



CIMAT

Centro de Investigación en Matemáticas, A.C.

Centro de Investigación en Matemáticas, A.C.

Desarrollo del Curso Introductorio: “Explorando Big Data a Través de Ejercicios Prácticos”.

REPORTE TÉCNICO

Que para obtener el grado de

**Maestro en Ingeniería de
Software**

P r e s e n t a

Octavio Duarte Vázquez

Directores de Reporte Técnico

Alejandro García Hernández

José Guadalupe Hernández Reveles

Índice

Introducción	1
Motivación	1
Objetivo	1
Contenido.....	1
Antecedentes	3
¿Qué es Big Data?.....	3
¿Cuál es la importancia de Big Data en el Mundo?	3
¿Cuál es el tamaño del mercado de Big Data?.....	4
¿Cuál es la relación entre los conceptos: Big Data, Científico de Datos e Ingeniero de Datos?	5
¿Por qué quisiéramos convertirnos en un Científico de Datos?.....	6
¿Cuál es la importancia de los profesionales de Big Data en la Industria?.....	7
¿Cuáles son las alternativas de capacitación actuales como Científico de Datos?	9
Con la gran cantidad de alternativas de capacitación, ¿por qué es necesario otro curso de Big Data?.....	9
¿Y después de realizar el curso qué sigue?	12
Experimento.....	13
Metodología.....	13
Perfil de la muestra y aplicación del instrumento.	14
Resultados	17
Resultados de la encuesta inicial.....	17
Resultados encuesta final.....	23
Resumen de resultados.....	26
Oferta Educativa	29
Conclusiones	41
Trabajo Futuro	43
Referencias.....	45
Anexo A: Explorando Big Data a través de ejercicios prácticos	459
Anexo B: Oferta educativa en Big Data.....	83
Anexo C: Autorización de publicación en formato electrónico de reporte técnico	101

Índice de Tablas

Tabla 4.1. Resumen de resultados de la Encuesta Inicial.	26
Tabla 4.2. Resumen de resultados de la Encuesta Final	27
Tabla 5.1. Oferta educativa por grado académico	31
Tabla 5.2. Tópicos de Databases.....	32
Tabla 5.3. Duración doctorado.....	34
Tabla 5.4. Duración maestría.	35
Tabla 5.5. Duración licenciatura	35
Tabla 5.6. Certificación.....	35
Tabla 5.7. Curso	36
Tabla 5.8. Oferta educativa por país.	36
Tabla 5.9. Costos por grado académico e institución.	38
Tabla 6.1. Alternativas de oferta educativa en Español.....	42

Índice de Figuras

Figura 2.1. Gráfica de distribución de ingresos de Big Data.....	5
Figura 2.2. Gráfica en el tiempo de empleos de Big Data (indeed, 2014).....	6
Figura 2.3. Salarios de Profesionales de Big Data (Piatetsky, 2014).....	7
Figura 2.4. Grado de escolaridad de Profesionales de Big Data.....	8
Figura 2.5. Procedencia de Profesionales de Big Data.....	8
Figura 2.6. Género de Profesionales de Big Data.....	9
Figura 2.7. Camino para convertirse en un Científico de Datos (Chandrasekaran, 2013)	10
Figura 2.8. Habilidades del Profesional de Big Data de acuerdo a su rol.....	11
Figura 4.1. Nivel de retroalimentación encuesta inicial	17
Figura 4.2. Nivel de conocimiento de Big Data.....	18
Figura 4.3. Conocimiento de oferta educativa	18
Figura 4.4. Conocimiento del rol de Científico de Datos	19
Figura 4.5. Alumnos que quieren adentrarse en el tema de Big Data.....	19
Figura 4.6. Conocimiento de bases NoSQL.....	20
Figura 4.7. Conocimiento real de bases NoSQL.....	20
Figura 4.8. Conocimiento de ingreso económico de Científicos de Datos	21
Figura 4.9. Conocimiento de aplicaciones reales de Big Data	21
Figura 4.10. Conocimiento de áreas de estudio un Científico de Datos	22
Figura 4.11. Conocimiento de áreas en las que se involucra Big Data.....	22
Figura 4.12. Nivel de retroalimentación encuesta final	23
Figura 4.13. Nivel de aceptación del curso.....	23
Figura 4.14. Nivel de convencimiento del curso	24
Figura 4.15. Nivel de motivación del curso.....	24
Figura 4.16. Alumnos con las habilidades base para convertirse en un Científico de Datos	25
Figura 4.17. Habilidades base con las que cuentan los alumnos	25
Figura 4.18. Porcentaje de las habilidades base con las que cuentan los alumnos	26
Figura 5.1. Oferta educativa por modalidad	30
Figura 5.2. Modalidad.....	30
Figura 5.3. Oferta educativa por grado académico	31

Introducción

Motivación

De acuerdo con el informe de McKinsey Global Institute para 2018 sólo en Estados Unidos se requerirán de 140,000 a 190,000 personas especialistas en Big Data, así como 1.5 millones de gerentes y analistas para analizar grandes volúmenes y tomar decisiones basados en datos (Manyika et al., 2011).

Adicionalmente la revista Forbes nos recomienda no esperar a que las universidades ofrezcan programas de Big Data. Sugiere empezar ahora mismo a especializarnos por nuestra cuenta, ya que el mercado de hoy en día ya demanda a este tipo de especialistas (Groenfeldt, 2013).

Desde el punto de vista de Gartner se crearán 1.9 millones de puestos para profesionales de Big Data en Estados Unidos para el año 2015 (Beyer, 2012).

Objetivo

Dadas las cifras anteriores este reporte técnico tiene como objetivos:

1. Entender cuál es la importancia y alcance de Big Data en el mundo a nivel social y económico.
2. La creación de un Curso Introductorio a Big Data que motive a los asistentes a tomar la decisión de convertirse en un Profesional de Big Data dándoles a conocer en qué consiste, cuáles son algunas tareas representativas que hace y las ventajas y desventajas que se presentaran al tomar dicha decisión.
3. El tercer y último objetivo es presentar las alternativas de la oferta educativa actual dirigida a formar Profesionales de Big Data.

Contenido

Este reporte técnico se encuentra dividida en tres grandes secciones:

- Primera: La cual incluye el capítulo 2 Antecedentes y da respuesta a los objetivos 1 y 3.
- Segunda: Abarca los capítulos 3. Experimento y 4. Resultados
- Tercera: Capítulos 5. Discusión y 6. Conclusiones y responde al objetivo 2.

La primera sección, el capítulo de Antecedentes, responde las siguientes preguntas: ¿Qué es Big Data?, ¿Cuál es la importancia de Big Data en el Mundo?, ¿Cuál es el tamaño del mercado de Big Data?, ¿Cuál es la relación entre los conceptos: Big Data, Científico de Datos e Ingeniero de Datos?, ¿Por qué quisiéramos convertirnos en un Científico de Datos?, ¿Cuál La importancia de los profesionales de Big Data en la Industria? , ¿Cuáles son las alternativas de capacitación actuales como Científico de Datos?, Con la gran cantidad de alternativas de capacitación, ¿por qué es necesario otro curso de Big Data? Y después de realizar el curso. ¿Qué sigue?

La segunda sección comprende la impartición del curso y se presenta en los capítulos del Experimento donde se presenta las situaciones en las que se impartió el curso así como el perfil de asistentes. Esta sección también está conformada por el capítulo 4. Resultados.

La tercera sección comprende los capítulos de 5. Discusión en el cual se dan los puntos de vista personales en base a las dos secciones anteriores y del capítulo 6. Conclusiones y Trabajo futuro.

Adicionalmente este reporte técnico cuenta con dos anexos sobre conceptos básicos en Big Data, el Anexo A: Tutorial del curso, para realizar la práctica del Experimento y el Anexo B que está conformado por la investigación de la oferta educativa a detalle.

Antecedentes

¿Qué es Big Data?

Big Data ha llegado aquí para quedarse y está teniendo un profundo efecto en la sociedad y los negocios. Big Data tiene un significado para cada tipo de personas, organizaciones e industrias. A continuación se mencionan cuatro definiciones comunes.

- **Wikipedia:** “Big Data es un término general para colecciones de datos tan grandes y complejas que son difíciles de procesar con el uso de herramientas de procesamiento de datos tradicionales.”, (Wikipedia, 2009)
- **Microsoft:** “Big Data es un término cada vez más utilizado para describir el proceso de aplicación de alta potencia de cómputo, machine learning¹ y de inteligencia artificial a información masiva y a menudo de gran complejidad.”, (Microsoft, 2012)
- **Mayer-Schönberger & Cuckier:** “Big Data se refiere a nuestra capacidad creciente hacer cálculos a vastas colecciones de información, analizarla instantáneamente y sacar conclusiones profundas de ellas.”, (Viktor Mayer-Schönberger, 2013)
- **IBM:** “Big Data está siendo generado por todo lo que nos rodea en cada momento. Cada proceso digital e intercambio de medios sociales lo produce. Sistemas, sensores y dispositivos móviles lo transmiten. Big Data está llegando desde múltiples fuentes a una velocidad alarmante, volumen y variedad.”, (IBM, 2014)

Una definición más común y aceptada de Big Data es “Un ambiente de datos en el cual los datos tengan las siguientes características o también llamadas las 3 V, Velocity, Variability, Volumen” (Sicular, 2013). En resumen que los datos provengan de distintas fuentes con distintos formatos, que tengan el orden de petabytes (1 PB = 10¹⁵ byte = 10¹² kB = 10⁹ MB = 10⁶ GB = 10³ TB (Wikipedia, 2014) y sigan creciendo aceleradamente. “Datos tan grandes, de diferentes fuentes y creciendo aceleradamente que no se pueden procesar en un solo equipo”, (Wikipedia, 2009).

La **Variación** la podemos explicar con este tipo de datos de distintos tipos y formatos: Transacciones, Logs, Usuario, Sensor, Social, Médica, Media.

La **Velocidad** y **Volumen** lo podemos entender por las siguientes situaciones:

- Walmart maneja más de 1 millón de transacciones con clientes cada hora (SAS Institute Inc, 2013).
- Google procesa más de 20 petabytes de información por día (Google, 2014).
- En YouTube se suben 100 horas de video cada minuto (YouTube, 2014).
- En Facebook se comparten más 30 billones de contenido cada mes (kissmetrics, 2014).

Una vez comprendido el concepto de Big Data se explicará cuál es su importancia en el Mundo.

¿Cuál es la importancia de Big Data en el Mundo?

Estando en la era de los datos y duplicando el tamaño de ellos cada 2 años (McGaughey, 2011), los datos equivalen a dinero, pero datos consolidados, que se entiendan y hablen entre ellos; ya

¹ Machine Learning. Es un campo de estudio que ofrece a las computadoras la capacidad de aprender sin ser programadas explícitamente (Samuel, 1959).

que el valor de grandes cantidades de datos que se tratan de manera independiente es mucho menor a tratar en conjunto la relación entre ellos, a este fenómeno se le llama “The Iceberg Problem” (DataStax, 2014).

Desde carros auto-manejados hasta drones que entregan paquetes a la puerta de nuestra casa, son solo el comienzo de las aplicaciones de Big Data (Rijmenam, 2013a).

La revolución de Big Data no solo se refiere al exponencial crecimiento del crecimiento de los datos, también recae en el mejoramiento de los métodos estadísticos y computacionales. La capacidad de cómputo se dobla cada 18 meses según la Ley de Moore, pero eso es nada a comparación de un algoritmo con una serie de reglas que puede ser usado para resolver un problema miles de veces más rápido que un método computacional convencional (Shaw, 2014). He aquí la importancia en el mundo académico.

En marketing algunos usos familiares son “sistemas de recomendación” que compañías como Facebook, Amazon, Netflix usan para recomendarnos o sugerirnos algún producto basado en intereses anteriores propios y de otros millones o billones de clientes.

El Institute for Quantitative Social Science de Harvard tiene por propósito ayudar a resolver problemas sociales a través de datos, de los que existen muchos ejemplos. Uno muy interesante que se aplicó en México, donde se detectó que 4 millones de familias se arruinaron al año por no tener un seguro de médico. Así surgió el Seguro Popular (Harvard, 2014). Con más datos podemos hacer más cosas y las posibilidades son ilimitadas.

Big Data ayudará a tomar las decisiones del futuro basadas en datos, ayudará a predecir el futuro basado en el poder de los algoritmos pero lo más importante nos ayudará a comprender mejor nuestro mundo como un todo y quedará en nosotros aplicarlo de forma correcta.

Una frase que resume todo lo anterior es: “Big Data nos ayuda a ver de nuevas formas, nos ayuda a ver mejor, nos ayuda a ver diferente” (Cukier, 2014). Big Data jugará un rol de gran importancia en la sociedad, en el sector empresarial y en los gobiernos.

¿Cuál es el tamaño del mercado de Big Data?

En cuanto a la importancia en números se expone que de acuerdo al estudio realizado por la firma Market Watch (Watch, 2014), en 2014 los vendedores de Big Data ganaron casi US \$ 30 mil millones desde el hardware, el software y los ingresos por servicios profesionales. Se espera que la inversión de Big Data crezca a una tasa compuesta anual de casi el 17% durante los próximos 6 años, lo que con el tiempo representará US \$ 76 mil millones a finales de 2020.

Desde el punto de vista de Wikibon community of IT practitioners (Kelly, 2014), obtenemos este punto que resulta muy interesante para los profesionales de Big Data el cual es que el mayor porcentaje el 40% del total del mercado se va a los servicios profesionales que prestan los profesionales de Big Data como lo muestra la figura 2.1.

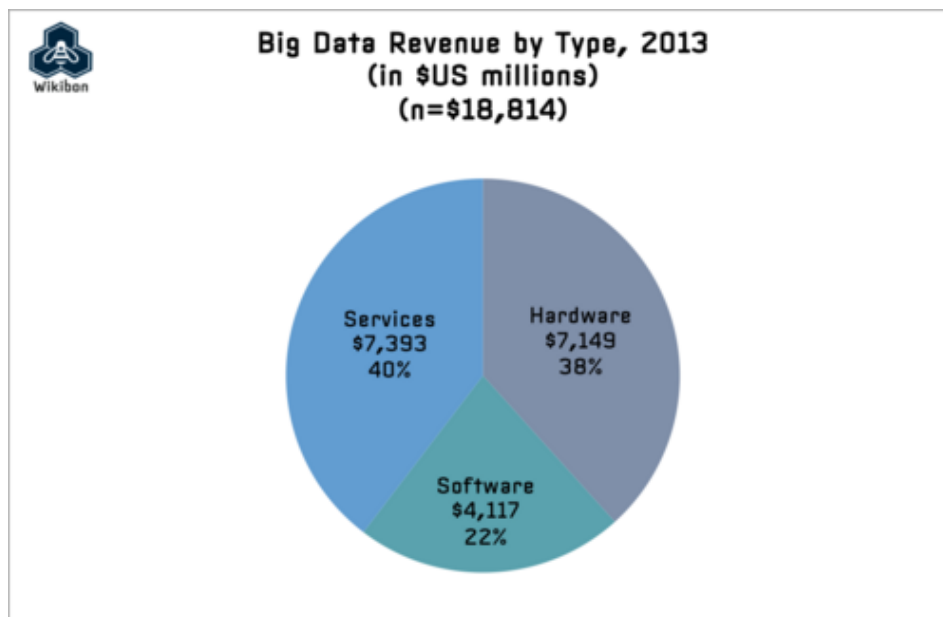


Figura 2.1. Gráfica de distribución de ingresos de Big Data

Finalmente LinkedIn nos dice que el mercado mundial de Big Data valía USD 6,3 mil millones en 2012 y se espera que llegue a USD 48,3 mil millones en 2018, a una tasa compuesta anual del 40,5% desde 2012 hasta 2018 (Collins, 2014).

Como se observa los tres estudios están de acuerdo en que el mercado de Big Data crecerá de manera constante, con valores distintos pero muy cercanos y con la misma tendencia.

¿Cuál es la relación entre los conceptos: Big Data, Científico de Datos e Ingeniero de Datos?

Ahora que el concepto de Big Data ha sido explicado, nos surgen las siguientes preguntas ¿qué tipo de personas trabajan el Big Data?, ¿existe alguna carrera profesional específica?

Existen varios roles en el ámbito de Big Data pero en este reporte técnico nos vamos a enfocar al rol del Científico de Datos y al rol del Ingeniero de Datos como actores principales, muchos autores segmentan los roles en más partes o les dan diferentes nombres lo que no se tocará en esta sección (Ariker Matt, 2013; BigData-Startups, 2013).

El rol del Científico de Datos es el más importante en cuanto a la interpretación de los datos, diseño de algoritmos y análisis predictivos, es el que aplica métodos matemáticos y estadísticos a los datos para obtener valor de ellos, adicionalmente aplica conocimientos y metodologías de distintas áreas a los datos como machine learning, deep learning, inteligencia artificial, el científico de datos es multidisciplinario (Rijmenam, 2013c).

En cuanto al Ingeniero de Datos él es el que diseña e implementa la solución de Big Data para almacenar, consumir, analizar, visualizar los datos, es el encargado de decidir las tecnologías de hardware y software que se adaptan mejor a la situación que se está tratando para obtener el mayor beneficio y valor de los datos. El Ingeniero de datos está muy relacionado con lenguajes de programación orientados a análisis científicos como Python, reconoce cuando utilizar Hadoop y

cuando utilizar bases de datos NoSQL, tiene el conocimiento para definir flujos de datos para conjuntos de datos del orden de los petabytes (Rijmenam, 2013b).

En este reporte técnico vamos a referirnos de ahora en adelante al científico de datos y al ingeniero de datos como a una sola entidad Científico de Datos, es decir que de ahora en adelante al referirnos al Científico de Datos vamos a abarcar estos dos roles principales. De igual manera se usará el término Profesional de Big Data cuando se requiera todos los roles relacionados con Big Data.

¿Por qué quisiéramos convertirnos en un Científico de Datos?

Recuerdan cuando nuestros padres nos decían de pequeños, ¡tienes que ser Doctor! o ¡Abogado! o la profesión que era o parecía la que tenía una mayor remuneración económica, pues tenemos otra buena opción ahora nosotros le deberíamos decir a nuestros, hijos o alumnos, tienen que ser un ¡Científico de Datos! o un ¡Ingeniero de Datos!

Big Data con el panorama actual catapulta a los científicos de datos como otra muy buena opción de carrera profesional y sobre todo bien remunerada. Ya que el Big Data es una herramienta clave para las empresas para ganar competitividad, tomar decisiones basadas en datos. Esto ha incrementado de manera exponencial la demanda laboral para profesionales del Big Data como se muestra en la figura 2.2.

"big data" Job Trends



Figura 2.2. Gráfica en el tiempo de empleos de Big Data (indeed, 2014)

Ligado con la alta demanda de científicos de datos los sueldos son muy atractivos como lo muestra la encuesta hecha por KD Nuggets (Piatetsky, 2014) en donde nos muestra el salario por rol, por tipo de empleador y región con una muestra de 240 profesionales de Big Data. Lo podemos ver más claro en la figura 2.3

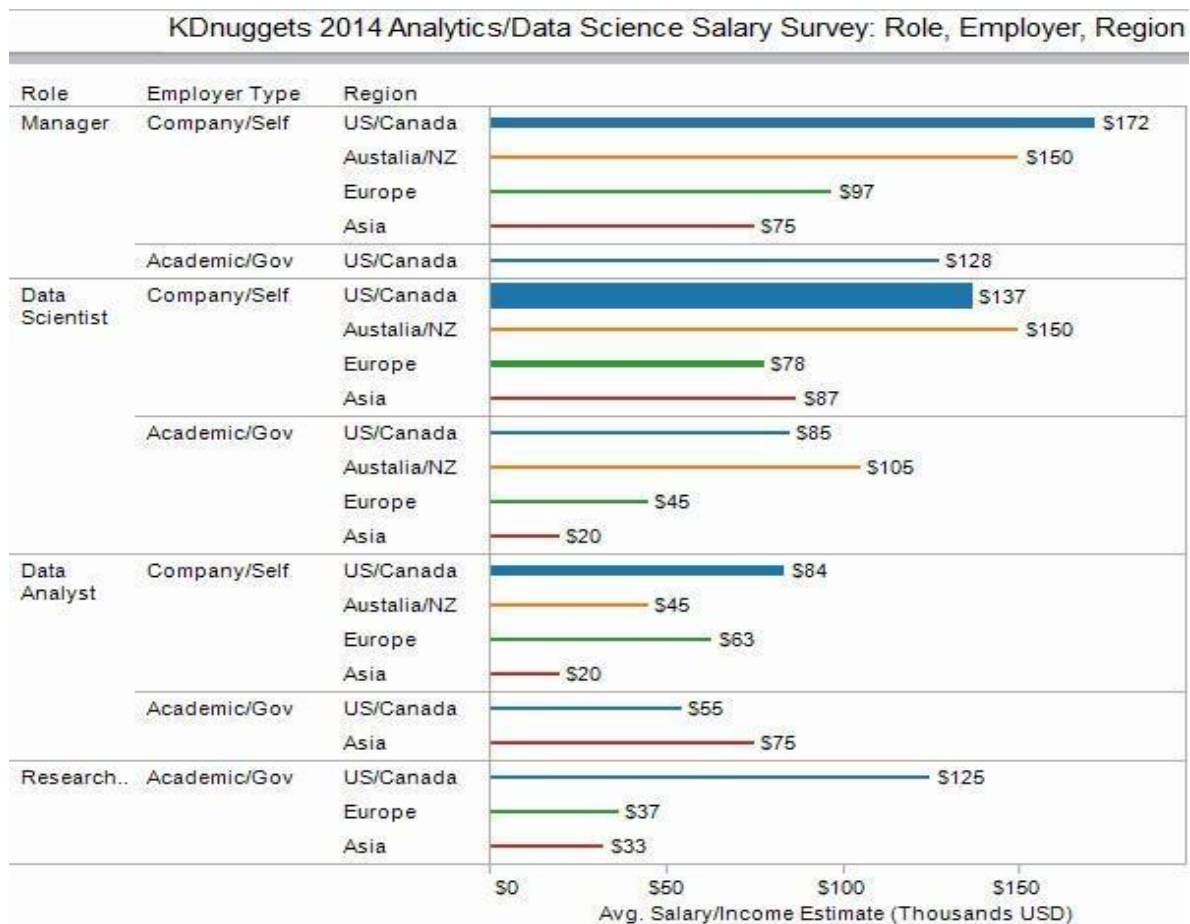


Figura 2.3. Salarios de Profesionales de Big Data (Piatetsky, 2014)

Basados en los puntos expuestos previamente al tomar la decisión de convertirse en un profesional del Big Data nos proporcionará un sueldo y una carrera profesional muy interesante. Además de eso la cultura o el ámbito en el que se desarrolla un profesional del Big Data es muy variada, ya que puede trabajar en la mayorías de las áreas, apoyándose en conocedores del dominio del tema, lo que nutre de conocimiento de manera incremental al profesional del Big Data, no será un experto en cada tema pero conocerá de él, en lo que sí será experto es en encontrar valor en los datos.

¿Cuál es la importancia de los profesionales de Big Data en la Industria?

Un aspecto muy importante es que los científicos de datos, no sólo se desarrollan como personas técnicas, es decir no están aislados en el área de sistemas y de allí no tienen interacción con el resto de la empresa a la que pertenecen, sino todo lo contrario, los científicos de datos van de la mano de la toma de decisiones de las empresas, interactúan con la mayoría de las áreas para obtener datos valiosos y saber cómo interpretarlos, es decir los científicos de datos están tomando decisiones o están al lado de los tomadores de decisiones. Decisiones que cambian o alinean el rumbo de empresas y tienen como fin obtener ganancias económicas y sacar ventaja ante los competidores esta es la razón por la que el salario promedio es alto.

Otros datos interesantes que nos muestra el estudio realizado por la firma Burtch Works con una muestra de 2,845 profesionales en Big Data son los siguientes (Burtch, 2013):

Los profesionales en Big Data están mejor preparados académicamente ya que:

- 46% tienen el grado de Doctor
- 42% una maestría
- 11% una licenciatura
- 1% sin grado

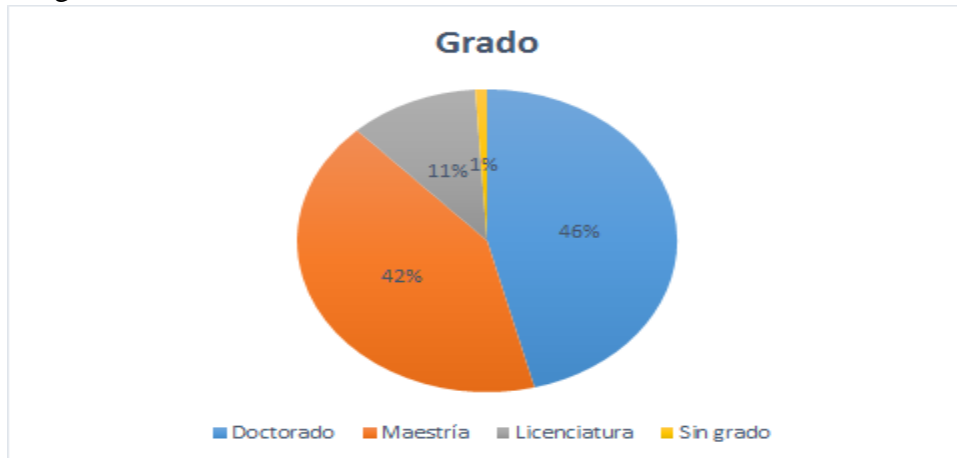


Figura 2.4. Grado de escolaridad de Profesionales de Big Data

Proceden de las siguientes áreas:

- Matemáticas y Estadística
- Ciencias de la Computación
- Ingeniería.

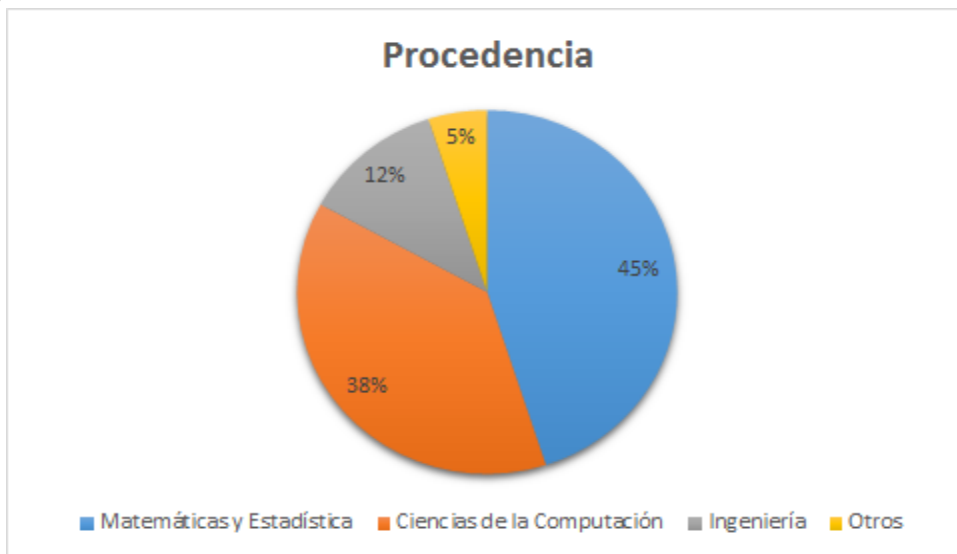


Figura 2.5. Procedencia de Profesionales de Big Data

En cuanto al género

- 75% hombres

- 25% mujeres



Figura 2.6. Género de Profesionales de Big Data

Pero no solo eso se necesita para convertirse en un profesional de Big Data, además de tener alguna maestría o doctorado se necesita tener **habilidades de comunicación** ya que como se mencionó los científicos de datos tienen que estar en contacto con la mayoría de áreas de las empresas y por ende saber comunicarse con conocedores del dominio a tratar para sacar el mayor valor a los datos, se necesita **un alto grado de curiosidad** y tener una **comprensión de lo que son negocios reales**, deben de saber que una mala decisión tiene consecuencias reales en las empresas.

¿Cuáles son las alternativas de capacitación actuales como Científico de Datos?

Como se mencionó anteriormente las alternativas actuales están enfocadas a cubrir temas específicos o el camino completo para convertirse en un profesional de Big Data.

Los resultados de la investigación de la oferta educativa actual se pueden consultar en el capítulo **Oferta Educativa**, de igual manera el detalle de cada programa se encuentra en el apéndice A de este reporte técnico:

Con la gran cantidad de alternativas de capacitación, ¿por qué es necesario otro curso de Big Data?

En resumen, **se asume que hay un interés** y conocimiento generalizado sobre Big Data y que los interesados son expertos, tienen claro el camino para desarrollarse en este ámbito y saben de antemano que existen perfiles de profesionales de Big Data. A continuación se detallan cada una de estas suposiciones.

Se asume que todos los interesados en Big Data son expertos y tienen habilidades y conocimientos necesarios para el tema.

- Tienen el conocimiento de la gran variedad de temas y problemas que se pueden resolver, o del valor que se puede generar a alguna industria con Big Data, que puede ser aplicado a

temas de salud, agricultura, educación, automotriz, aeroespacial, comercio electrónico, seguridad, etc., La lista es interminable

- Tienen las habilidades que son necesarias para entrar a el tema de Big Data, por ejemplo las Matemáticas, la Estadística y Programación como bases y están conscientes de que estas habilidades son sólo lo mínimo necesario comprender los temas posteriores.
- Los científicos de datos deben tener el sentido de negocio, debe pensar como un empresario, debe tener la visión de crear aplicaciones basadas en datos que obtengan o provoquen valor en forma de ganancias monetarias o en forma de conocimiento. Siendo esta característica una de las más difícil de encontrar en personas que vienen del área de estadística, matemáticas, ciencias de la computación e ingeniería.
- El científico de datos debe de ser curioso y autodidacta y buscar el porqué de las cosas. Finalmente debe ser un hacedor que no solo se quede en la idea sino que la aplique como un emprendedor.

Se asume que los interesados en Big Data tienen una noción de los diferentes aspectos de Big Data que deben de dominar para convertirse en profesionales del tema y saben que tienen un largo camino por recorrer y que no es nada fácil. Como se nos muestra en la figura 2.7.

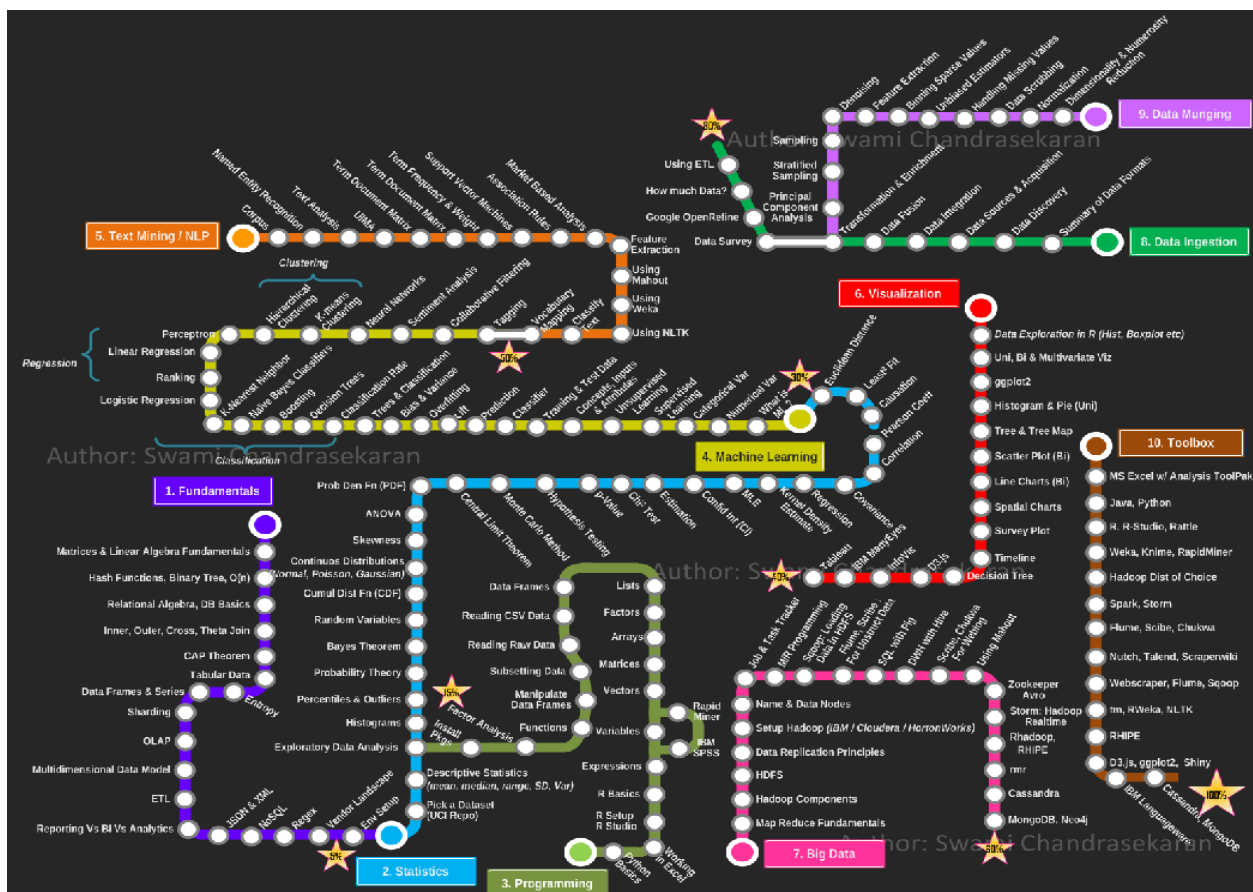


Figura 2.7. Camino para convertirse en un Científico de Datos (Chandrasekaran, 2013)

Se asume que los interesados en Big Data entienden que hay diferentes perfiles de un profesional del tema.

- Los profesionales de Big Data son multidisciplinarios ya que el camino para llegar a convertirse en un científico de datos comprende varias áreas de investigación como la figura 2.7 lo expone y la figura 2.8 presenta la relación entre las áreas de investigación y el rol del profesional de Big Data.

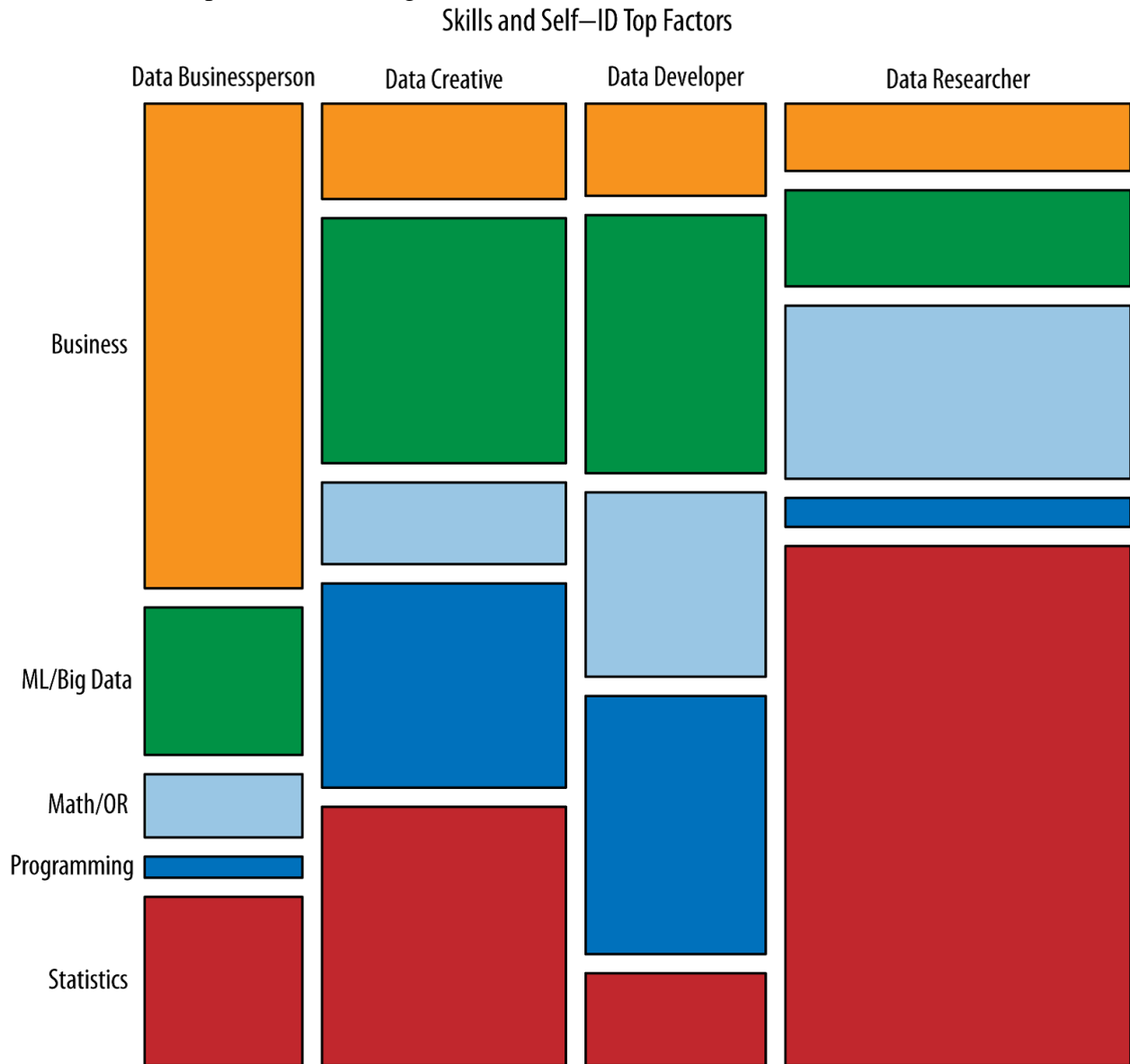


Figura 2.8. Habilidades del Profesional de Big Data de acuerdo a su rol (Vorhies, 2014)

El problema con asumir todo lo anterior, es que no se toma en consideración la motivación del interesado o su grado de conocimiento del tema. Para comenzar en este camino, se propone la realización de un curso introductorio práctico que presentar de una manera general y atractiva Big Data y que permita realizar ejercicios prácticos y así para motivar al interesado a convertirse en un Científico de Datos o en un Profesional de Big Data.

¿Y después de realizar el curso qué sigue?

Para diseñar este curso se analizaron una gran cantidad de alternativas de cursos introductorios, así como libros relacionados con el tema, dando como resultado que al final del curso se les recomienda a los alumnos convencidos de entrar al mundo de Big Data tres libros y dos cursos online.

Los alumnos que tomen el curso son instruidos mediante una ponencia la cual les explica los siguientes puntos a detalle:

1. El primer punto es la pregunta ¿Qué es Big Data y sus orígenes? esta pregunta es respondida mediante definiciones de grandes actores en la industria de los datos como IBM, Microsoft, Google, además de mencionarles la importancia que tuvo Google en el nacimiento de Big Data mediante la publicación de sus artículos Google File System (Ghemawat, Gobioff, & Leung, 2003) y Map Reduce (Dean & Ghemawat, 2004), adicionalmente mostrándoles el funcionamiento de Hadoop Distributed File System(HDFS) y Map Reduce.
2. El segundo punto ¿Por qué estamos en la Era del Datos? esto se explica mediante ejemplos de la vida diaria, como los gadgets de corredores de hace 10 años y los de ahora.
3. EL tercer punto muestra como grandes empresas como Coca Cola, Master Card, Walmart, Amazon, T-mobile están aplicando Big Data.
4. Como cuarto punto siendo el objetivo principal de este reporte técnico, se explica a los alumnos las habilidades y destrezas con las que se tiene que contar para aspirar a entrar al mundo de Big Data, adicionalmente se les advierte que el camino no es fácil ni corto, pero también dejándoles claro que la recompensa es muy alta a nivel económico, personal y cultural.
5. Finalmente el quinto punto explica es el ciclo de vida de los datos y el proceso de análisis de datos.

En la sección práctica del curso los alumnos verán 3 ejemplos de los más simples y comunes en el análisis de datos, así como el paso a paso de cómo resolverlos, donde el alumno se dará cuenta que para problemas pequeños las respuestas no son triviales y necesitan de bastante razonamiento. Como punto clave se les explica a los alumnos que conocimientos o áreas de estudio se necesitan para convertirse en un científico de datos y en ese momento se dan cuenta si son capaces de afrontarlo y por qué no, de disfrutar dichas áreas de estudio.

Aclarando este curso no está enfocado a aprender un tema en específico necesario para convertirse en un científico de datos ya que para eso se necesitan más de 4 horas para un solo tema y no sería suficiente. Tampoco está enfocado en cubrir alguna de las herramientas actuales.

Es por eso que este curso es diferente ya que te dice los beneficios y retos que hay que pasar para convertirse en un científico de datos, evitando así que se queden a medio camino. El presente Reporte Técnico y el material generado pueden utilizarse como curso propedéutico para una especialidad en Big Data.

Experimento

Metodología

Basados en el área de oportunidad expuesta en la sección anterior, se diseñó un curso de 4 horas que se impartió de manera presencial el cual tiene finalidad de alentar a los participantes para entrar en el mundo de Big Data y convertirse en un Profesional de Big Data.

El curso cuenta con cuatro partes principales:

1. Encuesta de entrada. Para saber el nivel inicial de conocimiento de los asistentes.
2. Presentación por parte del instructor. Para proveer los fundamentos teóricos del área.
3. Ejercicios Prácticos. Para hacer el conocimiento significativo.
4. Encuesta de salida. Para verificar si se logró el objetivo de aumentar el conocimiento de big data y motivar a algunos de los participantes a convertirse en Profesionales de Big Data.

La primera parte consta de una encuesta inicial de 5 minutos en la cual el participante responde una serie de preguntas enfocada a averiguar cuáles son sus conocimientos acerca de Big Data. La encuesta se encuentra en: <http://goo.gl/20HTJF>.

Estas son las preguntas

- ¿Qué es Big Data para ti?
- ¿Sabes de algún curso, certificación, licenciatura o maestría, relacionado con Big Data?
- ¿Sabes qué áreas se involucran en Big Data?
- ¿Sabes qué es un Científico de Datos?
- ¿Quieres dedicarte a Big Data?
- ¿Qué se necesita para ser un Científico de Datos?
- ¿Sabes que es NoSQL?
- Menciona 3 bases de datos NoSql Que conozcas
- ¿Sabes cuánto es el salario promedio de un Científico de Datos?
- Menciona alguna aplicación real que conozcas relacionada con Big Data y justifícala

La segunda parte es una presentación por parte del instructor que explica las siguientes preguntas y temas:

- ¿Qué es Big Data?
- ¿Por qué estamos en la Era del Datos
- Los Orígenes de Big Data
- ¿Qué es Hadoop y Map Reduce?
- Ejemplos reales de Big Data en la Industria
- ¿Porque querer convertirse en un científico de datos?
- Habilidades necesarias para empezar el camino hacia Big Data
- Ciclo de vida de los datos
- Proceso de Análisis de Datos

Al final de la presentación se les recomiendan tres libros y dos cursos en línea gratuita para los que se interesaron el tema.

- Libros
 - Practical Data Analysis, por Héctor Cuesta, 2014.
 - Agile Data Science, por Russell Journey, 2014.
 - Mining the Social Web, Second Edition por Matthew A. Russell, 2014
- Cursos
 - Big Data University (<http://bigdatauniversity.com/>)
 - Udacity Data Science (<https://www.udacity.com/courses#!/data-science>)

Esto les da a los alumnos una gran ventaja la cual consiste en el ahorro de tiempo en buscar el camino correcto hacia el mundo de Big Data.

La tercera parte del curso consta de una práctica en la cual se realizan los tres ejercicios siguientes:

1. Twitter. Análisis de Sentimiento “En este ejercicio se clasifican los tweets del tema de interés de los participantes en positivos o negativos, esto para entender el sentimiento de los tuiteros acerca del tema”.
2. Facebook. “En este ejercicio se aplicarán métodos matemáticos para obtener grado de distribución, centralidad del grupo de amigos del participante”.
3. Predicción del precio del Oro. “En este ejercicio se realiza la predicción del precio del oro basados en valores históricos aplicando métodos matemáticos y estadísticos como por ejemplo regresiones no lineales”.

En el Anexo A pueden consultar el detalle de los 3 ejercicios.

Esta práctica se realiza en Python3 y librerías de este lenguaje, se proporciona a los participantes una máquina virtual ya preparada con todo lo necesario para realizar la práctica; además de un tutorial con instrucciones paso a paso para realizar los ejercicios. Se puede descargar el material en el siguiente link <http://goo.gl/qr4OdV>.

En la cuarta parte se aplica una encuesta final de 5 minutos en la cual los participantes dan la evaluación y retroalimentación del curso, además de evaluarse ellos mismos y detectar si el objetivo del curso se cumplió el cual es motivar a los participantes a convertirse en profesionales de Big Data sabiendo que el camino es largo y complicado pero vale la pena el esfuerzo. El link es el siguiente <http://goo.gl/czCR0a>.

Las preguntas que se realizaron fueron las siguientes:

- ¿Qué te pareció el curso?
- ¿Cómo mejorarías el curso?
- ¿Cuáles de las habilidades bases para ser científico de datos ya tienes?
- ¿Dirías que este curso te motivo a ser un científico de datos?
- ¿Te inscribirías a un curso para convertirte en un Científico de Datos?

Perfil de la muestra y aplicación del instrumento.

El curso se impartió en cuatro ocasiones y a continuación se describen los lugares y el tipo de asistentes.

1. 26 de Septiembre de 2014

- a. CIMAT Zacatecas
 - b. 10 participantes alumnos de la Maestría en Ingeniería de Software. Se contó con alumnos desde primer a cuarto semestre y la Dr. Alejandra García de la UAZ.
- 2. 10 de Octubre de 2014
 - a. CIMAT Zacatecas
 - b. 14 participantes entre los cuales se encontraban alumnos por graduarse del CIMAT en la Maestría en Ingeniería de Software, participantes del CIMPS 2014, trabajadores en la industria del software, alumnos de ingeniería de software de la UAZ, Dos Doctores de la UAZ la Dra. Perla Eugenia Velasco Elizondo y el Dr. Sodel Vázquez Reyes.
- 3. 21 de Octubre de 2014
 - a. ITSZO
 - b. 23 participantes estudiantes de primer y tercer semestre de la Ingeniería en sistemas computacionales

Resultados

Los resultados del Experimento, para este caso la impartición del “Curso introductorio a Big Data”, están basados en las encuestas aplicadas a los participantes al inicio y al final del curso. En total el experimento se aplicó a **3** grupos para un total de **47** participantes.

Antes de revisar los resultados es necesario recalcar que solo 34 alumnos de los 47 contestaron la encuesta inicial o lo que equivale al 72.34%. El grupo de ITSZO, a nivel licenciatura mostró menos interés como se muestra en la figura 4.1.

Resultados de la encuesta inicial

La encuesta inicial tiene el fin de detectar el conocimiento previo de los participantes respecto a Big Data. A continuación se presentan los resultados.

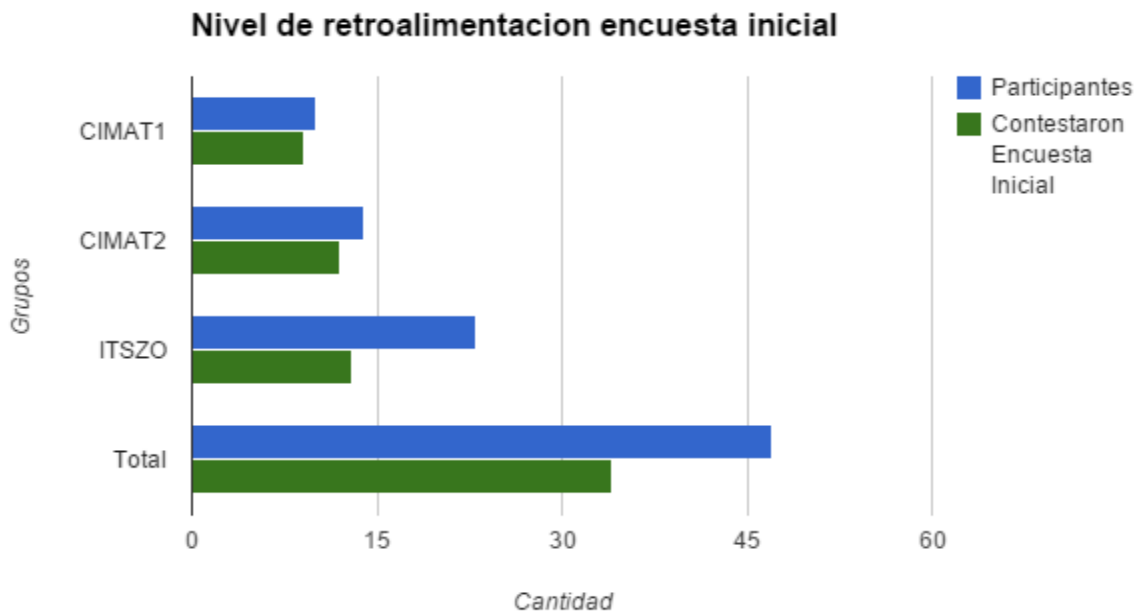


Figura 4.1. Nivel de retroalimentación encuesta inicial

Observamos que sólo el **13.95%** conoce y tiene claro el concepto de Big Data. Es preocupante que a nivel licenciatura no se tiene el conocimiento del tema como se observa en la figura 4.2 en el grupo 3 ITSZO.

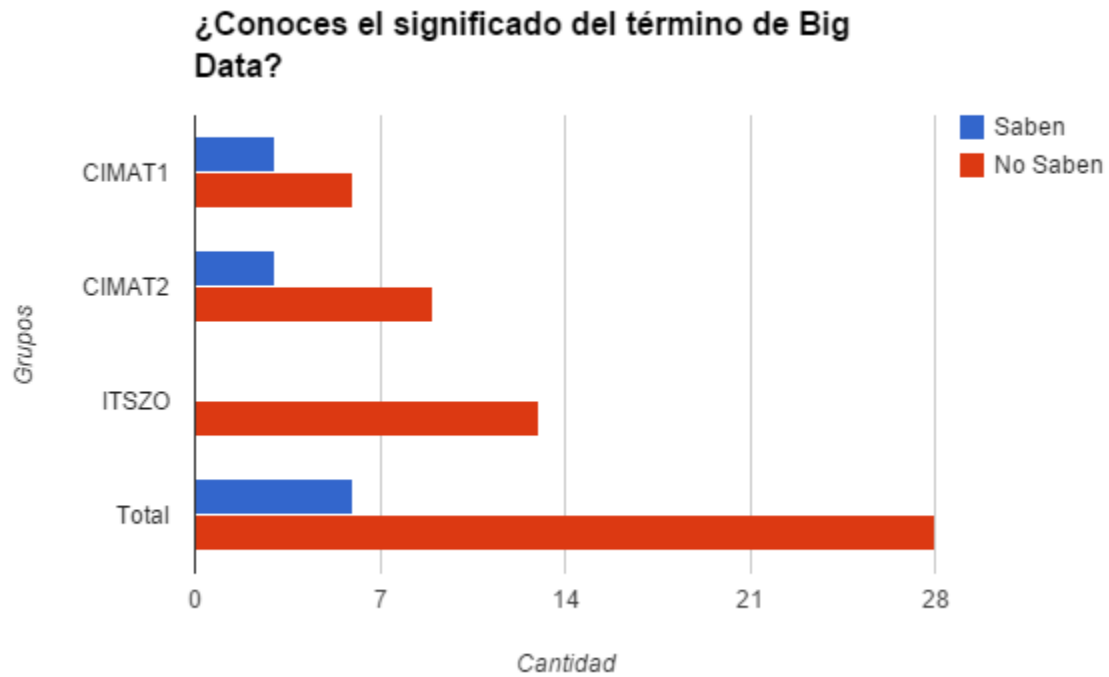


Figura 4.2. Nivel de conocimiento de Big Data

Solo el **5.88%** conoce de alguna oferta educativa, ello confirma el desconocimiento del tema como se muestra en la figura 4.3.

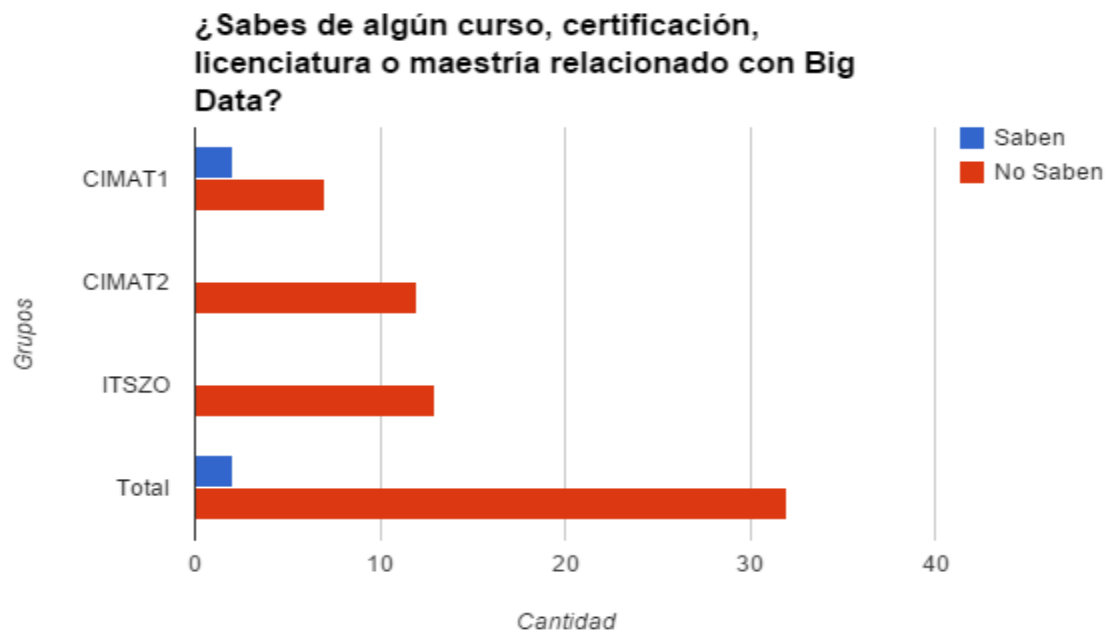


Figura 4.3. Conocimiento de oferta educativa

Solo el 23.53% sabe que es un científico de datos con una definición muy básica del mismo, también observamos que los alumnos del tercer grupo repite la tendencia de no estar enterados del tema como se muestra en la figura 4.4.

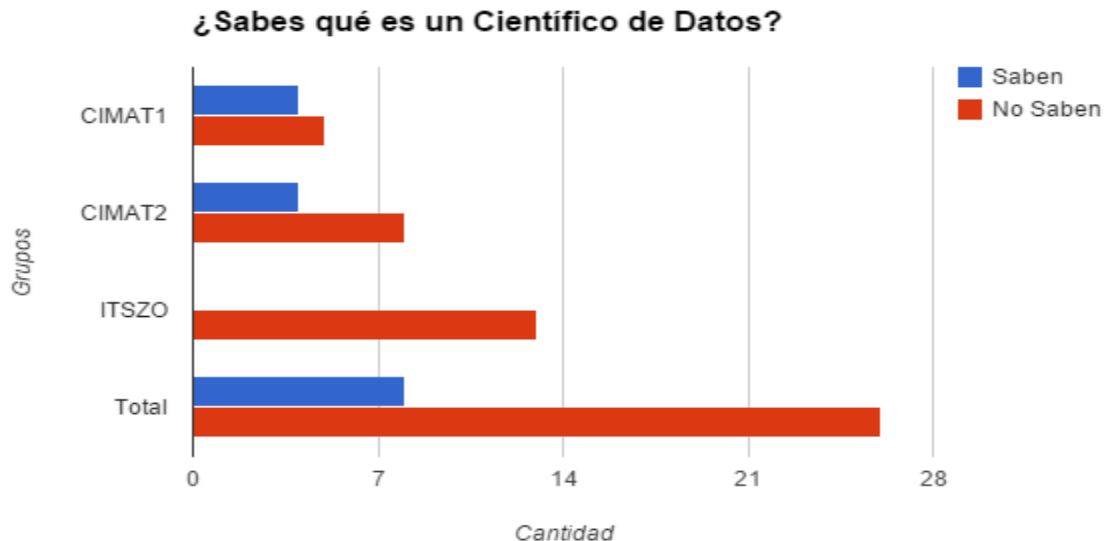


Figura 4.4. Conocimiento del rol de Científico de Datos

La figura 4.5 muestra algo interesante que es que el **67.65% de los alumnos quieren dedicarse a Big Data** aunque no sepan exactamente lo que es, ni por dónde empezar, ya que como muestra la figura 4.2 solo el 13.95% lo sabe.

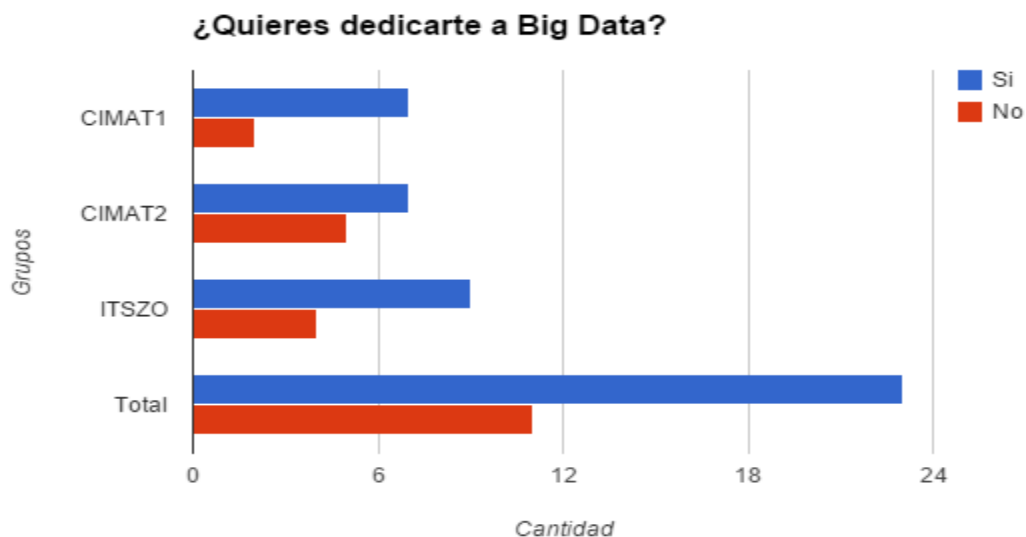


Figura 4.5. Alumnos que quieren adentrarse en el tema de Big Data

El tema de bases de datos NoSQL está muy relacionado con Big Data ya que aproximadamente **el 80% de los datos que generamos carecen de una estructura relacional** (IBM, 2014), es por eso que la siguiente pregunta les pregunta de manera directa si saben o no saben que es NoSQL solamente tomando en cuenta la respuesta de “Si” o “No” dando como resultado que **el 44.12% dice que sabe que es NoSQL**. Como se aprecia en la figura 4.6 nuevamente el tercer grupo ITSZO que corresponde a licenciatura son los alumnos menos relacionados con el tema.

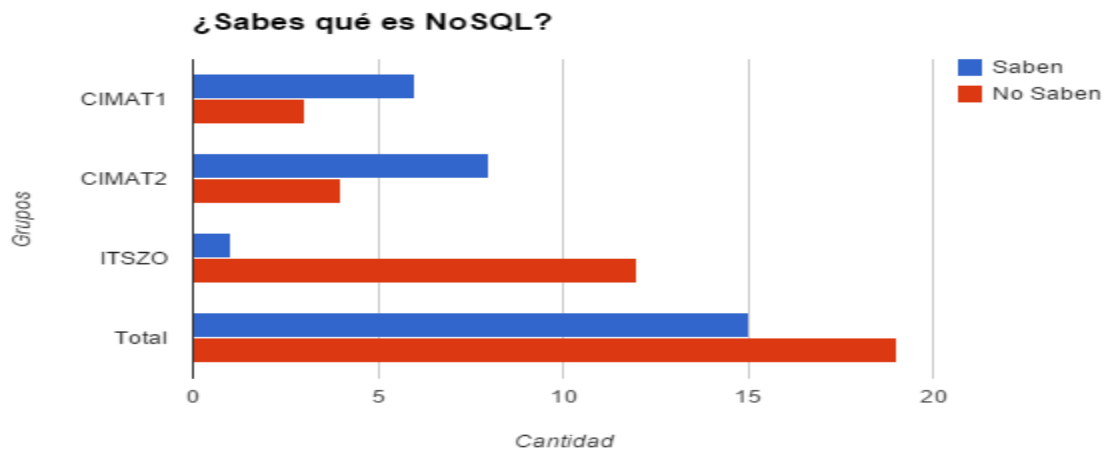


Figura 4.6. Conocimiento de bases NoSQL

Respecto a la pregunta, mencione 3 bases de datos NoSQL, se comprobó que realmente solo el 35.29% conoce las bases de datos NoSQL y no el 44.12%. Con una variación aproximada del 9% que son alumnos que creen saber que es una base NoSQL pero en realidad no lo saben como se muestra en la figura 4.7.

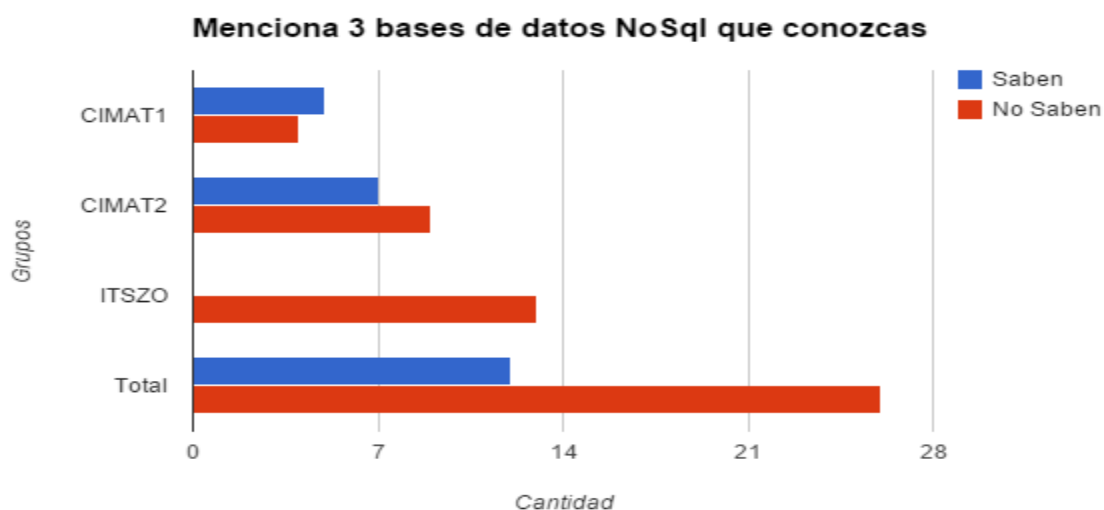


Figura 4.7. Conocimiento real de bases NoSQL

Un factor motivante en la decisión de tomar el camino para ser un profesional de Big Data es el rango del salario que se puede obtener y como se observa en la figura 4.8 el conocimiento de este aspecto es casi nulo con solo un 5.88%.

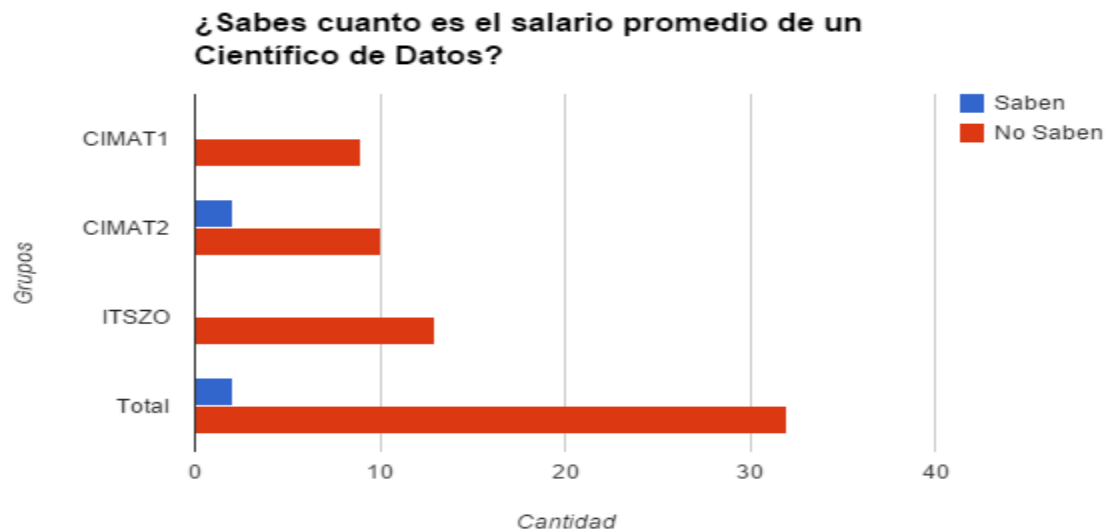


Figura 4.8. Conocimiento de ingreso económico de Científicos de Datos

En cuanto a la aplicación de Big Data en el mundo se les pidió mencionar un ejemplo real que conocieran con lo cual vemos que la tendencia del conocimiento se mantiene con un **20.59% contra un 17.65% que saben de manera correcta lo que es Big Data**, con estos resultados podemos deducir un 2.94% también saben lo que es Big Data pero no supieron expresarlo los resultados se ven en la figura 4.9.

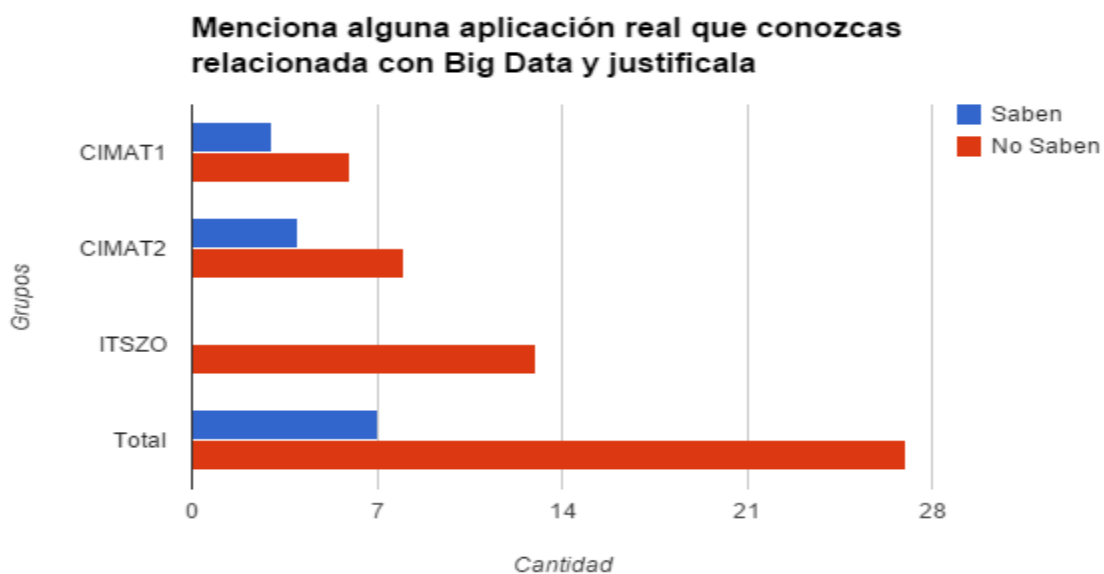


Figura 4.9. Conocimiento de aplicaciones reales de Big Data

Los siguientes dos resultados son diferentes a los anteriores ya que el tipo de pregunta que se realizó es de opción múltiple, en este tipo de preguntas ver las respuestas beneficia al alumno a poder escoger alguna respuesta aunque esto no se tomó como un factor que influye en los resultados. En la pregunta “¿Qué se necesita para ser un Científico de Datos?” se insertaron el 37.5% de opciones erróneas y se obtuvo que el 38.24% de los alumnos seleccionaron tres características o más que se necesitan para ser un científico de datos sin seleccionar ninguna opción incorrecta, esta medida se tomó como base para calificar su respuesta como que “Saben” observamos los resultados en la figura 4.10.

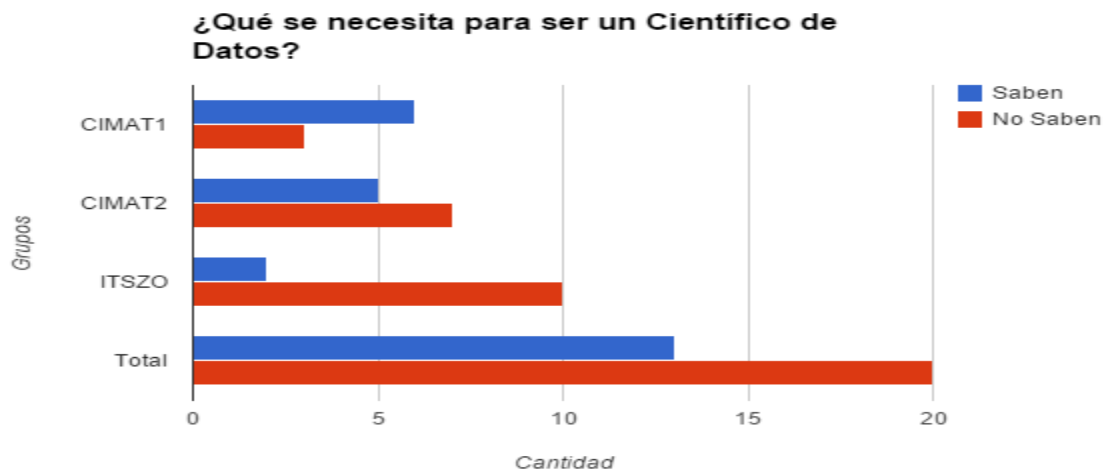


Figura 4.10. Conocimiento de áreas de estudio un Científico de Datos

En la siguiente pregunta “¿Sabes qué áreas se involucran en Big Data?” no se insertaron opciones erróneas pero se calificó como respuesta válida si se seleccionaron por lo menos 4 de las 7 opciones es decir un porcentaje mayor al 57% lo cual nos arroja un resultado del 44.12% de alumnos saben qué áreas están involucradas como se observa en la figura 4.11.

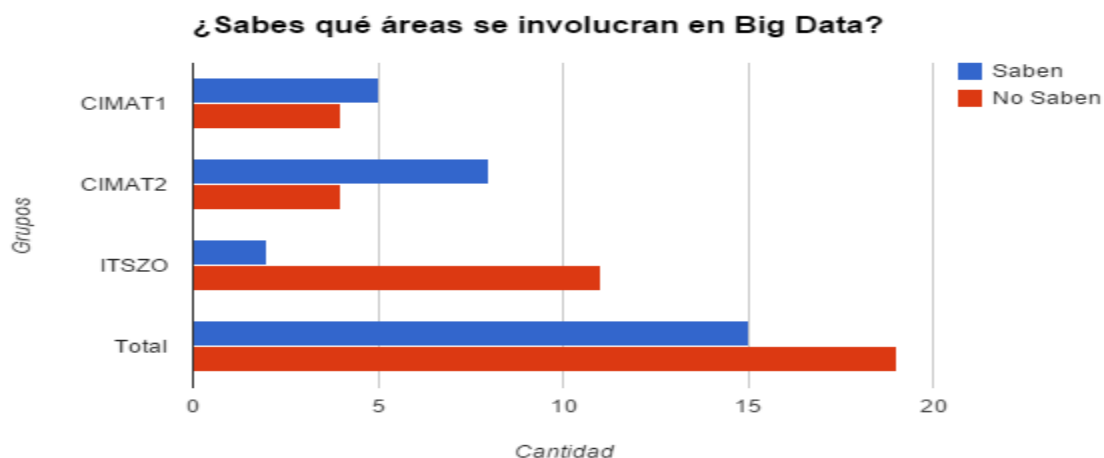


Figura 4.11. Conocimiento de áreas en las que se involucra Big Data

Resultados encuesta final

La segunda encuesta tiene el objetivo de obtener la retroalimentación del alumno acerca del curso en la cual 32 de los 47 alumnos llenaron la encuesta notando la misma tendencia que en la encuesta inicial como se muestra en la figura 4.12.

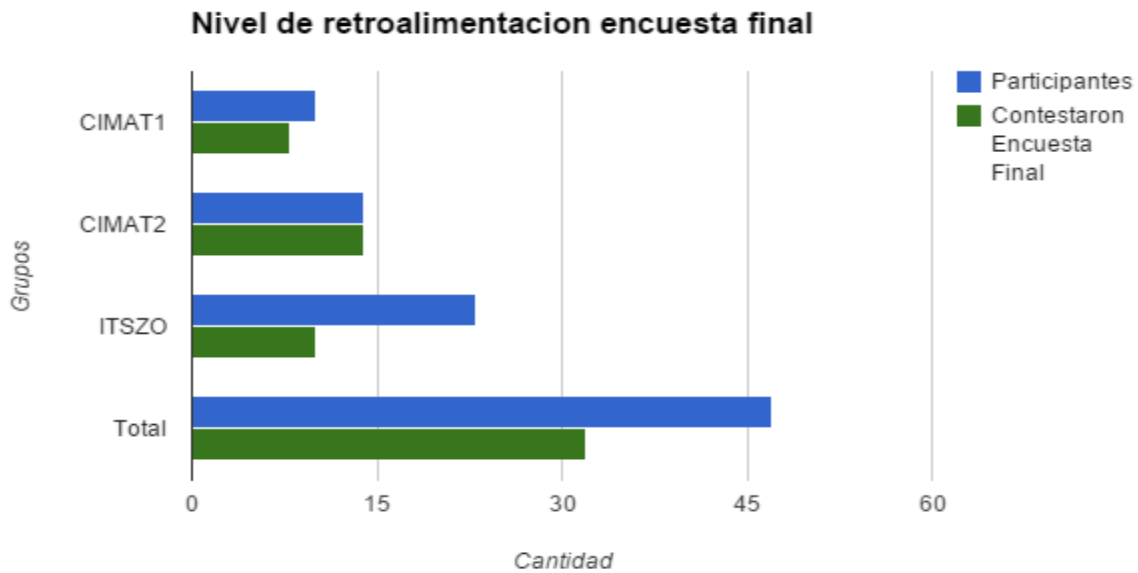


Figura 4.12. Nivel de retroalimentación encuesta final

En cuanto al agrado del curso se obtuvo un 100% de aceptación como lo muestra la figura 4.13.

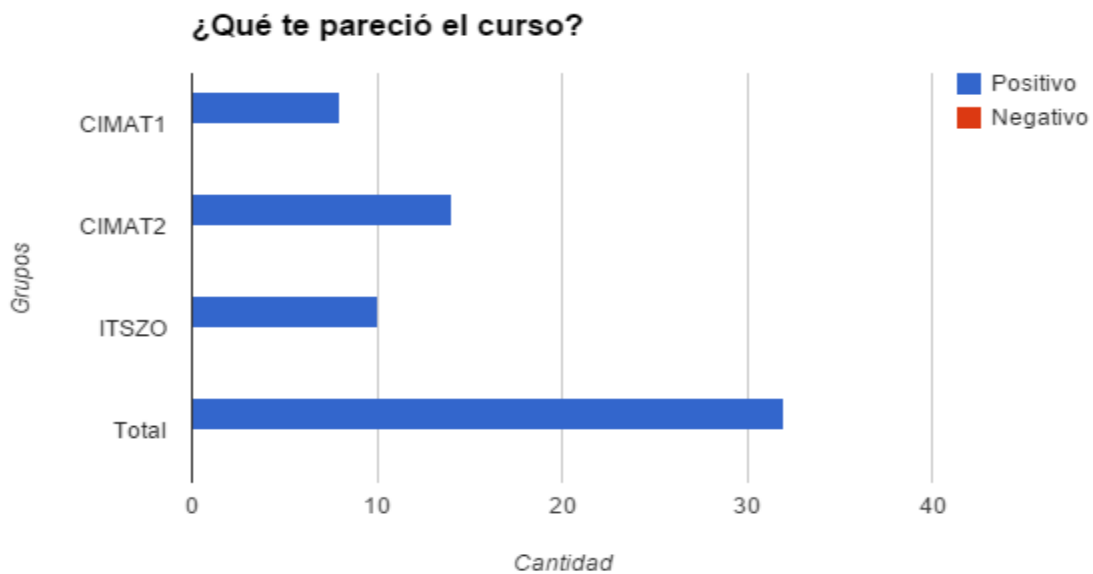


Figura 4.13. Nivel de aceptación del curso

Las siguientes dos preguntas que se realizaron están relacionadas con la primera. El 100% de alumnos que participaron en el curso creen que es un área de oportunidad importante el convertirse en un científico de datos y el 93.75% de personas se inscribirán a un curso para convertirse en uno, como lo muestran las figuras 4.14 y 4.15.



Figura 4.14. Nivel de convencimiento del curso

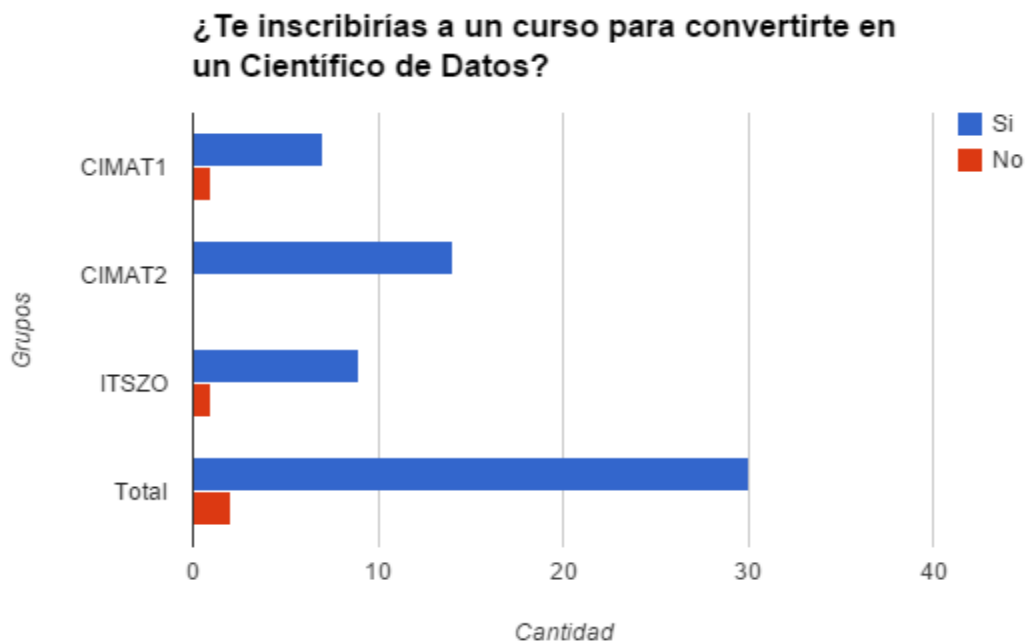


Figura 4.15. Nivel de motivación del curso

Finalmente se pidió a los alumnos identificar las habilidades base con las que cuentan para convertirse en un científico de datos. Como lo muestra la figura 4.16 solo el 21.87% cuentan con más del 50% de las habilidades base necesarias. Ello es preocupante pues esto agrega al ya largo camino otro tramo más.

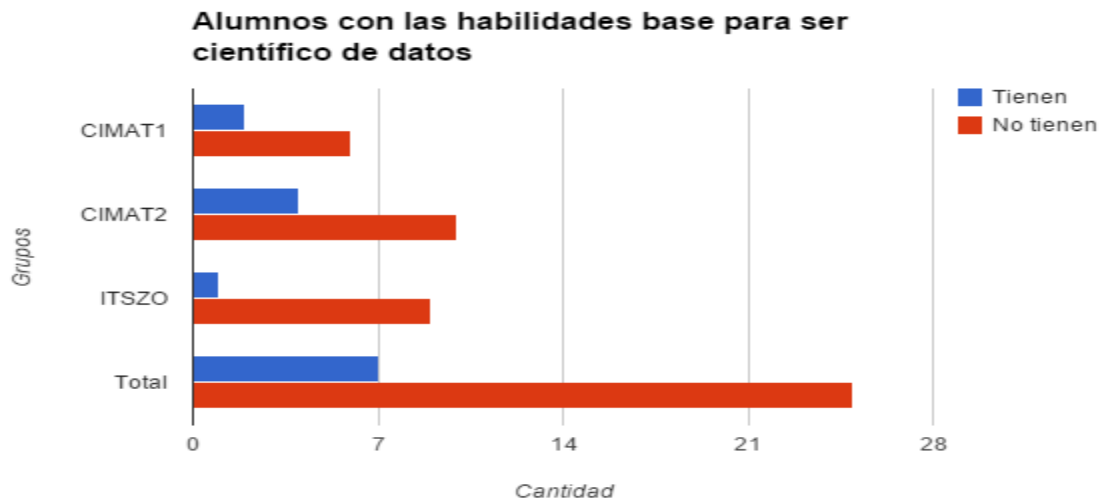


Figura 4.16. Alumnos con las habilidades base para convertirse en un Científico de Datos

En la figura 4.17 se presenta las habilidades que los alumnos afirmaron tener. Como se observa en los tres grupos la tendencia es similar. La curiosidad y las matemáticas predominan mientras que machine learning e inteligencia artificial son muy bajas.

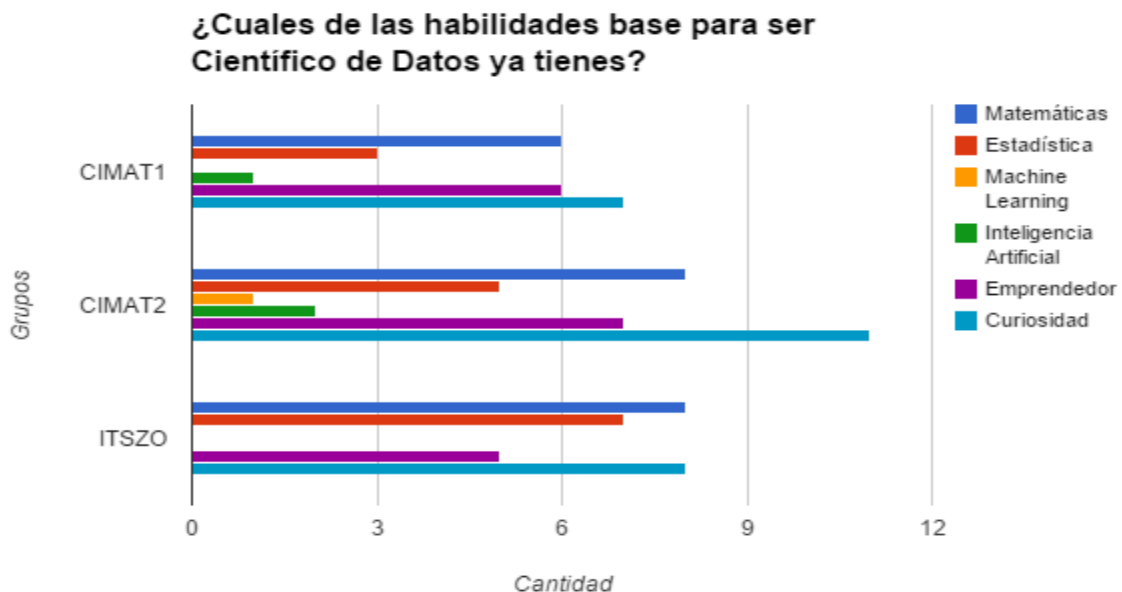


Figura 4.17. Habilidades base con las que cuentan los alumnos

En general vemos la distribución de habilidades de los alumnos en la siguiente figura 4.18.

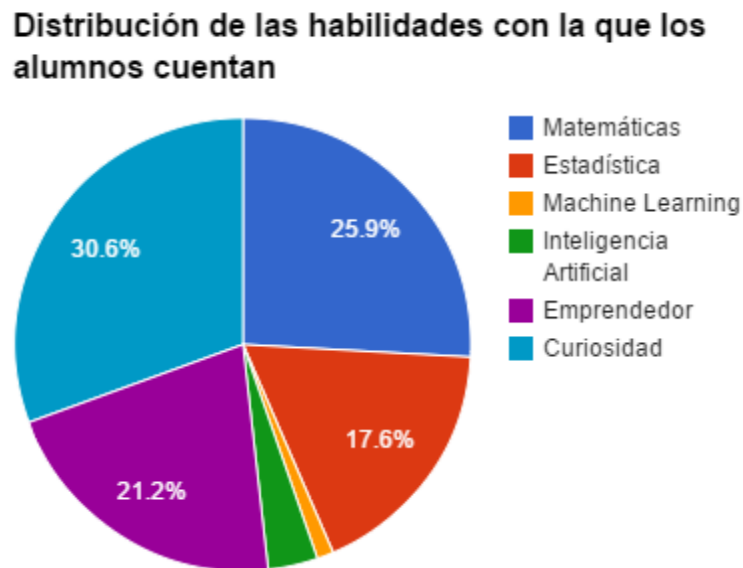


Figura 4.18. Porcentaje de las habilidades base con las que cuentan los alumnos

Resumen de resultados

La siguiente tabla muestra el resumen de resultados de la primera encuesta en la que se aprecia que el conocimiento del tema de Big Data es bajo, que es uno de los motivos de la creación del experimento.

Tabla 4.1. Resumen de resultados de la Encuesta Inicial.

Pregunta	Saben	No Saben
¿Conoces el significado del termino de Big Data?	18%	82%
¿Sabes de algun curso, certificacion, licenciatura o maestria relacionado con Big Data?	6%	94%
¿Sabes qué áreas se involucran en Big Data?	44%	56%
¿Sabes qué es un Científico de Datos?	24%	76%
¿Qué se necesita para ser un Científico de Datos?	38%	62%
¿Quieres dedicarte a Big Data?	68%	32%
¿Sabes que es NoSQL?	35%	76%
¿Sabes cuanto es el salario promedio de un Científico de Datos?	6%	94%
Menciona alguna aplicación real que conozcas relacionada con Big Data y justificala	21%	79%

En la tabla de resumen de resultado de la encuesta final se aprecia que el curso tuvo una gran aceptación y convenció a los alumnos a entrar al mundo de Big Data, de igual manera detectó que la mayoría de los alumnos no cuentan con las bases necesarias para comenzar.

Tabla 4.2. Resumen de resultados de la Encuesta Final

Pregunta	Saben	No Saben
¿Qué te pareció el curso?	100.00%	0.00%
¿Te inscribirías a un curso para convertirte en un Científico de Datos?	93.75%	6.25%
¿Cuales de las habilidades bases para ser Científico de Datos ya tienes?	21.88%	78.13%
¿Dirías que esté curso te motivo a ser un Científico de Datos?	100.00%	0.00%

Oferta Educativa

En la investigación de la oferta educativa se analizaron 95 ofertas que presentan programas relacionadas a Big Data, la metodología de la investigación fue la siguiente:

Se realizó una búsqueda masiva de oferta educativa relacionada con Big Data en universidades, empresas del giro, sitios dedicados a la impartición de cursos y la Web misma. Una vez identificada la oferta educativa se realizó un proceso manual en cual consistió en ingresar a el sitio web de cada una de las ofertas y obtener los siguientes datos:

1. Institución que imparte el curso
2. URL
3. Título
4. Modalidad. Forma en que se imparte el curso teniendo las siguientes categorías:
 - a. En Línea
 - b. Presencial
 - c. Mixto
5. Grado. Clasificación del valor curricular con los siguientes valores:
 - a. Licenciatura
 - b. Maestría
 - c. Doctorado
 - d. Certificación
 - e. Curso
6. Temario
7. País
8. Duración en meses
9. Costo en dólares
10. Idioma
11. Software utilizado en el curso
12. Lenguajes de programación utilizados en el curso

Una vez recolectada la información se obtuvieron los siguientes resultados:

En cuanto a modalidad se aprecia en las figuras 5.1 y 5.2 que predominan las impartición “Presencial” con el 57.89%, siguiendo con la modalidad de “En Línea” con el 30.53% en la cual la tendencia es la mitad del porcentaje de “Presencial” y en último lugar “Mixto” con solo el 11.58%.

Oferta educativa por modalidad

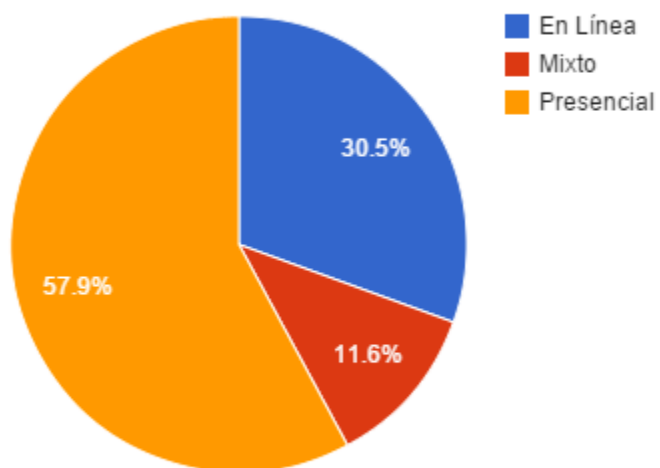


Figura 5.1. Oferta educativa por modalidad

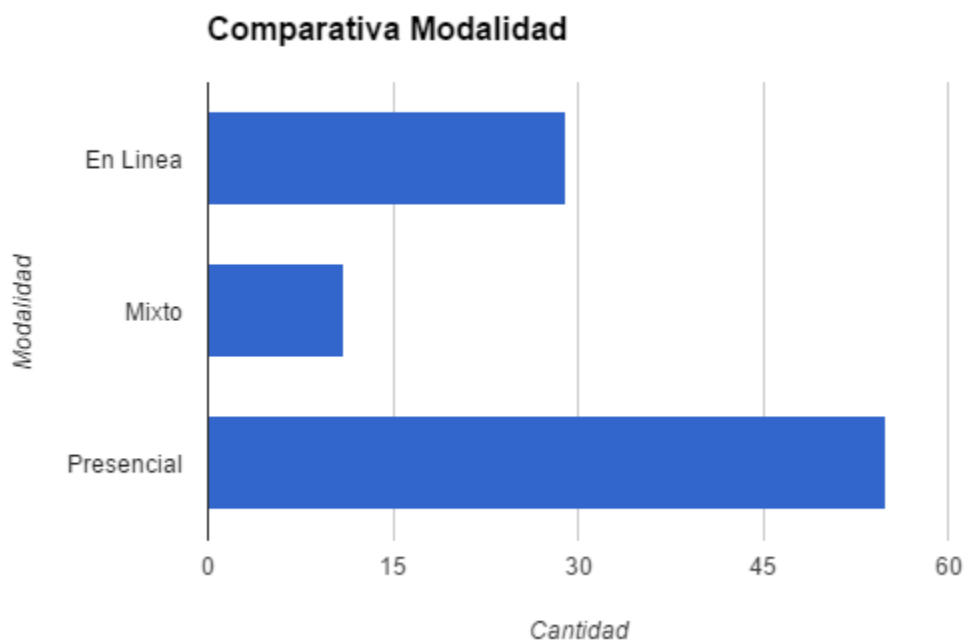


Figura 5.2. Modalidad

El grado que obtuvo la mayoría de la oferta educativa es “Maestría” con un 55.79%. En la tabla 5.1 y figura 5.3 observamos la distribución completa en cuanto al grado:

Tabla 5.1. Oferta educativa por grado académico

Grado	Cantidad	Porcentaje
Certificación	21	22.11%
Curso	5	5.26%
Doctorado	13	13.68%
Licenciatura	3	3%
Maestría	53	55.79%

Oferta educativa por grado académico

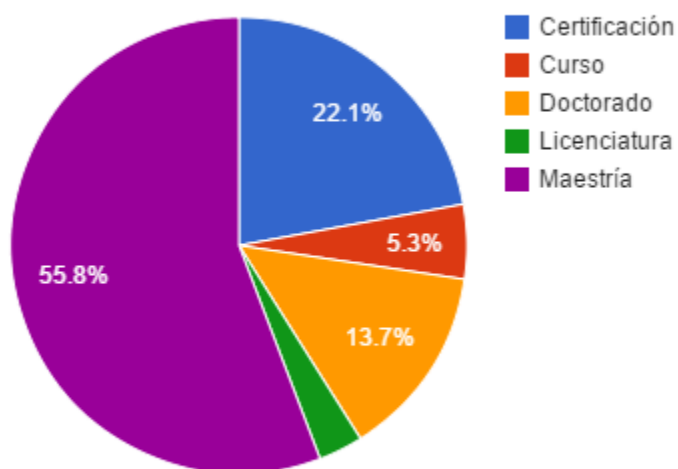


Figura 5.3. Oferta educativa por grado académico

En cuanto al temario de cada oferta educativa se realizó un proceso de homogeneización y análisis de frecuencia de materias el cual tiene como fin presentar las materias que son más frecuentemente impartidas para la formación de profesionales de Big Data. Siete de las instituciones no tienen el temario disponible, todos ellos son de nivel “Doctorado”. Esto se debe a que los temas de investigación son muy variados y no tienen un temario específico definido. Partiendo entonces de 88 cursos se tiene un total de 910 materias distintas. Para realizar el trabajo de homogeneización se realizó una exploración de los títulos de las materias lo cual llevó a la identificación de materias dedicadas a Databases o Machine Learning por ejemplo, y sus posibles variantes que tienen adicionalmente un adjetivo o complementos al título. Ejemplo de este patrón se muestra a continuación en la tabla 5.2 para el caso de **Databases**:

Tabla 5.2. Tópicos de Databases

Databases
Advanced Databases
Advanced Data Bases
Database Design
Database Management
Database Design and Administration
Database Systems Architecture
Database Systems.
Introduction to Relational Databases
Fundamentals of Database Systems
Scientific Databases

Como se aprecia en la tabla 5.2 todos los temas están relacionados se pueden englobar en el tema común que es Databases por lo tanto todo las combinaciones anteriores se homogenizaron a Databases y así para cada tema o área en común logrando una reducción de materias de 910 a 175.

```

891 Virtual and Augmented Realities
892 Visual Analytics & Applications
893 Visual Basic for Applications
894 Visual Data Mining
895 Visual Intelligence
896 Visualización de la Información
897 Visualization
898 Visualization for Analytics
899 Visualization in R with ggplot2
900 Visualization of Information CSE/Creat
901 Web Analytics
902 Web Analytics Site Optimization
903 Web and Social Media Analytics
904 Web Mining
905 Web Mining and Analytics
906 Web Services
907 Working with the Rattle Data Mining pa
908 Workshop Business Intelligence
909 Written Communication I
910 XML and Web Technologies

```

Figura 5.4. Materias Cruda Sin Duplicados.txt

```

Stochastic Processes
Strategy
System Management
Systems Desing
Systems Development
Technology
Telecommunications
Test Design
Time Series
Ubiquitous Computing
Uncertainly Modelling
Urban Sustainability
Vectors
Viability of Business Project
Visualization
Web Analytics
Web Design
Web Services
Web Technologies
Written Communication

```

Figura 5.5. Materias Homogeneizadas.txt

La figura 5.6 se muestra el top 25 del análisis de frecuencia de cada una de las materias homogeneizadas:

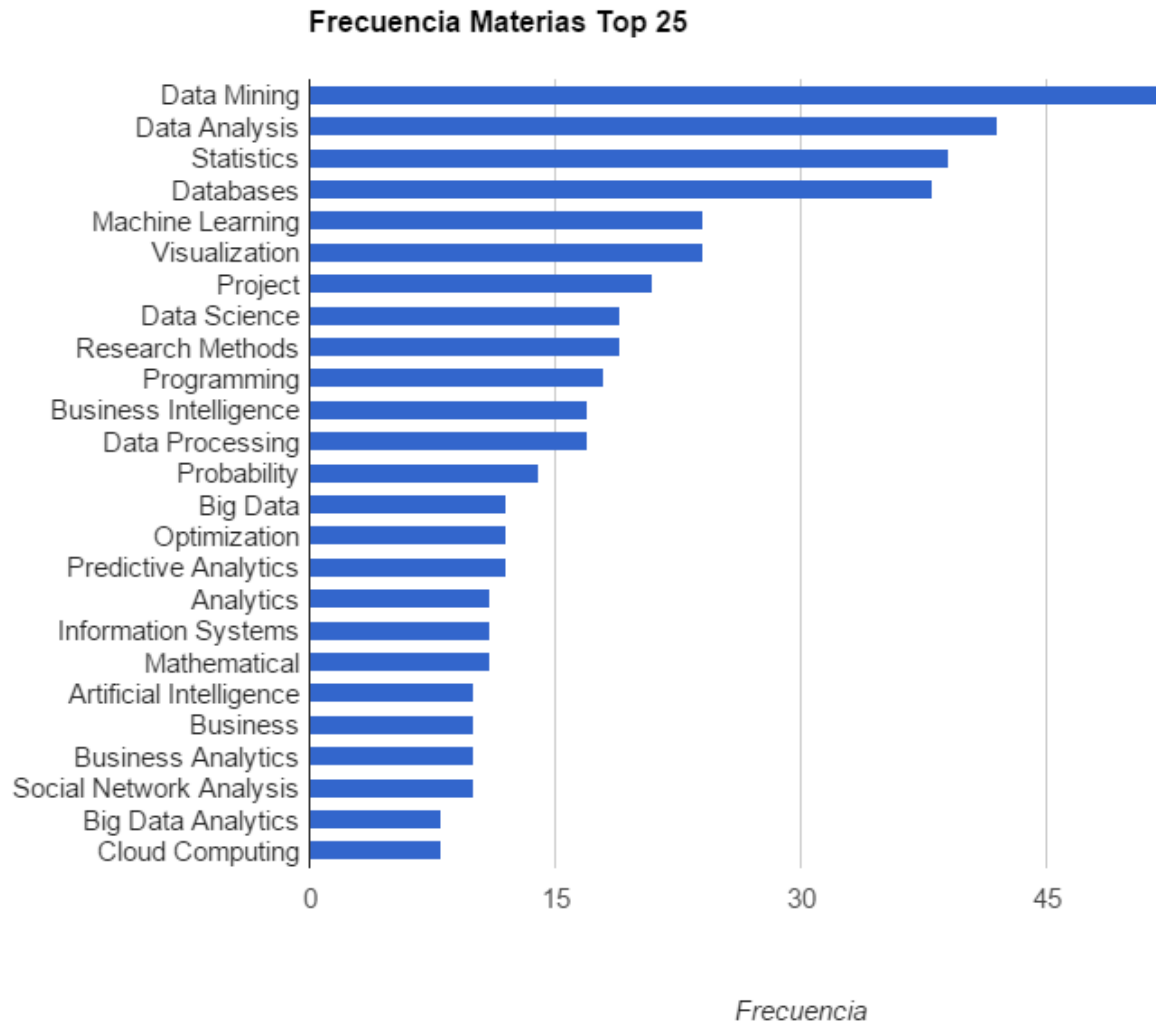


Figura 5.6. Top 25 Frecuencia de materias

Se puede encontrar la lista completa en el siguiente link <http://goo.gl/rWDx10> . Como se observa Data Mining es área más popular impartida en 55 cursos, después le sigue un grupo de 3 áreas: Data Analysis con 42, Statistics con 39 y Databases con 39. De allí en adelante la frecuencia disminuye. Con el top 25 nos damos cuenta de la importancia de cada materia o área según la oferta educativa actual.

La duración de cada programa es diferente pero se detectó la tendencia la cual consiste en que la duración de acuerdo al grado es similar por lo cual los resultados del promedio de duración se presentan por grado, para este cálculo se tomaron en cuenta 85 ofertas educativas ya que 10 no tenían la duración disponible en su sitio web. La figura 5.7 muestra el resumen del promedio por

programa:



Figura 5.7. Promedio de duración del programa por nivel de estudios

El detalle de la duración se muestra a continuación en el cual se observa la tendencia por nivel de grado de estudios.

En el grado académico **Doctorado** la duración fluctúa entre 2 y 4 años, predominando la duración de 4 años como se aprecia en la tabla 5.3.

Tabla 5.3. Duración doctorado

Grado	Duración	Cantidad
Doctorado	24	2
	36	2
	48	5
	ND	4
Doctorado Promedio		40

A nivel **Maestría** como se observa en la tabla 5.4 la duración es de los 9 meses hasta los dos años, esto en tiempo promedio pero algunas maestrías tienen la posibilidad de cursarlas en una modalidad de medio tiempo extendiendo la duración hasta 4 años. Predominando la duración de 2 años.

Tabla 5.4. Duración maestría.

Grado	Duración	Cantidad
Maestría	9	2
	11	1
	12	14
	16	1
	18	4
	20	1
	24	29
	ND	1
Maestría Promedio	19.25	

En el grado académico de **Licenciatura** se tienen pocos datos y la duración es de 4 y 5 años como se muestra en la tabla 5.5.

Tabla 5.5. Duración licenciatura

Grado	Duración	Cantidad
Licenciatura	48	1
	60	1
	ND	1
Licenciatura Promedio	54	

Las certificaciones tienen un rango muy amplio ya que entre ellas se cuentan con certificaciones específicas a un software o lenguaje como Pig que se puede tomar en un solo día. El rango de duración comprende desde 1 día a 2 años como se muestra en la tabla 5.6.

Tabla 5.6. Certificación

Grado	Duración	Cantidad
Certificación	0.04	1
	1.5	1
	6	3
	7	1
	10	1
	12	4
	15	1
	24	5
	ND	4
Certificación Promedio	12.91	

Finalmente los cursos tienen las mismas características que las certificaciones. El rango de duración comprende desde 1 día a 21 meses como se muestra en la tabla 5.7.

Tabla 5.7. Curso

Grado	Duración	Cantidad
Curso	0.04	3
	6	1
	21	1
Curso Promedio	5.42	

Como se observa en la tabla anterior los grados que presentan mayor variación de duración son los cursos y las certificaciones ya que estos se enfocan a temas específicos por otra parte el resto de los grados tienen un comportamiento similar y se podría decir que respetan un rango constante en lo que refiere a la duración.

La demanda de profesionales de Big Data es mundial pero Estados Unidos lleva la delantera en cuanto a lo que la oferta educativa se refiere como lo vemos en la siguiente tabla 5.8 y gráfica 5.8:

Tabla 5.8. Oferta educativa por país.

País	Cantidad
Alemania	1
Argentina	2
Australia	2
Austria	2
Bélgica	1
Canada	1
España	2
EUA	55
Francia	4
India	1
Internacional	7
Irlanda	3
Israel	1
Portugal	1
Reino Unido	11
Singapore	1

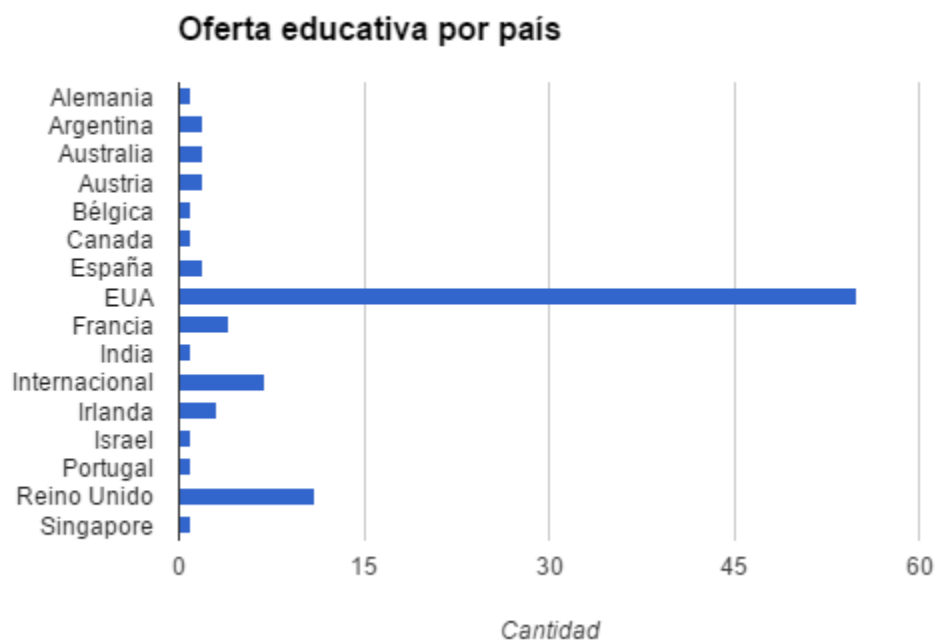


Figura 5.8. Oferta educativa por país.

Con la misma tendencia del país predominante su idioma predomina en la oferta educativa, el idioma Inglés tiene 95.8% de alternativas, existen tres alternativas que representan el 3.2% en idioma Español y finalmente con el 1.1% el Francés como se muestra en la figura 5.9:

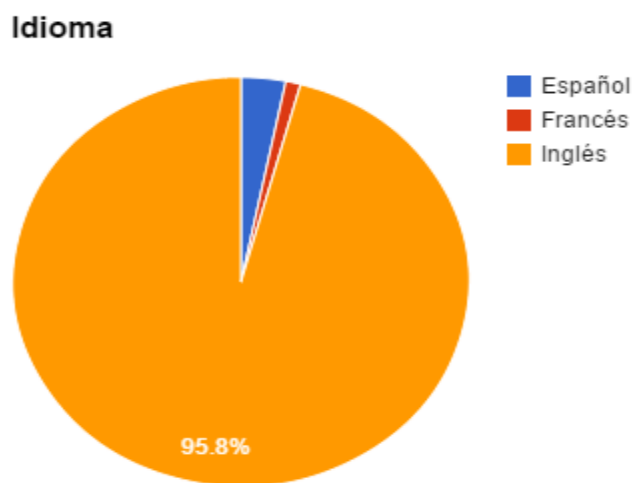


Figura 5.9. Distribución de idioma.

En cuanto al costo no existe ninguna relación en cuanto a modalidad o grado. El 58% de las instituciones no publicaban su costo es por eso que la tabla 5.9 sólo muestra los costos del 42% restante:

Tabla 5.9. Costos por grado académico e institución.

Certificación	\$0.00	Big Data University
	\$125.00	TDWI
	\$200.00	Cloudera
		Revolution Analytics
	\$600.00	EMC2 PROVEN PROFESSIONAL
	\$795.00	DIGITAL ANALYTICS ASSOCIATION
	\$2,500.00	University of Wisconsin Milwaukee, Sheldon B. Lubar School of Business
	\$3,852.00	Statistics.com, The Institute for Statistics Education
	\$7,449.00	NJIT ONLINE
	\$9,333.00	Statistics.com, The Institute for Statistics Education
	\$9,789.92	Indian School of Business ISB, Executive Education
	\$11,880.00	Standord Center for Professional Development
	\$19,800.00	Standord Center for Professional Development
Curso	\$0.00	Big Data University
	\$2,200.00	Udacity
Doctorado	\$18,965.39	Newcastle University
	\$57,000.00	Colorado Technical University
Maestría	\$2,222.00	The University of Tennessee
	\$4,967.43	Universidad Complutense de Madrid, Universidad Politécnica de Madrid
	\$7,005.35	Universidade Nova
	\$8,615.33	University of Glasgow
	\$10,121.00	Coventry University
	\$16,425.49	Goldsmiths University of London
	\$17,831.80	Barcelona Graduate School of economics
	\$18,468.65	Telecomm ParisTech
	\$19,646.17	De Montfort University Leicester
	\$20,379.20	Erasmus Mundus France(University of Pierre and Marie Curie Paris
	\$25,000.00	Thomas Edison State College
	\$25,137.44	The University of Sheffield
	\$26,320.00	Deakin University Australia
	\$26,570.64	University of Reading
	\$45,000.00	Singapore Management University
		The Chinese University of Hong Kong
	\$50,000.00	The George Washington University
	\$52,824.00	Southern Methodist University
	\$61,320.48	York University Schulich School of Business
	\$67,500.00	NYU STERN

Finalmente en relación a los lenguajes de programación o software utilizado en los cursos las instituciones no los mencionan, a excepción de cursos especializados en algún software como R, Pig o Hadoop es por eso que se determina que no existe alguna relación directa con lenguajes o softwares en la oferta educativa, esto también se corrobora en la figura 5.6 “ Top 25 Frecuencia de materias” en la cual no aparece ningún lenguaje o software como materia.

El resultado a detalle de la oferta educativa se puede consultar en el Anexo B.

Conclusiones

A continuación se presentan el alcance logrado respecto a los objetivos establecidos para la presente investigación y además el área de oportunidad.

El primer objetivo establecido consistió en **“Entender cuál es la importancia y alcance de Big Data en el mundo a nivel social y económico”** se desarrolla en el capítulo 2 Antecedentes, en el cual podemos apreciar que Big Data está aquí para quedarse por un largo tiempo teniendo una gran importancia a nivel global, no es algo que solo se aplique a áreas en específico sino que abarca o se puede aplicar a la mayoría, en el ámbito social, económico, político, etc. Por este motivo dar un ejemplo en específico en determinada área puede limitar nuestro criterio en cuanto a la amplitud y la importancia de Big Data. Valores más generales que nos ayudan a determinar la importancia son la gran cantidad de especialistas que se necesitan y necesitarán en los próximos años. Las estimaciones van desde 140,000 a 190,000 profesionales de Big Data para 2018. Otro aspecto importante es el valor del mercado de la industria de Big Data el cual generará USD \$ 76 mil millones de dólares para finales de 2020. Los datos anteriores han provocado una revolución a nivel laboral y por lo tanto en la demanda educativa. Cerraremos el primer objetivo con una frase que da respuesta al mismo. “Big Data cambiará la forma en que vemos, pensamos y vivimos”

El segundo objetivo planteado fue **“La creación de un Curso Introductorio a Big Data que motive a los asistentes a tomar la decisión de convertirse en un Profesional de Big Data dándoles a conocer en qué consiste, cuáles son algunas tareas representativas que hace y las ventajas y desventajas que se presentaran al tomar dicha decisión”** se logró mediante la creación del curso “Explorando Big Data a través de ejercicios prácticos” el cual esta descrito en el capítulo 3 Experimento, el resultado del experimento arroja dos datos contundentes: el 100% de los alumnos que tomaron el curso resultaron motivados a convertirse en un científico de datos y el 93.75% se inscribirán en un curso para lograrlo.

El tercer y último objetivo fue **“presentar las alternativas de la oferta educativa actual dirigida a formar Profesionales de Big Data”** se cumplió y se cubre en el Capítulo Oferta Educativa, en cual se analizan 95 opciones las cuales se clasifican por Modalidad, Grado, Duración, País e Idioma siendo estas dos últimas características una área de oportunidad en México ya que en nuestro país no se encontró ninguna oferta además de que sólo se encontraron 3 alternativas en idioma español , aunado de que el país que más ofrece alternativas es nuestro vecino Estados Unidos y su vez el que más va a demandar profesionales de Big Data por lo tanto nuestra ubicación geográfica tanto para instruirnos como para trabajar nos da más oportunidades que a otros países de habla hispana.

La tabla C1 muestra las tres alternativas de programas en idioma Español las cuales son maestrías y están impartidas de manera presencial:

Tabla 6.1. Alternativas de oferta educativa en Español.

Institución	url	Título	Duración (Meses)	País	Costo (\$US)
Universidad Politécnica de Madrid	http://www.mat.ucm.es/teci/wp/?page_id=85	Master Universitario en Tratamiento Estadístico Computacional de la Información	12	España	\$4,967.43
Universidad de Buenos Aires	http://triton.exp.dc.uba.ar/datamining/	Maestría en Explotación de Datos y Descubrimiento del Conocimiento	24	Argentina	ND
Universidad Central de Venezuela	http://www.matematica.ciens.ucv.ve/modulos/	Maestría en Modelos Aleatorios	24	Venezuela	ND

En el aspecto educativo **Data Mining** es el tema más recurrente en las alternativas de cursos, con lo que se determina que se sigue en búsqueda de más datos, seguido del análisis de los datos recolectados con la materia **Data Analysis**, continuando con las técnicas estadísticas para el análisis representadas por la materia **Statistics** y finalmente con la forma en que vamos a almacenar y procesar esos datos con **Databases**. Por lo tanto estas cuatro materias forman el esqueleto de los programas en los diferentes niveles.

La oferta educativa es amplia pero reciente, ha crecido y seguirá creciendo de manera explosiva. La información detallada de cada oferta se presenta en el Anexo B. En este punto se dan por cumplidos los tres objetivos de este reporte técnico de manera satisfactoria.

La importancia de Big Data en el mundo es más que clara. Agregando la gran oportunidad brindada por ubicarnos en esta zona geográfica es preocupante que en América Latina no exista oferta educativa al respecto, y no solo eso, ya que las alternativas que están por crearse como cursos en universidades que son difundidos localmente no tienen la capacidad de escalar a nivel masivo, no están diseñados para satisfacer la gran demanda que se aproxima ya que son impartidos en forma tradicional. Por otra parte, los jóvenes que no tienen el dominio del inglés están excluidos de la oferta educativa actual sin mencionar que el idioma Inglés será un factor decisivo para acceder al mercado laboral de Estados Unidos por tanto se puede marcar como requisito obligatorio para todos los profesionales de Big Data.

De la mano a la nula oferta educativa en México y Latinoamérica es notoria la falta de proactividad por parte de docentes e instituciones del sector público y privado en cuanto a la difusión de esta gran oportunidad y beneficios para acceder al mundo de Big Data siendo más profundo a nivel licenciatura en el cual se puede decir que no se tiene el conocimiento del tema. Este fue un factor determinante del éxito del curso impartido el cual informa de la oportunidad y beneficios pero a su vez de los retos mediante la exploración de Big Data por medio de ejercicios prácticos.

La creciente cantidad de datos se debe de manejar de forma responsable, aplicando normas éticas y atacando los grandes problemas sociales, políticos, ecológicos y de salud que tenemos actualmente siendo áreas de oportunidad tremendas para profesionales de Big Data. Por las razones anteriores que muestran la carencia de difusión y oferta educativa es que se propone el siguiente trabajo futuro.

Trabajo Futuro

Se proponen dos vertientes de trabajo futuro

1. Curso en línea. Interactivo, práctico y en español con las siguientes características:
 - a. El material del curso estará disponible en línea para consultarse a cualquier hora del día para que el alumno lleve su propio ritmo de aprendizaje.
 - b. Cada concepto teórico debe de ir acompañado de un ejercicio práctico.
 - c. Un ejercicio general por cada unidad, con datos de prueba generados aleatoriamente, esto para que cada alumno tenga resultados diferentes.
 - d. Entorno de desarrollo web accesible desde cualquier computadora mediante un navegador web con la finalidad que el alumno realice sus ejercicios en cualquier momento y lugar.
 - e. Métricas. El curso obtendrá métricas de los alumnos por ejemplo tiempos en cada ejercicio y tema, temas más complicados, orden del curso con lo cual el curso debe ir mejorando continuamente.
 - f. Gamificación. Lo cual consiste en motivar a los alumnos a competir entre sí, al detectar que se tiene un periodo sin entrar al curso mandarle un recordatorio motivante para que siga adelante, con capacidad de hacer grupos de alumnos para tener una supervisión por parte de un tercero del avance de sus interesados.
 - g. El entorno de producción del curso sea escalable a medida que la demanda del curso crezca.
2. Curso presencial. El cual debe de contar con las herramientas del curso en línea pero con el control del instructor con las siguientes características.
 - a. Explicación a fondo de los conceptos teóricos con una variedad más amplia de ejemplos.
 - b. Resolución de dudas a nivel personal.
 - c. Niveles de especialización. Los cuales no se pueden impartir mediante una plataforma online, con esto nos referimos a que su aprendizaje es más valioso de forma presencial.
 - i. Instalación y configuración de la infraestructura para problemas específicos por ejemplo Hadoop, OpenStack, ElasticSearch, MongoDB, etc.
 - ii. Software y lenguajes para problemas específicos. Spark, Pig, R, etc.

Referencias

- Ariker Matt, M. T. (2013). Five Roles You Need on Your Big Data Team. Retrieved October 16, 2014, from <http://blogs.hbr.org/2013/07/five-roles-you-need-on-your-bi/>
- Beyer, M. (2012). *Gartner Says Solving “Big Data” Challenge Involves More Than Just Managing Volumes of Data*. Gartner. Retrieved from <http://www.gartner.com/it/page.jsp?id=1731916>
- BigData-Startups. (2013). The Skill Sets Required For Different Big Data Jobs. Retrieved October 15, 2014, from <http://www.bigdata-startups.com/big-data-job-descriptions/>
- Burtch, L. (2013). *Salaries for Big Data Professionals*.
- Chandrasekaran, S. (2013). The Long Road To Become a Big Data Scientist - Infographic. Retrieved October 16, 2014, from <http://www.bigdata-startups.com/BigData-startup/long-road-big-data-scientist-infographic/>
- Collins, W. (2014). Big Data Market Size - 2018. Retrieved October 16, 2014, from <https://www.linkedin.com/today/post/article/20140709114739-340063519-big-data-market-size-2018>
- Cuesta, H. (2014). *Practical Data Analysis* (p. 360).
- Cukier, K. (2014). Big data is better data. Retrieved October 15, 2014, from http://www.ted.com/talks/kenneth_cukier_big_data_is_better_data
- DataStax. (2014). Big Data Challenges. Retrieved October 19, 2014, from <http://www.datastax.com/big-data-challenges>
- Dean, J., & Ghemawat, S. (2004). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 1–13. Retrieved from <http://dl.acm.org/citation.cfm?id=1327492>
- Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. *ACM SIGOPS Operating Systems Review*, 37(5), 29. doi:10.1145/1165389.945450
- Google. (2014). Google Annual Search Statistics | Statistic Brain. Retrieved October 15, 2014, from <http://www.statisticbrain.com/google-searches/>
- Groenfeldt, T. (2013). Data Scientists -- Don't Wait For Universities, Grow Your Own - Forbes. Retrieved February 02, 2015, from <http://www.forbes.com/sites/tomgroenfeldt/2013/06/21/data-scientists-dont-wait-for-universities-grow-your-own/>
- Harvard. (2014). IQSS. Retrieved October 18, 2014, from <http://www.iq.harvard.edu/>
- IBM. (2014). *IBM What is big data? — Bringing big data to the enterprise*. www.ibm.com. Retrieved from <http://www.ibm.com/big-data/us/en/>

- indeed. (2014). Job Search | one search. all jobs. Indeed.com. Retrieved October 16, 2014, from <http://www.indeed.com/>
- Kelly, J. (2014). Big Data Vendor Revenue And Market Forecast 2013-2017. Retrieved October 16, 2014, from http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017
- kissmetrics. (2014). Facebook Statistics. Retrieved October 15, 2014, from <https://blog.kissmetrics.com/facebook-statistics/>
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung, A. (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved February 02, 2015, from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
- McGaughey, K. (2011). Worl' Data More Than Doubling Every Two Years. Retrieved October 18, 2014, from <http://www.emc.com/about/news/press/2011/20110628-01.htm>
- Microsoft. (2012). The Big Bang: How the Big Data Explosion Is Changing the World | News Center. Retrieved October 15, 2014, from <http://news.microsoft.com/2013/02/11/the-big-bang-how-the-big-data-explosion-is-changing-the-world/>
- Piatetsky, G. (2014). Data Scientists Salary Survey: US, Canada, Australia lead. Retrieved October 16, 2014, from <http://www.kdnuggets.com/2014/03/data-scientists-salary-survey-us-canada-australia-lead.html>
- Rijmenam, V. M. (2013a). Self-driving Cars Will Create 2 Petabytes of Data a Year. Retrieved October 18, 2014, from <http://www.bigdata-startups.com/self-driving-cars-create-2-petabytes-data-annually-big-data-opportunities-automotive-industry/>
- Rijmenam, V. M. (2013b). What Does It Take To Become a Big Data Engineer. Retrieved October 16, 2014, from <http://www.bigdata-startups.com/job-description-big-data-engineer/>
- Rijmenam, V. M. (2013c). What Does It Take To Become a Big Data Scientist. Retrieved October 16, 2014, from <http://www.bigdata-startups.com/job-description-big-data-scientist/>
- Samuel, A. (1959). Machine Learning. Retrieved from http://en.wikipedia.org/wiki/Machine_learning
- SAS Institute Inc. (2013). Big Data Meets Big Data Analytics.
- Shaw, J. (2014). Understanding big data leads to insights, efficiencies, and saved lives. Retrieved October 18, 2014, from <http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal>
- Sicular, S. (2013). Gartner's Big Data Definition. Retrieved October 16, 2014, from <http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/>

Viktor Mayer-Schönberger, K. C. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (p. 256). Retrieved from <http://www.amazon.com/Big-Data-Revolution-Transform-Think/dp/0544002695>

Vorhies, W. (2014). How to Become a Data Scientist. Retrieved February 04, 2015, from <http://www.datasciencecentral.com/profiles/blog/show?id=6448529%3ABlogPost%3A199501>

Watch, M. (2014). The Big Data Market: 2014 - 2020 - Opportunities, Challenges, Strategies, Industry Verticals and Forecasts - MarketWatch. Retrieved October 16, 2014, from <http://www.marketwatch.com/story/the-big-data-market-2014-2020-opportunities-challenges-strategies-industry-verticals-and-forecasts-2014-07-17>

Wikipedia. (2009). Big data.

Wikipedia. (2014). Petabyte.

YouTube. (2014). Estadísticas: YouTube. Retrieved October 15, 2014, from <https://www.youtube.com/yt/press/es-419/statistics.html>

Anexo A: Explorando Big Data a través de ejercicios prácticos

El objetivo de este tutorial es presentar los conceptos básicos del proceso de análisis de datos mediante tres ejemplo prácticos que son:

- Twitter. Análisis de Sentimiento. “En este ejercicio se clasifican los tweets del tema de interés de los participantes en positivos o negativos, esto para entender el sentimiento de los tuiteros acerca del tema”.
- Facebook. Análisis del grafo de amigos en Facebook. “En este ejercicio se aplicarán métodos matemáticos para obtener grado de distribución, centralidad de el grupo de amigos del participante”.
- Predicción del precio del Oro. “En este ejercicio se realiza la predicción del precio del oro basados en valores históricos aplicando métodos matemáticos y estadísticos como por ejemplo regresiones no lineales”.

Requerimientos

Para poder realizar esta práctica se necesita lo siguiente:

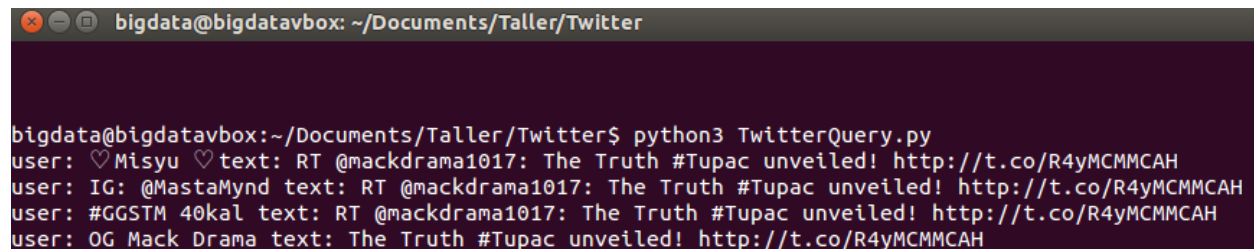
- Laptop con al menos 4 gb de ram.
- Contar con VirtualBox instalado.
- Contar con conocimientos básicos de programación. Python de preferencia.
- Conexión a internet.

Se proporciona la máquina virtual genérica de nombre CIMPS.ova cuyo password es “cimps2014” que ya está configurada y lista para correr los tres ejercicios, de igual manera se proporciona el código fuente de los tres ejercicios. Se puede descargar los archivos de la siguiente dirección <http://goo.gl/qr4OdV>.

De forma alternativa a la máquina virtual al final de este tutorial en la sección de instalación de requerimientos se presenta una guía del procedimiento para instalar las librerías necesarias en el SO Ubuntu.

Nota

Todo el código generado en las máquinas virtuales deberá correrse con Python 3 como se muestra a continuación:



```
bigdata@bigdatavbox: ~/Documents/Taller/Twitter
bigdata@bigdatavbox:~/Documents/Taller/Twitter$ python3 TwitterQuery.py
user: ♥ Misyu ♥ text: RT @mackdrama1017: The Truth #Tupac unveiled! http://t.co/R4yMCMMAH
user: IG: @MastaMynd text: RT @mackdrama1017: The Truth #Tupac unveiled! http://t.co/R4yMCMMAH
user: #GGSTM 40ka1 text: RT @mackdrama1017: The Truth #Tupac unveiled! http://t.co/R4yMCMMAH
user: OG Mack Drama text: The Truth #Tupac unveiled! http://t.co/R4yMCMMAH
```

Figura A1 Ejecutar script con Python 3

Ejercicio 1. Análisis de sentimientos en twitter

En este ejercicio se utilizan librerías como Twython y NLTK y se utiliza el API de twitter para saber si los tweets que hagan referencia a una cuenta de twitter o un hashtag o trending topic son positivos o negativos. Todo se basa en naturallanguage.org.

Anatomía de los datos de twitter

Como sabemos twitter es un micro blog para compartir mensajes de hasta 140 caracteres (tweet). De twitter podemos obtener una gran variedad de datos como los mismos tweets, seguidores, mensajes directos y tendencias en temas.

Sin embargo, podemos obtener más información que el simple texto del mensaje como día y hora, links, mención de los usuarios (@), hashtags (#), número de retweets, lenguaje, numero de favoritos y geocode.

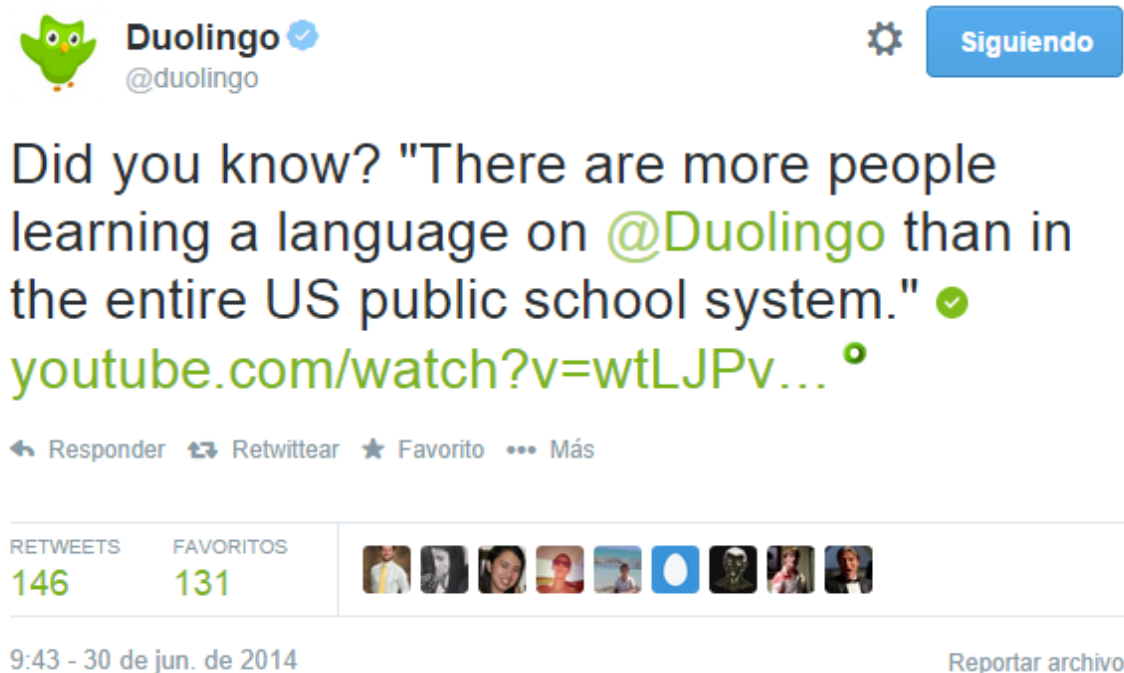


Figura A2 Tweet

Seguidores Followers

Los usuario en twitter pueden seguir a otros usuarios con una grafo directo². Las relaciones no son mutuas. Con esto por ejemplo podríamos calcular quién es el individuo con más influencia.

Tendencias Trends

Las tendencias son palabras o hashtags con alta popularidad entre usuarios en un momento o lugar específico. Con estos datos se podríamos detectar y predecir futuras tendencias. Las tendencias que se muestran a los usuarios están basadas en su locación y personas de las que son seguidores.

Twitter API

El API de twitter utiliza OAuth que es un estándar abierto para autorización de acceso.

Ahora se procederá a obtener cuenta para usar la API de Twitter, dirigiéndonos a el siguiente link <https://apps.twitter.com/> el cual presentara una página como la siguiente figura:

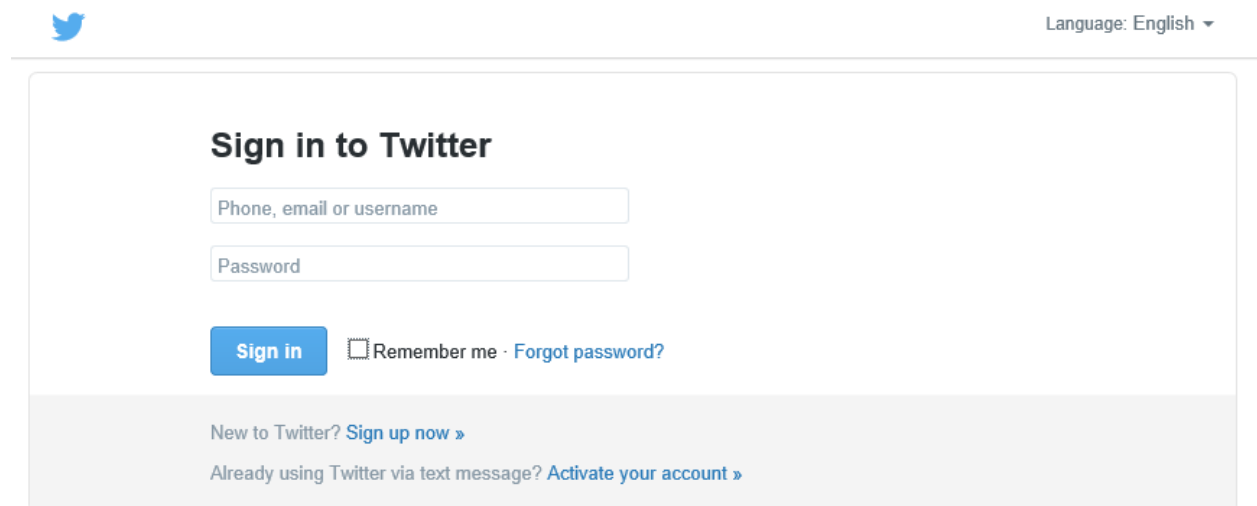
² Grafo Directo. En la relación entre nodos que va en una sola dirección.

Twitter Apps

Please [sign in](#) with your Twitter Account to create and maintain Twitter Apps.

Figura A3 Administrador de aplicaciones de Twitter

Accedemos con nuestra cuenta de twitter, si no tenemos creamos una.



The image shows the Twitter sign-in page. At the top left is the Twitter logo, and at the top right is the language selector "Language: English". The main content area is titled "Sign in to Twitter". It contains two input fields: "Phone, email or username" and "Password". Below these fields is a blue "Sign in" button, a checkbox for "Remember me", and a link for "Forgot password?". At the bottom of the form, there are two links: "New to Twitter? Sign up now »" and "Already using Twitter via text message? Activate your account »".

Figura A4 Inicio de sesión de Twitter



Creamos una nueva aplicación.



The image shows the Twitter Apps creation page. At the top left is the Twitter logo and the text "Application Management". At the top right is a small profile picture icon. The main content area is titled "Twitter Apps". On the right side of the page, there is a button labeled "Create New App".

Figura A5 Creación de nueva aplicación de Twitter

Llenamos los campos necesarios, el nombre de la aplicación es único por lo tanto algunos nombres están ya usados por otras aplicaciones.

 Application Management

Create an application

Application details

Name *

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

Figura A6 Datos de nueva aplicación de Twitter

Aceptamos las reglas del desarrollador.

Developer Rules of the Road

3. Respect user privacy

Your Service must display and comply with a privacy policy that is presented before download, installation or sign up (as applicable) that clearly discloses what you are doing with information you collect from users. If your Service supports cookies, your privacy policy must disclose that third parties may be placing and reading cookies on the systems of your users in the course of providing content to them. Your privacy policy should also provide information about user options for cookie management and the [Do Not Track](#) setting in supporting web browsers.

Clearly disclose when you are adding location information to a user's Tweets, whether as a geotag or annotations data. Be clear about whether you are adding a place or specific coordinates. If your application allows users to Tweet with their location be sure that it complies with the best practices found [here](#).

You should not solicit another developer's consumer keys or consumer secrets especially if they will be stored or used for actions outside of that developer's control. Keys and secrets that are compromised will be reset by Twitter. For example, online services that ask for these values in order to provide a "tweet-branding" service are not allowed.

Do not facilitate or encourage the publishing of private or confidential information.

Do not store Twitter passwords.

Do not store non-public Twitter Content except at the explicit direction of a Twitter end user.

As a reminder, Twitter's services are not directed to persons under 13. If you operate a commercial website or online service that is

☒ Yes, I agree

Create your Twitter application

Figura A7 Reglas del desarrollador

Y obtenemos el siguiente resultado.

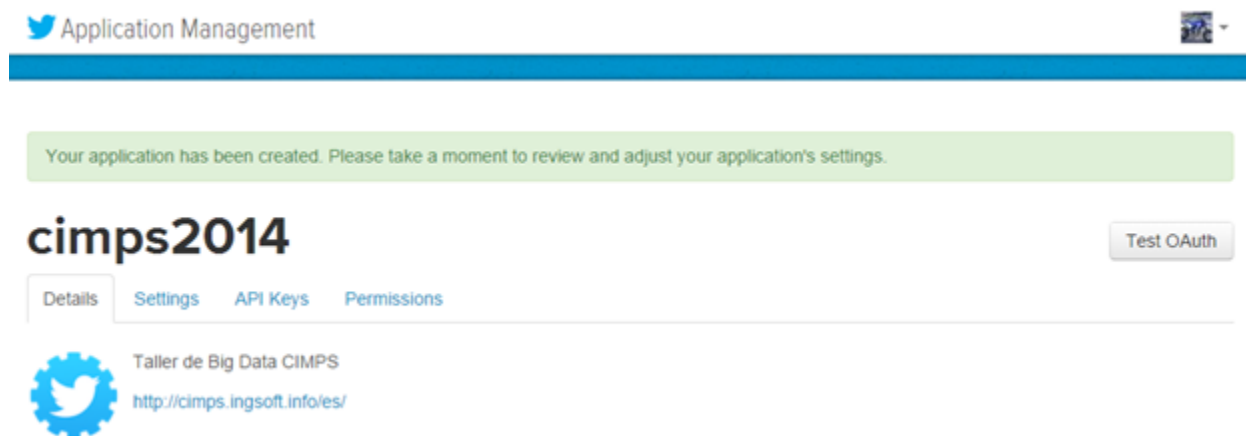


Figura A8 Aplicación de Twitter creada

Ahora solicitaremos los accesos a la aplicación. Comenzaremos en la pestaña de API Keys y solicitaremos el token de acceso.

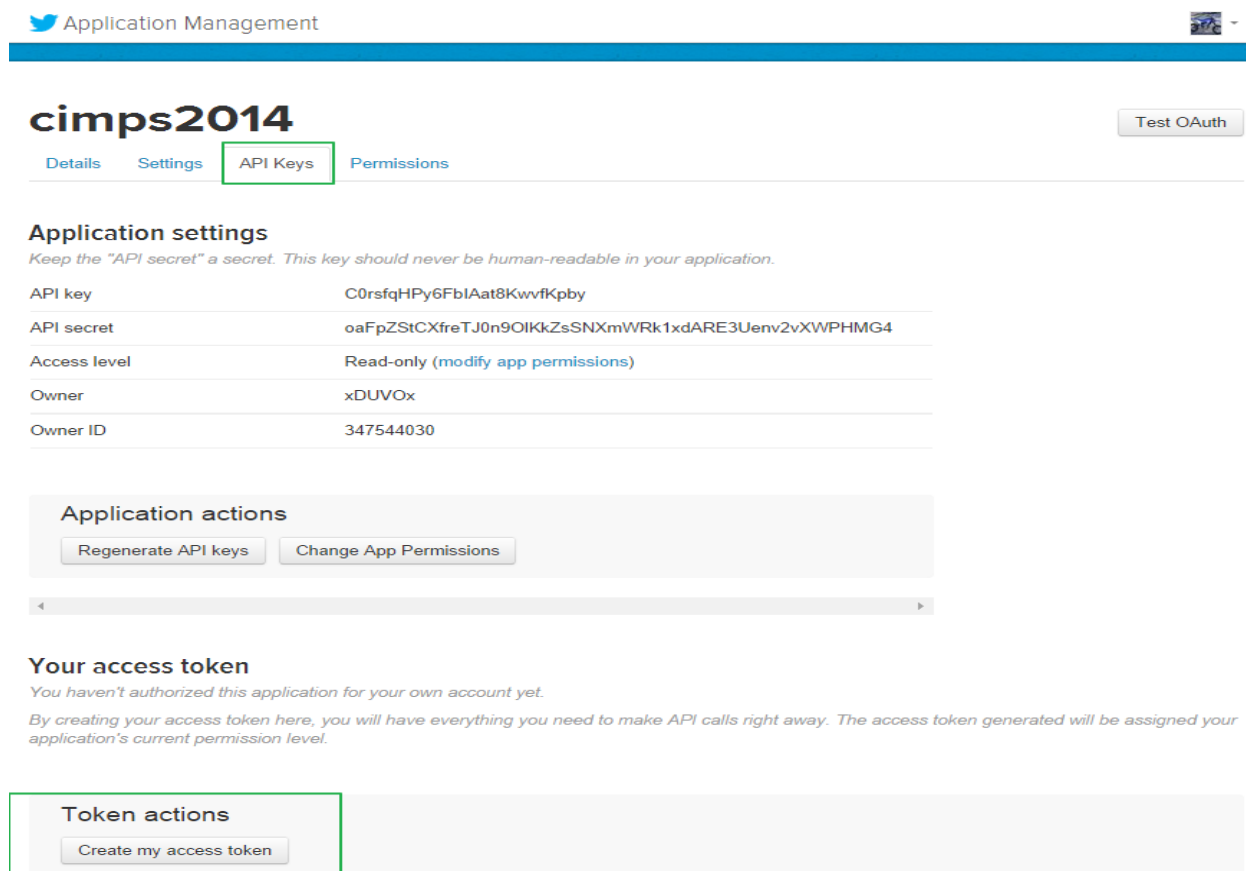


Figura A9 Solicitud de acceso

Obteniendo el siguiente resultado

cimps2014

DetailsSettingsAPI KeysPermissions

Application settings

Keep the "API secret" a secret. This key should never be human-readable in your application.

API key	C0rsfqHPy6FblAat8KwvfKpby
API secret	oaFpZStCXfreTJ0n9OIKkZsSNXmWRk1xdARE3Uenv2vXWPHMG4
Access level	Read-only (modify app permissions)
Owner	xDUVOx
Owner ID	347544030

Application actions

Regenerate API keysChange App Permissions

Your access token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access token	347544030-DNuWXctF78ICleMAMThRsdlwBHANWFveymlwPSVB
Access token secret	J7zIY1Eng7LkOVDNJce66QDoExcl8YMjYz174RpO4c1Jm
Access level	Read-only
Owner	xDUVOx
Owner ID	347544030

Figura A10 Accesos a la aplicación

Los datos que necesitamos para usar el api de twitter son los siguientes: **API key, API secret, Access token, and Access token secret.**

Debemos de tener en cuenta que tenemos un límite de 180 consultas cada 15 minutos. Manos a la obra, comenzaremos realizando algunos ejemplos para relacionarnos con la API de Twitter.

En este taller usaremos Twython una librería de python para conectarnos con twitter. Existe una variedad de librerías para usar la API de Twitter, aquí encontraremos la lista oficial. <https://dev.twitter.com/overview/api/twitter-libraries>

Realizaremos una búsqueda simple en la cual imprimimos tal cual el resultado de los tweets del resultado.

```
1 from twython import Twython
2
3 ConsumerKey = "XXXXXXXXXXXX6wScmptzCPXBjM"
4 ConsumerSecret = "XXXXXXXXXXXXi5eTX0018RuSakVJ2gR4g4D28D09OQ0"
5 AccessToken = "XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXZPX7DMv"
6 AccessTokenSecret = "XXXXXXXXXXXX575ikgr0UITSvXXXXXXXXXXXXjr_jynouU"
7
8 twitter = Twython(ConsumerKey,
9                   ConsumerSecret,
10                  AccessToken,
11                  AccessTokenSecret)
12
13
14 result = twitter.search(q="cimps")
15
16 for status in result["statuses"]:
17     print(statuses)
```

Twython convierte el JSON enviado por Twitter a un objeto nativo de python.

```
{'id': 513742841374191616, 'metadata': {'iso_language_code': 'es', 'result_type': 'recent'}, 'in_reply_to_user_id': None, 'coordinates': None, 'truncated': False, 'contributors': None, 'id_str': '513742841374191616', 'in_reply_to_screen_name': None, 'in_reply_to_status_id': None, 'favorited': False, 'created_at': 'Sun Sep 21 17:33:32 +0000 2014', 'text': 'Quedan solo 9 días para el CIMPS2014, Información y registro en: http://t.co/xOziwOuWB7 http://t.co/i5XOkU4A46', 'retweet_count': 0, 'place': None, 'user': {'profile_sidebar_fill_color': 'F3F3F3', 'protected': False, 'id': 234809116, 'time_zone': 'Mexico City', 'name': 'Cimat Zacatecas', 'utc_offset': -18000, 'lang': 'es', 'profile_sidebar_border_color': 'DFDFDF', 'follow_request_sent': False, 'id_str': '234809116', 'listed_count': 4, 'created_at': 'Thu Jan 06 16:35:34 +0000 2011', 'profile_background_color': 'EBEBEB', 'statuses_count': 117, 'profile_use_background_image': True, 'location': 'Zacatecas', 'profile_background_image_url': 'http://pbs.twimg.com/profile_background_images/353351399/logo_fondo_cimat.gif', 'verified': False, 'profile_background_tile': True, 'default_profile': False, 'entities': {'url': {'urls': [{'expanded_url': None, 'url': 'http://ingsoft.mx'}]}, 'indices': [0, 17]}}, 'description': {'urls': []}, 'friends_count': 36, 'default_profile_image': False, 'favourites_count': 0, 'profile_image_url': 'http://pbs.twimg.com/profile_images/1310438581/IngSoft-PNG_v0.1-FINAL_normal.png', 'profile_link_color': '990000', 'notifications': False, 'followers_count': 272, 'contributors_enabled': False, 'screen_name': 'cimat_zacatecas', 'is_translation_enabled': False, 'profile_background_image_url_https': 'https://pbs.twimg.com/profile_background_images/353351399/logo_fondo_cimat.gif', 'following':
```

True, 'profile_text_color': '333333', 'description': 'La Unidad CIMAT Zacatecas se dedica a la generación, transmisión y aplicación de conocimientos en el área de TICs. [biografía completa en <http://ingsoft.mx/>]', 'geo_enabled': False, 'is_translator': False, 'url': 'http://ingsoft.mx', 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/1310438581/IngSoft-PNG_v0.1-FINAL_normal.png', 'possibly_sensitive': False, 'retweeted': False, 'lang': 'es', 'entities': {'hashtags': [], 'symbols': [], 'urls': [{'indices': [65, 87], 'expanded_url': 'http://cimps.ingsoft.info/', 'url': 'http://t.co/xOziwOuWB7', 'display_url': 'cimps.ingsoft.info'}, {'indices': [88, 110], 'expanded_url': 'http://fb.me/1BhAehC0Z', 'url': 'http://t.co/i5XOkU4A46', 'display_url': 'fb.me/1BhAehC0Z'}], 'user_mentions': []}, 'in_reply_to_user_id_str': None, 'favorite_count': 0, 'in_reply_to_status_id_str': None, 'source': 'Facebook', 'geo': None}

Podemos restringir la salida navegando por la estructura del resultado en JSON

```
for status in result["statuses"]:
    #print(statuses)
    print("user: {0} text: {1}".format(status["user"]["name"], status["text"]))
```

Código. TwitterQuery.py

La salida se verá de la siguiente forma:

user: Cimat Zacatecas text: Quedan solo 9 días para el CIMPS2014, Información y registro en: <http://t.co/xOziwOuWB7> <http://t.co/i5XOkU4A46>

user: Temo Lemus text: RT @ConferenceCIMPS: Invitamos a las comunidades, académica, gobierno y negocios a participar en #CIMPS'14. Visita <http://t.co/yRs2Bgb0BQ> h...

user: Conference CIMPS text: Divulgación del congreso CIMPS 2014 en Universidad Politécnica de Zacatecas UPZ. <http://t.co/snDmmTvMoe>

user: Conference CIMPS text: Generación de conocimiento a nivel internacional gracias a los trabajos enviados al congreso CIMPS. <http://t.co/nPOLEktusx>

user: Conference CIMPS text: Entrevista realizada por CNN a Deepak Daswani, él estará presente como expositor en CIMPS 2014. <http://t.co/PVTcHRHi7C>

user: Fercho..!! text: RT @ConferenceCIMPS: Invitamos a las comunidades, académica, gobierno y negocios a participar en #CIMPS'14. Visita <http://t.co/yRs2Bgb0BQ> h...

user: Sayoko McKinney text: CIMPS <http://t.co/VUCjBwWN1x>

user: CIMAT text: RT @isragaytan: Estare dando una conferencia de #BigData en el CIMAT el día 2 de Octubre más info en <http://t.co/NvK1UyddtK>

Navegar por resultado en JSON nos ayuda a solo obtener la información que deseamos.

Línea de tiempo (TimeLine)

Ahora obtendremos nuestro timeLine, o la de otros usuarios con la función `get_user_timeline`.


```
time = twitter.get_user_timeline(screen_name = "duolingd", count = 15)
for tweet in time:
    print(" User: {0} \n Created: {1} \n Text: {2} "
          .format(tweet["user"]["name"],
                  tweet["created_at"],
                  tweet["text"]))
```

Código. LineaDeTiempo.py

El resultado será como se muestra a continuación:

User: Duolingo

Created: Mon Sep 22 00:10:32 +0000 2014

Text: @MuseWhy2 Ooo! Please apply here: <http://t.co/hMJp0stTka>. Thank you!!

User: Duolingo

Created: Fri Sep 19 00:02:42 +0000 2014

Text: RT @shitduosays: Paul drinks wine before the cat. - Ólann Pól fíon roimh an gcat.

User: Duolingo

Created: Thu Sep 18 23:36:39 +0000 2014

Text: @damacri86 ¿Cuál es tu nombre de usuario?

User: Duolingo

Created: Thu Sep 18 19:43:24 +0000 2014

Text: @ladyinblue24 Danish for English speakers is in beta on the web :)

User: Duolingo

Created: Thu Sep 18 16:44:11 +0000 2014

Text: Video: @Luisvonahn on Making Language Lessons Available to Everyone and Becoming Your New Digital Tutor <http://t.co/RgoY5GJz1l> -@bigthink

Seguidores (Followers)

En este ejemplo analizaremos cómo obtener la lista de seguidores de un usuario en específico. Usaremos el método `get_followers_list` usando el `screen_name` o el `user_id`

```
followers = twitter.get_followers_list(screen_name="xduvox")
for follower in followers["users"]:
    print(" {0} \n ".format(follower))
```

Código. SeguidoresPorNombre.py

Cada usuario se seguidor de la siguiente manera:

```
{'profile_text_color': '333333', 'muting': False, 'profile_use_background_image': True,
'profile_sidebar_fill_color': 'DDEEF6', 'id_str': '38707297', 'entities': {'url': {'urls': [{'indices': [0, 22], 'url':
'http://t.co/gzsXPxDtxH', 'expanded_url': 'http://novador.blogspot.com', 'display_url':
'novador.blogspot.com'}]}}, 'description': {'urls': []}}, 'protected': False, 'is_translation_enabled': False,
'notifications': False, 'created_at': 'Fri May 08 17:36:39 +0000 2009', 'profile_link_color': '0084B4',
'is_translator': False, 'name': 'Pepe Hernández', 'follow_request_sent': False, 'followers_count': 385,
'geo_enabled': True, 'following': False, 'statuses_count': 3237, 'profile_sidebar_border_color': 'CODEED',
'default_profile': True, 'contributors_enabled': False, 'verified': False, 'status': {'in_reply_to_user_id':
None, 'id_str': '513694900810825728', 'entities': {'user_mentions': [], 'hashtags': [{'indices': [104, 113],
'text': 'nikeplus'}]}, 'media': [{'sizes': {'small': {'w': 340, 'resize': 'fit', 'h': 340}, 'large': {'w': 640, 'resize': 'fit',
'h': 640}, 'medium': {'w': 600, 'resize': 'fit', 'h': 600}, 'thumb': {'w': 150, 'resize': 'crop', 'h': 150}},
'media_url': 'http://pbs.twimg.com/media/ByECyjiCIAEKRag.png', 'display_url':
'pic.twitter.com/gKrwYkGXsZ', 'id_str': '513694900051648513', 'id': 513694900051648513, 'indices':
[114, 136], 'url': 'http://t.co/gKrwYkGXsZ', 'expanded_url':
'http://twitter.com/pphdez76/status/513694900810825728/photo/1', 'media_url_https':
'https://pbs.twimg.com/media/ByECyjiCIAEKRag.png', 'type': 'photo'}]}, 'urls': [{'indices': [81, 103], 'url':
'http://t.co/Ht8s44Byjw', 'expanded_url': 'http://go.nike.com/0882oerq', 'display_url':
'go.nike.com/0882oerq'}]}, 'symbols': [], 'possibly_sensitive': False, 'in_reply_to_status_id_str': None,
'contributors': None, 'in_reply_to_status_id': None, 'favorited': False, 'in_reply_to_screen_name': None,
'coordinates': None, 'text': 'Para aguantarle el ritmo a los chiquillos... Acabo de correr 12.0 km con Nike+.
http://t.co/Ht8s44Byjw #nikeplus http://t.co/gKrwYkGXsZ', 'created_at': 'Sun Sep 21 14:23:03 +0000
2014', 'geo': None, 'id': 513694900810825728, 'lang': 'es', 'truncated': False, 'retweeted': False, 'place':
None, 'favorite_count': 0, 'retweet_count': 0, 'source': '<a href="http://www.apple.com"
rel="nofollow">iOS</a>', 'in_reply_to_user_id_str': None}, 'profile_background_image_url_https':
'https://abs.twimg.com/images/themes/theme1/bg.png', 'location': 'Zacatecas, México', 'friends_count':
505, 'listed_count': 11, 'id': 38707297, 'default_profile_image': False, 'description': 'Aficionado a correr.
Hogareño. Mis hobbies: la educación, la gestión de proyectos de TICs y el software. Zacatecas...', 'lang':
'es', 'time_zone': 'Central Time (US & Canada)', 'profile_background_color': 'CODEED',
'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme1/bg.png',
'profile_background_tile': False, 'profile_image_url':
'http://pbs.twimg.com/profile_images/1794973756/DSC00071_normal.JPG', 'profile_image_url_https':
'https://pbs.twimg.com/profile_images/1794973756/DSC00071_normal.JPG', 'url':
'http://t.co/gzsXPxDtxH', 'utc_offset': -18000, 'favourites_count': 1977, 'screen_name': 'pphdez76'}
```

Ahora solo imprimimos el usuario, el nombre y el número de sus tweets

```
followers = twitter.get_followers_list(screen_name="xduvox")

for follower in followers["users"]:
    print(" user: {0} \n name: {1} \n Number of tweets: {2} "
          .format(follower["screen_name"],
                  follower["name"],
                  follower["statuses_count"]))
```

Código. SeguidoresPorNombre.py

Se mostrará como sigue:

user: pphdez76

name: Pepe Hernández

Number of tweets: 3237

user: EzraPenland

name: Ezra Penland

Number of tweets: 4635

user: KirkDBorne

name: Kirk Borne

Number of tweets: 26265

user: hardwarehackmx

name: Hardware Hacking Mx

Number of tweets: 1484

user: celuactivo

name: Celuactivo Unefon

Number of tweets: 25952

user: ConferenceCIMPS

name: Conference CIMPS

Number of tweets: 9

user: LordThranguil

name: Benjamin Ruedas

Number of tweets: 2434

user: elidiomaro

name: Elidio Marquina

Number of tweets: 1497

user: jmcerdeiraa

name: Jose Manuel Cerdeira

Number of tweets: 1188

Lugares y Tendencias

En este ejemplo se obtendrán las tendencias cercanas a un lugar. El API de twitter usa WOEID (Yahoo! Where On Earth ID) como relación del ID de un lugar. Por ejemplo el ID de zacatecas lo podemos obtener de la siguiente manera:

En el siguiente link <https://developer.yahoo.com/yql/console/> escribiremos la siguientes consulta `select * from geo.places where text="zacatecas"` dando como resultado 23424900 como se muestra en la siguiente figura.

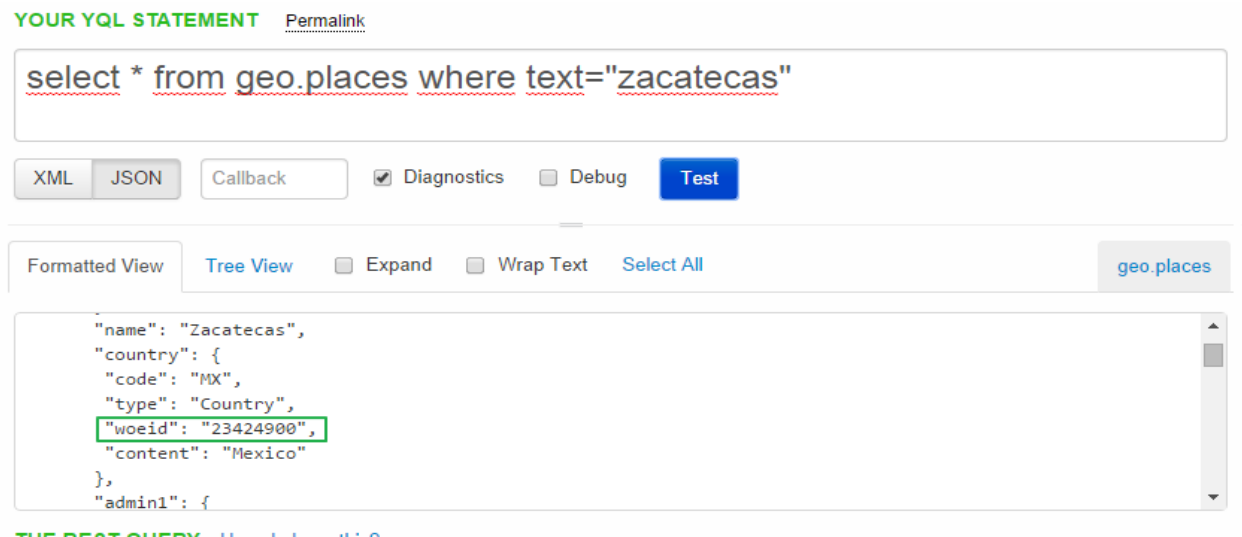


Figura A11 Yahoo WOEID

Se usará el método `twitter.get_place_trends` y el id del lugar.

```
result = twitter.get_place_trends(id = 23424977)

if result:
    for trend in result[0].get("trends", []):
        print("{0} \n".format(trend["name"]))
```

Código. LugarTrends.py

Las tendencias en Zacatecas serían:

#UsosParaElOmnilife
#DENvsSEA
Manning
#VoyASoñar
#TiempoDeAyudar
#DinahJanexWetSeal
Bustos
Salcido

Ronaldinho
Othoniel Arce

Clasificación de sentimientos

En este ejercicio se clasificarán los tweets como positivos o negativos, es decir obtendrá el sentimiento como un valor tangible que puede ser ampliamente utilizado para obtener el pensamiento de las personas acerca de productos, personas, servicios, políticos, etc.

Como sabemos los tweets están limitados a 140 caracteres, tienen un lenguaje casual y en algunos casos los tweets tienen mucho ruido como nombres de usuario, links, letras repetidas y emoticons. Los siguientes tweets son un ejemplo de la clasificación.

"Photoshop, I hate it when you crash " - Negative

"@Ms_HipHop im glad ur doing weeeell " - Positive

En general para la clasificación de un tweet se debe realizar este proceso, empezamos obteniendo las palabras de tweets definidos como de prueba, después tenemos que entrenar un clasificador con una bolsa de palabras (Bag of Words) que es una lista de palabras con su frecuencia en el texto de los tweets de prueba. Por ejemplo la palabra "genial" aparecerá en los tweets positivos. Una vez entrenado el clasificador haremos el query a twitter y los clasificaremos.

En el siguiente diagrama describiremos el proceso de la clasificación de sentimientos y usaremos el clasificador Naive Bayes implementado en la librería NLTK (Natural Language Toolkit) para la clasificación de un tweet como positivo o negativo.

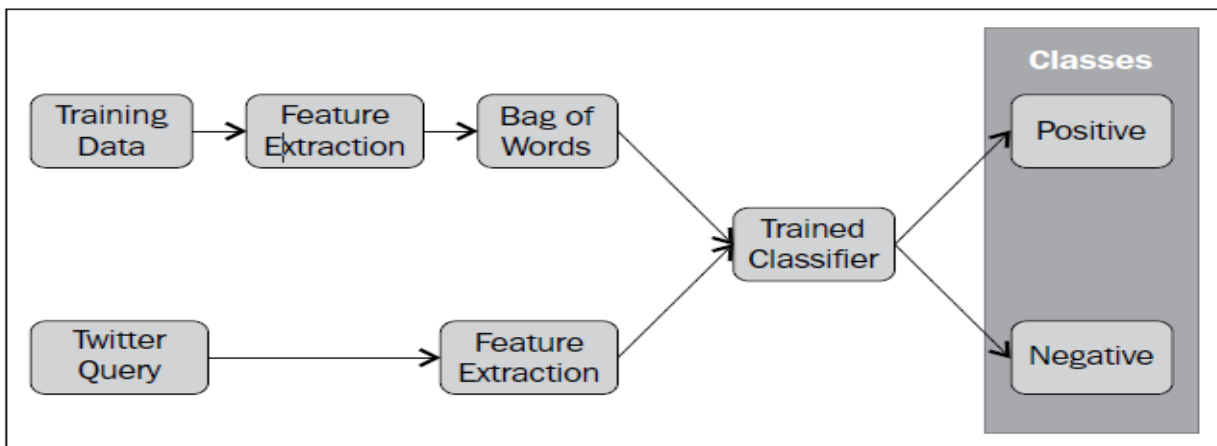


Figura A11 Clasificación de Sentimientos (Cuesta, 2014)

NLTK es una poderosa librería de python que incluye potentes algoritmos para tokenización de texto, análisis, razonamiento semántico, y la clasificación de texto.

Bag of words

Bag of words es un modelo que se utiliza para convertir un documento en una lista de palabras sin ordenar, comúnmente utilizado para la clasificación de los textos mediante la frecuencia de una palabra en un documento. Vamos a utilizar la frecuencia como una característica para el entrenamiento del clasificador. En NLTK, contamos con métodos tales como `nltk.word_tokenize` y `nltk.FreqDist`.

En el siguiente código, podemos ver cómo importar NLTK y el uso del método `nltk.word_tokenize`

```
>>> import nltk
```

```
>>> nltk.word_tokenize("Busy day ahead of me. Also just remembered that I left peah slices in the fridge  
at work on Friday. ")
```

```
['Busy', 'day', 'ahead', 'of', 'me.', 'Also', 'just', 'remembered', 'that', 'I', 'left', 'peah', 'slices', 'in', 'the',  
'fridge', 'at', 'work', 'on', 'Friday', '.']
```

Naïve Bayes

Naive Bayes es el algoritmo de clasificación más simple entre los métodos de clasificación bayesiana. En este algoritmo, simplemente tenemos que aprender las probabilidades haciendo la suposición de que los atributos A y B son independientes, es por eso que este modelo se define como un modelo de producción independiente. Naïve Bayes se utiliza ampliamente en la clasificación de texto porque el algoritmo puede ser entrenado fácil y eficientemente. En Naïve Bayes podemos calcular la probabilidad de una condición A dado B (descrito como $P(A | B)$), si ya sabemos que la probabilidad de B dado A (descrito como $P(B | A)$), y, además, la probabilidad de A (descrito como $P(A)$) y la probabilidad de B (descrito como $P(B)$) de forma individual, como se muestra en el Teorema de Bayes anterior.

NLTK incluye una implementación del algoritmo de Naive Bayes.

En el siguiente código vamos a implementar el algoritmo de Naive Bayes y lo vamos a utilizar para clasificar los tweets de una consulta sencilla utilizando la API de Twitter.

Primero Importamos la Librería.

```
import nltk
```

Después creamos las funciones para la bagofwords y para la extracción de la frecuencia de cada palabra en los tweets `wordFeatures` y `getFeatures`.

```
def bagOfWords(tweets):  
    wordsList = []  
    for (words, sentiment) in tweets:  
        wordsList.extend(words)  
    return wordsList  
  
def wordFeatures(wordList):  
    wordList = nltk.FreqDist(wordList)  
    wordFeatures = wordList.keys()  
    return wordFeatures  
  
def getFeatures(doc):  
    docWords = set(doc)  
    feat = {}  
    for word in wordFeatures:  
        feat['contains(%s)' % word] = (word in docWords)  
    return feat
```

Código. ClasificacionDeSentimientos.py

Usaremos un corpus de Sentimen140 un sitio dedicado al análisis de sentimientos de dicho sitio usaremos 200 tweets positivos y 200 negativos previamente clasificados que utilizaremos para entrenar a nuestro clasificador, estos tweet tendrán el siguiente formato.

```
positiveTweets = [('...', 'positive'),  
('...', 'positive'), ... ]  
negativeTweets = [('...', 'negative'),  
('...', 'negative'), ...]
```

A continuación, vamos a crear el corpus, combinar los aspectos positivos y los negativos de los tweets, y extraer la lista de palabras usando el método `nltk.word_tokenize` sólo excluyendo las palabras con menos de tres caracteres:

```
tweets = []  
for (words, sentiment) in positiveTweets + negativeTweets:  
    words_filtered = [e.lower() for e in nltk.word_tokenize(words) if len(e) >= 3]  
    tweets.append((words_filtered, sentiment))
```

Código. ClasificacionDeSentimientos.py

Ahora, vamos a obtener las características de todas las palabras:

```
wordFeatures = wordFeatures(bagOfWords(tweets))
```

Código. ClasificacionDeSentimientos.py

A continuación, vamos a obtener el conjunto de entrenamiento utilizando el método `nltk.classify.apply_features`:

```
training_set = nltk.classify.apply_features(getFeatures, tweets)
```

Código. ClasificacionDeSentimientos.py

Por último, vamos a entrenar el algoritmo Naive Bayes como se muestra en el siguiente código:

```
classifier = nltk.NaiveBayesClassifier.train(training_set)
```

Código. ClasificacionDeSentimientos.py

Podemos obtener las características más informativas de nuestro clasificador utilizando el método `show_most_informative_features`. Podemos ver el resultado en la siguiente captura de pantalla. Esta lista muestra las palabras más frecuentes o informativas utilizadas por el clasificador:

```
print(classifier.show_most_informative_features(32))
```

Código. ClasificacionDeSentimientos.py

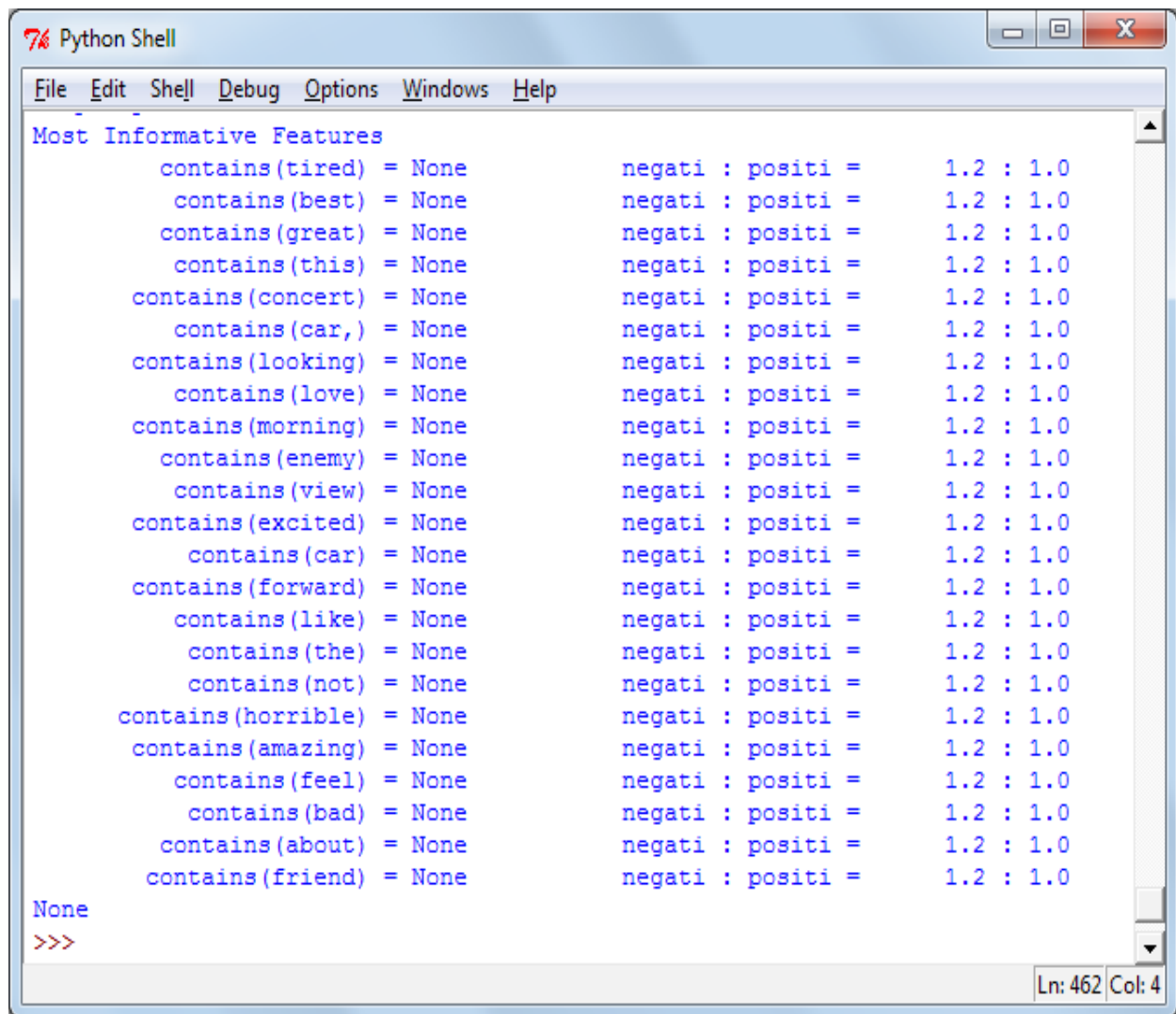


Figura A12 Clasificador

Ahora vamos a realizar una búsqueda de Twitter de la palabra “Duolingo” y vamos a clasificar a cada tweet como positivo o negativo con el método de `classifier.classify`

```
for status in result["statuses"]:
    print("Tweet: {0} \n Sentiment: {1} \n"
          .format( status["text"]
                    , classifier.classify(getFeatures( status["text"].split()))))
```

Código. ClasificacionDeSentimientos.py

La salida debe ser como la siguiente:

Tweet: @duolingo can you make this app help Arabic speakers to learn English and more languages

Sentiment: negative

Tweet: Ma soeur est accro à Duolingo à cause de moi ahah.

Sentiment: positive

Tweet: @outoftheazul @TheEconomist True! So spread the word about other resources like #duolingo

Sentiment: negative

Tweet: Arkadaş arkadaşın dil öğrenmek için programlara para dökmesine izin vermez.
<http://t.co/JRJ8WTWYvI>'da ücretsiz olarak dil öğrenin.

Sentiment: positive

Tweet: @AmirZiadeh9 3m jarreb edross almani bass 3m esta3mel duolingo app!

Sentiment: positive

Tweet: The duolingo app is actually really good

Sentiment: positive

Tweet: @LuisvonAhn Hey Luis, fellow CMU-alum here. Would be awesome 2 meet with you on #edtech, more specifically on #Duolingo and @LeksiEducation

Sentiment: positive

Tweet: I am wondering how the @duolingo profile page / news feed has gotten worse

Sentiment: positive

Tweet: @ginnygoodweed I've been trying to learn more on the duolingo app but it expects me to know way more than I remember from school lol

Sentiment: positive

Tweet: @neymarjr @AleSirenita ya aprenderé a hablar portugués con duolingo y vas a ver las frases que voy a escribir!!! Parabens! jajaja

Sentiment: positive

Tweet: I am probably the last person to discover duolingo. Now to learn some french.

Sentiment: positive

Tweet: @duolingo Status update on Swedish? Release date keeps getting bumped another day...

Sentiment: positive

Tweet: RT @pants: the duolingo language learning app feeding me a hard truth <http://t.co/xQZOyMVPS9>

Sentiment: positive

Tweet: the duolingo language learning app feeding me a hard truth <http://t.co/xQZOyMVPS9>

Sentiment: positive

Tweet: @Imjftniall <http://t.co/XizsjUpAsN>

Sentiment: positive

Retos Actuales en el análisis de sentimiento

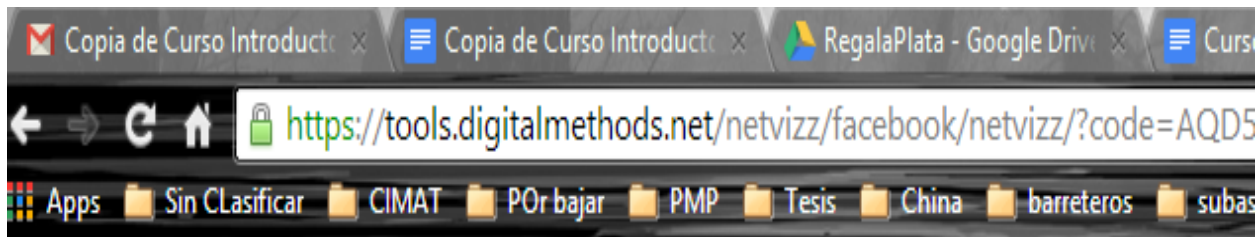
- Clasificación de tweets objetivos y subjetivos
- Manejo de la negación más precisa
- Manejo de comparaciones
- Detección de sarcasmo
- Clasificación del asunto los tweets
- Detectar si el sentimiento de un tweets está relacionado a alguna entidad

Ejemplo 2 Análisis del grafo de amigos en Facebook

En este Ejemplo se descargara la red de amigos en forma de grafo y pondremos en práctica algunos análisis como, ¿es hombre o mujer?, ¿cómo están conectados mis amigos entre ellos?, entre otros. Este ejercicio despertará tu curiosidad de saber más de esta gran red. Analizando datos de esta gran red, como por ejemplo ¿le va al América o a el Atlante?, ¿de cuál de sus amigas esta enamorado?, etc.

En esta ocasión no trabajaremos con el API de Facebook directamente sino que utilizaremos una app llamada netvizz.

Bajaremos nuestra información para analizarla localmente. Dirigiéndonos a la url <https://apps.facebook.com/netvizz/>, la información personal como se muestra a continuación:



netvizz v1.01

Netvizz is a tool that extracts data from different sections of the Facebook platform (personal profile, groups, pages) for research p

For questions, please consult the [FAQ](#) and [privacy](#) sections. Non-commercial use only.

New: there is now an [overview video](#) that introduces the different modules and other things to consider.

Big networks may take some time to process. **Be patient and try not to reload!**

Developing and hosting netvizz costs time and money. If the tool is useful for you, please consider to

[Donate](#)

The following modules are currently available:

personal network - extracts your friends and the friendship connections between them

personal like network - creates a network that combines your friends and the objects they liked in a bipartite graph

group data - creates networks and tabular files for both friendships and interactions in groups

page like network - creates a network of pages connected through the likes between them

page data - creates networks and tabular files for user activity around posts on pages

Figura A12 Aplicación Netvizz

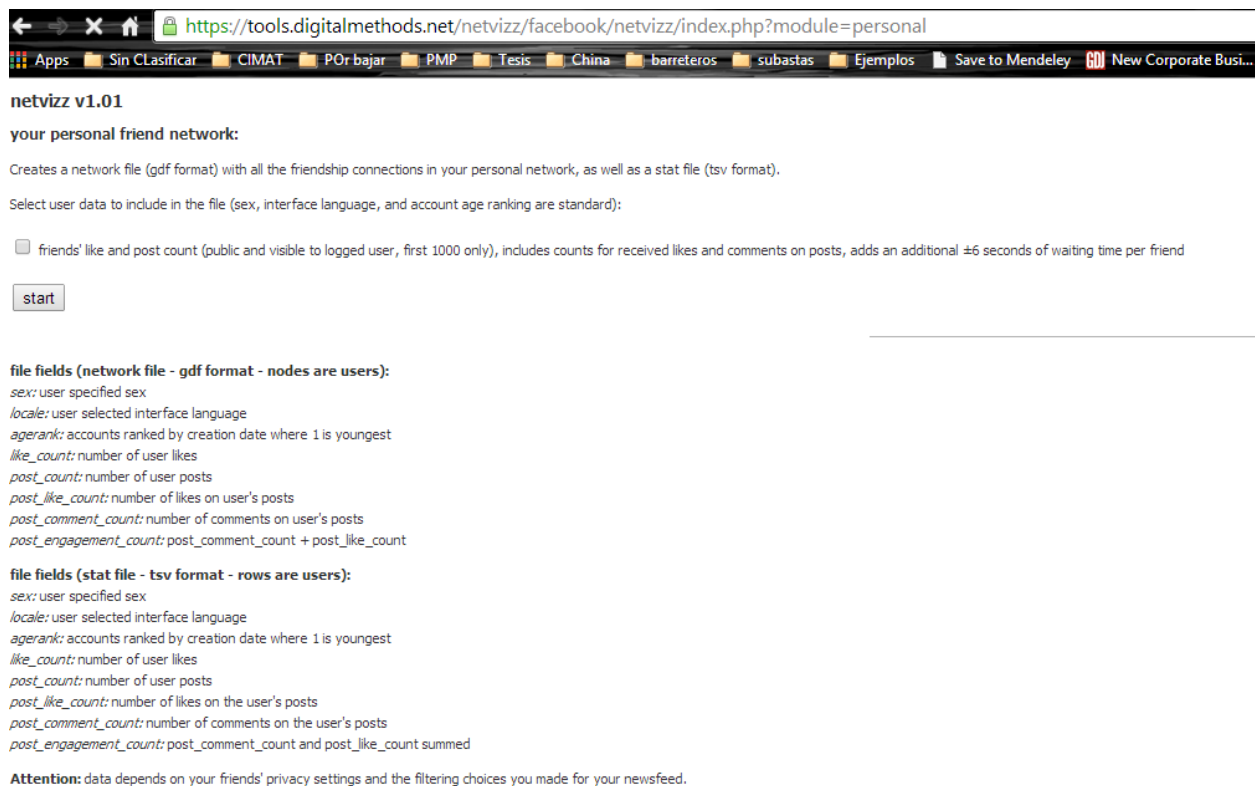


Figura A13 Información personal mediante Netvizz

Descargamos el archivo gdf

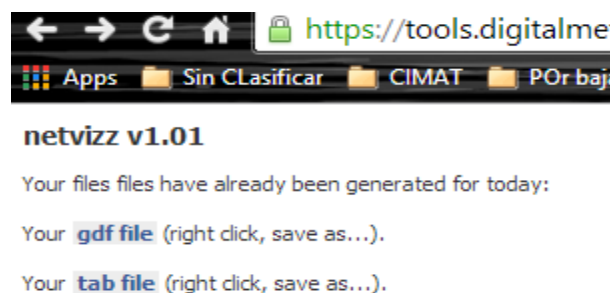


Figura A14 Archivo con información personal GDF

Ahora para visualizar y ver nuestra red de manera visual utilizaremos la herramienta gephi.

```
bigdata@bigdata-VirtualBox:~/Downloads/Requerimientos/gephi/bin$ ls
gephi gephi64.exe gephi.exe
bigdata@bigdata-VirtualBox:~/Downloads/Requerimientos/gephi/bin$ ./gephi
```

Figura A15 Iniciando Gephi

Figura A16 Representación de relacion de amigos en Gephi

Ratio Hombre – Mujer

Primero separaremos el archivo en dos, nodes.csv y links.csv, deberán lucir como las siguientes imágenes.

```
1  nodedef>name VARCHAR,label VARCHAR,sex VARCHAR,locale VARCHAR,agerank INT
2  516935740,Andrea Martinez Devia,female,es_LA,166
3  524068147,Luis Alonso Talavera Trevejo,male,es_LA,165
4  524466435,Alberto Vázquez,male,es_LA,164
5  527546182,Emmanuel Loya,male,es_LA,163
6  531334703,Margarita Berumen,female,es_LA,162
7  550508928,Karina Becerra Rico,female,es_LA,161
8  551438039,Esmeralda Perez de Bañuelos,female,es_LA,160
9  564594856,Olmo Rod,male,es_LA,159
10 564824452,Fernando Martinez,male,es_LA,158
11 565964871,Rene Xruz,male,es_LA,157
```

Figura A17 Archivo nodes.csv

1. Primero, necesitamos importar las librerías requeridas.

```
import numpy as np
import operator
from pylab import *
```

Código. HMratio.py

2. La función `numpy.genfromtext`, obtendrá solo la columna de genero del archivo `nodes.csv`.

```
nodes = np.genfromtxt("nodes.csv", dtype=str, delimiter=',', skip_header=1, usecols=(2))
```

Código. HMratio.py

3. usaremos la función `countOf` para contar cuántos hombres tenemos en la lista de nodos.

```
counter = operator.countOf(nodes, 'male')
```

Código. HMratio.py

4. Calculamos el porcentaje.

```
male = (counter * 100) / len(nodes)  
female = 100 - male
```

Código. HMratio.py

5. Lo dibujamos.

```
figure(1, figsize=(6,6))  
ax = axes([0.1, 0.1, 0.8, 0.8])  
  
labels = 'Male', 'Female'  
fracs = [male, female]  
explode=(0, 0.05)  
  
pie(frac, explode=explode, labels=labels, autopct='%1.1f%%', shadow=True, startangle=90)  
title('Male to Female Ratio', bbox={'facecolor':'0.8', 'pad':5})  
  
show()
```

Código. HMratio.py

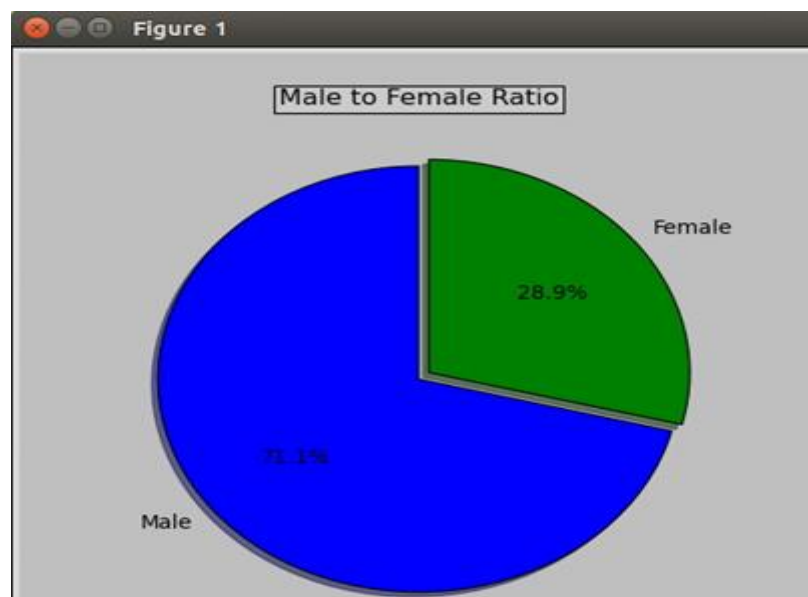
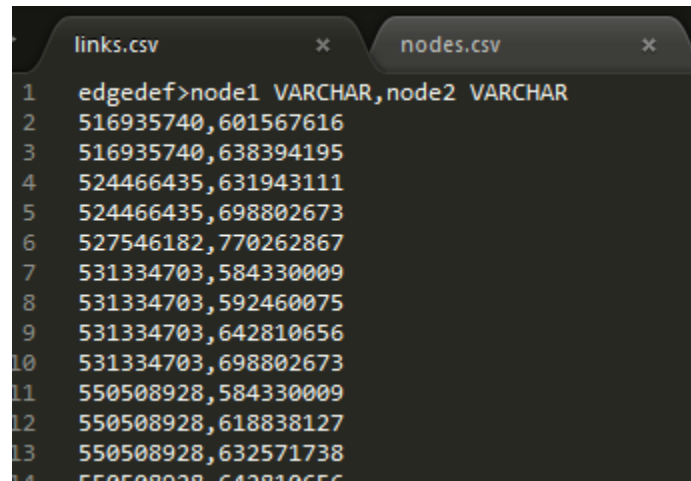


Figura A18 Ratio Hombre-Mujer

Grado de Distribución

El grado de un nodo es el número conexiones con otros nodos. Por lo tanto analizaremos cuantas veces aparece cada nodo. En el siguiente ejercicio obtendremos el nodo fuente y destino del archivo links.csv



```
links.csv  x  nodes.csv  x
1  edgedef>node1 VARCHAR,node2 VARCHAR
2  516935740,601567616
3  516935740,638394195
4  524466435,631943111
5  524466435,698802673
6  527546182,770262867
7  531334703,584330009
8  531334703,592460075
9  531334703,642810656
10 531334703,698802673
11 550508928,584330009
12 550508928,618838127
13 550508928,632571738
14 550508928,642810656
```

Figura A19 Archivo nodes.csv

El código siguiente realiza el cálculo y lo gráfica.

```
import numpy as np
import matplotlib.pyplot as plt
import operator

links = np.genfromtxt("links.csv", dtype=str, delimiter=',', skip_header=1, usecols=(0,1))

dic = {}
#Node Degree , reshape number is double of your links
for n in sorted(np.reshape(links,2158)):

    if n not in dic:
        dic[n] = 1
    else:
        dic[n] += 1

size = len(dic)

# Degree
plt.bar(range(size),list(dic.values()))
plt.xticks(range(size), list(dic.keys()), rotation=90)
plt.show()
```

Código. GradoDeDistribucion.py

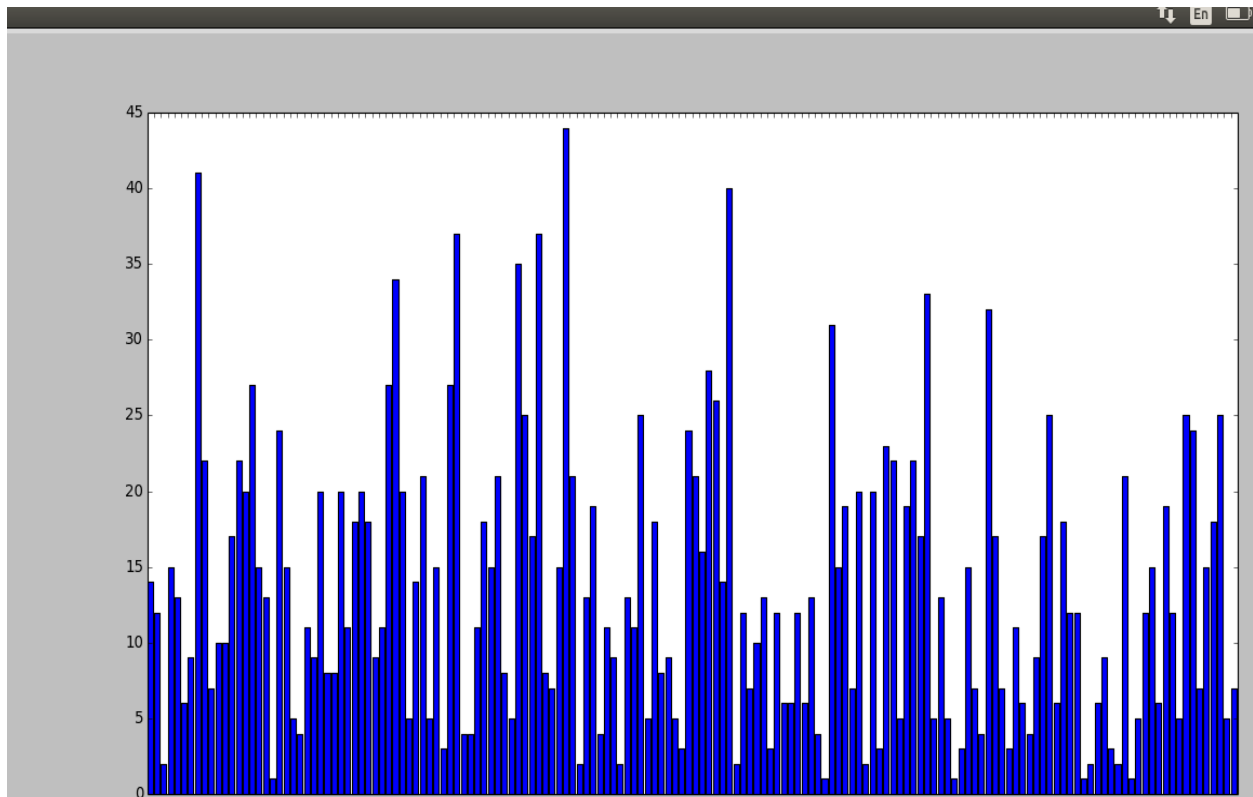


Figura A20 Grado de distribución

Histograma de la Gráfica

Ahora obtendremos la cantidad de nodos que tienen el grado 1, cuántos de 2, 3, etc.

```
#Histogram
histogram = {}
#your high degree + 1 in range()
for n in range(45):
    histogram[n] = operator.countOf(list(dic.values()), n)

plt.bar(range(45), list(histogram.values()))
plt.show()
```

Código. GradoDeDistribucion.py

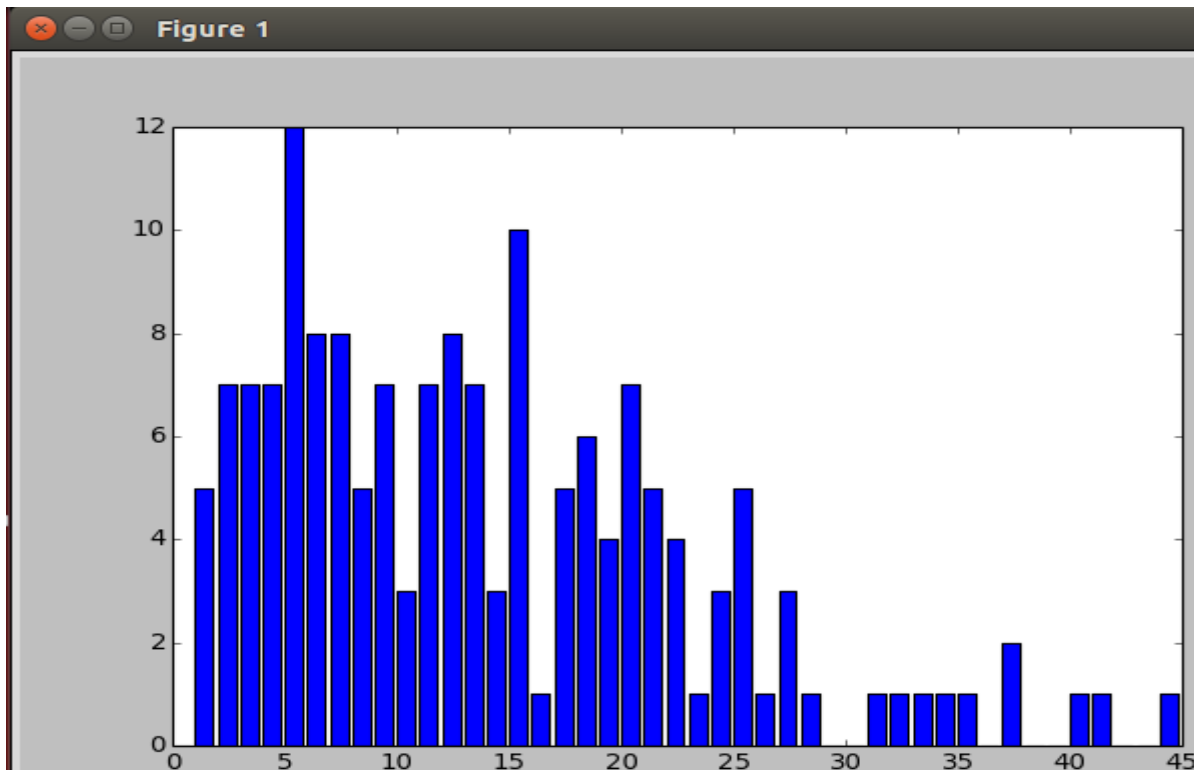


Figura A21 Cantidad de nodos en grado de distribución

Centralidad

Si queremos entender la importancia de un nodo, debemos definir su centralidad, que es una medida relativa de la importancia de un nodo. Existen diferentes maneras de calcular la centralidad como cercanía (closeness) o intermediación (betweenness). En este ejemplo definiremos la centralidad como el nodo más fuertemente conectado.

Nosotros podemos crear nuestro algoritmo de centralidad no solo tomando en cuenta el grado o número de conexiones de un nodo, por ejemplo también tomando en cuenta el número de contenido compartido y likes a cierta publicación.

Para obtener la centralidad ordenaremos los nodos en orden descendente de acuerdo a su grado de distribución.

```
# Centrality
sort = sorted(dic.items(), key=lambda x: x[1], reverse=True)
print(len(sort))
```

Código. GradoDeDistribucion.py

El resultado debe de ser parecido a el siguiente:

```
[('698802673', 44),
 ('100000979886185', 41),
```



```
('700772917', 40),  
('694419007', 37),  
('1132876308', 37),  
('100004017410400', 35),  
('100000236637985', 34),  
.....]
```

Gephi es una excelente herramienta para obtener resultados fáciles y rápidos. Sin embargo si queremos presentar resultados gráficos en una página interactiva necesitamos implementar un método de visualización.

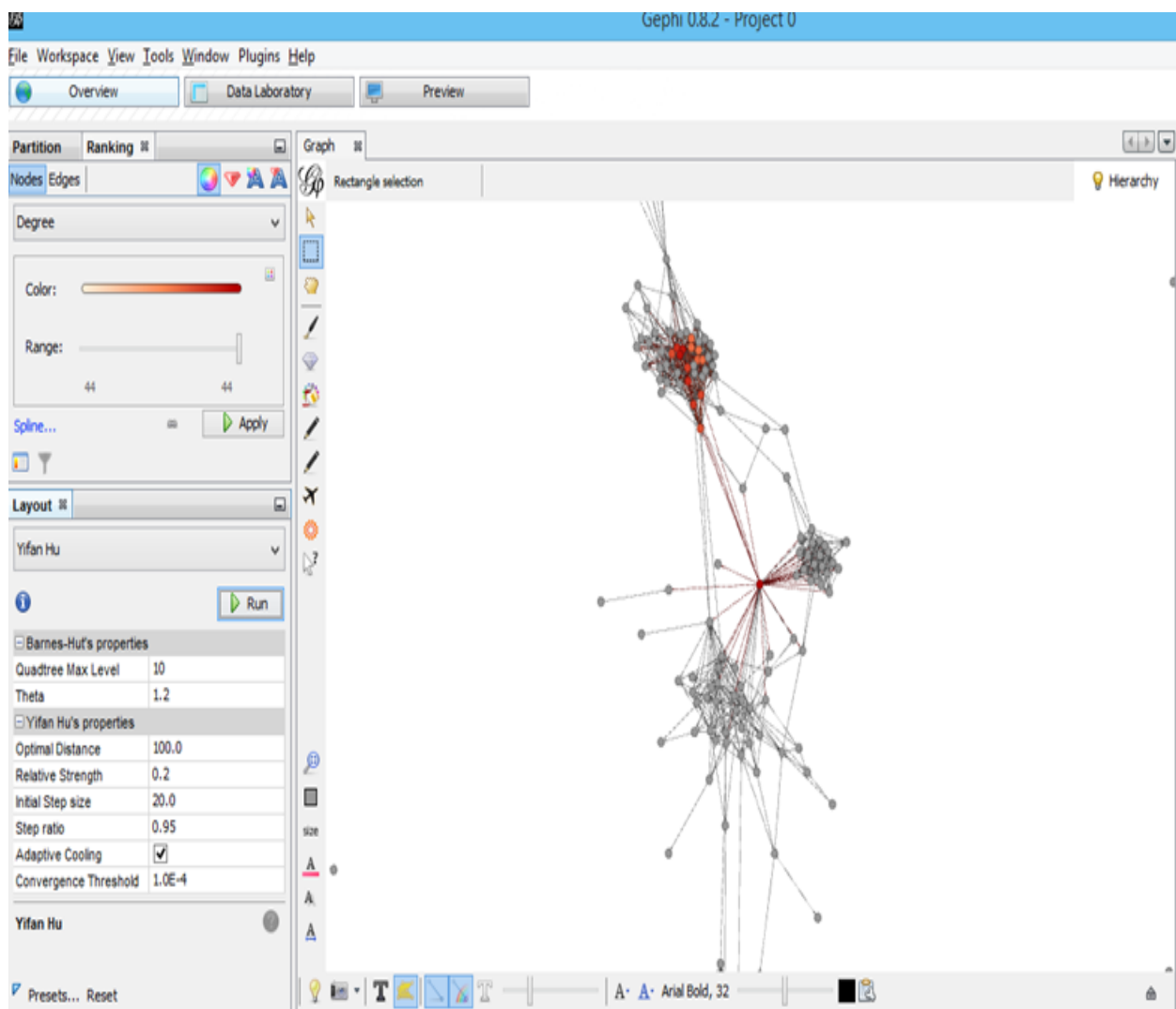


Figura A22 Detección de centralidad mediante Gephi

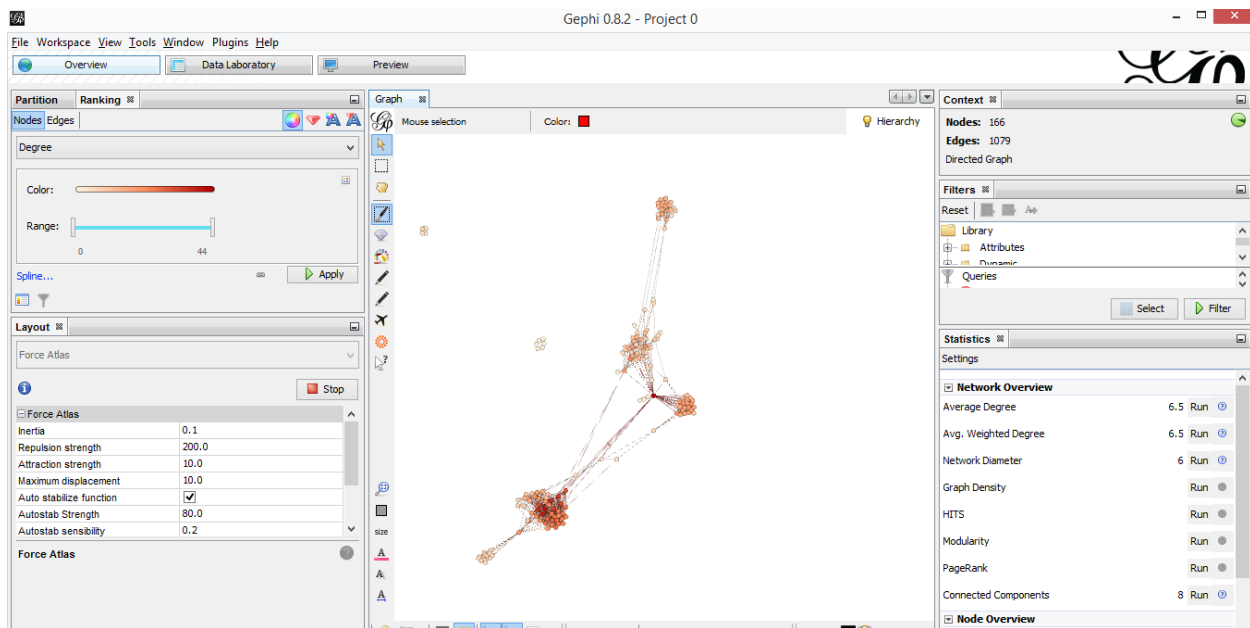
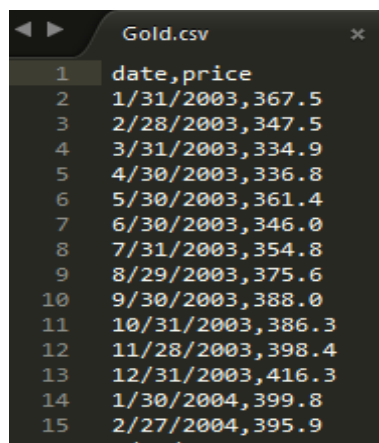


Figura A23 Detección de centralidad mediante Gephi

Ejemplo3. Predicción del precio del Oro.

En este ejercicio se dará una introducción a los conceptos básicos de regresión y datos en series de tiempo. Distinguiremos algunos de los conceptos básicos, como tendencia, estacionalidad y el ruido. Luego se introduce el precio del oro histórico en serie de tiempos y también obtendremos una visión general sobre cómo realizar una predicción mediante regresión kernel ridge. Y por último se presenta una regresión utilizando la serie de tiempo suavizada como entrada.

El análisis de regresión es una herramienta estadística para entender la relación entre las variables dependientes e independientes. En este ejercicio, vamos a implementar una regresión no lineal para predecir el precio del oro en base a los precios del oro histórico. Para este ejemplo, vamos a utilizar los precios históricos de oro desde enero 2003 hasta mayo 2013 en un rango mensual, con datos obtenidos de www.gold.org. Por último, vamos a predecir el precio del oro para junio de 2013 y vamos a contrastarlo con el precio real de una fuente independiente.



	date,price
1	1/31/2003,367.5
2	2/28/2003,347.5
3	3/31/2003,334.9
4	4/30/2003,336.8
5	5/30/2003,361.4
6	6/30/2003,346.0
7	7/31/2003,354.8
8	8/29/2003,375.6
9	9/30/2003,388.0
10	10/31/2003,386.3
11	11/28/2003,398.4
12	12/31/2003,416.3
13	1/30/2004,399.8
14	2/27/2004,395.9
15	3/27/2004,403.5

Figura A24 Valores históricos del Oro

En este ejemplo, vamos a implementar una regresión Kernel Ridge con la serie original y la serie de tiempo suavizada, para comparar las diferencias en la salida.

Regresión no lineal

Estadísticamente hablando la regresión no lineal es un tipo de análisis de regresión para la estimación de las relaciones entre una o más variables independientes en una combinación no lineal. En este capítulo, vamos a utilizar el `mlpy` biblioteca de Python y su aplicación regresión ridge Kernel.

Regresión Kernel Ridge(KRR)

Primero, necesitamos importar las librerías `numpy`, `mlpy`, y `matplotlib`:

```
import numpy as np
import mlpy
from mlpy import KernelRidge
import matplotlib.pyplot as plt
```

Código. KRR.py

Definimos una semilla para la generación de números aleatoria.

```
np.random.seed(10)
```

Código. KRR.py

Después necesitamos cargar los valores históricos del Oro del archivo Gold.csv y guardarlos en targetValues.

```
targetValues = np.genfromtxt("Gold.csv",  
                             skip_header=1,  
                             dtype=None,  
                             delimiter=',',  
                             usecols=(1))
```

Código. KRR.py

A continuación vamos a crear una nueva matriz con 125 puntos de capacitación, uno para cada registro de los targetValues que representan el precio mensual de oro desde enero 2003 hasta mayo 2013:

```
trainingPoints = np.arange(125).reshape(-1, 1)
```

Código. KRR.py

A continuación, vamos a crear otra matriz con 126 puntos de prueba que representa a los 125 puntos originales en targetValues y que incluye un punto extra para nuestro valor predicho para Jun 2013:

```
testPoints = np.arange(126).reshape(-1, 1)
```

Código. KRR.py

Ahora, creamos la matriz kernel de formación (knl) y la matriz kernel de pruebas (knlTest).

KRR dividirá al azar los datos en subconjuntos de igual tamaño, luego procesará un estimador KRR independiente para cada subgrupo. Por último, promediar las soluciones locales en un predictor global:

```
#training kernel matrix  
knl = mlp.kernel_gaussian(trainingPoints, trainingPoints, sigma=1)  
#testing kernel matrix  
knlTest = mlp.kernel_gaussian(testPoints, trainingPoints, sigma=1)
```

Código. KRR.py

Ahora instanciamos la clase mlp.KernelRidge en el Objeto knlRidge:

```
knlRidge = KernelRidge(lmb=0.01, kernel=None)
```

Código. KRR.py

El método learn calcula los coeficientes de regresión, usando la matriz de entrenamiento Kernel y los targetValues como parámetros:

```
knlRidge.learn(knl, targetValues)
```

Código. KRR.py

El método pred() calcula la predicción de la respuesta, usando la matriz de pruebas Kernel como entrada:

```
resultPoints = knlRidge.pred(knlTest)
```

Código. KRR.py

Finalmente, graficamos las series de tiempo de los targetValues y los puntos de resultado:

```
fig = plt.figure(1)
plot1 = plt.plot(trainingPoints, targetValues, 'o')
plot2 = plt.plot(testPoints, resultPoints)
plt.show()
```

Código. KRR.py

En la siguiente figura, se puede observar los puntos que representan los valores de destino, los valores conocidos y la línea que representan los puntos de resultado del método pred. Podemos observar que el último segmento de la línea, que es el valor previsto para junio de 2013.

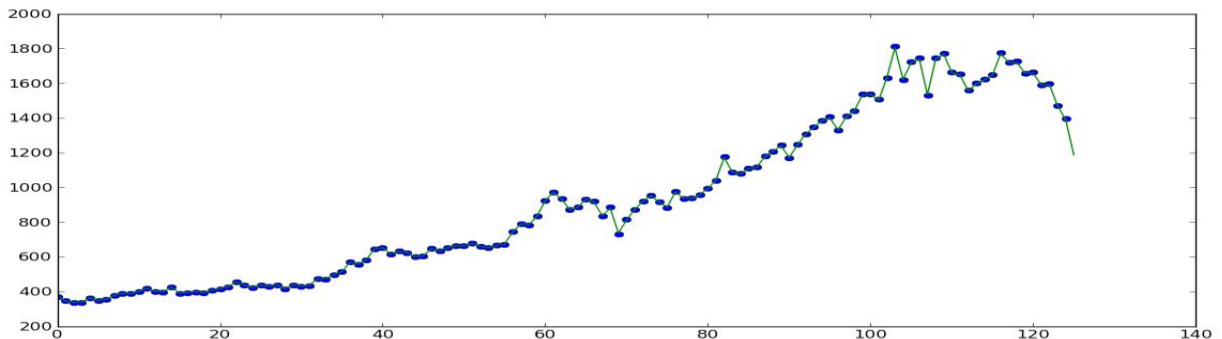


Figura A25 Gráfica de valores históricos del Oro y valor predicho

En la siguiente captura de pantalla, podemos observar los puntos resultantes de la knlRidge.

El método pred() y el último valor (1186.16129538) es el valor previsto para junio de 2013.

bigdata@bigdata-VirtualBox: ~/Documents/Taller/Oro				
391.62343215	394.791291	398.05559853	393.50543841	409.32662915
417.50171719	427.93499065	454.80317134	437.5283018	424.616125
436.6521699	430.36247932	436.53558434	417.53065005	438.30353695
431.01372734	435.84844666	473.97659268	473.78612405	496.07270974
516.08479561	568.65737968	558.48777252	582.89804434	645.39511935
653.16022859	615.82759093	632.69461022	625.24207034	599.98618762
606.01209391	646.59151906	634.04149999	651.1214154	665.15735846
663.10132029	677.60321678	660.15898497	651.91884393	665.88156173
673.71863778	743.35933566	789.68723037	784.41590816	834.1866912
923.0937462	971.22403064	932.85811851	872.26301423	884.80316882
931.99401118	914.54055006	838.42617861	878.51579443	737.31822862
810.95678601	872.40937545	917.89168433	953.01999998	914.64443849
886.38731898	971.98444789	937.11897103	936.7718605	957.51016466
992.90306715	1043.00379077	1170.80325892	1089.47935716	1076.97972427
1107.14368599	1116.03303904	1176.36610067	1208.7325989	1239.51413576
1171.27046153	1242.32094619	1307.12838129	1343.38595492	1383.85097939
1400.21052961	1329.35732296	1405.38790921	1440.13464059	1530.38444854
1535.76768304	1501.3482692	1629.64604751	1803.83015783	1622.14643238
1717.76197169	1739.11892089	1534.37590801	1736.64798768	1767.27956776
1660.76742321	1646.36358188	1557.34365693	1595.2421219	1618.05827517
1648.59106968	1768.32721459	1720.07173547	1718.51077847	1658.84824093
1657.43062521	1590.26281453	1591.23005356	1470.40013319	1389.83693915
1186.16129538]				

Figura A26 Valores históricos del Oro y valor predicho

Suavizando las serie en el tiempo del precio del Oro

Como podemos ver en la serie tiempo del precio del oro son ruidosas y es difícil de detectar una tendencia o patrones con una apreciación directa. Así que para hacerlo más fácil podemos suavizar la serie de valores del oro en el tiempo. En el siguiente código, que suavizar la serie de precios del oro en el tiempo:

```

import matplotlib.pyplot as plt
import numpy as np
import dateutil.parser as dparser
from pylab import *
import datetime

def smooth(x,window_len):

    s=np.r_[2*x[0]-x[window_len-1::-1],x,2*x[-1]-x[-1:-window_len:-1]]
    w = np.hamming(window_len)

    y=np.convolve(w/w.sum(),s,mode='same')
    return y[window_len:-window_len+1]

x = np.genfromtxt("Gold.csv", dtype=str, delimiter=',', skip_header=1,usecols=(0))
xx = []

for xs in x:
    print(xs)
    month, day, year = xs.split('/')
    xx.append(datetime.date(int(year),int(month),int(day)))

y = np.genfromtxt("Gold.csv",
                  skip_header=1,
                  dtype=None,
                  delimiter=',',
                  usecols=(1))

y2 = smooth(y, len(y))

print(y2)

plt.step(xx, y2)
plt.step(xx, y, 'co')
plt.show()

```

Código. smoothTS.py

En la siguiente figura, se observa la serie histórica de los precios del oro históricas (la línea de puntos) y podemos ver la serie de tiempo suavizada (la línea) usando la ventana de Hamming:

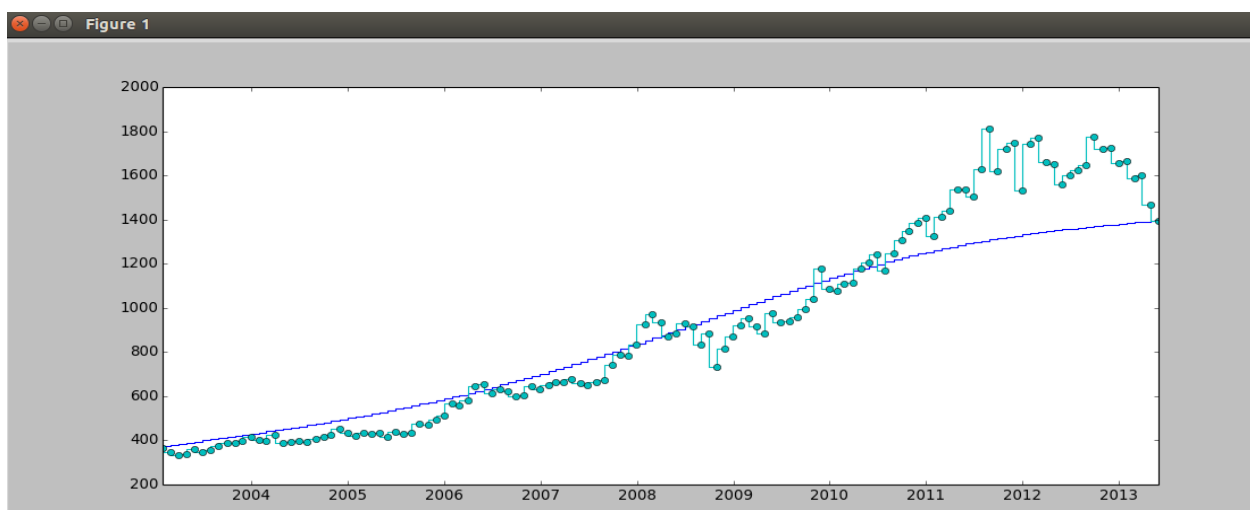


Figura A27 Gráfica suavizada de valores históricos del Oro y valor predicho

Prediciendo el precio suavizado del oro

Por último, ponemos todo junto e implementamos la regresión Kernel Ridge para la serie de tiempo del precio del oro suavizado. Podemos encontrar el código completo del KRR como sigue:

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3 import dateutil.parser as dparser
4 from pylab import *
5 import mlp
6
7 def smooth(x,window_len):
8
9     s=np.r_[2*x[0]-x[window_len-1:-1],x,2*x[-1]-x[-1:-window_len:-1]]
10    w = np.hamming(window_len)
11
12    y=np.convolve(w/w.sum(),s,mode='same')
13    return y[window_len:-window_len+1]
14
15
16 y = np.genfromtxt("Gold.csv",
17                 skip_header=1,
18                 dtype=None,
19                 delimiter=',',
20                 usecols=(1))
21
22
23 targetValues = smooth(y, len(y))
24
25 np.random.seed(10)
26
27 trainingPoints = np.arange(125).reshape(-1, 1)
28 testPoints = np.arange(126).reshape(-1, 1)
29
30 knl = mlp.kernel_gaussian(trainingPoints, trainingPoints, sigma=1)
31 knlTest = mlp.kernel_gaussian(testPoints, trainingPoints, sigma=1)
32
33 knlRidge = mlp.KernelRidge(Lmb=0.01, kernel=None)
34 knlRidge.learn(knl, targetValues)
35 resultPoints = knlRidge.pred(knlTest)
36
37 print(resultPoints)
38
39 plt.step(trainingPoints, targetValues, 'o')
40 plt.step(testPoints, resultPoints)
41 plt.show()
```

Código. smoothKRR.py

En la siguiente figura, se puede observar la línea de puntos que representa la serie de tiempo suavizada de los precios históricos del oro y la línea que representa la predicción para el precio del oro en junio de 2013:

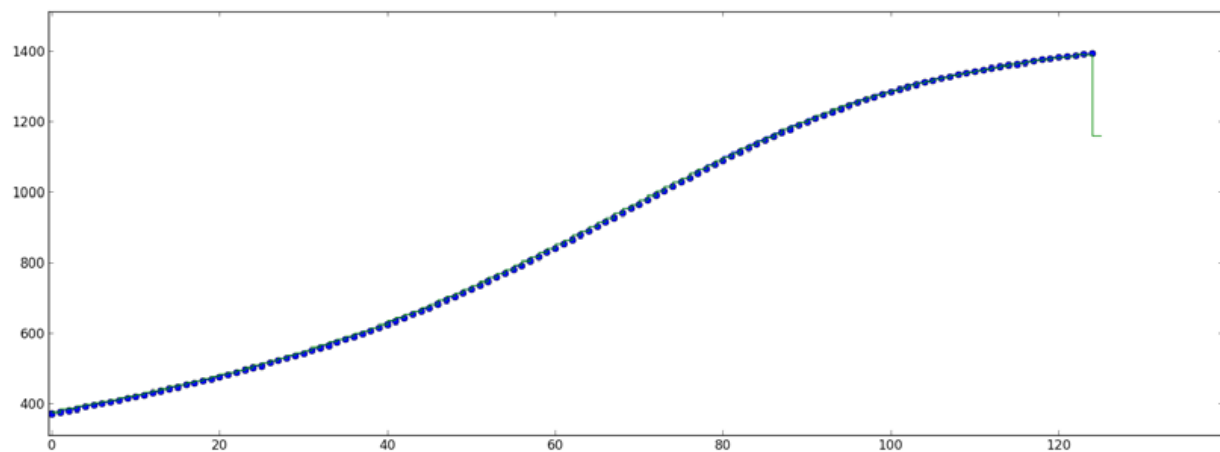


Figura A28 Gráfica de valores históricos del Oro y valor predicho suavizados

En la siguiente captura de pantalla, podemos ver los valores previstos para la serie de tiempo suavizada. Esta vez se puede observar que los valores son mucho menores que las predicciones originales:

```
bigdata@bigdata-VirtualBox: ~/Documents/Taller/Oro
[ 374.51133782 375.42728144 383.20978715 386.55925943 392.46531865
396.8304114 402.08036522 406.64963103 411.61891271 416.54736937
421.68226564 426.89298681 432.13320761 437.39687851 442.8277588
448.26856991 453.80700511 459.40349131 465.12155923 471.01763701
477.16764456 483.29753088 489.55962234 495.95533425 502.39775747
508.99669838 515.76327523 522.62405761 529.50269351 536.564622
543.83066213 551.22299912 558.78854779 566.51553608 574.26896919
582.27331468 590.45078922 598.88464414 607.45457256 616.14616297
625.16174122 634.50669063 643.78120412 653.32348913 663.03902494
672.68603314 682.74256612 693.00507801 703.35589709 713.83712636
724.39386317 735.19319542 746.17419283 757.35101254 768.84375678
780.43479607 792.16744732 803.98556052 815.98165738 828.02764284
840.22490923 852.43778715 864.71824328 877.13144778 889.59198403
902.02890179 914.55809069 927.05772476 939.63584676 952.18020148
964.72720589 977.21402856 989.83832244 1002.45977268 1015.04520552
1027.63727974 1040.05240733 1052.3297269 1064.41247888 1076.40420589
1088.49204982 1100.18449972 1111.71219714 1123.15663634 1134.16932013
1145.13983676 1156.05105186 1166.69205995 1177.07052312 1187.29986588
1197.29078841 1207.11631934 1216.54910612 1225.72275768 1234.64789851
1243.27697587 1251.68807657 1259.86157756 1267.63529686 1275.09002695
1282.30403757 1289.27215968 1295.89100453 1302.25556121 1308.37414243
1314.1067606 1319.73793319 1325.20466654 1330.46088301 1335.4613124
1340.25195252 1344.77634816 1349.18125634 1353.38699135 1357.33420754
1361.23729824 1364.98019319 1368.80631479 1372.33661519 1376.00064877
1378.7428981 1382.82640402 1384.20059902 1389.86822295 1388.63246369
1159.23545044]
bigdata@bigdata-VirtualBox:~/Documents/Taller/Oro$
```

Figura A29 Valores históricos del Oro y valor predicho suavizados

Contrastando el valor predicho

Por último, vamos a buscar una fuente externa para ver si nuestra predicción es realista. En la siguiente figura podemos observar un gráfico de The Guardian / Thomson Reuters para junio de 2013 El precio del oro fluctuó entre 1.180,0 y 1.210,0, con un promedio oficial de 192.0 en el mes. Nuestra predicción para la regresión contraída Kernel con datos completos es 1186.0, lo cual no es malo en absoluto. Podemos ver las cifras exactas en la siguiente tabla:

Fuente	Junio 2013
The Guardian/Thomson Reuters	1192
Kernel ridge regression - modelo predictivo	1186.161295
Kernel ridge regression suavizado - modelo predictivo	1159.23545044

Tabla A1 Análisis de valores

Una buena práctica cuando queremos construir un modelo predictivo es probar diferentes enfoques para el mismo problema. Si desarrollamos más de un modelo, es posible comparar los resultados de las pruebas en contra de la otra y seleccionar el mejor modelo. Para este ejemplo particular, el valor predicho utilizando los datos completa es más preciso que el valor predicho utilizando los datos suavizados.

In words of the mathematician named George E. P. Box "All models are wrong, but some are useful".

Creación de tu propio ambiente de trabajo.

Una alternativa a la máquina virtual proporcionada es la creación de tu propio ambiente de trabajo, a continuación se presenta la guía paso a paso para crearlo.

Los requerimientos de software se pueden instalar de la siguiente manera en SO Ubuntu.

Instalando Python3

En ubuntu por default la versión de python es la 2.7 pero nosotros trabajaremos con la 3

```
$ sudo apt-get install python3
```

```
$ sudo apt-get install idle3
```

Instalando PIP

```
$ sudo apt-get install python3-pip
```

Instalando NumPY

```
$ sudo apt-get install python3-numpy
```

```
$ sudo apt-get install python3-nose
```

```
$ pip3 install nose
```

Probando la libreria numpy.

```
>>> import numpy
```

```
>>> numpy.test()
```

Instalando SciPy

SciPy es un software de código abierto para las matemáticas, la ciencia y la ingeniería.

```
$ sudo apt-get install python3-scipy
```

Probado la libreria scipy

```
>>> import scipy
```

```
>>> scipy.test()
```

Instalando mlp

```
$ sudo apt-get install python3-matplotlib
```

mlp necesita el paquete GNU Scientific Library (GSL) development package que lo podemos instalar del Ubuntu Software Center como se muestra en la siguiente figura.

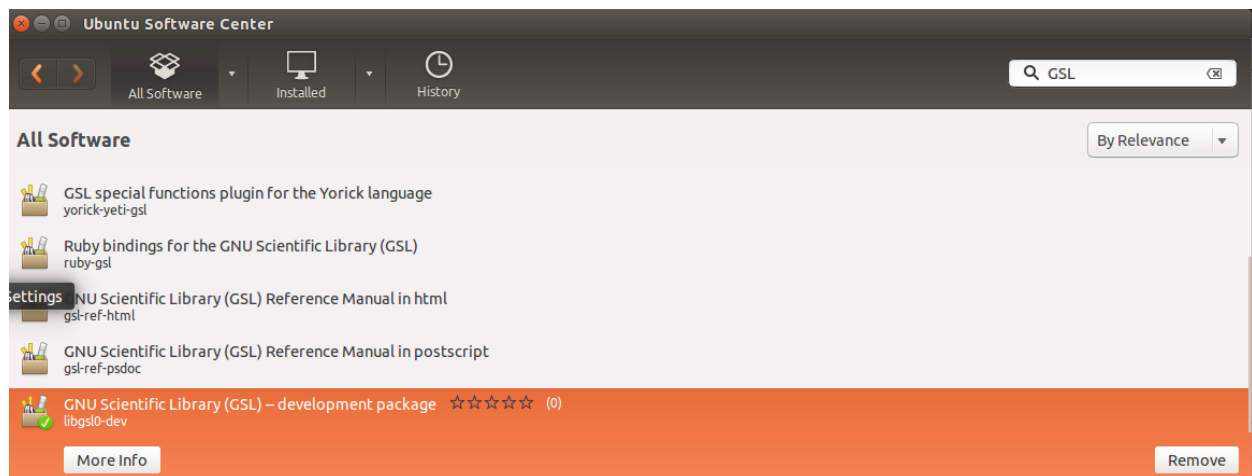


Figura A31 Instalación de la librería GSL

Instalando Gephi

Descargamos Gephi 0.8.2 desde la página oficial <https://gephi.github.io/users/download/>

wget <https://launchpad.net/gephi/0.8/0.8.2beta/+download/gephi-0.8.2-beta.tar.gz>

Después extraemos y abrimos una terminal en el directorio, a continuación, escribimos `./bin/gephi` para iniciar.

Instalando NLTK

Primero descargamos la versión para de NLTK para python 3 de la página oficial.

<http://www.nltk.org/nltk3-alpha/> , `nltk-3.0a4.tar.gz`., descomprimos el archivo y lo instalamos dela siguiente manera:

```
$ sudo python3 setup.py install
```

Librerías adicionales y Datos

```
sudo python -m nltk.downloader -d /usr/share/nltk_data all
```

Instalando twython

```
sudo pip3 install twytho
```

Anexo B: Oferta educativa en Big Data

En este anexo se encuentra el detalle de la oferta educativa relacionada con Big Data.

<p>Institución: Universidade Nova url: http://www.isegi.unl.pt/MAA/ Modalidad: Presencial Grado: Maestría Título: Master Degree Advanced Analytics Temario: Descriptive Analytics Databases Data Mining Modeling Data Analytics Business Intelligence Predictive Analytics Research Methods Dissertation Duración(Meses): 24 País: Portugal costo(\$US): \$7,005.35 Lenguaje: ND Software: ND Idioma: Ingles</p>	<p>Institución: Ben Gurion University of the Negev url: http://in.bgu.ac.il/en/international-studies/Pages/Data_Mining_BI.aspx Modalidad: Presencial Grado: Maestría Título: Data Mining and Business Intelligence MSc Temario: Data Mining Databases Decision Analytics Data Acquisition Information Systems Artificial Intelligence Business Intelligence Networks Predictive Analysis Machine Learning Data Analysis Security Duración(Meses): 24 País: Israel costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles</p>
<p>Institución: University College Cork, Ireland url: http://www.ucc.ie/en/compsci/postgraduatecourses/analyticsdisc/ Modalidad: Presencial Grado: Maestría Título: Msc Data Science & Analytics Temario: Data Mining Databases Statistics Linear Models Optimization Complex Systems Systems Development Programming Cloud Computing Research Methods Decision Science Data Analysis Duración(Meses): 12 País: Irlanda costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles</p>	<p>Institución: Erasmus Mundus Belgium Université Libre de Bruxelles ULB France Université Francois Rabelais Tours UFRT, France Ecole Central Paris ECP, Germany Technische Universität Berlin TUB and Spain Technical Univesity of Catalonia. url: http://it4bi.univ-tours.fr/ Modalidad: Presencial Grado: Maestría Título: Master Programme In Information Technologies for Business Intelligence Temario: Databases Business Process Decision Analytics Ethics Data Mining Web Technologies Business Intelligence Web Services Viability of Business Project Big Data Analytics Distributed Systems Project Duración(Meses): 24 País: Francia costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles</p>
<p>Institución: Berkeley School of Information url: http://datascience.berkeley.edu/ Modalidad: Mixto Grado: Maestría Título: Master of Information and Data Science Temario: Research Methods Data Processing Databases Data Mining Statistics Ethics</p>	<p>Institución: Thomas Edison State College url: http://mba.tesc.edu/mba-data-analytics/ Modalidad: Presencial Grado: Maestría Título: Master Business Administration in Data Analytics Temario: Data Mining Data Analysis Visualization</p>

Presentation Skills Duración(Meses): 20 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Statistics Business Predictive Analytics Duración(Meses): 18 País: EUA costo(\$US): \$25,000.00 Lenguaje: ND Software: ND Idioma: Ingles
Institución: Dublin City University url: http://www.computing.dcu.ie/postgraduate/mcm/msc-computing-mcm Modalidad: Presencial Grado: Maestría Título: Master of Science Data Studies Temario: Professional Skills Computer Architecture Data Management Visualization Statistics Data Analysis Mathematical Cloud Computing Duración(Meses): 12 País: Irlanda costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Institución: Univesiteit Gent url: https://sites.google.com/site/mastatugent/ Modalidad: Presencial Grado: Maestría Título: Master of Statistical Data Analysis Temario: Data Analysis Missing Data Computational Biology Data Mining Business Risk Analysis Time Series Project Duración(Meses): 24 País: Bélgica costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles
Institución: Bournemouth University url: http://courses.bournemouth.ac.uk/courses/postgraduate-degree/applied-data-analytics/msc/1975/ Modalidad: Presencial Grado: Maestría Título: Data Analytics MSc Temario: Data Mining Research Methods Project Business Intelligence Analytics Big Data Cloud Computing Duración(Meses): 24 País: Reino Unido costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Institución: Universidad de Buenos Aires url: http://triton.exp.dc.uba.ar/datamining/ Modalidad: Presencial Grado: Maestría Título: Maestría en Explotación de Datos y Descubrimiento del Conocimiento Temario: Machine Learning Data Analysis Data Mining Statistics Business Data Science Databases Mathematical Artificial Intelligence Visualization GIS Project Duración(Meses): 24 País: Argentina costo(\$US): ND Lenguaje: ND Software: ND Idioma: Español
Institución: Dublin Institute of Technology url: http://www.comp.dit.ie/DT217/ Modalidad: Presencial Grado: Maestría Título: Msc In Computing Knowledge Management Temario: Knowledge Management Systems Development Databases Computer Architecture Probability Research Methods Project Business Intelligence GIS Computer Systems Machine	Institución: Erasmus Mundus France(University of Pierre and Marie Curie Paris 6, University of Lyon Lumière 2, Polytec' Nantes), Romania(University Polithenica of Bucharest), Italy (University of East Piedmont) and Spain (Technical Univesity of Catalonia). url: http://www.em-dmkm.eu/ Modalidad: Presencial Grado: Maestría

<p>Learning Security Ubiquitous Computing Design Web Design Man and Machine Computer Science Data Privacy Law Strategy</p> <p>Duración(Meses): 24</p> <p>País: Irlanda</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>	<p>Título: Master Course in Data Mining and Knowledge Managment an European Master</p> <p>Temario: Logic Optimization Probability Data Analysis Databases Software Engineering French Machine Learning Research Methods Data Mining Bioinformatics Science Data Processing Modeling Ontology Engineering Visualization Bayesian Methods Language Statistics Multivariate Modeling Natural Language Processing Semantic Web Graphical Models Project</p> <p>Duración(Meses): 24</p> <p>País: Francia</p> <p>costo(\$US): \$20,379.20</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>
<p>Institución: NYU STERN</p> <p>url: http://www.stern.nyu.edu/programs-admissions/global-degrees/business-analytics/</p> <p>Modalidad: Presencial</p> <p>Grado: Maestría</p> <p>Título: Master of Science in Business Analytics</p> <p>Temario: Social Network Analysis R Data Science Predictive Analysis Data Analysis Network Analytics Decision Models Analytics Visualization Business Strategy Project</p> <p>Duración(Meses): 12</p> <p>País: EUA</p> <p>costo(\$US): \$67,500.00</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>	<p>Institución: Saint Joseph's University</p> <p>url: http://online.sju.edu/programs/business-intelligence-curriculum.asp</p> <p>Modalidad: En Línea</p> <p>Grado: Maestría</p> <p>Título: Master of Science in Business Intelligence & Analytics</p> <p>Temario: Databases Business Analytics DSS Modeling Big Data Business Process Data Mining Management Predictive Analytics Business Intelligence</p> <p>Duración(Meses): 24</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>
<p>Institución: Saint Peter's University</p> <p>url: http://www.saintpeters.edu/data-science-and-business-analytics/#image-2</p> <p>Modalidad: Presencial</p> <p>Grado: Maestría</p> <p>Título: Data Science Master of Science in Data with a contration in Business Analytics.</p> <p>Temario: Data Science Statistics Databases Business Analytics Predictive Analytics Business Intelligence Data Mining Big Data Analytics Machine Learning Data Visualization</p> <p>Duración(Meses): ND</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: R</p>	<p>Institución: University of Southern California</p> <p>url: http://gapp.usc.edu/graduate-programs/masters/computer-science/data-science</p> <p>Modalidad: Mixto</p> <p>Grado: Maestría</p> <p>Título: Master of Science in Computer Science – Data Science</p> <p>Temario: Algorithms Databases Artificial Intelligence Statistics Data Science Data Systems Data Processing Data Mining Probability Big Data Analysis Optimization</p> <p>Duración(Meses): 24</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p>

Software: Weka SAS Idioma: Ingles Institución: The Chinese University of Hong Kong url: http://www.sta.cuhk.edu.hk/Dept/PostG/DBS_MSC.html Modalidad: Presencial Grado: Maestría Título: Master of Science in Business Analytics Temario: Data Science Regression Analysis Data Analysis Data Mining Statistics Probability Simulation Time Series Actuarial Principles Bayesian Methods Risk Analysis Duración(Meses): 24 País: EUA costo(\$US): \$45,000.00 Lenguaje: ND Software: ND Idioma: Ingles	Software: ND Idioma: Ingles Institución: The Catholic University of America url: http://msba.cua.edu/ Modalidad: Presencial Grado: Maestría Título: Master's Degree in Business Analysis Temario: Management Data Analysis Management Accounting Leadership Career Development Financial Management Entrepreneurship Research Methods Business Project Duración(Meses): 12 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles
Institución: Technische Universität Dortmund Facultät statistik url: http://www.statistik.tu-dortmund.de/703.html Modalidad: Presencial Grado: Maestría Título: Master of Science Data Studies Temario: Statistics Logic Analysis Data Structures Vectors Duración(Meses): 24 País: Alemania costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Institución: Lewis University url: http://online.lewisu.edu/msds/data-science.asp Modalidad: En Línea Grado: Maestría Título: Online Master of Science in Data Science Temario: Statistics Mathematical Security Big Data Systems Duración(Meses): 24 País: EUA costo(\$US): ND Lenguaje: R Java C++ Python Software: MatLab Idioma: Ingles
Institución: University of Bristol url: http://www.cs.bris.ac.uk/Teaching/MachineLearning/ Modalidad: Presencial Grado: Maestría Título: MSc in Advanced Computing Temario: Machine Learning High Performance Computing Cloud Computing Programming Robotic Systems Computational Genomics Uncertainty Modelling Bioinformatics Artificial Intelligence Server Software Research Methods Statistics Duración(Meses): 12 País: Reino Unido costo(\$US): ND Lenguaje: ND	Institución: Carnegie Mellon University Heinz College url: http://in.bgu.ac.il/en/international-studies/Pages/Data_Mining_BI.aspx Modalidad: Presencial Grado: Maestría Título: Master of Information Systems Management degree with a Business Intelligence and Data Analytics concentration Temario: Management Business Analytics Databases Telecommunications Technology Agile Methods Data Science Business Business Intelligence Business Process Presentation Skills Data Mining Data Structures Decision Analytics Data Privacy Law Data Processing Distributed

Software: ND Idioma: Ingles	Systems E-commerce Computer Architecture Entrepreneurship Ethics Data Analysis Visualization Finalcial Accounting Duración(Meses): 12 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles
Institución: Elmhurst College url: http://public.elmhurst.edu/data_science Modalidad: En Línea Grado: Maestría Título: M.S in Data Science Temario: Databases Data Analysis Data Mining Programming Analytical Methods Research Methods Duración(Meses): 24 País: EUA costo(\$US): ND Lenguaje: C++ Java Python Software: MatLab Idioma: Ingles	Institución: The George Washington University url: http://www.gwanalytics.org/ Modalidad: En Línea Grado: Maestría Título: MS in Business Analytics Temario: Business Analytics Databases Programming Probability Statistics Data Mining Data Processing Optimization Risk Analysis Marketing Business Data Analysis Social Network Analysis Analytics Business Process Visualization Duración(Meses): 12 País: EUA costo(\$US): \$50,000.00 Lenguaje: ND Software: ND Idioma: Ingles
Institución: Royal Hollowy University of London url: https://www.royalholloway.ac.uk/computerscience/prospectivestudents/postgraduatetaught/bigdata.aspx Modalidad: Presencial Grado: Maestría Título: MSc in Data Science and Analytics Temario: Data Analysis Programming Data Mining Big Data Systems Duración(Meses): 24 País: Reino Unido costo(\$US): ND Lenguaje: ND Software: Matlab,MongoDB, Cassandra, HBase, Hadoop, Pig Idioma: Ingles	Institución: Singapore Management University url: http://www.smu.edu.sg/programmes/postgraduate/mitb Modalidad: Presencial Grado: Maestría Título: Master of IT in Business – Analytics (MITB-AT) Temario: Analytics Data Analysis Visualization Social Network Analysis Data Analytics Business Analytics Management Modeling Project Technology Finalcial Accounting Management Accounting Strategy Duración(Meses): 24 País: Singapore costo(\$US): \$45,000.00 Lenguaje: ND Software: ND Idioma: Ingles
Institución: Central Connecticut State University url: http://web.ccsu.edu/datamining/ Modalidad: En Línea Grado: Maestría	Institución: Telecom ParisTech url: http://www.telecom-paristech.fr/formation-continue/masteres-specialises/big-data.html Modalidad: Presencial

Título: Data Mining Course Temario: Data Mining Clustering Predictive Analytics Artificial Intelligence Machine Learning Statistics SAS Data Analysis Duración(Meses): 18 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Grado: Maestría Título: Masterè Spécialisé Big Data Temario: Security Statistics Databases Data Privacy Law Hadoop Big Data Visualization Distributed Systems Machine Learning Data Mining Duración(Meses): 24 País: Francia costo(\$US): \$18,468.65 Lenguaje: ND Software: ND Idioma: Francés
Institución: University of Glasgow url: http://www.gla.ac.uk/postgraduate/taught/datascience/ Modalidad: Presencial Grado: Maestría Título: DataScience MSc Temario: Big Data Distributed Systems Data Mining Machine Learning Profesional Skills Research Methods Project Operating Systems Artificial Intelligence Computer Architecture Computer Vision Methods and Applications Programming Security Cloud Computing Software Engineering Functional Programming Human Computer Interaction Information Systems Real-time systems Duración(Meses): 24 País: Reino Unido costo(\$US): \$8,615.33 Lenguaje: ND Software: ND Idioma: Ingles	Institución: Illinois Institute of Technology url: http://science.iit.edu/programs/graduate/master-data-science Modalidad: En Línea Grado: Maestría Título: Master of Data Science Temario: Mathematical Data Mining Machine Learning Statistics Databases Data Processing Ethics Project Management Software Engineering Duración(Meses): 24 País: EUA costo(\$US): ND Lenguaje: C++ Java Python Software: MatLab Idioma: Ingles
Institución: Universidad Complutense de Madrid, Universidad Politécnica de Madrid url: http://www.mat.ucm.es/teci/wp/?page_id=85 Modalidad: Presencial Grado: Maestría Título: Master Universitario en Tratamiento Estadístico Computacional de la Información Temario: Mathematical Optimization Software Engineering Data Analysis Predictive Analytics Neural Networks Data Mining Time Series Bayesian Methods Calculus Numerical Methods Statistics Modeling Social Network Analysis Networks Duración(Meses): 12 País: España costo(\$US): \$4,967.43 Lenguaje: ND	Institución: Arizona State University url: https://wpcarey.asu.edu/masters-programs/business-analytics Modalidad: Mixto Grado: Maestría Título: Master of Business Analytics Temario: Analytics Data Mining Business Analytics Regresion Models Data Management Analytical Decision Making Tools Project Duración(Meses): 16 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles

Software: ND Idioma: Español	
Institución: De Montfort University Leicester url: http://www.dmu.ac.uk/study/courses/postgraduate-courses/business-intelligence-systems-and-data-mining-msc.aspx Modalidad: Presencial Grado: Maestría Título: Business Intelligence Systems and Data Mining MSc Temario: Business Intelligence Databases Research Methods Data Mining Information Systems Systems Design Neural Networks Project Duración(Meses): 12 País: Reino Unido costo(\$US): \$19,646.17 Lenguaje: ND Software: SAS Idioma: Ingles	Institución: The University of Tennessee url: http://bas.utk.edu/academic-programs/masters/business-analytics/default.asp Modalidad: Presencial Grado: Maestría Título: Master's in Business Analytics Temario: Professional Skills Decision Analytics Data Analysis Statistics System Management Management Accounting Regression Models Data Mining Business Analytics Chain Management Simulation Data Processing Time Series SAS Programming Duración(Meses): 24 País: EUA costo(\$US): \$2,222.00 Lenguaje: Gurobi R MySQL Software: Excel, AMPL, JMP, Tableau Idioma: Ingles
Institución: Université Lumière Lyon2, Université de Lyon url: http://dea-e.cd.univ-lyon2.fr/ Modalidad: En Línea Grado: Maestría Título: Master of Science Data Studies Temario: Data Mining Databases Data Analysis Machine Learning Duración(Meses): 12 País: Francia costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Institución: Elmhurst College url: http://public.elmhurst.edu/data_science Modalidad: En Línea Grado: Maestría Título: MS in Data Science Temario: Data Analysis Databases Data Mining Programming Analytical Methods Research Methods Duración(Meses): 24 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles
Institución: Coventry University url: http://www.coventry.ac.uk/course-structure/2014/faculty-of-engineering-and-computing/postgraduate/data-science-and-computational-intelligence-msc/ Modalidad: Presencial Grado: Maestría Título: Data Science and Computational Intelligence MSc Temario: Neural Networks Machine Learning Fuzzy logic Data Mining Business Intelligence Cloud Computing Project Management Systems Development Duración(Meses): 12	Institución: Southern Methodist University url: http://datascience.smu.edu/ Modalidad: En Línea Grado: Maestría Título: Master Business Administration in Data Analytics Temario: Statistics Databases Visualization Security Data Mining Data Science Economics Data Analysis Duración(Meses): 24 País: EUA costo(\$US): \$52,824.00 Lenguaje: ND Software: ND

País: Reino Unido costo(\$US): \$10,121.00 Lenguaje: ND Software: ND Idioma: Ingles	Idioma: Ingles
Institución: Heriot Watt University url: http://www.macs.hw.ac.uk/cs/pgcourses/ds.htm Modalidad: Presencial Grado: Maestría Título: MSc IN Data Science Temario: 3D Modelling and Animation Biologically Inspired Computation Security Data Mining Visualization Software Engineering Interaction Design Big Data Distributed Systems Research Methods Duración(Meses): 9 País: Reino Unido costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Institución: Danube University Krems url: http://www.donau-uni.ac.at/en/studium/data-studies/index.php Modalidad: Mixto Grado: Maestría Título: Master of Science Data Studies Temario: Data Analysis Data Processing Humanities Data Privacy Law Project Information Systems Research Methods Interactive and Integrated Collaboration Duración(Meses): 24 País: Austria costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles
Institución: University of Reading url: http://www.reading.ac.uk/sse/masters/sse-mscadvancedcomputerscience.aspx Modalidad: Presencial Grado: Maestría Título: MSc Advanced Computer Science Temario: Mathematical Statistics Research Methods Data Analysis Data Mining Big Data Analytics Cloud Computing GPU Computing Project Brain Computer Interface Neural Networks Artificial Intelligence Image Processing Visualization Entrepreneurship Social Network Analysis Data Privacy Law Duración(Meses): 24 País: Reino Unido costo(\$US): \$26,570.64 Lenguaje: ND Software: ND Idioma: Ingles	Institución: Deakin University Australia url: http://www.deakin.edu.au/buslaw/information-business-analytics/courses/business-analytics.php Modalidad: Presencial Grado: Maestría Título: Master of IT in Business – Analytics (MITB-AT) Temario: Information Management Descriptive Analytics Visualization Databases Predictive Analytics Business Intelligence Decision Analytics Duración(Meses): 18 País: Australia costo(\$US): \$26,320.00 Lenguaje: ND Software: ND Idioma: Ingles
Institución: Barcelona Graduate School of economics url: http://www.barcelonagse.eu/master-data-science.html Modalidad: Presencial Grado: Maestría Título: Master in Data Science Temario: Statistics Optimization Databases	Institución: City University of New York School of Professional Studies url: http://sps.cuny.edu/programs/ms_dataanalytics Modalidad: En Línea Grado: Maestría Título: Online Master's Degree in Data Analytics (MS) Temario: Programming Simulation Mathematical

<p>Microeconomics Machine Learning Econometrics Probability Visualization Project Big Data Analytics Social Network Analysis Data Mining Data Analysis Public Police</p> <p>Duración(Meses): 9 País: España costo(\$US): \$17,831.80 Lenguaje: ND Software: ND Idioma: Ingles</p>	<p>Statistics Data Acquisition Project Innovation and Strategy Web Analytics Business Analytics Data Mining Databases Data Analysis Information Systems Technology Urban Sustainability</p> <p>Duración(Meses): 18 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles</p>
<p>Institución: Lancaster University url: http://www.lancaster.ac.uk/data-science/ Modalidad: Presencial Grado: Maestría Título: MSc Data Science Computing Specialism Temario: Data Science Data Mining Programming Statistics Distributed Systems Business Intelligence Dissertation Duración(Meses): 12 País: Reino Unido costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles</p>	<p>Institución: The University of Auckland url: https://cdn.auckland.ac.nz/assets/science/about/our-faculty/prospectuses-handbooks/pdfs/2013-data-science-flyer.pdf Modalidad: Presencial Grado: Maestría Título: Master of Professional Studies in Data Science (MProfStuds) Temario: Programming Computer Science Practical Computing Computer Systems Algorithms Data Structures Mathematical Software Development Software Engineering Computer Architecture Networks Distributed Systems Operating Systems Human Computer Interaction Databases Artificial Intelligence Image Processing Duración(Meses): 12 País: Australia costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles</p>
<p>Institución: University of Virginia url: http://dsi.virginia.edu/academics Modalidad: Presencial Grado: Maestría Título: Master of Science in Data Science Temario: Programming Statistics Linear Models Algorithms Data Science Ethics Data Mining Cloud Computing Machine Learning Modeling & Simulation Data Analysis Risk Analysis Optimization Probability Databases Time Series Stochastic Processes Resampling Methods Duración(Meses): 11 País: EUA costo(\$US): ND Lenguaje: ND Software: ND</p>	<p>Institución: York University Schulich School of Business url: http://www.schulich.yorku.ca/client/schulich/Schulich_LP4W_LND_WebStation.nsf/page/MScBA+Program+Length+and+Curriculum?OpenDocument Modalidad: Presencial Grado: Maestría Título: Master of Business Analytics Temario: Predictive Modeling Data Management Programming Professional Skills Research Methods Analytics Presentation Skills Data Analysis Project Duración(Meses): 24 País: Canada costo(\$US): \$61,320.48 Lenguaje: ND Software: ND</p>

Idioma: Ingles Institución: The University of Sheffield url: http://www.sheffield.ac.uk/is/pgt/courses/data_science Modalidad: Presencial Grado: Maestría Título: MSc Data Science Temario: Data Science Data Analysis Data Mining Databases Research Methods Information Systems Business Intelligence Data Processing Dissertation Duración(Meses): 12 País: Reino Unido costo(\$US): \$25,137.44 Lenguaje: ND Software: ND Idioma: Ingles	Idioma: Ingles Institución: Universidad Centrar de Venezuela url: http://triton.exp.dc.uba.ar/datamining/ Modalidad: Presencial Grado: Maestría Título: Maestría en Modelos Aleatorios Temario: Probability Statistics Computacional Methods Random Process Project Simulation Mathematical Neural Networks Test Design Sampling Bayesian Methods Data Analysis Time Series Data Mining Information Systems Duración(Meses): 24 País: Argentina costo(\$US): ND Lenguaje: ND Software: ND Idioma: Español
Institución: Danube University Krems url: http://www.donau-uni.ac.at/en/studium/data-studies/index.php Modalidad: Mixto Grado: Maestría Título: Master of Science Data Studies Temario: Data Analysis Data Processing Humanities Research Methods Data Privacy Law Project Information Systems Interactive and Integrated Collaboration Duración(Meses): 24 País: Austria costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Institución: Worcester Polytechnic Institute url: http://www.wpi.edu/academics/datascience/degree-requirements.html Modalidad: En Línea Grado: Maestría Título: MS Degree Program in Data Science Temario: Data Science Statistics Regresion Models Multivariate Analysis Databases Big Data Data Mining Machine Learning Big Data Systems Business Intelligence Duración(Meses): 24 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles
Institución: Goldsmiths University of London url: http://www.gold.ac.uk/pg/msc-data-science/ Modalidad: Presencial Grado: Maestría Título: MSc in Data Science Temario: Machine Learning Big Data Data Science Neural Networks Natural Language Processing Semantic Web Data Compression Artificial Intelligence Open Data Data Mining Duración(Meses): 24 País: Reino Unido costo(\$US): \$16,425.49 Lenguaje: Python R Pig Software: Matlab, Spark,SPSS,Hadoop	Institución: College of Charleston url: http://www.cofc.edu/academics/majorsandminors/data-science.php Modalidad: Presencial Grado: Licenciatura Título: Data Science Temario: Data Science Data Management Calculus Linear Algebra Data Structures Statistics Duración(Meses): ND País: EUA costo(\$US): ND Lenguaje: ND Software: ND

<p>MongoDB,Hadoop MongoDB</p> <p>Idioma: Ingles</p>	<p>Idioma: Ingles</p>
<p>Institución: University of San Francisco College of Arts and Sciences</p> <p>url: http://www.usfca.edu/artsci/bsds/</p> <p>Modalidad: Mixto</p> <p>Grado: Licenciatura</p> <p>Título: The Major in Data Science</p> <p>Temario: Programming Algorithms Data Structures Visualization Calculus Linear Algebra Mathematical Probability Statistics Macroeconomics Software Development Databases Data Mining</p> <p>Duración(Meses): 60</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>	<p>Institución: Northern Kentucky University</p> <p>url: http://informatics.nku.edu/departments/computer-science/programs/datascience.html</p> <p>Modalidad: Presencial</p> <p>Grado: Licenciatura</p> <p>Título: Bachelor of Science in Data Science.</p> <p>Temario: Calculus Programming Probability Culture and Creativity Written Communication Big Data Analysis Data Structures Linear Algebra Business Process Natural Science Data Mining Project Management Data Analysis Culture and Creativity Data Science Cultural Pluralism</p> <p>Duración(Meses): 48</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>
<p>Institución: Penn State University</p> <p>url: http://bdss.psu.edu/education</p> <p>Modalidad: Presencial</p> <p>Grado: Doctorado</p> <p>Título: Social Data Analytics</p> <p>Temario: ND</p> <p>Duración(Meses): 24</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>	<p>Institución: University Technology Sydney</p> <p>url: http://www.uts.edu.au/future-students/find-a-course/courses/c03051</p> <p>Modalidad: Presencial</p> <p>Grado: Doctorado</p> <p>Título: Analytics</p> <p>Temario: ND</p> <p>Duración(Meses): 48</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>
<p>Institución: Aarhus University</p> <p>url: http://icoa.au.dk/news/single/artikel/seeking-candidate-for-industrial-phd/</p> <p>Modalidad: Presencial</p> <p>Grado: Doctorado</p> <p>Título: Industrial phd in Big Data Analysis</p> <p>Temario: ND</p> <p>Duración(Meses): ND</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>	<p>Institución: University College London</p> <p>url: http://www.ucl.ac.uk/statistics/news/news-010212</p> <p>Modalidad: Presencial</p> <p>Grado: Doctorado</p> <p>Título: Studentships in Statistics Available: "Big Data" and its Applications</p> <p>Temario: ND</p> <p>Duración(Meses): 36</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>

Institución: University of Washington url: http://escience.washington.edu/blog/new-phd-tracks-big-data Modalidad: Presencial Grado: Doctorado Título: Track's in "Big Data" Temario: Big Data Machine Learning Visualization Statistics Duración(Meses): ND País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Institución: Brown Univesyti url: http://cs.brown.edu/~kraskat/phd.html Modalidad: Presencial Grado: Doctorado Título: Biga Data Temario: ND Duración(Meses): ND País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles
Institución: University of Southern California url: http://www.marshall.usc.edu/phd/fields/iom/requirements Modalidad: En Línea Grado: Doctorado Título: Data Sciences & Operation Temario: Information Systems Economics Stochastic Processes Programming Presentation Skills Project Networks Written Communication Duración(Meses): 48 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Institución: Gerge Mason University url: http://spacs.gmu.edu/category/academics/graduate-programs/ Modalidad: Presencial Grado: Doctorado Título: Informatics or Data Science Temario: Numerical Methods Databases Computer Science Visualization Duración(Meses): 48 País: EUA costo(\$US): ND Lenguaje: C C++ Fotran Software: ND Idioma: Ingles
Institución: Colorado Technical University url: http://www.coloradotech.edu/degrees/doctorates/computer-science/big-data-analytics Modalidad: En Línea Grado: Doctorado Título: Doctor in Computer Science Big Data Analitics Temario: Big Data Analytics Computer Systems Data Processing General Concentration Duración(Meses): 36 País: EUA costo(\$US): \$57,000.00 Lenguaje: ND Software: ND Idioma: Ingles	Institución: Newcastle University url: http://teaching.ncl.ac.uk/ccfbd-cdt/ Modalidad: Presencial Grado: Doctorado Título: Cloud Computing for Big Data Temario: ND Duración(Meses): 48 País: EUA costo(\$US): \$18,965.39 Lenguaje: ND Software: ND Idioma: Ingles
Institución: University of Washington url:	Institución: University of Rochester url: http://www.rochester.edu/data-

http://www.stat.washington.edu/graduate/programs/machinelearning/ Modalidad: Presencial Grado: Doctorado Título: Machine Learning and Big Data Temario: Statistics Machine Learning Graphical Models Visualization Databases Optimization Duración(Meses): ND País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	science/degrees/index.html Modalidad: Presencial Grado: Doctorado Título: Data Science Temario: ND Duración(Meses): 48 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles
Institución: Carnegie Mellon University url: http://www.ml.cmu.edu/prospective-students/joint-phd-mlstat.html Modalidad: En Línea Grado: Doctorado Título: Join PhD Program Statistics & Machine Learning Temario: Statistics Machine Learning Regression Analysis Probability ADA Graphical Models Optimization Databases Algorithms Duración(Meses): 24 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Institución: Big Data University url: http://bigdatauniversity.com/bdu-wp/bdu-course/big-data-fundamentals/ Modalidad: En Línea Grado: Curso Título: Big Data Fundamentals Temario: Big Data Data Analysis Duración(Meses): 0.04 País: Internacional costo(\$US): \$0.00 Lenguaje: ND Software: Linux VMWare Idioma: Ingles
Institución: Big Data University url: http://bigdatauniversity.com/bdu-wp/bdu-course/data-mining-with-r-let-r-rattle-you/ Modalidad: En Línea Grado: Curso Título: Data Mining with R – Let R ‘rattle’ you Temario: R Duración(Meses): 0.04 País: Internacional costo(\$US): \$0.00 Lenguaje: R Software: ND Idioma: Ingles	Institución: Connecticut College url: http://www.conncoll.edu Modalidad: Presencial Grado: Curso Título: Data Mining Course Temario: Machine Learning Classification Evaluation Business Data Processing Clustering Data Mining Visualization Data Analysis Duración(Meses): 6 País: EUA costo(\$US): ND Lenguaje: ND Software: Weka Idioma: Ingles
Institución: Udacity url: https://www.udacity.com/courses#!/data-science Modalidad: En Línea Grado: Curso Título: Data Science	Institución: Big Data University url: http://bigdatauniversity.com/bdu-wp/bdu-course/introduction-to-pig/ Modalidad: En Línea Grado: Curso

Temario: Computer Science Hadoop Data Science R Statistics Machine Learning NoSQL Duración(Meses): 21 País: EUA costo(\$US): \$2,200.00 Lenguaje: Python R MapReduce Software: Hadoop MongoDB Idioma: Ingles	Título: Introduction to Pig Temario: Pig Duración(Meses): 0.04 País: EUA costo(\$US): \$0.00 Lenguaje: ND Software: Pig Idioma: Ingles
Institución: Nova Southeastern University url: http://scis.nova.edu/masters/certificate_bi.html Modalidad: Mixto Grado: Certificación Título: Graduate Certificate In Business Intelligence/ Analytics Temario: Decision Analytics Databases Data Mining Duración(Meses): 7 País: EUA costo(\$US): ND Lenguaje: ND Software: ND Idioma: Ingles	Institución: DIGITAL ANALYTICS ASSOCIATION url: http://www.digitalanalyticsassociation.org/certification Modalidad: Presencial Grado: Certificación Título: Digital Analytics Association Web Analyst Certification Program Temario: Web Analytics Marketing Business Duración(Meses): ND País: Internacional costo(\$US): \$795.00 Lenguaje: ND Software: ND Idioma: Ingles
Institución: University of Illinois at Urbana-Champaign url: http://mias.illinois.edu/index.php?q=page/core-course Modalidad: En Línea Grado: Certificación Título: Foundations of Data Sciences Temario: Probability Statistics Entropy Information Systems Optimization Clustering Data Processing Linear Models Duración(Meses): 1.5 País: EUA costo(\$US): ND Lenguaje: Java Software: SNow JLIIS Idioma: Ingles	Institución: Cloudera url: http://cloudera.com/content/cloudera/en/training/certification/ccp-ds/essentials.html Modalidad: En Línea Grado: Certificación Título: Cloudera Certified Professional: Data Scientist Temario: Data Acquisition Data Analysis Data Processing Machine Learning Clustering Classification Analytics Modeling Probability Visualization Optimization Duración(Meses): 6 País: Internacional costo(\$US): \$200.00 Lenguaje: ND Software: ND Idioma: Ingles
Institución: University of Washington url: http://www.pce.uw.edu/certificates/data-science.html Modalidad: Mixto Grado: Certificación Título: Data Science Temario: Data Science Data Analysis Descriptive	Institución: Stanford url: http://scpd.stanford.edu/public/category/courseCategoryCertificateProfile.do?method=load&certificateId=10555807 Modalidad: Mixto Grado: Certificación Título: Mining Massive Data Sets Graduate

<p>Analytics</p> <p>Duración(Meses): 10</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>	<p>Certificate</p> <p>Temario: Social Network Analysis Machine Learning Data Mining</p> <p>Duración(Meses): 24</p> <p>País: EUA</p> <p>costo(\$US): \$19,800.00</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>
<p>Institución: University of Washington</p> <p>url: http://www.pce.uw.edu/certificates/data-science.html</p> <p>Modalidad: Mixto</p> <p>Grado: Certificación</p> <p>Título: Certificate in Data Science</p> <p>Temario: Data Science Data Analysis Descriptive Analytics</p> <p>Duración(Meses): 24</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>	<p>Institución: Statistics.com, The Institute for Statistics Education</p> <p>url: http://www.statistics.com/programming-for-data-science</p> <p>Modalidad: En Línea</p> <p>Grado: Certificación</p> <p>Título: Programming for Data Science Certificate Program</p> <p>Temario: Predictive Analytics Data Mining Hadoop Python Programming R Databases Natural Language Processing Data Analysis GIS</p> <p>Duración(Meses): 24</p> <p>País: EUA</p> <p>costo(\$US): \$3,852.00</p> <p>Lenguaje: R Python SQL</p> <p>Software: CrowANALYTIX Hadoop Ggplot2 SAS</p> <p>Idioma: Ingles</p>
<p>Institución: Statistics.com, The Institute for Statistics Education</p> <p>url: http://www.statistics.com/analytics-for-data-science</p> <p>Modalidad: En Línea</p> <p>Grado: Certificación</p> <p>Título: Analytics for Data Science Certificate Program</p> <p>Temario: Data Processing Visualization Optimization Social Network Analysis Predictive Analytics Risk Analysis Analytics Decision Models Linear Algebra Natural Language Processing Data Analytics GIS</p> <p>Duración(Meses): 12</p> <p>País: EUA</p> <p>costo(\$US): \$9,333.00</p> <p>Lenguaje: ND</p> <p>Software: Data Mining add for Excel XLMiner ArcGIS</p> <p>Idioma: Ingles</p>	<p>Institución: Standord Center for Professional Development</p> <p>url: http://scpd.stanford.edu/public/category/courseCategoryCertificateProfile.do?method=load&certificateId=10555807</p> <p>Modalidad: En Línea</p> <p>Grado: Certificación</p> <p>Título: Mining Massive Data Sets Graduate Certificate</p> <p>Temario: Social Network Analysis Statistics Data Mining</p> <p>Duración(Meses): 24</p> <p>País: EUA</p> <p>costo(\$US): \$19,800.00</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>
<p>Institución: Standord Center for Professional Development</p>	<p>Institución: NJIT ONLINE</p> <p>url: http://online.njit.edu/programs/certs/datamining-</p>

<p>url:http://scpd.stanford.edu/public/category/courseCategoryCertificateProfile.do?method=load&certificateId=1209602</p> <p>Modalidad: En Línea</p> <p>Grado: Certificación</p> <p>Título: Data Mining and Applications Graduate Certificate</p> <p>Temario: Data Mining Statistics Paradigms for Computing with Data</p> <p>Duración(Meses): 24</p> <p>País: EUA</p> <p>costo(\$US): \$11,880.00</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>	<p>cert.php</p> <p>Modalidad: En Línea</p> <p>Grado: Certificación</p> <p>Título: Certificate in Data Mining</p> <p>Temario: Computer Science Computer Architecture Operating Systems Networks Databases Calculus Probability</p> <p>Duración(Meses): 12</p> <p>País: EUA</p> <p>costo(\$US): \$7,449.00</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>
<p>Institución: SPEARS School of Business Oklahoma State University</p> <p>url: http://analytics.okstate.edu/</p> <p>Modalidad: Mixto</p> <p>Grado: Certificación</p> <p>Título: SAS and OSU Data Mining Certificate Program</p> <p>Temario: Databases Data Mining SAS Programming Research Methods Analytics Business Intelligence Statistics</p> <p>Duración(Meses): 6</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: Python R</p> <p>Software: MapReduce/Hadoop</p> <p>Idioma: Ingles</p>	<p>Institución: RockHurst University</p> <p>url: http://analytics.okstate.edu/</p> <p>Modalidad: Presencial</p> <p>Grado: Certificación</p> <p>Título: Data Science Certificate</p> <p>Temario: Business Intelligence Data Mining Visualization Predictive Analytics Data Processing Social Network Analysis Analytics</p> <p>Duración(Meses): 15</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: R Python MySQL Oracle</p> <p>Software: Tableau MicroStrategy Google Analytics RStudio</p> <p>Idioma: Ingles</p>
<p>Institución: TDWI</p> <p>url: http://tdwi.org/cbip</p> <p>Modalidad: Presencial</p> <p>Grado: Certificación</p> <p>Título: Certified Business Intelligence Professional</p> <p>Temario: Business Analytics Data Management Data Analysis Leadership Information Systems Databases</p> <p>Duración(Meses): ND</p> <p>País: Internacional</p> <p>costo(\$US): \$125.00</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>	<p>Institución: Indian School of Business ISB, Executive Education</p> <p>url: http://executive-education.isb.edu/certificate-programmes/certificate-programme-in-business-analytics.html</p> <p>Modalidad: Presencial</p> <p>Grado: Certificación</p> <p>Título: Certificate Programme in Business Analytics</p> <p>Temario: Big Data Business Analytics Data Analysis</p> <p>Duración(Meses): 6</p> <p>País: India</p> <p>costo(\$US): \$9,789.92</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>
<p>Institución: Revolution Analytics</p>	<p>Institución: EMC2 PROVEN PROFESSIONAL</p>

<p>url: http://www.revolutionanalytics.com/academyr-specialist-certification</p> <p>Modalidad: Presencial</p> <p>Grado: Certificación</p> <p>Título: Academy R-Specialist Certification</p> <p>Temario: Big Data Analytics R Big Data Statistics Big Data Analysis</p> <p>Duración(Meses): ND</p> <p>País: Internacional</p> <p>costo(\$US): \$200.00</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>	<p>url: https://education.emc.com/guest/campaign/data_science.aspx</p> <p>Modalidad: En Línea</p> <p>Grado: Certificación</p> <p>Título: Data Science and Big Data Analytics</p> <p>Temario: Big Data Analytics Data Science Big Data Analytics Key Roles for a Succesful Analytic Project Project Lifecicle Developing core deliverables for stakeholders R Statistics Bayesian Methods Clustering Predictive Analytics Classification Linear Algebra Time Series Data Analysis Hadoop Databases Project Management Visualization Big Data Systems</p> <p>Duración(Meses): ND</p> <p>País: Internacional</p> <p>costo(\$US): \$600.00</p> <p>Lenguaje: ND</p> <p>Software: ND</p> <p>Idioma: Ingles</p>
<p>Institución: University of California Irvine Extension (UCI)</p> <p>url: http://unex.uci.edu/areas/it/predictive_analytics/</p> <p>Modalidad: En Línea</p> <p>Grado: Certificación</p> <p>Título: Predictive Analytics Certificate Program Online</p> <p>Temario: Predictive Analytics Data Processing Algorithms Business Intelligence Data Analysis Data Mining Big Data Analytics Ensemble Methods Risk Analysis R Hadoop</p> <p>Duración(Meses): 12</p> <p>País: EUA</p> <p>costo(\$US): ND</p> <p>Lenguaje: R</p> <p>Software: Hadoop</p> <p>Idioma: Ingles</p>	<p>Institución: University of Wisconsin Milwaukee, Sheldon B. Lubar School of Business</p> <p>url: http://www4.uwm.edu/business/programs/certificates/business-analytics.cfm</p> <p>Modalidad: En Línea</p> <p>Grado: Certificación</p> <p>Título: Graduate Certificate in Business Analytics</p> <p>Temario: Machine Learning Data Analysis Data Mining Statistics Business Data Science Databases Mathematical Artificial Intelligence Visualization GIS Project</p> <p>Duración(Meses): 12</p> <p>País: EUA</p> <p>costo(\$US): \$2,500.00</p> <p>Lenguaje: Python R SQL Server BI</p> <p>Software: OLAP Business Objects Excel Data Mining Client BW SPSS SQL Server BI Development Studio SAS SAP</p> <p>Idioma: Ingles</p>
<p>Institución: Big Data University</p> <p>url: http://bigdatauniversity.com/</p> <p>Modalidad: En Línea</p> <p>Grado: Certificación</p> <p>Título: Hadoop Fundamentals I V3</p> <p>Temario: Hadoop</p> <p>Duración(Meses): 0.04</p> <p>País: EUA</p>	

costo(\$US): \$0.00 Lenguaje: ND Software: Linux VMWare Idioma: Ingles	
---	--

**Anexo C: Autorización de publicación en formato electrónico de
reporte técnico.**



CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS, A.C.

**BIBLIOTECA
AUTORIZACION
PUBLICACION EN FORMATO ELECTRONICO DE TESIS**

El que suscribe

Autor(es) de la tesis: -----

Título de la tesis: -----

Institución y Lugar: -----

Grado Académico: Licenciatura () Maestría () Doctorado () Otro ()

Año de presentación: -----

Área de Especialidad: -----

Director(es) de Tesis: -----

Correo electrónico: -----

Domicilio: -----

Palabra(s) Clave(s): -----

Por medio del presente documento autorizo en forma gratuita a que la Tesis arriba citada sea divulgada y reproducida para publicarla mediante almacenamiento electrónico que permita acceso al público a leerla y conocerla visualmente, así como a comunicarla públicamente en la Página WEB del CIMAT.

La vigencia de la presente autorización es por un periodo de 3 años a partir de la firma de presente instrumento, quedando en el entendido de que dicho plazo podrá prorrogar automáticamente por periodos iguales, si durante dicho tiempo no se revoca la autorización por escrito con acuse de recibo de parte de alguna autoridad del CIMAT

La única contraprestación que condiciona la presente autorización es la del reconocimiento del nombre del autor en la publicación que se haga de la misma.

Atentamente

Nombre y firma del tesista

CALLE JALISCO S/N MINERAL DE VALENCIANA APDO. POSTAL 402
C.P. 36240 GUANAJUATO, GTO., MÉXICO

TELÉFONOS (473) 732 7155, (473) 735 0800 EXT. 49609 FAX. (473) 732 5749

E-mail: biblioteca@cimat.mx

