendo un comentario ser hecho hacia un post directamente (idPost) o es un comentario a una respuesta (idAnswer), por lo que estos campos son opcionales teniendo que ser obligatoriamente uno de los dos. También se guardan en esta tabla el id del comentario (idComment), el texto (textComment), la fecha que se realizó (dateComment) y el identificador del usuario que lo escribió (idUser).

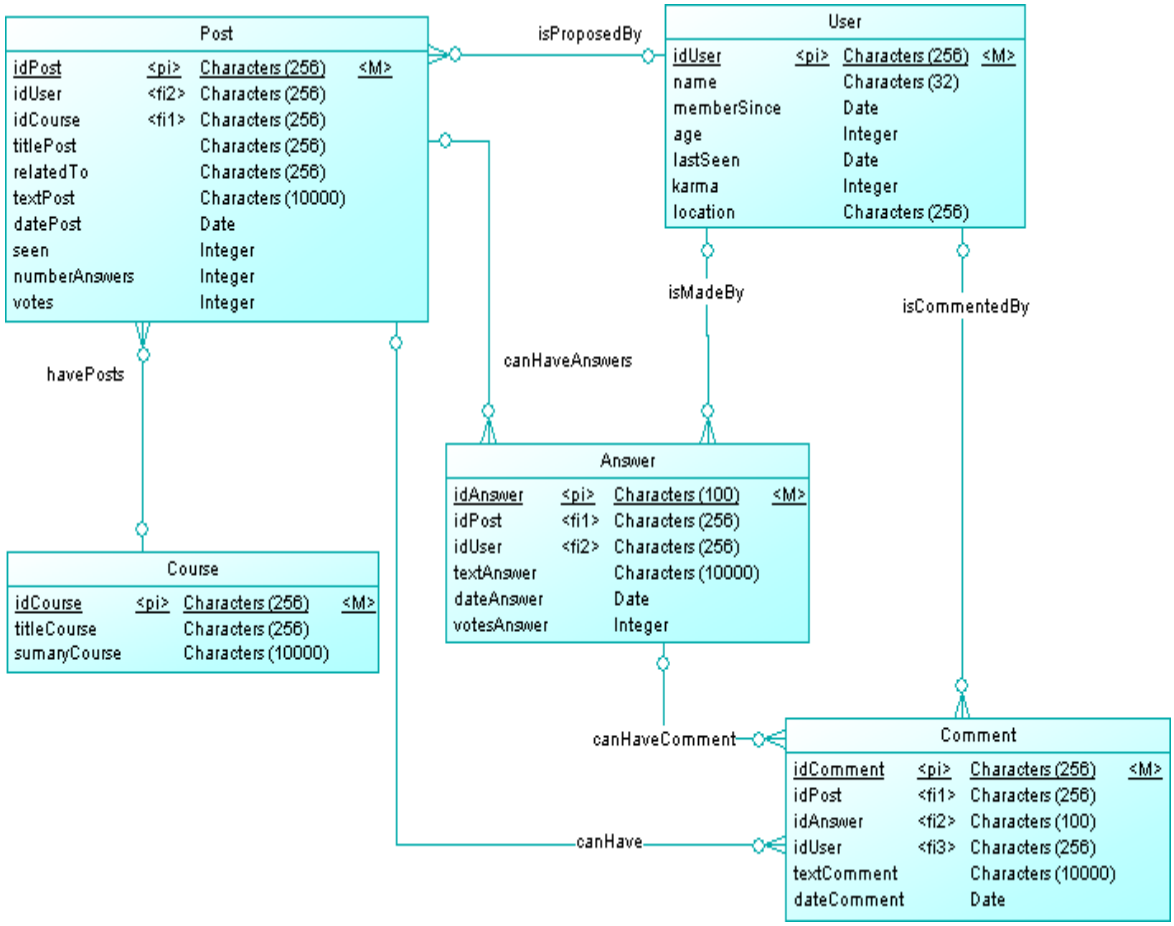


Figura 2. Modelo Relacional para guardar datos de Foros Udacity

3.3. Resultados de Scrapy en Udacity

Los resultados de los scripts obtenidos se realizaron por última vez el día 16 de marzo del 2015, por lo que la información que se haya actualizado a partir de tal fecha no constará en los datos que se extrajo. La base de datos completa de los datos que se extrajo hasta esta fecha se encuentra en el repositorio "https://github.com/snrepele/scrapyUdacity" con el nombre "udacity.sql"

3.3.1. Script #1. Capturar Todos los cursos

Objetivo: Obtener los todos los cursos con una descripción del curso y su identificador único y guardarlos en formato Json y en la tabla curso de la base de datos. En la tabla 4 se evidencia que se extrajo el 100% de direcciones url de los cursos. También se puede acceder a la codificación en Python en el repositorio https://github.com/snrepele/scrapyUdacity en el archivo "CursoP0.py"

Entradas: Página Catálogo de todos los cursos que se dictan en Udacity

Proceso: En este primer script se tiene que ir recogiendo las direcciones url de cada curso que se encuentra en el catálogo de Udacity, para luego ingresar a estas urls y obtener 3 campos en específico, el título del Curso, su identificador único, y un resumen del curso.

Método de Pruebas: Se realizó un conteo manual del número de cursos desde la página de Udacity para comparar con los resultados de script #1

Resultados:

Tabla 4. Resultados Script#1

Total de Ítems Extraídos:	72 items de 72 existentes
Porcentaje de Extracción.:	100%
Ingresados a Base de Datos:	72 items
Guardados en Archivo Json:	72 items

3.3.2. Script #2. Capturar links de hilos de discusión

Objetivo: Obtener los links de cada hilo de discusión para el curso de Introducción a la Psicología: La tabla 5 evidencia que se guardo todos los links de los hilos de discusión del curso y que no se los guardo en la base de datos, solo en un archivo temporal para el siguiente paso del scraping. Se puede acceder a este script en el repositorio "https://github.com/snrepele/scrapyUdacity" con el nombre "ExtraerLinksPost(Paso1).py".

Método de Pruebas: Se realizó el conteo del número de páginas diferentes del foro multiplicado por el número de hilos de discusión por página, y se le resto si la página no completa el máximo de hilos por página. El número de hilos debería ser de 1827

Entradas: Página principal de foros en el curso de Introducción a la Psicología y el número de total de páginas diferentes.

Proceso: En cada página diferente para los foros de este curso se guardo la dirección url de cada hilo de discusión

Resultados:

Tabla 5 .Resultados Script#2

Total de Ítems Extraídos:	1827 items de 1827
Porcentaje de Extracción:	100%
Ingresados a Base de Datos:	No se requiere guardado en Bd
Guardados en Archivo Json:	1857 items

3.3.3. Script #3. Capturar datos asociados a cada hilo de discusión

Objetivo: Obtener para cada hilo de discusión del curso Introducción a la Psicología las respuestas asociadas, los comentarios, los usuarios que participan y demás datos asociados. Este script se encuentra en el repositorio "https://github.com/snrepele/scrapyUdacity" con el nombre "entradaForo(Paso2).py"

Método de Pruebas:

a.- Se comprobó en primera instancia que todas las páginas a buscar respondan con un estado: HTTP 200 a las solicitudes de las arañas (comprobación mediante archivos LOG), luego se realizó una tabla con 10 ejecuciones para saber si el error era consistente o era aleatorio al estado del servidor. La tabla 6 muestra que en este paso se perdieron algunas páginas y algunos mensajes no pudieron ser extraídos por motivo que no se encontró los links en el servidor de Udacity.

b.- Una vez corregidos los errores por páginas con respuesta HTTP 404 en segunda instancia se comprobó con 10 consultas en la base de datos local con datos tomados de manera manual de la página de foros del Curso Introducción a la Psicología (Tabla 7):

Resultados Prueba (a):

Tabla 6.Comprobación Servidor Script #3

Número de Ejecución	Páginas Ingresadas (HTTP 200)	Páginas Error (HTTP 404)	Discusiones Extraídas	Respuestas Extraídas	Comentarios Extraídos
1	1851	2	1825	2665	1008
2	1845	8	1819	2645	994
3	1849	4	1823	2660	1004
4	1850	3	1824	2650	1004
5	1851	2	1825	2665	1008
6	1848	5	1822	2648	998
7	1849	4	1823	2660	1004
8	1851	2	1825	2665	1008
9	1852	1	1826	2665	1008
10	1847	6	1800	2645	998

El número de páginas que no responden entre las ejecuciones (1-10) es de 1 a 8 por lo que se procedió a realizar ajustes al código para realizar solicitudes repetitivas si las páginas no responden correctamente a las solicitudes.

Los resultados obtenidos luego de haber realizado dichos ajustes al código son que el 100% de páginas respondieron correctamente (HTTP 200) para poder ser extraídas.

Resultados Prueba (b)

Tabla 7. Datos Extracción Hilos Discusión

#Prueba	Datos a Comprobar	Encontrado en Base de Datos Local	Encontrado en Archivo Json
1	Discusión: ATTN: STAFF: Lesson 2 Problem Set - Problem 13 - Correct Answer Not Being Accepted	Si	Si
2	Respuesta: Still nothing. Why does it take so long for anyone from the staff to respond and fix accordingly? It shouldn't be so hard to fix. Change the structure of the code or whatever. Just make it work or please update to let us know why the trouble?	Si	Si
3	Discusion Id: 100118876	Si	Si
4	Comment Id: 100098587	Si	Si
5	Respuesta: First try: 0.219 seconds	Si	Si
6	Respuesta: Muchas gracias!!! desde Argentina! I loved the course, only I still couldn´t download my certificate of completion, why is that???? Val.	Si	Si
7	Comentario ID a Respuesta : 100077340	Si	Si
8	Respuesta Id: 100183243	Si	Si
9	Discusión: These questions are really tedious	Si	Si
10	Comentario : I had to pay I believe $150 for the course?	Si	Si

En esta prueba se extrajo variables aleatorias del foro entre hilos de discusión, comentarios y respuestas para comprobarlas en la base de datos local y en el archivo Json

La tabla 8 muestra que se pudieron extraer todas las discusiones, respuestas y comentarios de cada hilo de discusión y que se guardaron en el modelo relacional de la base de datos y en un archivo json.

Resultados:

Tabla 8. Resultados Script#3

Total de Ítems Extraídos:	1872 items de 1872
Número de Discusiones:	1827
Número de Respuestas:	2814
Número de Comentarios:	1021
Porcentaje de Extracción:	100%
Ingresados a Base de Datos:	1827 discusiones, 2814 respuestas y 1021 comentarios
Guardados en Archivo Json:	1872 items

3.3.4. Script #4. Obtener perfiles de usuarios

Objetivo: Obtener información del perfil de cada uno de los usuarios que realizó algún tipo de participación en los foros del curso de Psicología como generación de hilos de discusión, comentarios a estos, y respuestas a discusiones. No se toma en cuanta participación como puntos si el usuario no ha participado de forma textual en los foros. La tabla 10 muestra el porcentaje de cumplimiento de este objetivo y el número de perfiles extraidos. Se puede acceder al script en lenguaje Python en el repositorio "https://github.com/snrepele/scrapyUdacity" con el nombre "usuarios (Paso3).py".

Método de Pruebas: Se eligió 10 usuarios al azar de manera visual en la página de cursos y se comprobó su información con la base de datos local luego de que realizará la extracción de un total de 1650 perfiles de usuarios. La tabla 9 muestra el método de comprobación de los perfiles de usuarios a las que se accedió mediante scraping.

Entradas: Links de usuarios guardados en archivo json, mediante el script 3

Proceso: Redirigir a cada link de usuario la araña para que obtenga los datos del perfil de los usuarios en udacity

Resultados

Tabla 9. Datos Extracción Perfil Usuarios

#Prueba	Datos a Comprobar	Encontrado en Base de Datos Local	Encontrado en Archivo Json
1	idUser: users/100000133/vaishaks	Si	Si
2	Benjamin Keep	Si	Si
3	idUser:users/100017881/mario-haus	Si	Si
4	idUser: users/100033726/joseph-8	Si	Si
5	Tyler Eagan	Si	Si
6	idUser:users/100044229/kirankumar-vasudev-sripati	Si	Si
7	user:Robert Hustwick bio: I am 31 years old... trying to make good use	Si	Si
8	user: Paula Franzini: locatin: Quebec	Si	Si
9	User: snowpolar	Si	Si
10	User: Mudassir: bio: I am an undergraduate student of Electrical Engineering	Si	Si

Tabla 10. Resultados Script#4

Total de Ítems Extraídos:	1650 items de 1650
Número de Usuarios:	1650
Porcentaje de Extracción:	100%
Ingresados a Base de Datos:	1650
Guardados en Archivo Json	1650

3.4. Experimentación sobre Plataforma MIriadaX con WebScraper

En el presente proyecto se realizó una experimentación para la extracción de datos de los foros de otra plataforma que constaba entre las opciones principales para realizar la extracción de foros, como lo es MIriadaX, por ser la única plataforma que sus cursos son dictados en idioma castellano. Se realizó la extracción con la extensión para el explorador Chrome WebScraper, por ser una extensión con una línea de aprendizaje corta y haber estado también dentro del análisis de tecnologías para extracción de datos

3.4.1. Modelo gráfico en Web Scraper para foros en MIriadaX

La figura 3 muestra el modelo que se configuro en web scraper para la extracción de datos del foro de MIriadaX, está gráfica la genera el propio webscraper permitiendo depurar en resumen cómo debe ir la araña buscando primero en las categorías de módulos, luego a cada página correspondiente para obtener el link de cada discusión, una vez que obtiene el link, esta elige para cada entrada el texto de la respuesta, el nombre del usuario que la escribió y a la discusión que corresponde.

La araña programable en webscraper se mantiene en un ciclo infinito hasta que no haya más páginas siguientes en los foros de cada módulo, para luego seguir al siguiente módulo a realizar la misma extracción

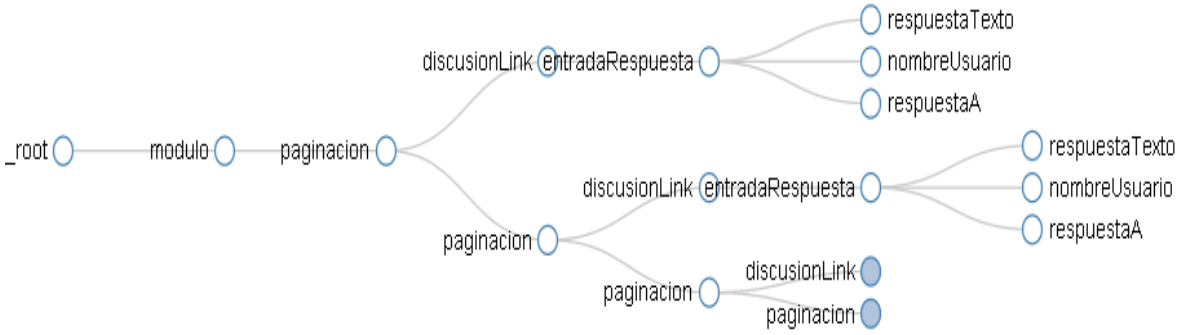


Figura 3 Modelo de Araña realizada con WebScraper para foros de Miriadax

El modelo de foros de la plataforma MiríadaX es sencillo de manejar y entender pues se compone de categorías entre las que se aprecia las categorías propias del cuerpo del conocimiento de la materia y categorías ajenas como presentaciones personales o preguntas relacionadas a la plataforma, por tanto los datos son fáciles de filtrar siempre y cuando un usuario no haya posteado de manera incorrecta en otra categoría a la que se refiere su discusión.

La araña configurada no tiene problemas para la extracción de datos, lo realiza con un intervalo de 2 segundos cada página, con un retraso de espera de 500 milisegundos por cada nueva página, esta extracción la realiza mediante el mismo navegador web Chrome que muestra visualmente cada página a la que se va ingresando.

Una de las ventajas observadas por ser una herramienta que realiza la extracción por navegador es que utiliza los mismos datos de sesión iniciada para acceder a las páginas que no se puede acceder sin estar autentificado en Miriadax, la utilización de cookies permite que al realizar la extracción no se bloquee la dirección ip desde la que se está realizando pues trabaja como si un humano estuviese dando clic en los enlaces página a página.

La estructuración de los datos extraídos es un inconveniente con webScraper pues solo permite la extracción de datos a un formato csv , en el cual se debe realizar cualquier nuevo cambio a su estructura que se desee implementar

3.4.2. Formato de Salida de Datos de Foros de MIriadaX

El formato (Tabla 11) que se diseño para extraer la información en una tabla con extensión csv consta de un campo “modulo” que se refiere dentro del curso al conjunto de contenidos que se asocian a contenidos específicos previamente diseñados por los creadores del curso; el segundo campo “discusionLink” se refiere al título que dan a cada hilo de discusión creado;el tercer campo “respuestaTexto” es el todo el texto escrito en la entrada del foro; en el cuarto campo se guarda al usuario al que corresponde dicho mensaje; y el último campo se extrae si el mensaje es una respuesta a otro mensaje que fue escrito con anterioridad y que se relaciona al campo “discusionLink”

Tabla 11. Formato de Salida Foros Miriadax

módulo	discusionLink	respuestaTexto	nombreUsuario	respuestaA

En la tabla 12 se puede observar un extracto de la información recuperada del curso MOOC de bioestadística con R desde la plataforma MIriadaX donde se pudo recuperar 337 hilos de discusión agrupadas en 8 categorías y con un total de 927 respuestas

Tabla 12. Extracto de Datos de Foro de Curso Bioestadística con R en MIriadaX

módulo	discusionLink	respuestaTexto	nombreUsuario	respuestaA
Módulo 2	Diagrama de barras	No logro que para valores discretos por ejemplo apgar1 en neonatos sobre el eje x aparezcan cada uno de los valores de la variable. Revisé el video explicativo del profesor, pero no lo aclara simplemente lo dice por simple inspeccion.	FERNANDO JOSE VIVAS MORALES	Diagrama de barras
Módulo 2	TEST CORREGIDO	Bueno, siguiendo los ejemplos resueltos durante las clases, se puede hacer de forma similar para resolver los del test. Si tienes algun problema con alguna pregunta concreta del test, dime y te digo como lo he hecho, sino te sale.	belen luque	RE: TEST CORREGIDO
Módulo 1	Introducción a la Estadística. Video s/sonido	Tiene un mensaje que dice que se ha silenciado el audio por reclamos de copyright en una canción contenida en el video... Me encontré el mismo video pero con sonido:	OSMAN DAVID ACOSTA ORTEGA	RE: Introducción a la Estadística. Video s/sonido

Resultados

La tabla 13 muestra un resumen del total de hilos de discusión que se extrajo mediante webscraper en miriadax, siendo un número menor de discusiones de las que se extrajo en la plataforma Udacity.

Tabla 13 Resultados Scrapy a Curso en Miriadax

Curso:	Curso Práctico de Bioestadística con R(Primera parte) 3 ra edición
Número de Discusiones	337 discusiones
Número de Respuestas	927
Categorías	8

CAPÍTULO IV.

APLICACIÓN DE TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL

4.1. Herramientas Elegidas

Para la aplicación de técnicas de Procesamiento de Lenguaje Natural a los datos extraídos del foro del curso, se utilizó el lenguaje de programación Python con la librería de código abierto “Natural Language Toolkit (NLTK)” que fue creada en el año 2001 como parte de un curso de computación lingüística por el departamento de computación y ciencias de la información de la Universidad de Pennsylvania con la que se pretendía enseñar a los estudiantes las bases del procesamiento de lenguaje natural.

4.1.1. Python

Python es un lenguaje de programación creado por Guido van Rossum a principios de los años 90 cuyo nombre está inspirado en el grupo de cómicos ingleses “Monty Python. (Gonzáles, 2011). Python ofrece excelentes funcionalidades con una curva de aprendizaje corta, una sintaxis y semántica transparente y al ser un lenguaje interpretado facilita la exploración interactiva. Cuenta con una sintaxis simple, clara y sencilla, con una gran cantidad de librerías disponibles, que incluyen procesamiento gráfico, conectividad web y procesamiento numérico por lo que se transforma en un lenguaje fuertemente usado en la industria, en la investigación científica y la educación alrededor del mundo. (Bird, Klein, & Loper, 2009)

4.1.2. NLTK 2.0 (Natural Language Toolkit)

Esta librería consta de algunos módulos que permiten realizar tareas de procesamiento de texto, clasificación, interpretación semántica, evaluación de métricas, probabilidades, fragmentación de oraciones, descubrimiento de alineaciones en oraciones, acceso a corpus ya entrenados, etiquetado de palabras entre otras funciones. (Bird et al., 2009)

Desde su creación en el 2001, se ha ido incrementando sus funcionalidades con ayuda de contribuyentes hasta llegar a una versión 3.0 en la actualidad que es utilizada en decenas de universidades en varios proyectos de investigación. (Bird, et al., 2009). Su librería incluye extensa documentación, demostraciones gráficas, tutoriales que explican las tareas soportadas y un conjunto de textos previamente entrenados. (Bird, 2005)

Según (Loper, 2004) NLTK tiene un diseño que cumple con requerimientos de facilidad de uso, consistencia, extensibilidad, documentación, simplicidad y modularidad, todos estas características listadas en orden decreciente de importancia

Contrastando con estas, los creadores de la herramienta aclaran en el libro oficial de NLTK 2.0, que aunque se provee de un rango extenso de funciones, no es una enciclopedia, ni

tampoco un sistema y tampoco está altamente optimizado para un alto rendimiento o algoritmos complejos de bajo nivel.

4.2. Análisis Exploratorio de Texto de Hilos de Discusión del Curso "Introduction to Psychology" de Udacity

4.2.1. Tokenización

La tokenización es el proceso en que se separan las palabras de un texto para formar un conjunto de palabras y signos de puntuación por separado llamados tokens (Manning & Schütze, 1999). Existen varias técnicas para tokenizar como separar el texto por espacios, separar por puntos o separar por expresiones regulares, según en el idioma en que se realice esta tarea y el ámbito de los datos, se utilizan unas u otras de estas técnicas

Cada token dentro del toolkit NLTK representa una unidad simple de texto y es definida para maximizar la interoperabilidad entre diferentes módulos, por lo que una simple clase "Token class" se encarga de realizar la tokenización del texto por diferentes métodos. (Loper, 2004)

Se puede pensar que la tarea de separar las palabras de un texto de manera correcta es una actividad trivial que se resuelve fácilmente con programación, sin embargo resulta bastante complejo si se considera la variedad de lenguajes existentes y los sistemas de escritura diversos (Dale, Moisi, & Somers, 2000).

La librería NLTK contiene la función "word tokenize" que realiza la tokenización por medio de espacios y signos de puntuación, en este proyecto se utilizó esta función para el texto formado por los mensajes principales en discusiones de lo que obtuvo 92242 tokens entre palabras y signos.

En una observación rápida de los tokens se pudo verificar que existen inconvenientes con realizar la tokenización con este método para el idioma inglés pues separa tokens también si están escritos con contracción. Por ejemplo para la frase "I can't get the answer" se muestra la siguiente lista de tokens "I, ca, n't, get, the, answer ", como se puede observar la palabra can't es separada en dos tokens y al ser "ca" un token podría no ser analizado con eficacia.

Otro de los inconvenientes sucede con las direcciones urls, pues al existir puntuación dentro de una dirección como dos puntos, puntos, y otras, estas son separadas incorrectamente, al tratarse de direcciones url se debería tomarlas como un token único dentro de un escenario perfecto. Por ejemplo en algún foro se cita la dirección

“http://en.wikipedia.org/wiki/Monocular_vision”, y el tokenizador de Nltk devuelve 3 tokens diferentes: “http”, “:” y “//en.wikipedia.org/wiki/Monocular_vision”.

Algunos otros tokenizadores del mismo paquete de Nltk disponibles son punktWordTokenizer y wordPunctTokenizer, sin embargo al realizar pruebas con estos el primero separa la palabra “can`t” en dos tokens: “can” y “`t” , lo que tampoco resulta conveniente. El segundo tokenizador realiza una separación por medio de cualquier signo de puntuación, por lo que para el mismo caso devuelve 3 tokens: “can”, “ ` “ y “t” , un análisis aún menos eficiente.

Un proceso de tokenización realizado mediante espacios en blanco realizado sobre el texto devuelve 80647 tokens, un valor menor a los anteriores pruebas, sin embargo se puede evidenciar que da solución al inconveniente de las contracciones y de las direcciones urls dentro del texto, devolviendo un solo token en ambos casos. El inconveniente es que devuelve tokens con signos de puntuación debido a que las palabras no siempre están rodeadas de espacios en blanco, a menudo se coloca signos de puntuación como comas, punto y coma, puntos, guiones, que se convierten en un inconveniente cuando forman períodos como al marcar una abreviatura como en “etc. O Calif” (Manning & Schütze, 1999).

Otra de las formas de tokenizar es mediante expresiones regulares programables lo que permite tener más control sobre este proceso (Bird et al., 2009). Aplicando algunos patrones de expresiones regulares con el objetivo de eliminar el inconveniente de signos de puntuación, separar palabras con guiones, controlar números, abreviaciones de entidades y controlar los paréntesis, se obtuvo una lista de tokens de 97127. Esta aplicación produjo que las contracciones se separen en 3 tokens y que las direcciones urls se separen en varios tokens, sin embargo este proceso produce tokens limpios que son manejables de forma fácil.

La tokenización no es una tarea fácil y resulta más complicada de lo esperado, por lo que ninguna solución funciona bien en todos los ámbitos (Bird et al., 2009), debido a que existen elementos propios de cada lenguaje que inducen a fallas en la separación correcta de palabras, también tipos de datos que se mezclan, la mala escritura y el uso incorrecto de puntuación.

Se presenta en resumen las técnicas de tokenización aplicadas al corpus y el número de tokens obtenidos, en la tabla 14.

Tabla 14. Resumen Técnicas Tokenización Aplicadas

Tipo Tokenización	Word Tokenization	PunktWord Tokenizer	WordPunct Tokenizer	Mediante Espacios en Blanco	Mediante Patrones de Texto
Número de Tokens	92242	91787	97159	80647	97127

Una vez aplicados varios métodos de tokenización se decidió utilizar la tokenización mediante la librería "Word Tokenize" de NLTK, que utiliza un modelo que separa por espacios en blanco y por puntuación.

4.2.2. Normalización de Texto

Normalizar el texto involucra la fusión de diferentes formas escritas de una ficha en una forma normalizada canónica (Indurkhya & Damerau, 2010) con el objetivo de realizar un análisis del texto con la mayor homogenización posible, lo que conlleva a descapitalizar palabras, eliminar signos de puntuación, números, utilizar sinónimos, y también la utilización de diccionarios de palabras raíz.

Para la normalización del texto en este proyecto se utilizó la descapitalización de palabras, la eliminación de signos de puntuación y la eliminación de formas numéricas en el texto, obteniendo un nuevo conjunto de 76044 tokens.

4.2.3. Eliminar Palabras vacías

Se denomina palabras vacías, palabras de parada o stop words en idioma inglés a las palabras que se usan muy frecuentemente en un lenguaje y que tienen poco contenido léxico, o su presencia en el texto es indiferente. (Bird et al, 2009).

La remoción de palabras vacías es importante por dos razones, la primera debido a que permite relacionar las consultas y los documentos con palabras clave de manera más directa, y la segunda porque reduce el tamaño del conjunto de palabras a analizar en un rango del 30 al 50%. (Indurkhya & Damerau, 2010)

Por lo tanto la eliminación de palabras vacías significa eliminar del análisis palabras sin contenido directo como los pronombres, adverbios, modificadores y otros que dentro de un lenguaje se consideren superfluos o dentro de un ámbito en específico se consideren sin importancia, de este hecho depende que el conjunto de palabras vacías sea de unas pocas palabras o de una lista robusta. (Indurkhya & Damerau, 2010)

Dentro de la librería de NLTK están definidas algunas listas de palabras vacías para diferentes idiomas, para el idioma inglés están definidas 127 palabras. Luego de realizar la eliminación de estos tokens dentro de nuestro corpus se obtuvo aproximadamente menos del 50 % de palabras, con un total de 36790 tokens.

4.2.4. Análisis de Frecuencias de Palabras

El análisis de frecuencias de palabras, es el conteo de las palabras dentro del corpus para conocer cuáles son las que se han escrito mayor número de veces, por lo que se considera un dato estadístico sobre el corpus analizado.

Nltk cuenta con la función “FreqDist” para realizar este tipo de análisis previamente se tuvo que pasar a texto de tipo “nltk Las 35 palabras con mayores probabilidades de ser escritas y que han sido escritas más de 100 veces que otras se muestran en la tabla 15:

Tabla 15. Palabras que se usan con más frecuencia en el foro

1	'would'	383	19	'please'	136
2	'answer'	278	20	'even'	129
3	'one'	277	21	'person'	127
4	'people'	260	22	'video'	122
5	'like'	243	23	'correct'	118
6	'course'	235	24	'say'	115
7	'think'	226	25	'see'	114
8	'question'	209	26	'help'	109
9	'know'	202	27	'something'	109
10	'also'	171	28	'first'	105
11	'get'	170	29	'right'	105
12	'could'	163	30	'someone'	105
13	'time'	162	31	'brain'	104
14	'really'	151	32	'problem'	104
15	'psychology'	143	33	'thanks'	104
16	'lesson'	138	34	'want'	103
17	'way'	138	35	'http'	102
18	'different'	136			

Como se puede observar en el extracto la tokenización separo las direcciones url en “http” y los demás por lo que tenemos el token “http” como una palabra también.

La figura 4 muestra estas mismas palabras en una gráfica de dispersión donde se evidencia que existe mayor probabilidad de que los estudiantes de este curso utilicen esas palabras a otras palabras.

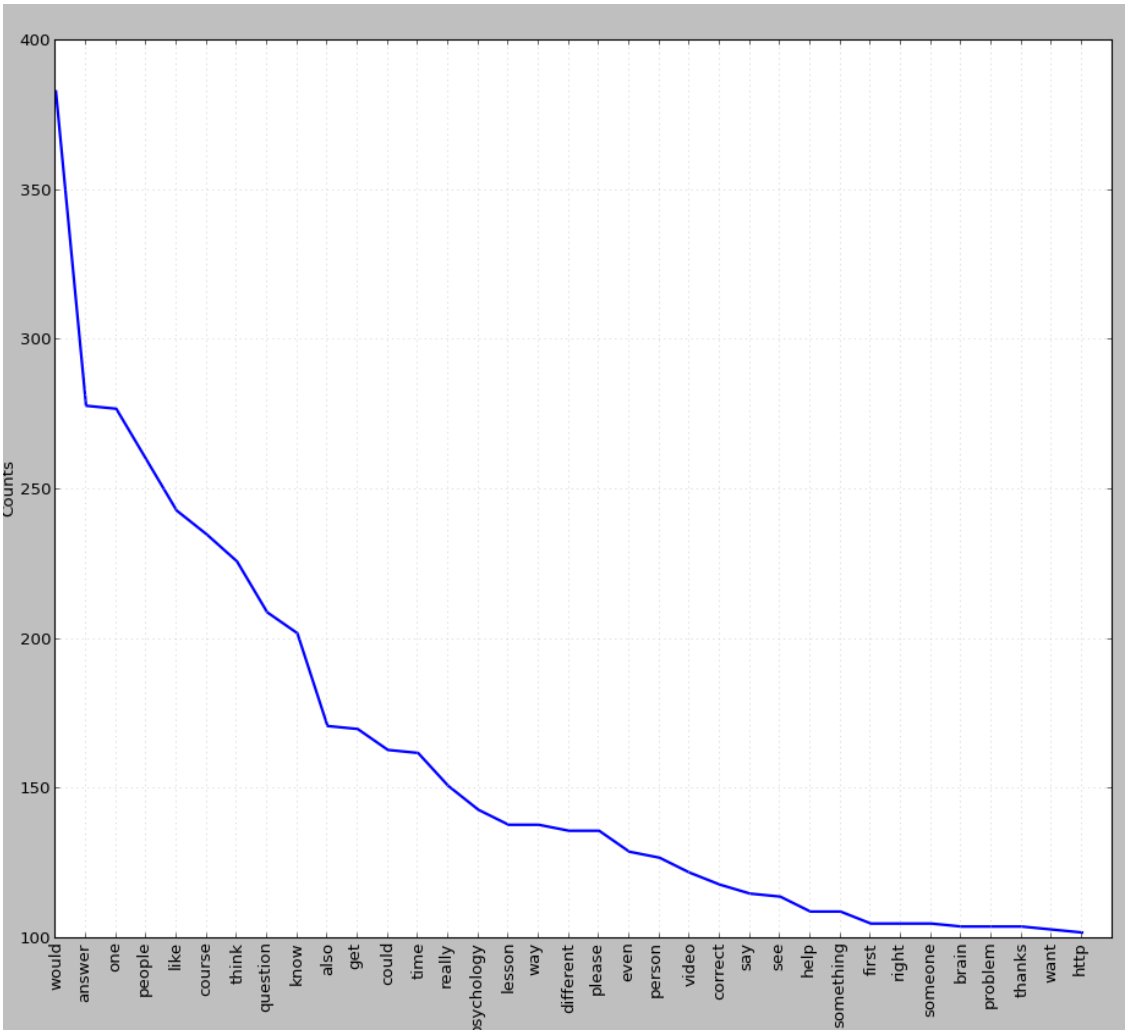


Figura 4. Diagrama de Frecuencias de Palabras más Usadas en el foro

4.2.5. Palabras escritas solo una vez

La función “haxapases”en nltk se la utiliza para una distribución de frecuencias, para obtener las palabras que fueron escritas tan solo una vez. El número de estas es igual a 3272. En la tabla 16 se muestra un extracto de 30 palabras menos usadas

Tabla 16. Palabras menos usadas en el foro

'aare'	'academy'	'accomodate'	'accustomed'	'acknowledging'
'aba'	'accelerator'	'accompanied'	'acheive'	'acronyms'
'abnormal'	'accesable'	'accompaniment'	'aches'	'acth'
'abnormality'	'accessed'	'accomplished'	'achieving'	'acting'
'absorb'	'accessibility'	'accounted'	'acidity'	'actings'
'abusers'	'acclimation'	'accross'	'acido'	'actively'

4.2.6. Diccionario de Palabras

Una vez realizada este conteo de palabras se puede realizar un diccionario de palabras que contendrá a cada palabra dentro del corpus sin repetirse. El diccionario formado contiene un total de 6897 palabras diferentes. Es importante realizar esta tarea pues a futuro permite decidir el número de palabras que se necesitarán para aplicar algoritmos de clasificación de palabras.

4.2.7. Palabras en el mismo contexto

Las palabras en el mismo contexto se refieren a las dos palabras que aparecen junto a la palabra que se quiere analizar, la primera que aparece antes y la otra palabra que aparece después. En la tabla 17 se analiza las dos primeras palabras con mayor probabilidad descritas previamente en la tabla 15.

Tabla 17 Palabras en el mismo contexto para las 2 palabras más comunes

Palabra Analizar	Palabra anterior	Palabra siguiente
Would	Abusive	breakup
	Acknowledging	definitely
	act	like
	act	situation
	adults	say
	affects	best
	age	like
	aggressive	really
	ago	really
	agree	good
	almost	appreciated
	also	drug
	also	know
	annoying	categorize
	answer	answered
	answer	behavioral
	answer	form
	answer	trucks
	answers	think
	antoine	words
Answer	able	correct
	accept	question
	accept	would
	accepted	tried
	accepting	even
	according	clapping
	actual	study

	amp	help
	answer	came
	answers	issue
	answers	previously
	anyway	number
	applied	many
	attempted	closest
	attention	suggests
	aware	placed
	behind	comes
	better	second
	better	would
	blocks	given

4.2.8. Análisis de Resultados

A pesar de aplicar la eliminación de palabras vacias por medio de la librería NLTK los resultados aún muestran indicios de palabras vacias como “even” , “see”, “want” y otras. La librería de palabras vacías de nltk resulto muy básica para el análisis de este corpus por lo que se procedió a ajustar este proceso mediante el uso de una lista personalizada de palabras vacías mucho más robusta compuesta por 667 palabras

4.2.9. Resultados con lista de palabras vacías personalizada

Aplicando la eliminación de palabras vacías con la nueva lista de palabras vacías el resultado fue de 26001 tokens, luego aplicando nuevamente la eliminación de palabras con la librería de NLTK se obtuvo 25997 tokens. Este resultado muestra una optimización considerable con una diferencia de 10993 palabras que se reconocen como válidas para realizar un análisis de frecuencias de palabras, y con un diccionario de 6478 palabras diferentes.

Los resultados con una distribución de frecuencias con las 20 palabras más usadas se muestran en la tabla 18.

Tabla 18. Palabras más usadas con lista de palabras vacias personalizada

'answer'	278	'problem'	104
'people'	260	'http'	102
'question'	209	'test'	92
'time'	162	'study'	88
'psychology'	143	'find'	81
'lesson'	138	'understand'	81
'person'	127	'answers'	79
'video'	122	'good'	78
'correct'	118	'language'	75
'brain'	104	'wrong'	75

Los resultados muestran que se eliminaron algunas palabras que no interesaban dentro de este análisis como “are”, “see” o “would”, o one”. Lo que nos indica que la nueva lista de palabras vacias permite filtrar estos verbos que son bastante relativos en cuanto a su significado morfológico. La figura 5 muestra la distribución de frecuencias de las palabras más usadas ya con la lista de palabras vacías personalizada.

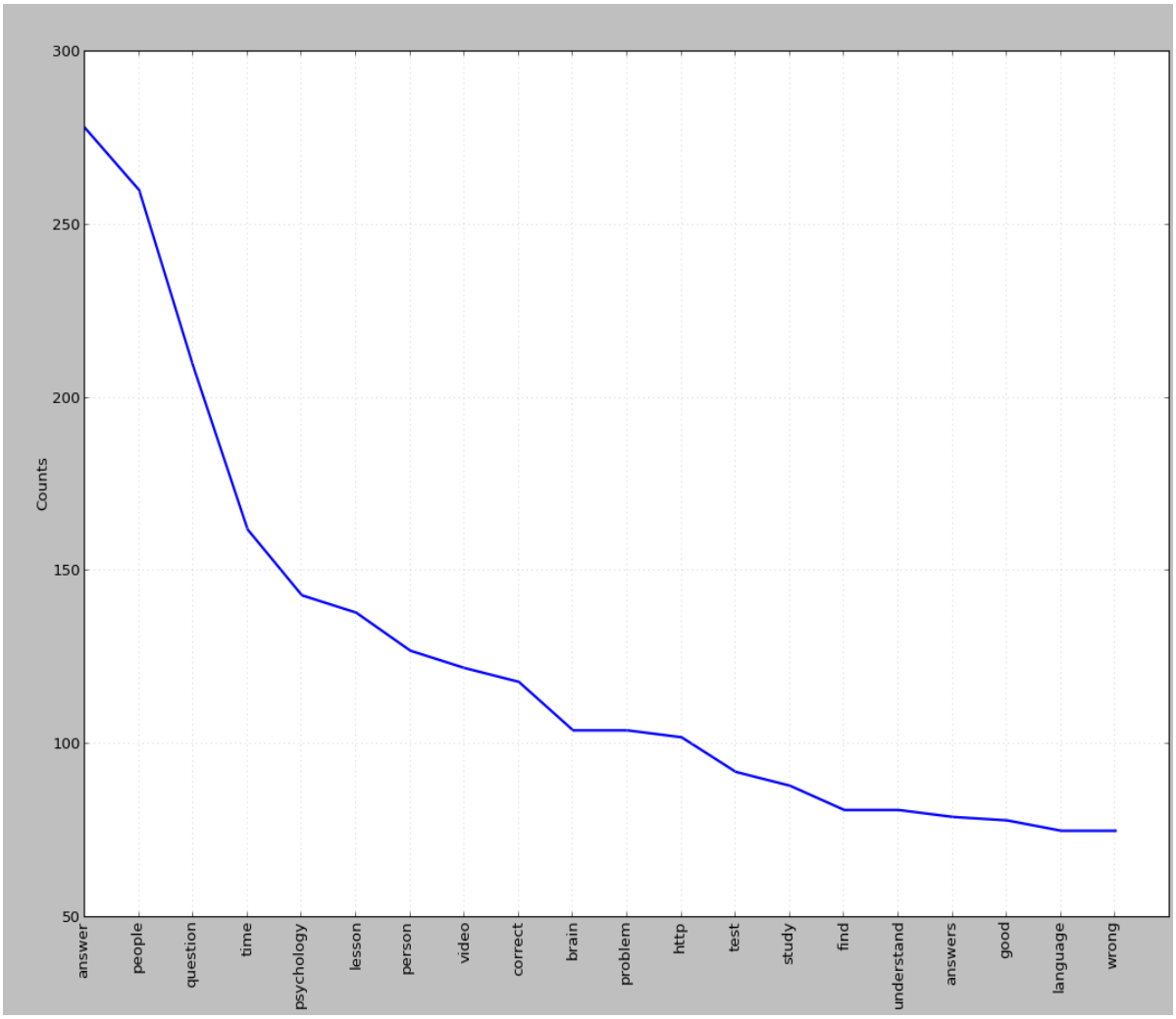


Figura 5 Diagrama de Frecuencias para palabras más usadas con lista de palabras vacías personalizada

4.2.10. Análisis de Bigramas

Los bigramas son palabras conformadas por dos tokens que se forman para dar un sentido único a la formación como por ejemplo en el idioma inglés: “ice cream”, “homo sapiens” o “nervous system”. Es posible extraer estas colocaciones de palabras si se repiten una y otra vez dentro de un texto, por lo que se concluye que deben ser palabras que deben ir juntas.

Usando la función collocations de Nltk se puede ubicar algunos de estos bigramas en nuestro corpus. La tabla 19 muestra un extracto de algunos bigramas de palabras que son nombradas con mayor frecuencia en el foro

Tabla 19. Bigramas más nombrados en el foro

julien caussin;	introduction psychology;
regards julien;	even though;
best regards;	reaction time;
ice cream;	long term;
problem set;	genetic relatedness;
nervous system;	final exam;
please help;	sample size;
correct answer;	frontal lobe;
anyone else;	get certificate;
answer question	caussin hello;

Utilizando una distribución de frecuencias para los bigramas se puede explorar que palabras forman bigramas con otras, como por ejemplo para la palabra “social” (tabla 20). En la tabla se muestra que para la palabra social, es más probable que se hable de “social norms”, de “psychology social”, de “social loafing ” y de “social facilitation” cuando se utiliza la palabra “social” dentro del foro.

Tabla 20. Palabras que forman bigrams con la palabra social

('norms'	9	('also'	1
('psychology'	7	('encounters'	1
('loafing'	3	('environmental'	1
('facilitation'	2	('expectation'	1
('influence'	2	('fears'	1
('learning'	2	('functions'	1
('psychologist'	2	('historic'	1
('readjustment'	2	('however'	1
('support'	2	('intelligence'	1
('acceptance'	1	('interaction'	1

La figuar 6 muestra la distribución de frecuencias de bigrams para la palabra “social” lo que corrobora que para la palabra “social” es más probable que se considere también las palabras “norms” , “psychology” ,”loafing” o “facilitation”.

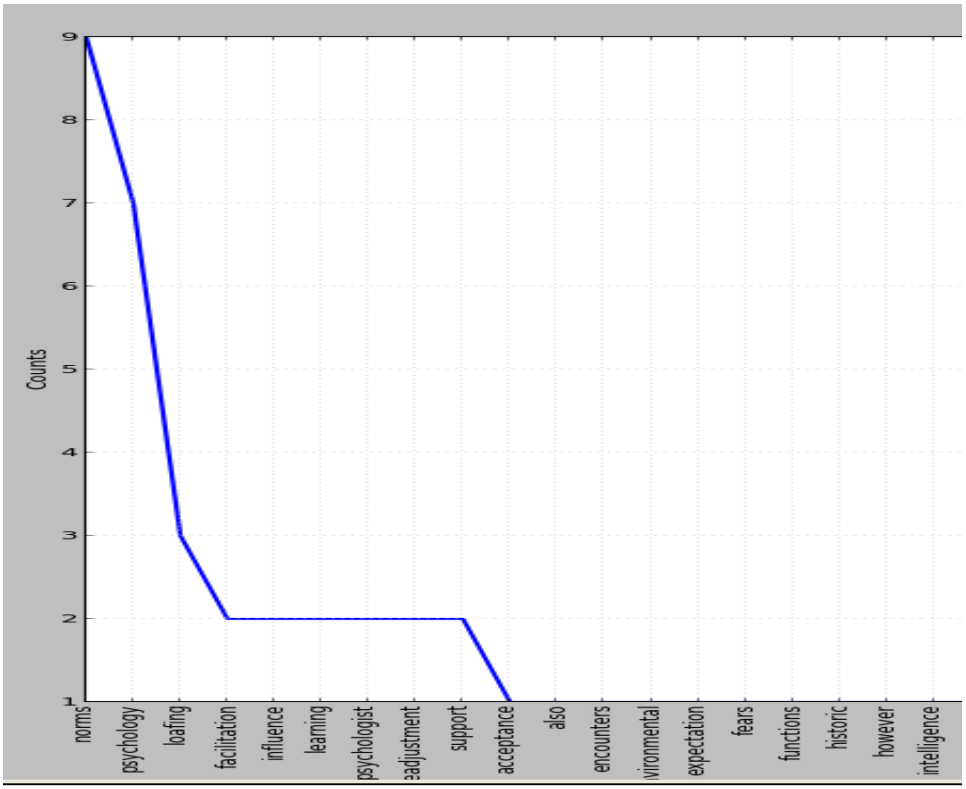


Figura 6. Diagrama de Frecuencias de Bigrams para palabra “social”

4.3. Reconocimiento de Entidades en Texto de Hilos de Discusión del Curso “Introduction to Psychology” de Udacity

4.3.1. Definición de Entidades

Los nombres de entidades son sustantivos propios que se refieren a individuos, organizaciones, localizaciones, fechas y tiempos (Bird et al, 2009). Dependiendo del objetivo del análisis se usa en general para identificar todas las menciones de un nombre de una entidad en un texto, con la importancia de poder predecir sobre que se está hablando en las oraciones.

4.3.2. Importancia del Reconocimiento de Entidades

Como un ejemplo de su aplicación está un humano que le pregunta a una máquina ¿Quién invento la teoría del Psicoanálisis? , la máquina buscaría entre el corpus palabras asociadas a la pregunta lo que lo llevaría a la siguiente oración en el corpus.

<<Sigmund Freud,fue el científico que creó el Psicoanálisis. Jodorowsky fue el artista que creó la Psicomagia, que no emplea el arte como terapia, sino la terapia como arte.>>

Realizando un análisis de entidades la máquina podría extraer a Freud y Jodorowsky, pero antes se debe realizar una segmentación de oraciones, pues al no segmentarlas se podría asignar la respuesta al científico equivocado en este caso Jodorowsky. Mediante el análisis de distancias léxicas entre palabras la máquina podría dar por entendido que es Sigmund Freud el creador del psicoanálisis.

A la luz de este ejemplo la búsqueda de Entidades es un aspecto importante en el procesamiento de lenguaje natural, podría dar a los buscadores ese sentido semántico que se busca y a los robots la capacidad de inducción y deducción.

Para este proyecto se realizó una extracción de Entidades en el corpus de hilos de discusión para extraer las entidades que son nombradas entre todo el corpus realizando para ello las tareas que se muestran en la gráfica de modo secuencial

4.3.3. Proceso previo al reconocimiento de Entidades

El reconocimiento de entidades requiere de un previo proceso de preprocesamiento de texto, y de reconocimiento morfológico de las palabras, por lo que es necesario utilizar las librerías de NLTK para la categorización de palabras. La figura 7 muestra el proceso previo al reconocimiento de entidades, de manera ordenada.

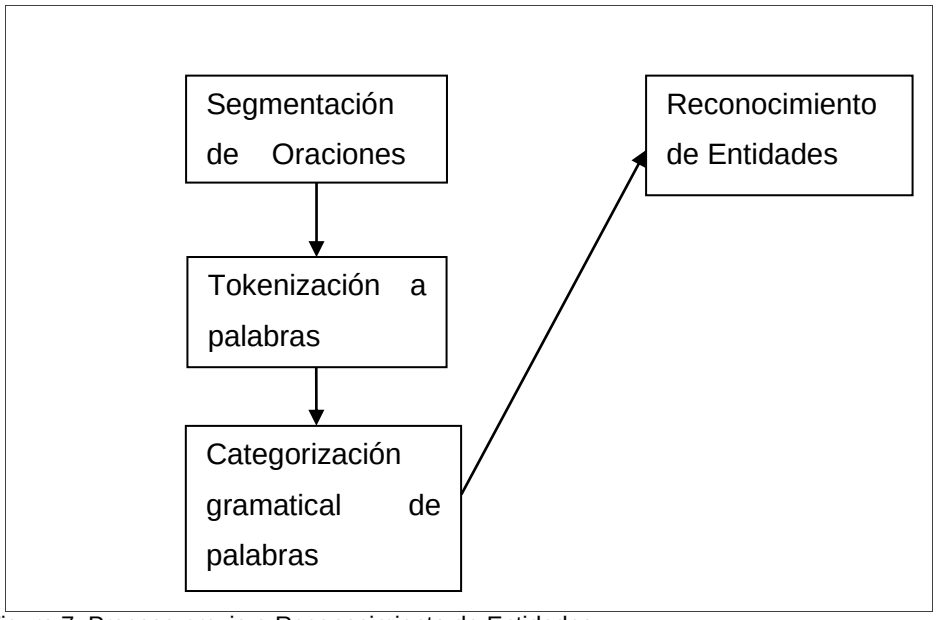


Figura 7. Proceso previo a Reconocimiento de Entidades

Para el primer proceso de segmentación de oraciones se utilizó la función “sent_tokenize” de NLTK con lo que nos devuelve una lista con 4883 oraciones. Luego se realizó la

tokenización de cada una de estas oraciones lo que nos devuelve una lista donde en cada posición están los tokens de cada oración.

Para el tercer paso se utilizó la función pos_tag que nos categoriza cada palabra de las oraciones de nuestro corpus basado en corpus ya categorizados. Por ejemplo para la oración

'I am enrolled in Psychology (PS001) starting June 3, 2013.'

La categorización de palabras nos devuelve las palabras cada una con su tipo gramatical

[('I', 'PRP'), ('am', 'VBP'), ('enrolled', 'VBN'), ('in', 'IN'), ('Psychology', 'NNP'), ('(', 'NNP'), ('PS001', 'NNP'), (')', 'NNP'), ('starting', 'VBG'), ('June', 'NNP'), ('3', 'CD'), (',', ','), ('2013', 'CD'), ('.', '.')]

En la tabla 21 se muestra el significado de cada sigla que la categorización automática realiza.

Tabla 21. Significado de Siglas dentro de NLTK en análisis gramátical

PRP: Pronombre
VBP: Verbo en tiempo presente
VBN: Verbo en tiempo pasado
IN: Preposición
NNP: Sustantivo Propio
VBG: Verbo en presente participio
CD: Digito

El siguiente proceso es reconocer las entidades dentro de las oraciones ya categorizadas sus palabras. Para realizar esta tarea se utilizó la función ne_chunk lo que nos devuelve un árbol con la categoría de cada palabra para cada oración del corpus. Por ejemplo para el ejemplo citado anteriormente se muestra en la figura 8.

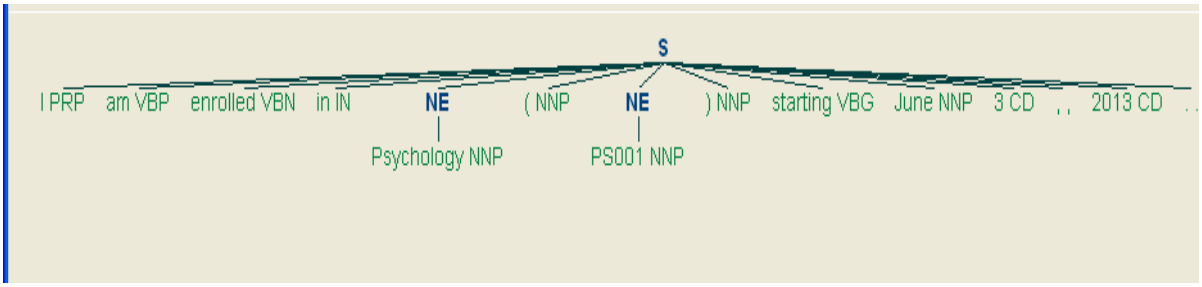


Figura 8. Ejemplo de Reconocimiento de Entidades en el texto de hilos de discusión

Una vez realizado un script para el análisis de todas las oraciones de todo el corpus se encontró un total de 817 entidades de un total de 1435 reconocimientos. La tabla 22 muestra un ejemplo de las entidades automáticamente reconocidas.

Tabla 22. Ejemplos de Entidades Reconocidas en el foro

Entidad Reconocida	Repeticiones dentro del Corpus
Please	43
Udacity	25
Can	24
Chomsky	9
MOOCs	7
Wikipedia	3
India	3
Julien Caussin	30
Germany	3
Andy	8
Neuron	6
Broca	3
Lauren Castellano	4
Biology	2
Japanese	2
Skinner	6
Canine Cognition CenterBrazil	1
Brain Games	1
Psychology	24
REM	4

La figura 9 muestra la distribución de frecuencias de estas entidades reconocidas evidenciando que es más probable que se hable de unas entidades que de otras.

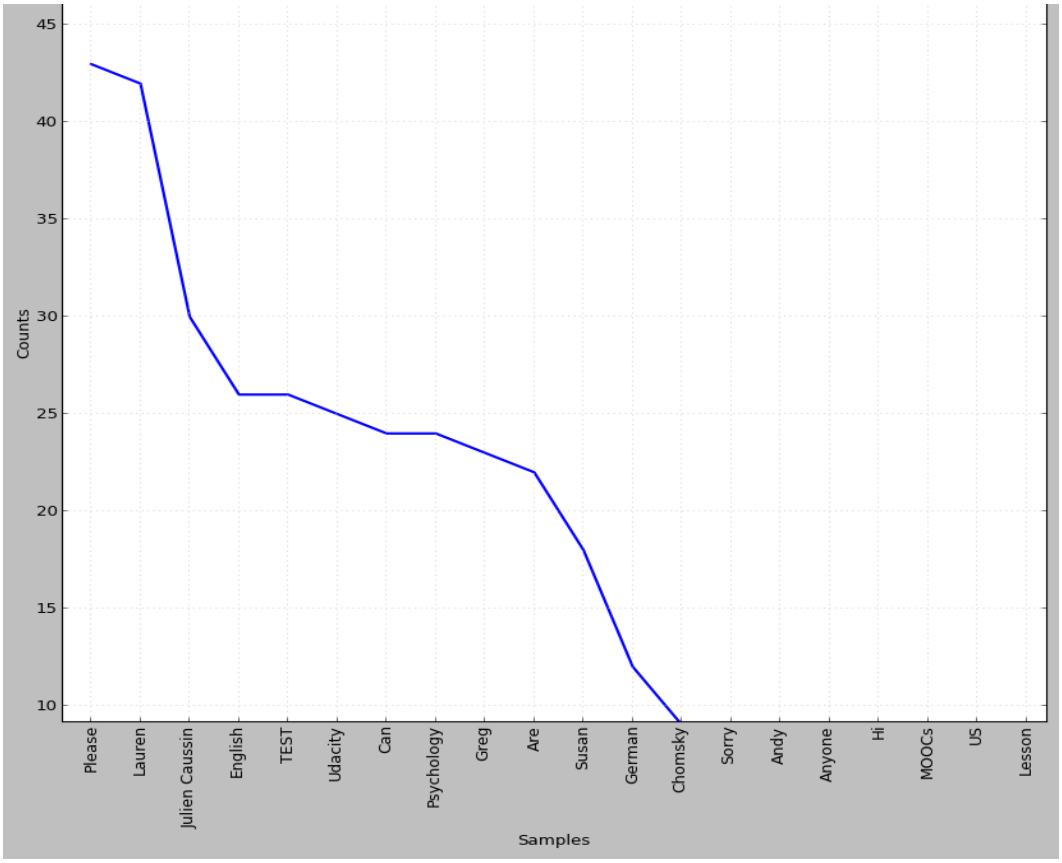


Figura 9. Diagrama de Frecuencias de Entidades Reconocidas

4.3.4. Aplicando Stop Words al reconocimiento de Entidades

Al observar que surgen palabras como verbos como “can” o “are” se evidenció que el modelo de reconocimiento de entidades no estaba funcionado correctamente por lo que se aplicó el reconocimiento de palabras vacías dentro del conjunto de entidades para su eliminación. Se realizó esta tarea con un conjunto de palabras vacías para el idioma inglés mucho más robustas de la que cuenta NLTK, igual a 667 palabras, con lo que se pretendía eliminar palabras como “Can, Are, Hi” dentro de las palabras que se reconocen como entidades.

La tabla 23 muestra un resumen de las entidades reconocidas y las veces que se repitió dentro del cuerpo de texto, en total se reconocieron 722 entidades de 1215 reconocimientos. El diagrama de frecuencias de la figura 10 evidencia que existe mayor probabilidad de que se hable de esas entidades que de otras entidades reconocidas.

Tabla 23. Entidades reconocidas con filtro de palabras vacías

Entidad Reconocidas	Repeticiones dentro del Corpus
Lauren	43
Julien Caussin	30
English	26
Test	26
Udacity	26
Psychology	24
Greg	23
Susan	18
German	12
Chomsky	9
Skinner	6
American	5
Bc	5
Dsm	5

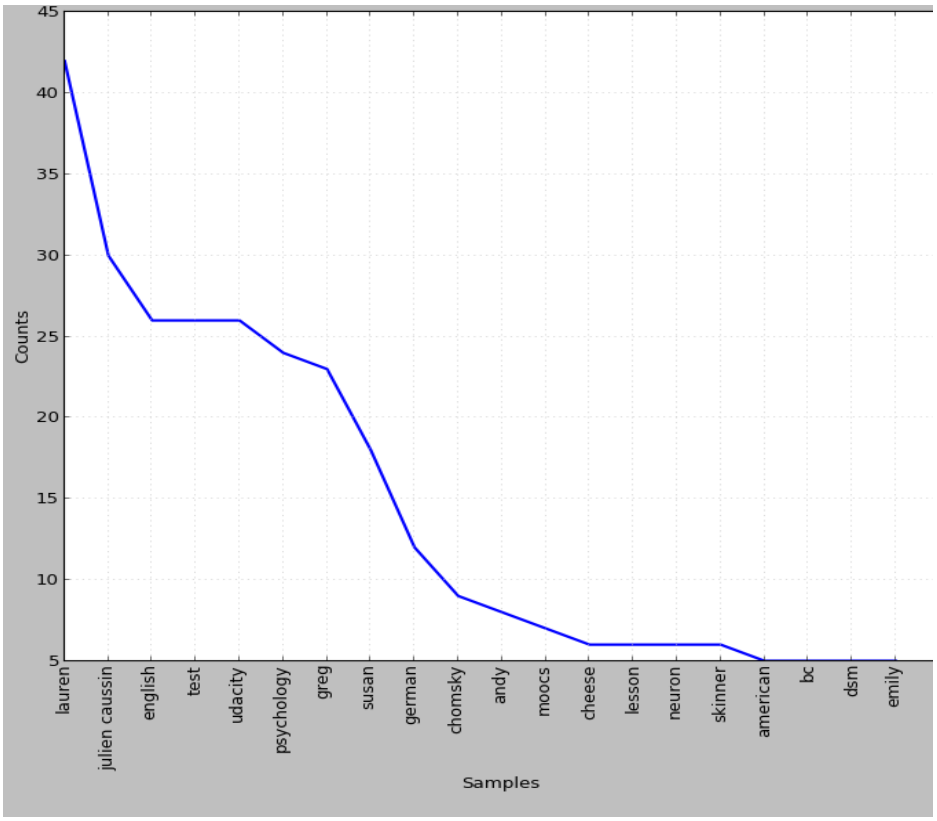


Figura 10.Diagrama de Frecuencias de Entidades Reconocidas luego de filtrar palabras vacías

4.3.5. Reconocimiento Según Tipo de Entidad

Dentro del mismo ámbito de entidades existen diferentes tipos. Por ejemplo podemos algunas de las más comunes entidades son las organizaciones, las personas, las localizaciones y las entidades geo-políticas que se refieren a las ciudades, provincias y países.

Se utilizó en primer instancia el reconocimiento de tipos de entidades por medio de la función ne_chunk de nltk para reconocer personas, organizaciones y localizaciones, pero los resultados no fueron óptimos en cuanto a clasificar las entidades dentro de estas categorías por lo que se utilizó la librería creada por la Universidad de Stanford llamada NER en su versión 3.5.1

Stanford NER es una implementación realizada en lenguaje Java , que funciona como un modelo condicional de campos aleatorios (CFR) que ha sido previamente entrenado mediante el corpus CoNLL, en 3 , 4 y hasta 7 clases incluyendo entidades tipo Dinero, Porcentaje, Fechas, Miscelaneas y Tiempo.

En el presente proyecto se realizó el análisis de entidades en 3 entidades: Localización, Persona y Organización teniendo en cuenta que son las de mayor importancia y que son indicadores de mayor relevancia.

Se hizo uso de este clasificador de Entidades mediante una interfaz para lenguaje Python cargada en la librería nltk para su versión 2.0 , teniendo inconvenientes en la ejecución de archivos con extensión .jar , pero que se pudieron solucionar añadiendo a python la dirección url de donde se encuentra el archivo java.exe del jkd necesario para ejecutar Stanford NER.

Los resultados se muestran un total de 536 unigramas con entidades tipo persona, 115 unigramas de tipo organización y 112 unigramas con entidades de tipo localización (tabla 24)

Tabla 24. Entidades reconocidas por Tipo

PERSONAS	ORGANIZACIONES	LOCALIZACIONES
['Susan', 'Snycerski', 'Greg', 'Feist', 'Lauren', 'Castellano', 'Freud', 'Louise', 'Juanita', 'Duke', 'Zetterburg', 'Julien', 'Caussin', 'Jared', 'Jared', 'Andy', 'Julien', 'Caussin', 'Jared', 'Julien', 'Caussin', '--', 'Andrew', 'Honestly', 'Lauren', 'Castellano', 'Lauren', 'Castellano', 'Manuel', 'Wernicke', 'Homo', 'Homo', 'Soma', 'Mike', 'Julien', 'Caussin', 'David', 'J', 'Julien	['Duke', 'University', 'Canine', 'Cognition', 'Center', 'YouTube', 'Firefox', 'Athena', 'SSRI', 'Google', 'CNS', 'NGO', 'NGO', 'BBC', 'CRC', 'NBC', 'news', 'Moss', 'Landing', 'Marine', 'Laboratories', 'Brain', 'Observatory', 'American', 'Academy', 'of', 'Sleep', 'Medicine', 'Bell', 'Google', 'Defense', 'Language', 'Institute', 'Udacity', 'General', 'Premise', 'ASL', 'Bells', 'Firefox',	['Bangalore', 'USA', 'Europe', 'Japan', 'Korea', 'China', 'India', 'Brazil', 'Russia', 'Broca', 'Neuroscience', 'Course', 'O.S', 'The', 'Netherlands', 'Germany', 'France', 'United', 'States', 'Texas', 'Germany', 'US', 'Monterey', 'CA', 'Korea', 'U.S.', 'Congo', 'US', 'Berlin', 'Earth', 'Atlanta', 'Atlanta', 'Atlanta', 'Selma', 'Selma', 'Broca', 'Germany', 'South', 'Korea', 'Japan', 'Korea', 'NZ', 'US', 'Italy', 'Hollywood', 'San', 'Jose', 'State', 'Germany', 'Hong', 'Kong', 'Hong', 'Kong',

Se podría ajustar los resultados de las columnas para saber si son bigrams de lo que se obtiene la tabla 25.

Tabla 25. Entidades tipo persona reconocidas por bigramas

Julien Caussin;	Amy Cuddy;
Caussin Julien;	Emily Rooney;
Datsyuk Zetterberg;	Mara Emily;
Egas Moniz;	Lauren Castellano;
Eric Haines;	Evans Eric;
Mubashir Omar;	Chomsky Skinner;
Rooney Mara;	Skinner Chomsky;
Togethertheyuncoveredamine Djurkovsaid;	Dear Susan;
Evans Prof;	Gru Julien;
Prof Evans;	David Evans

De la columna organizaciones no se obtiene bigramas. Y para la columna de Localizaciones se obtiene 2 bigrams: Hong Kong y Japan Korea.

4.4. Análisis de Sentimientos

El texto puede clasificarse en dos tipos principales, los hechos y las opiniones. Los hechos son expresiones objetivas sobre entidades y eventos que informan un carácter directo de un acontecimiento, al contrario de las opiniones que son una forma subjetiva de expresar una opinión, sentimiento, valoración hacia entidades, eventos y sus propiedades (Indurkhya & Damerau, 2010).

La web cambió dramáticamente la forma en que las personas expresas sus puntos de vista, debido a que las personas pueden valorar y expresarse directamente sobre cualquier cosa en internet, por lo que analizar y monitorizar las opiniones resulta una tarea formidable y útil sabiendo que las opiniones son tan importantes que nos permiten tomar decisiones en base al criterio de otros, un aspecto muy importante si se piensa en el bienestar de las organizaciones (Indurkhya & Damerau, 2010).

El análisis de sentimientos es el estudio computacional de las opiniones, sentimientos y emociones expresadas en un texto, siendo el principal problema identificar si una expresión es positiva (favorable), o si una expresión es negativa (desfavorable) acerca de una entidad o un producto (Nasukawa & Yi, 2003).

4.4.1. Algoritmo Elegido

El algoritmo que se escogió es el algoritmo Naïve Bayes por ser un algoritmo sencillo que obtiene resultados bastante útiles en cuanto a clasificación de texto utilizando como entrada una bolsa de palabras (que ya se obtuvo previamente mediante pln), este algoritmo comienza por el cálculo de la probabilidad a priori de cada etiqueta, que se determina marcando la frecuencia de cada etiqueta en el conjunto de entrenamiento. La contribución de cada característica se combina entonces con esta probabilidad previa, para llegar a una estimación de probabilidad para cada etiqueta. (Bird et al., 2009).

Como se describe en (Hernández, Ramírez, & Ferri, 2008) NaïveBayes se trata del modelo más simple de clasificación con redes bayesianas, asumiendo que todos los atributos son independientes, a pesar de aquello es uno de los más fuertes y más utilizados.

4.4.2. Características

Para poder enseñar a una máquina como realizar una clasificación se debe en primer lugar indicarle cuáles serán las características en las que se va a basar para realizar una categorización correcta. En segundo lugar se debe indicar al clasificador como codificar aquellas características que se ha elegido.

Escoger las correctas características y su codificación es el paso más importante que influye en la capacidad de aprender que tendrá el clasificador. Usualmente se elige esas características tomando en cuenta los aspectos relevantes a resolver dentro del problema. En este proyecto se utilizó un modelo de características basado en las 2000 palabras más frecuentes en el texto.

Los resultados muestran para palabras frecuentes dentro del corpus su polaridad (tabla 26), como por ejemplo para la palabra “regard” se dice que más de 10 veces se la ha nombrado como una palabra positiva que como una palabra negativa, caso contrario para “volumen” que se muestra que casi 5 veces más se la utiliza de forma negativa; esto podría deberse a reclamaciones por parte de los estudiantes sobre el volumen de los videos o algún reclamo concerniente al volumen

Tabla 26. Polaridad de características elegidas por clasificador

Características más importantes	Clasificada como	Valor
Regard	Positiva	10.5
Studies	Positiva	7.5
Symbols	Positiva	6.9
Behaviour	Positiva	6.9
Labeled	Positiva	6.3
Regards	Positiva	5.1
Located	Positiva	5.1
Everyday	Positiva	5.1
Btw	Negativa	5.1
Volumen	Negativa	4.8
Random	Negativa	4.8
Religion	Positiva	4.7
German	Positiva	4.5
Tool	Positiva	4.5
Awareness	Positiva	4.5
Aceptance	Positiva	4.5
Controversial	Positiva	4.3
Discussed	Positiva	4.1
Patterns	Positiva	4.1
Assumption	Negativa	4.1

4.4.3. Medidas de Rendimiento

La precisión mide la probabilidad de que si un sistema clasifica un documento en una cierta categoría el documento pertenezca a esta categoría**.** (Hernández, Ramírez, & Ferri, 2008)

Podemos extraer la precisión del etiquetado de polaridad de las palabras mediante las métricas de precisión que nos indican que porcentaje ha sido etiquetado correctamente del total de reconocidos y la memoria que nos indican una métrica de los valores que han sido correctamente identificados de un total de muestras. Mientras los valores más se acercan a 1 la clasificación logra una mejor aceptabilidad.

Los resultados obtenidos (tabla 27) muestran valores de precisión para los valores positivos de 0.85 , y para los valores negativos de 0.79. La métrica de memoria para los valores positivos es de 0.77 lo que significa que no se reconocieron algunos valores del total de muestras , y un 0.873 para los valores negativos que nos indica que un porcentaje mayor a los positivos de palabras si fue reconocida por el clasificador.

Tabla 27.Medidas de Rendimiento de Clasificador Naïve Bayes

Polaridad	Precisión	Memoria
Positivo	0.858731	0.772
Negativo	0.792915	0.873

El script codificado en Python se puede acceder en el repositorio público “https://github.com/snrepele/scrapyUdacity” en el archivo “clasificartexto.py”

CONCLUSIONES

Una vez terminadas todas las fases del presente proyecto se concluye que:

1.- El uso de la técnica scrapy particularmente usando el framework de Python es la mejor opción para obtener datos de portales web cuando estos no disponen de mecanismos de acceso directo a los datos tales como apis.

2.- Mediante el procesamiento de lenguaje natural se pudo rescatar a las entidades que se hace mayor referencia dentro del foro obteniendo los mejores resultados con la librería Named Entity Recognition de la Universidad de Stanford.

3.- Una adecuada, eficiente y consistente tarea de extracción de datos mediante la técnica de scraping garantiza el éxito de las futuras tareas de minería de texto.

4.- Mediante aprendizaje automático se pudo catalogar la polaridad en que se usan las palabras dentro del foro con un algoritmo Bayesiano simple, con datos previamente entrenados con medidas de rendimiento aceptables.

4.- Las técnicas de inteligencia artificial aplicadas al conjunto de datos nos muestran en general que una máquina es capaz de obtener resultados eficientes si el conjunto de datos de entrenamiento es grande y las reglas gramaticales de escritura se respetan en el conjunto de datos.

RECOMENDACIONES

Al terminar el presente proyecto y para proyectos futuros se recomienda lo siguiente:

1. Utilizar un modelo relacional para el guardado de los datos una vez que se haya podido acceder a la información de las páginas web a extraer, permite depurar fácilmente si existen fallas en el proceso de guardado de ítems y permite tener una visión más clara de la información que se está extrayendo para los posibles análisis a realizar.

2. Tener en cuenta al momento de programar las arañas, el tiempo en que cada robot accede a una página, para poner en consideración los fallos que los servidores puedan tener y que por lo tanto provoquen información faltante en el proceso de guardado de datos.

3.- Para la extracción de entidades es necesario tomar en consideración que la normalización del texto no sea exhaustiva, pues se corre el riesgo de que con texto demasiado normalizado, el porcentaje de entidades identificadas tiende a bajar.

4.- En el proceso de programación de scripts para extracción de datos mediante Scrapy Python se recomienda utilizar archivos log para guardar los eventos, pues al imprimirlos en consola, el tiempo de duración del script se alarga, y no se tiene una constancia en archivo de los eventos que no podamos controlar por nosotros mismos.

5.- Utilizar una lista de palabras vacías lo más robusta posible, si se puede que sea personalizada al ámbito de datos en donde se va aplicar; esto nos evitará de contar con resultados no deseados al final del proceso que nos hagan retroceder hasta el principio.

BIBLIOGRAFÍA

Alpaydin, E. (2004). *Introduction to Machine Learning.* MIT press.

Benítez, R., Escudero, G., Kanaan, S., & David, M. (2013). *Inteligencia Artificial Avanzada.* Barcelona: UOC.

Bird, S. (2005). NLTK-Lite : Efficient Scripting for Natural Language Processing. *In Proceedings of the 4th International Conference on Natural Language Processing (ICON)*, (pp. 11-18).

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly Media, Inc.

Brown, J. (2000). Growing up digital: How the Web changes work, education, and the ways people learn. *Change*, 11-20.

Cetina, V. (2012). *Aprendizaje por Refuerzo.*

Chafkin, M. (2013). Udacity's Sebastian Thrun, godfather of free online education, changes course. *Fast Company, 14*.

Collazos, A. (2014, 08 12). *Revista Educación Virtual*. Retrieved 09 11, 2014, from Revista Educación Virtual: http://revistaeducacionvirtual.com/archives/1278

Coursera. (2014). *Coursera*. Retrieved 10 27, 2014, from https://www.coursera.org/about/privacy

CURL. (2014). *Front Page*. Retrieved from http://curl.haxx.se/

CURL. (2015, 11 03). *CURL History*. Retrieved from http://curl.haxx.se/docs/history.html

Dale, R., Moisi, H., & Somers, H. (2000). *Handbook of Natural Language Processing.* Marcel Dekker, Inc.

Garrison. (2011). *E-learning in the 21st century: A framework for research and practice.* Taylor & Francis.

Garrison, D., & Anderson, T. (2010). *El e-learning en el siglo XX.* (A. F. Calle, Trans.) OCTAEDRO.S.L.

Gonzáles, R. (2011). *Python para todos.* España.

Gutiérrez, M. A., & Nava, C. D. (2014). Más allá de OCW: los cursos masivos abiertos en línea (MOOCs). *XI Encuentro de Didáctica de la Historia Económica.* Santiago de Compostela.

Harvard. (2012, 05 02). *MIT and Harvard announce edX*. Retrieved from MIT and Harvard announce edX: http://news.harvard.edu/gazette/story/2012/05/mit-and-harvard-announce-edx/

Hernández, J., Ramírez, J., & Ferri, C. (2008). *Introducción a la Minería de Datos.* Madrid: PEARSON.

Indurkhya, N., & Damerau, F. J. (2010). *HANDBOOK OF NATURAL LANGUAGE PROCESSING 2ND EDITION* (Vol. 2). CRC Press.

Json. (2015). *Introduction JSon*. Retrieved from http://www.json.org/

Kolowich, S. (2013). *The chronicle of higher education*. Retrieved 10 24, 2014, from http://chronicle.com/article/The-Professors-Behind-the-MOOC/137905/#id=overview

Kumar, E. (2011). *Natural Language Processing.* I.K International Publishing House.

Liyanagunawardena, T. A. (2013). MOOCs: A Systematic Study of the Published Literature 2008-2012. *International review of research in open and distance learning.*, 202-227.

Loper, E. (2004). NLTK: Building a Pedagogical Toolkit in Python. *PyCon DC* .

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Mit press.

Martínez, C. (2012). E-learning: un análisis desde el punto de vista del alumno. *RIED. Revista iberoamericana de educación a distancia*.

Mazoue, J. G. (2013). The MOOC Model: Challenging Traditional. *EDUCAUSE Review*.

McConnell, D. (2006). *E-Learning Groups and Communities.* Mc-Graw-Hill Education.

Mcgibbney, L. J. (n.d.). *Nutch Wiki*. Retrieved 02 27, 2015, from https://wiki.apache.org/nutch/FrontPage#What_is_Apache_Nutch.3F

miriadaX. (2014). *MiriadaX.net*. Retrieved 09 09, 2014, from https://www.miriadax.net/nuestra-filosofia

Mitchell, R. (2013). *Web Scraping with Java.* Packt Publishing Ltd.

Mooc.es. (n.d.). *Que es un mooc?* Retrieved 09 09, 2014, from Mooc.es: http://mooc.es/que-es-un-mooc/

Nasukawa, T., & Yi, J. (2003). Sentiment Analysis:Capturing Favorability Using Natural Language Processing. *Proceedings of the 2nd international conference on Knowledge capture*, (pp. 70-77).

Russell, S., & Norvig, P. (2004). *Inteligencia Artificial Un Enfoque Moderno Segunda Edición.* Madrid: PEARSON EDUCACIÓN, S.A.

ScrapeSentry. (2014). *ScrapeSentry Scraping Threat Report 2014.* ScrapeSentry.

Scrapy. (2015). *Doc Scrapy.org*. Retrieved 02 27, 2015, from http://doc.scrapy.org/en/0.22/intro/overview.html

Sierra, B. (2006). *Aprendizaje Automático: Conceptos básicos y avanzados.* Madrid, España: PEARSON EDUCACIÓN. S.A.

Udacity.com. (n.d.). *Udacity*. Retrieved 10 10, 2014, from Intro to Psychology: https://www.udacity.com/course/ps001

Udacity-About Us. (n.d.). Retrieved 09 10, 2014, from Udacity-About Us: https://www.udacity.com/us

Vázquez, E., & López, E. (2014). Los MOOC y la Educación Superior: la expansión del conocimiento.Editorial. *Profesorado: Revista de curriculum y formación del profesorado, 18*(1), 3-12.

Vázquez, E., López, E., & Sarasola, J. (2013). *La expansión del conocimiento en abierto : los MOOC.* (J. León, Ed.) Octaedro,S.L.

Vinader, R., & Abuín, N. (2013). Nuevos modelos educativos: los MOOCs como paradigma de la formación online. *Historia y Comunicación Social, 18*, 801-814.

Yuan, L., & Powell, S. (2013). MOOCs and open education: Implications for higher education. *CETIS, JISC*.

Zavando, S. (2011). Aplicación de e-learning en el proceso de enseñanza-aprendizaje. *VI Congreso de Educación a Distancia MERCOCUR/SUL 2002-UCN.*

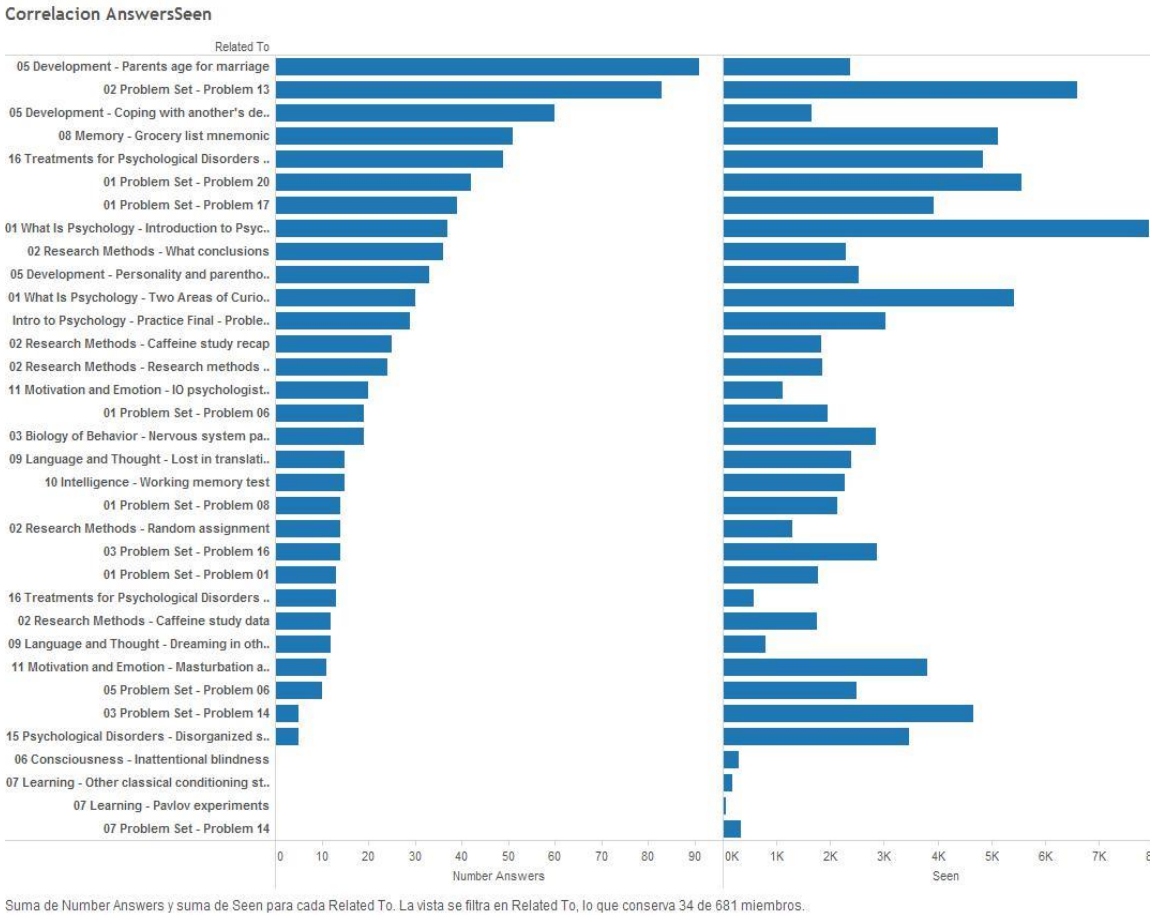
ANEXOS

Anexo 1. Temas más leídos en el foro

	Relacionado a	Visto
1	Null	54121
2	03 Biology of Behavior - Dividing the autonomic nervous system	9448
3	01 What Is Psychology - Introduction to Psychology (and Tickles too)	7939
4	02 Problem Set - Problem 13	6600
5	01 Problem Set - Problem 20	5560
6	01 What Is Psychology - Two Areas of Curiousity	5416
7	08 Memory - Grocery list mnemonic	5133
8	16 Treatments for Psychological Disorders - Congratulations you finished intro psych	4843
9	04 Sensation and Perception - Distance vision competition	4056
10	01 Problem Set - Problem 17	3925

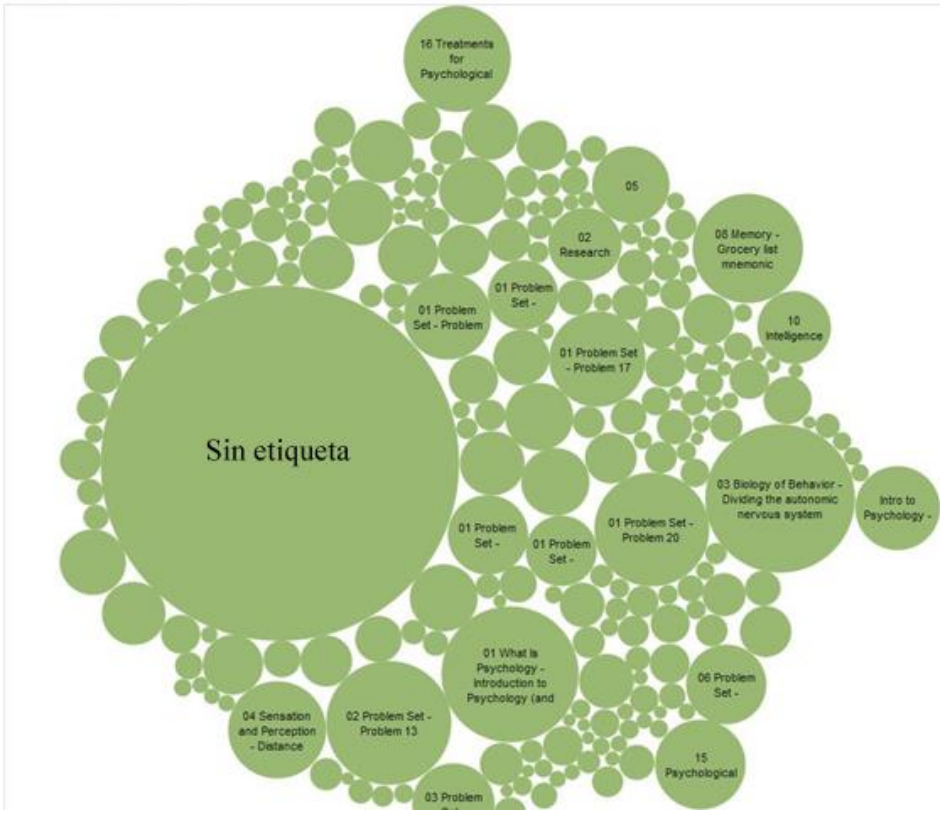
.

Anexo 2. Correlación entre Respuestas y Leídos en Temas del foro

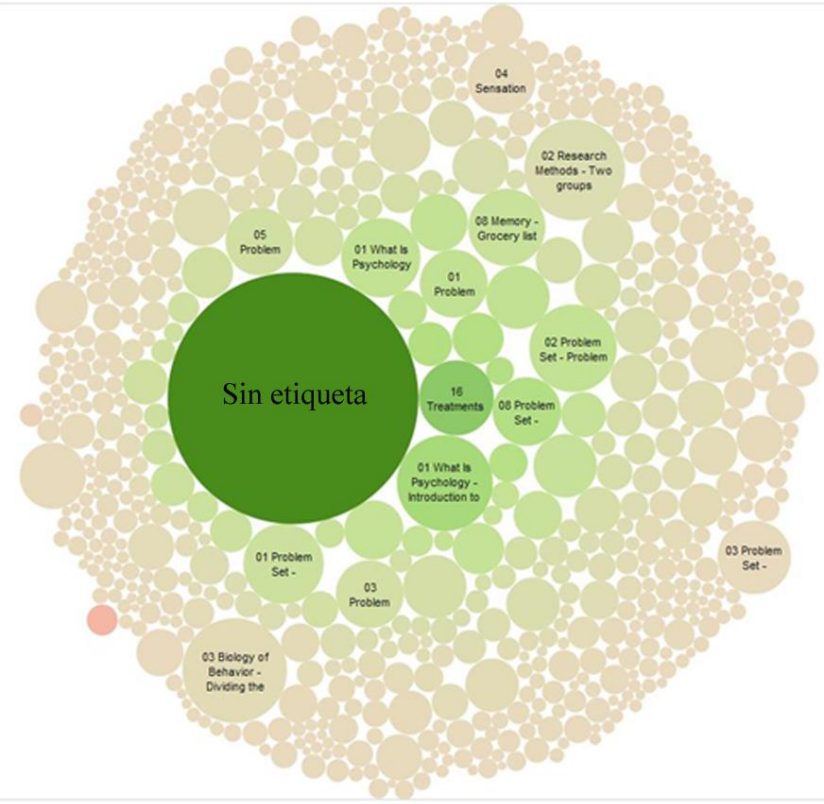


Suma de Number Answers y suma de Seen para cada Related To. La vista se filtra en Related To, lo que conserva 34 de 681 miembros.

Anexo 3. Gráfica de Temas que generan mayor interés por número de vistos



Anexo 4. Gráfica de Temas que generan mayor interés por número de vistos, respuestas y votos a favor



Anexo 5. Tabla de 10 Temas que generaron mayor interés en estudiantes por número de respuestas, vistos y votos a favor

	Related To	Number Answers	Seen	Votes
1	Null	364	54,121	181
2	05 Development - Parents age for marriage	91	2,370	13
3	02 Problem Set - Problem 13	83	6,600	22
4	05 Development - Coping with another's death	60	1,663	16
5	08 Memory - Grocery list mnemonic	51	5,133	17
6	16 Treatments for Psychological Disorders - Congratulations you finished intro psych	49	4,843	77
7	01 Problem Set - Problem 20	42	5,560	16
8	01 Problem Set - Problem 17	39	3,925	23
9	01 What Is Psychology - Introduction to Psychology (and Tickles too)	37	7,939	45
10	02 Research Methods - What conclusions	36	2,288	7

Anexo 6. Número de inscritos por Curso al 11 de Septiembre del 2014 en Udacity

	CURSO	NÚMERO INSCRITOS
1	Intro to Artificial Intelligence	71945
2	Data Analysis with R	18445
3	Applied Cryptography	62581
4	Artificial Intelligence for Robotics	60601
5	Computer Networking	6481
6	Data Wrangling with MongoDB	14803
7	Design of Computer Programs	88878
8	Developing Android Apps	74206
9	Developing Scalable Apps with Java	11150
10	Differential Equations in Action	35124
11	Functional Hardware Verification	15270
12	How to Build a Startup	224962
13	HTML5 Game Development	125714
14	Interactive 3D Graphics	53,131
15	Intro to Algorithms	70779
16	Intro to Algorithms	70779
17	Intro to Computer Science	384,676
18	Intro to Data Science	36573
19	Intro to Descriptive Statistics	5791
20	Intro to Hadoop and MapReduce	54445

21	Intro to Inferential Statistics	3707
22	Intro to Parallel Programming	60236
23	Intro to Point & Click App Development	26788
24	Intro to Statistics	148338
25	Intro to the Design of Everyday Things	55262
26	Intro to Theoretical Computer Science	26493
27	Machine Learning: Reinforcement Learning	4662
28	Machine Learning: Supervised Learning	19756
29	Machine Learning: Unsupervised Learning	6981
30	Make Your Own 2048	31249
31	Mobile Web Development	53206
32	Programming Foundations with Python	42328
33	Programming Languages	84281
34	Software Debugging	29912
35	Software Development Life Cycles	6877
36	Software Testing	58031
37	UX Design for Mobile Developers	17372
38	Web Development	191850
39	Website Performance Optimization	14,934
40	Intro to Java Programming	159620
	MEDIA	64741

Anexo 7. Número de mensajes y participantes en Cursos de Miriadax

Curso	Número de mensajes:	Número de participantes:
Estrategias Metodológicas para el Docente E-learning (2ª edición)	2385	764
Estadística descriptiva (4ª. edición)	598	167
Curso Fundamental de Microeconomía (3ª edición)	510	187
Matemáticas esenciales en los números reales y complejos	250	74
Probabilidad Básica	416	100
Matemáticas básicas (4ª. edición)	430	134
La 3ª edad de oro de la televisión	3541	617
MEDIA	1161	292