Plataforma Big Data para el análisis de métricas sociales mundiales

Jordi Contestí Llull

Máster en Inteligencia de Negocio y Big Data (2015 – 2017)

Big Data

Diego Miranda Saavedra

Josep Curto Díaz

Julio 2017

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Plataforma Big Data para el análisis de métricas sociales mundiales*
Nombre del autor:	*Jordi Contestí Llull*
Nombre del consultor/a:	*Diego Miranda Saavedra*
Nombre del PRA:	*Josep Curto Díaz*
Fecha de entrega (mm/aaaa):	07/2017
Titulación::	*Máster en Inteligencia de Negocio y Big Data (2015 - 2017)*
Área del Trabajo Final:	*Trabajo final de máster MIB-SI*
Idioma del trabajo:	*Español*
Palabras clave	*Big Data, psicometría, Twitter, World Happiness Report*

Resumen del Trabajo (máximo 250 palabras):

En la actualidad, existen muchas iniciativas diferentes destinadas a estudiar la psicometría humana e intentar comprender cómo evolucionan las sociedades desde el punto de vista de las propias personas y sus sentimientos.

En este sentido, la ONU publica anualmente un informe con la clasificación mundial de la felicidad de los países, en los que se tienen en cuenta indicadores de muchas fuentes diferentes.

Ahora bien, en la actualidad se está poniendo de manifiesto cada vez más que las redes sociales son un adecuado termómetro de las opiniones y sentimientos de parte de los ciudadanos.

Es por todo ello que este trabajo final de máster tiene como finalidad la implementación de un sistema Big Data que permita estudiar el nivel de felicidad de diferentes países a partir de los datos publicados en las redes sociales.

En concreto, el trabajo se centra en la importación de los mensajes publicados en la red social Twitter desde 8 países de habla hispana diferentes, entre los que se incluye España.

Posteriormente, se ha realizado un análisis de los datos recogidos y se han contrastado con los resultados obtenidos con el informe anual de la ONU, en el que se constata que existe una similitud entre los resultados de ambos sistemas.

Abstract (in English, 250 words or less):

Nowadays, there are many different initiatives designed to study human psychometry and try to understand how societies evolve from the point of view of people themselves and their feelings.

In this sense, the UN annually publishes a report with the world classification of countries' happiness, which takes into account indicators from many different sources.

However, it is increasingly becoming clear that social networks are an adequate thermometer of opinions and feelings on the part of citizens.

This is why, this final master's work has as its purpose the implementation of a Big Data system that allows to study the level of happiness of different countries from the data published in social networks.

Specifically, the project focuses on the import of messages published on the Twitter social network from 8 different Spanish-speaking countries, including Spain.

Subsequently, an analysis of the data collected has been carried out and contrasted with the results obtained with the annual report of the UN, in which it is verified that there is a similarity between the results of both systems.

Índice

Lista de figuras

Lista de tablas

1 Introducción

1.1 Contexto y justificación del trabajo

En los últimos años, desde diferentes instituciones como por ejemplo OECD (OECD, 2017) o Gallup (Gallup, 2017) se está intentando comprender la sociedad humana desde la perspectiva psicométrica. En estos casos, no solo se trata de comprender cómo evolucionan las sociedades humanas en el contexto de países, datos financieros o empresas, sino desde el punto de vista de las propias personas y sus sentimientos.

En este sentido, desde estas instituciones se han ido desarrollando índices vinculados a nivel de país que describen estados, desde la felicidad o bienestar hasta la calidad de vida de sus habitantes.

Todas estas iniciativas buscan proporcionar luz sobre un tema muy complejo. Sin embargo, actualmente la información por país que ofrecen este tipo de instituciones tiene una frecuencia anual e incluso en algunos casos mayor.

Además, toda esta información se publica de forma agregada anualmente, lo que dificulta su análisis posterior.

De este modo, resulta imposible para los analistas de este tipo de datos poder medir el impacto de acontecimientos puntuales, tales como las consecuencias de un acto terrorista o la publicación de unos resultados electorales, puesto que suceden de forma puntual.

Por otro lado, en la actualidad se están empezando a utilizar las redes sociales como termómetro de los sentimientos de un grupo de personas hacia determinadas empresas, productos o acontecimientos importantes. Del mismo modo, cabría la posibilidad de utilizar las redes sociales como fuente de información para valorar el sentimiento de felicidad o tristeza de los habitantes de un país.

Es por todo ello que sería posible complementar y contrastar la información ofrecida por todas estas instituciones con un análisis del sentimiento de felicidad capturado a partir de las redes sociales.

Actualmente, existen ejemplos como el conocido sitio web Hedonometer (University of Vermont Complex Systems Center; The MITRE Corporation, 2017), que muestra los datos de felicidad analizados en Twitter para EEUU. Ahora bien, esta web se centra únicamente en idioma inglés y está diseñada para cubrir los eventos que afectan a este país.

Con este propósito, este trabajo de fin de máster consiste en el diseño e implementación de un sistema de Big Data que permita recoger la información de los habitantes de diferentes países a partir de las redes sociales, analizar el sentimiento de felicidad a partir de todos estos datos para cada país y contrastar los resultados con los informes publicados por parte de las principales instituciones mundiales referentes al estudio psicométrico de diferentes países del mundo.

De este modo, será posible obtener información diaria del sentimiento de felicidad de un país. Toda esta información podría ser de utilidad no únicamente a las entidades que valoran el grado de felicidad mundial, como también a gobiernos, entidades públicas y empresas privadas que deseen reconocer patrones de felicidad entre los habitantes para mejorar los servicios prestados.

El análisis de los datos a partir de redes sociales abre la puerta a comprender la sociedad desde un punto de vista muy diferente del que hasta ahora han planteado muchas instituciones internacionales, basado en el análisis de los indicadores de países por expertos y la realización de encuestas. Las principales ventajas del método propuesto en este trabajo de fin de máster son las siguientes:

- Los datos de felicidad de un país podrían obtenerse en tiempo real.

- La realización de encuestas en los países resulta complejo. En cambio, desde este punto de vista la captura de datos de redes sociales es más sencilla.

- Tal como se explicará más adelante, los informes de todas estas instituciones contemplan multitud de parámetros asociados indirectamente a la felicidad, como por ejemplo, la corrupción. De todos modos, la percepción de felicidad es siempre algo muy subjetivo. En cambio, la utilización de las redes sociales permite estudiar los sentimientos de los ciudadanos directamente sin contemplar parámetros que quizá no se perciban como generadores de infelicidad por parte de la ciudadanía.

1.2 Objetivos del trabajo

El objetivo de este trabajo es el diseño e implementación de un sistema de Big Data que facilite la adquisición, el almacenamiento y la explotación de datos provenientes de la red social Twitter de diferentes países del mundo, valorar el sentimiento de felicidad o tristeza contenido en los datos capturados y analizar de forma combinada toda esta información con los informes de estado de felicidad y bienestar de las principales instituciones mundiales.

Por lo tanto, el trabajo podría descomponerse en los siguientes hitos:

1. En primer lugar, seleccionar un grupo de países a explorar. Para ello, se ha tomado la decisión de utilizar la posición actual en los informes de felicidad de diferentes países, de tal forma que la muestra contenga países de diferentes posiciones y así facilitar el análisis de datos posterior.

2. Después, diseñar un sistema de Big Data que permita capturar los tweets de diferentes países seleccionados de una forma simple, pero que permita ampliar cómodamente el proyecto a un número más alto de

países. Asimismo, el sistema debería ser lo más simple posible, sin perjudicar la capacidad de análisis de grandes cantidades de datos.

3. Estudiar y seleccionar los componentes que se usarán para el sistema de Big Data de forma que cubran todas las necesidades del proyecto.

4. Revisar el estado de arte de los componentes seleccionados para el sistema de Big Data, con el objetivo de seleccionar aquellos componentes más adecuados para el trabajo.

5. Implementar este sistema de Big Data de forma que permita la ingestión, el almacenamiento, el procesamiento y el análisis del dato, tanto de las fuentes actuales como posibles fuentes que se deseen integrar en el futuro.

6. Contrastar analíticamente la evolución diaria, semanal y general de la felicidad de los países respecto a su situación actual en los informes de felicidad de la ONU.

7. Dar respuesta las siguientes preguntas analíticas generales:

    a. Analizar el comportamiento diario de un país respecto a la felicidad comparado con su situación en la clasificación de felicidad a nivel mundial, ¿los países con una mejor nota tienen un comportamiento respecto a la felicidad más regular?, ¿los países peor valorados en los informes de felicidad tienen igualmente un índice de felicidad menor en las redes sociales?

    b. Detectar patrones comunes a todos los países analizados, como por ejemplo, ¿aumenta el nivel de felicidad durante los fines de semanas de forma general?

1.3 Enfoque y método seguido

La estrategia seguida en el desarrollo del trabajo ha sido la siguiente:

- Realizar una planificación de todos las tareas asociadas al trabajo, con la inclusión de los hitos de entregas previstos. Además, durante todo el proyecto se ha realizado un seguimiento de esta planificación y se han tomado las medidas necesarias para evitar el incumplimiento de la planificación.

- Definición de los requisitos funcionales y técnico del sistema Big Data a implementar.

- Estudiar las herramientas más recomendables a partir de los requisitos previamente definidos y aprender a utilizarlas. En este punto, se ha optado por adaptar productos existentes.

- Diseño del sistema Big Data, según las herramientas seleccionadas.

- Implementación del sistema Big Data y carga de los datos, con la realización de pruebas para poder validar que el sistema cumple con los requisitos funcionales y no funcionales acordados.

- Finalmente, diseño de visualizaciones y análisis de datos para responder a las preguntas analíticas planteadas.

Se ha descartado la implementación de un producto nuevo debido a que existen herramientas disponibles con licencias libres, así como por la gran complejidad que representaría implementar un sistema Big Data desde cero.

1.4 Planificación del trabajo

Para la planificación del trabajo se han tenido en cuenta los siguientes recursos:

- Dedicación semanal prevista del alumno responsable del trabajo. En todas las tareas se ha añadido un pequeño margen de tiempo para hacer frente a imprevistos.

- Servidor externo proporcionado por la UOC en el que albergar los componentes que formarán parte del sistema Big Data y de los datos asociados.

- PC de escritorio para la redacción de la memoria y la presentación.

En la siguiente imagen se puede observar la planificación temporal prevista mediante un diagrama de Gantt en el que se pueden encontrar los hitos parciales de entrega asociados al proyecto:

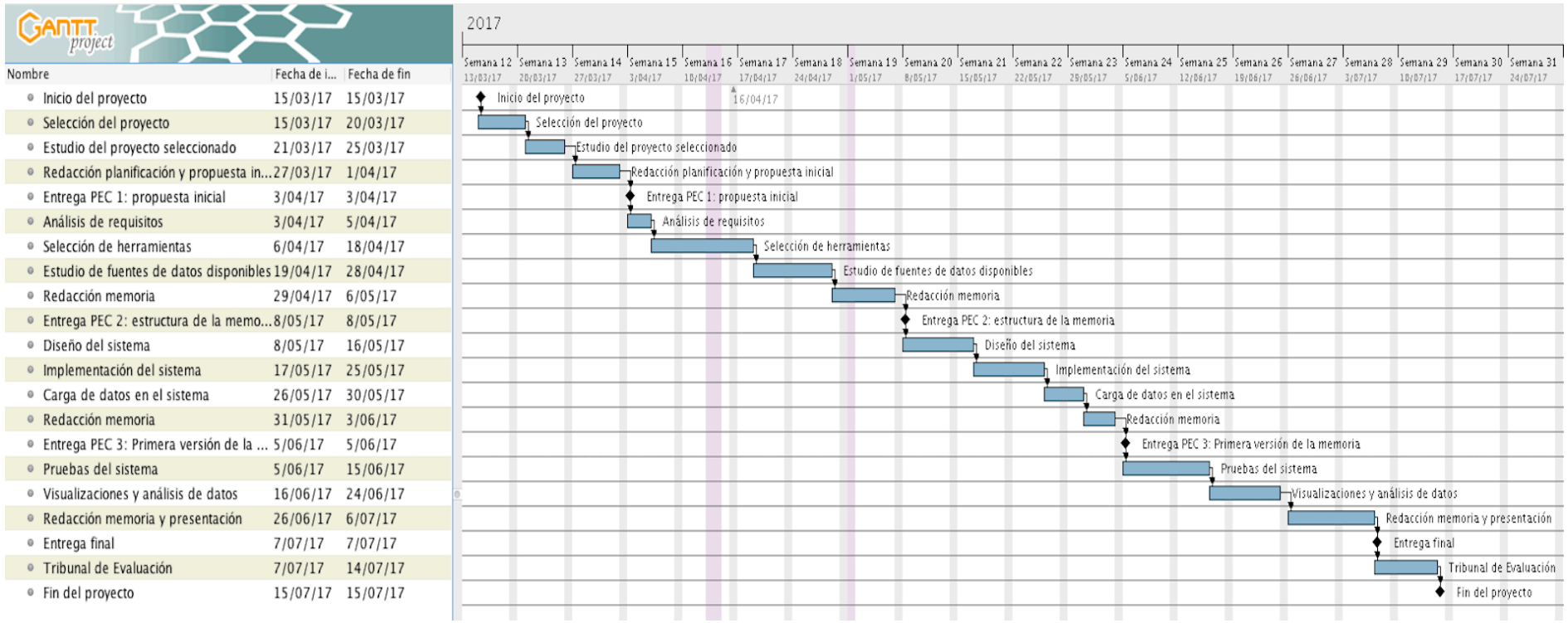

Figura 1. Planificación del proyecto

1.5 Breve sumario de productos obtenidos

Los productos obtenidos a partir de trabajo realizado son los siguientes:

- Listado de países seleccionados para el análisis de este proyecto.
- Sistema Big Data, formado por una selección de herramientas, una guía de instalación y configuración y los ficheros de código fuente de los programas preparados.
- Evolución diaria del sentimiento de felicidad de todos los países seleccionados.
- Visualizaciones y análisis de datos que permiten responder a las preguntas analíticas planteadas.
- Informes finales, que incluyen la presentación y esta memoria.

1.6 Breve descripción de los otros capítulos de la memoria

El resto de esta memoria se divide en capítulos explicados a continuación y que se incluyen a partir de este punto:

- Requisitos y alcance del proyecto. En este apartado se definen las necesidades técnicas y no funcionales contempladas en el proyecto desarrollado.
- Estado del arte y selección de herramientas. Por su parte, en este apartado se incluye un estudio del estado de actual de las herramientas necesarias para cubrir los requisitos del proyecto, las diversas opciones disponibles y la justificación de la herramienta seleccionada en cada caso.
- Diseño del sistema. Posteriormente, se especifica el diseño del sistema Big Data seleccionado, incluyendo las ventajas y desventajas del mismo.
- Fuentes de datos. Este apartado contiene la descripción de los países seleccionados y los motivos considerados para su selección.
- Implementación del sistema. Después, se explica el proceso de implementación seguido, restricciones encontradas y pruebas realizadas.
- Carga de datos en el sistema. En este apartado se incluye el proceso realizado para la carga de datos en el sistema Big Data y las dificultades encontradas.

- Análisis de datos y resultados. Por su parte, en este apartado se incluye la descripción del análisis de los datos ejecutado y los resultados obtenidos.

- Conclusiones. En este apartado se incluyen las conclusiones alcanzadas a partir de todo el trabajo realizado y también se incluyen trabajos que podrían realizarse en el futuro como ampliación y mejora del proyecto desarrollado.

- Bibliografía.

- Anexos.

2 Requisitos y alcance del proyecto

En este apartado se enuncian los requisitos que satisface el proyecto desarrollado.

Los requisitos del proyecto se han dividido en requisitos funcionales y requisitos técnicos, según se explican a continuación.

2.1 Requisitos funcionales

Los requisitos funcionales iniciales contemplados en este trabajo de fin de máster son los siguientes:

1. El sistema debe permitir la captura de tweets de todos los países configurados de forma diaria.

2. El sistema debe analizar los tweets recibidos y asociar un valor numérico que exprese el sentimiento de felicidad o tristeza del tweet.

3. No deben contemplarse los retweets en el análisis a realizar. El motivo es que se entiende que pueden no expresar un sentimiento personal como un tweet redactado directamente por una persona

4. En la medida de lo posible, debe procurarse descartar los usuarios que puedan corresponderse con bots de publicación de mensajes automáticos o spammers.

5. El sistema debe almacenar los siguientes datos como resultado del procesamiento de los tweets, para poder realizar el análisis de datos posterior:

    a. Identificador único del usuario.

    b. Identificador único del tweet.

    c. Texto del tweet.

    d. Valor numérico asociado a la felicidad.

    e. País asociado al tweet.

    f. Fecha de publicación del tweet: día, mes y año.

    g. Hora y minuto de la publicación del tweet.

    h. Idioma del tweet.

6. El análisis posterior debe contemplar la evolución diaria, semanal y mensual de la felicidad calculada por país y su comparación con las métricas internacionales de la ONU y OCDE. De esta forma, será posible contrastar la evolución de la felicidad de los países respecto a las métricas internacionales mencionadas.

2.2 Requisitos técnicos

Por su parte, los requisitos técnicos asociados a este trabajo de fin de máster se enumeran a continuación:

1. El sistema Big Data debe ser capaz de capturar de forma ininterrumpida todos los tweets publicados de forma diaria por cada uno de los países considerados en el análisis.

2. Las herramientas seleccionadas para el sistema Big Data deben ser Open Source, para facilitar la reutilización del resultado del trabajo y reducir los costes asociados.

3. El sistema debe poder recuperarse cuando la conexión con Twitter no esté disponible o devuelva algún tipo de error, de tal forma que el procesa sea relanzado de forma automática.

4. Deberá ser posible en el futuro ejecutar de nuevo un algoritmo de cálculo del sentimiento diferente sobre los mismos tweets, así como otros análisis adicionales. Por lo tanto, deben almacenarse los textos asociados a los tweets.

5. El análisis de los tweets debe realizarse mediante un procesamiento ejecutado de forma diaria, para poder obtener los resultados asociados en tiempo próximo al tiempo real.

6. El sistema debe permitir la adición de nuevos países fácilmente.

7. El sistema debe poder escalar fácilmente a volúmenes de datos mayores, tanto por la agregación de nuevos países, como por el aumento en la producción de tweets de los países configurados.

8. El sistema debe ser capaz de seguir funcionando correctamente a pesar de que puedan darse fallos en los nodos que componen el sistema.

3 Estado del arte y selección de herramientas

En este apartado se describe la situación actual de las herramientas necesarias para la implementación del sistema Big Data.

Todas las herramientas seleccionadas para este proyecto son Open Source.

Además, todas las aplicaciones seleccionadas permiten cubrir todas las necesidades asociadas al proyecto:

- En primer lugar, la captura de los tweets.
- Posteriormente, el procesamiento de los tweets.
- Almacenamiento del resultado del procesamiento de los tweets.
- Y finalmente el análisis y visualización de los datos almacenados.

En los siguientes apartados se explican los detalles de asociados al proceso de estudio y selección de herramientas realizado.

3.1 Captura de los tweets

Para la captura de los tweets, se han valorado las siguientes herramientas que podrían permitir la captura de los tweets:

- Utilización de un cliente básico implementado directamente mediante Python, que posteriormente enviaría los tweets recogidos al sistema de procesamiento.
- Apache Spark (The Apache Software Foundation, Spark), y en concreto, el módulo Spark Streaming.
- Apache Flume (The Apache Software Foundation, Flume).
- Y por último, Apache Kafka (The Apache Software Foundation, Kafka).

Para el proceso de captura de los tweets, la aplicación utilizada ha sido Apache Kafka, por los siguientes motivos:

- Para empezar, debemos descartar la implementación de un cliente básico mediante Python, puesto que se ha propuesto como requisito que el sistema sea capaz de escalar fácilmente a volúmenes de datos mayores.
- Por otro lado, Apache Spark es una herramienta que encaja muy bien en el procesamiento de los tweets. Por lo tanto, en esta situación el sistema de captura y procesamiento se ejecutaría sobre Apache Spark. Ahora bien, dado que se ha contemplado como requisito implementar un

sistema tolerante a fallos, esta arquitectura podría provocar que una caída de Apache Spark impida la captura de los tweets.

- Finalmente, Apache Flume y Apache Kakfa son dos herramientas que encajan en el proyecto, dado que permiten capturar los tweets correctamente y pueden ejecutarse en clúster. Apache Flume por su parte está diseñado para integrarse con las herramientas del ecosistema Apache Hadoop y en cambio Apache Kafka tiene un propósito mucho más general. Adicionalmente, Apache Kafka es un sistema con una gran capacidad de escalabilidad superior a Apache Flume. Puesto que se ha propuesto como requisito que el proyecto sea escalable fácilmente y se ha descartado Apache Hadoop, se ha decidido implementar el proceso de captura mediante Apache Kafka.

3.2 Procesamiento

Para el procesamiento, limpieza de los tweets y la ejecución del análisis de sentimiento de los tweets se ha seleccionado Apache Spark.

Apache Spark es una herramienta que goza actualmente de un alto nivel de implantación en los proyectos Big Data y exista mucha documentación sobre su configuración y utilización. Además, es una herramienta que durante el máster realizado se ha utilizado en diversas asignaturas y esto ha facilitado su utilización en este trabajo.

Adicionalmente, Apache Spark se integra correctamente con Apache Kafka. La integración entre Apache Spark y Apache Kafka se ha realizado mediante Python.

Después del tratamiento de los tweets mediante Apache Spark, los resultados procesados se guardan en una base de datos NoSQL, tal como se explica a continuación.

3.3 Almacenamiento

Para almacenar los resultados procesados se ha seleccionado el motor de base de datos NoSQL MongoDB (MongoDB, Inc.).

MongoDB es una base de datos NoSQL orientada a documentos y que puede ejecutarse de forma distribuida. MongoDB permite almacenar una gran cantidad de datos y es escalable al poder añadir más nodos en el caso que se desee, por lo que encaja adecuadamente con los requisitos del proyecto.

Además, al no tener un esquema fijo en el modelo de datos, permitirá que en el futuro sea posible modificar y ampliar la información registrada de forma sencilla.

Por otra parte es una base de datos que funciona muy bien para el almacenamiento de ficheros JSON, tal como se generan en el proyecto.

MongoDB se ha convertido en un estándar de las bases de datos NoSQL y es utilizada por grandes compañías en el mundo, tal como puede verse en su propia web. Además, existe una gran bibliografía referente a su uso y administración.

En MongoDB se almacenan los resultados de la captura diaria de los tweets desde Apache Kafka y procesados por Apache Spark.

La integración entre MongoDB y Apache Spark se ha realizado mediante Python.

3.4 Análisis y visualización de datos

Para el tratamiento de los tweets y el análisis posterior se ha seleccionado el lenguaje de programación Python, puesto que cuenta con numerosas librerías para el tratamiento y visualización de datos.

Concretamente, se ha seleccionado el sistema de notebooks Jupyter (Project Jupyter, 2017), que permite realizar análisis de datos y creación de gráficos mediante Python.

4 Diseño del sistema

En el siguiente esquema puede observarse una representación del sistema Big Data diseñado, en el que se incluyen todas las herramientas identificadas en el apartado anterior:

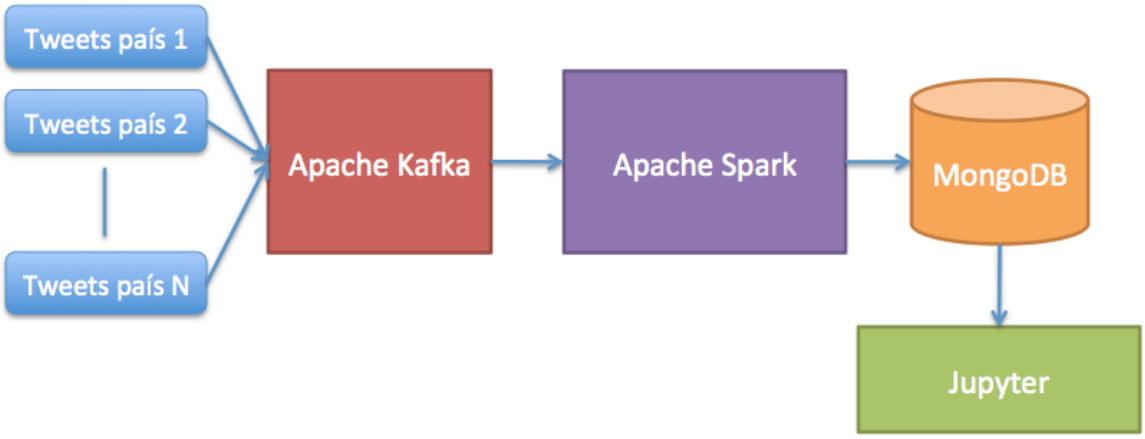

Figura 2. Diagrama del sistema Big Data

El proceso que se ha seguido es el siguiente:

1. Se ha preparado un cliente en Python que permite realizar la conexión con Twitter para la descarga de los tweets de los países seleccionados. Este cliente utiliza la librería tweepy (Roesslein, 2017) para conectarse a Twitter.

2. El cliente de Python envía a Apache Kafka todos los tweets de los diferentes países configurados.

3. Posteriormente, Apache Kafka se encarga de almacenar los tweets según la configuración preparada y durante un tiempo máximo configurado. En el caso de este trabajo, se ha configurado por un tiempo máximo de 7 días.

4. Después, automáticamente Apache Spark se encarga de leer los tweets almacenados en Apache Kafka y procesarlos convenientemente. Apache Spark carga los tweets en Resilient Distributed Datasets y los procesa de forma agrupada. En este momento, Apache Spark analiza el estado de felicidad de cada uno de los tweets publicados y extrae los datos necesarios según los requisitos indicados anteriormente.

5. Posteriormente, Apache Spark guarda todos los datos procesados en MongoDB, para que más adelante puedan realizarse las tareas de análisis deseadas.

6. Finalmente, mediante Jupyter y Python, se realiza la explotación y visualización de los datos cargados en MongoDB.

5 Fuentes de datos

Este apartado contiene la descripción de las fuentes de datos utilizadas en el sistema Big Data desarrollado.

Las fuentes de datos se dividen en los dos tipos siguientes:

- Redes sociales de diversos países seleccionados.
- Índices sobre la felicidad, bienestar y calidad de vida de los ciudadanos emitidos por algunas instituciones internacionales.

En los siguientes puntos se explican las fuentes de datos utilizadas en cada caso.

5.1 Redes sociales

Para la recogida de los datos de redes sociales se ha utilizado Twitter. Twitter es una red social que encaja perfectamente en los objetivos del proyecto, puesto que es una red social con muchos usuarios diversos y es muy usada en muchos países. Además, permite seleccionar los mensajes generados por localización geográfica, funcionalidad imprescindible para poder cargar los datos de los países analizados.

Los mensajes de Twitter recibidos a través del API de la red social se generan en formato JSON (Ecma International, 2013). Los campos de información que se necesitan para cumplir con el objetivo de este proyecto son los siguientes:

Nombre del campo	Descripción	Formato
id	Identificador del tweet	Numérico
user_id	Identificador del usuario	Numérico
text	Texto del tweet	Cadena de caracteres
country_code	Código del país	Cadena de caracteres
lang	Idioma del tweet	Cadena de caracteres
created_at	Fecha y hora de creación	Cadena de caracteres
followers_count	Número de followers	Numérico
friends_count	Número de usuarios	Numérico

	seguidos	
retweeted_status	¿Es un retweet?	Cadena de caracteres

Tabla 1. Campos de un tweet recogidos en la aplicación

Por otro lado, con el objetivo de seleccionar una muestra de países heterogénea, se han tomado las siguientes decisiones:

- Los países deberían estar situados en lugares diferentes dentro de los índices internacionales de felicidad, para poder tener una muestra lo más heterogénea de países posible. Específicamente, para realizar esta selección se ha utilizado el informe World Happiness Report de la ONU del año 2017.

- Habitualmente, los tweets del país deberían estar escritos en un idioma que sea posible procesar y valorar desde el punto de vista de felicidad. Hay que destacar que actualmente no existen diccionarios de sentimientos para cualquier idioma. Por todo ello, se ha seleccionado el idioma español como idioma a utilizar en el procesamiento de los tweets.

Los países seleccionados y la posición actual en el informe World Happiness Report de la ONU son los siguientes:

País	Posición
Costa Rica	12
Argentina	24
España	34
Ecuador	44
Bolivia	59
Paraguay	70
Venezuela	82
Honduras	91

Tabla 2. Países seleccionados en el estudio y posición en WHR

5.2 Índices de felicidad internacionales

Respecto a los índices de felicidad, los informes que se han utilizado para contrastar el análisis de los datos realizado son los siguientes:

Informe World Happiness Report (SDSN - ONU, 2017).

Este informe está financiado por la ONU y se fundamenta en los datos de la empresa Gallup. Actualmente está disponible todos los informes con los datos de la felicidad por países de forma anual, correspondientes a todos los años entre 2005 y 2016.

Todos los datos están disponibles en formato Microsoft Excel.

Se incluye información acerca 155 países, a los que se les asocia un coeficiente de felicidad numérico. Por lo tanto, desde el punto de vista de comparación con los datos de este trabajo de fin de máster, se tiene un coeficiente numérico asociado a cada país y diferente en función del año en cuestión.

Este coeficiente de felicidad se calcula a partir de numerosos indicadores recogidos por los investigadores del proyecto, tan variados como los siguientes:

- Sentimiento particular de felicidad, tales como el número de veces que un ciudadano ríe o la sensación de tristeza que experimenta durante el día.

- Salud, en el que se incluye la esperanza de vida, entre otros valores.

- Situación económica, en la que se incluye por ejemplo el nivel de desempleo, Producto Interior Bruto del país o volumen de donaciones por parte de los ciudadanos a la caridad, entre otros.

- Situación política y social, con la percepción de la corrupción, tipo de sistema político, nivel de confianza en el gobierno del país, sensación de inseguridad o confianza respecto a los conciudadanos.

Algunos datos se recogen a través de entrevistas realizadas a los ciudadanos de los respectivos países y otros se toman de informes internacionales de la situación de los países.

OECD Better Life Index (OECD, 2017), (OECD.Stat, 2016).

Este informe anual muestra datos sobre el nivel de vida de diferentes países, para los años comprendidos entre el 2013 y 2016.

Incluye los datos de todos los países que conforman la OECD, además de Brasil, Rusia y Sudáfrica.

Los formatos en los que se encuentran estos índices son los siguientes:

- Microsoft Excel.
- CSV.
- PC-Axis.
- XML - SDMX.

Los datos contenidos en este informe se componen de la siguiente información:

- Configuración de las viviendas: número de habitaciones por casa, gasto asociado a las viviendas y volumen de casa sin las condiciones mínimas para vivir.
- Ingresos de los ciudadanos por hogar.
- Empleo: nivel de empleo en los ciudadanos, grado de inseguridad en el mercado laboral, ratio de desempleo de larga duración e ingresos por ciudadanos.
- Comunidad: calidad de la red asistencial del país.
- Educación: años de educación y nivel de desempeño de los estudiantes.
- Entorno: contaminación del aire y calidad del agua.
- Compromiso civil: participación electoral y en la elaboración de reglamentos por parte de la ciudadanía.
- Salud: esperanza de vida.
- Seguridad: sentimiento de seguridad al caminar solo durante la noche y ratio de homicidios.
- Equilibrio entre vida personal y laboral: total de empleados que trabajen muchas horas al día y tiempo diario dedicado al ocio y al cuidado personal.

Finalmente, indicar que este informe se ha descartado por no ofrecer información de todos los países contemplados en este trabajo, puesto que no se encuentran dentro del grupo de países de la OECD.

6 Implementación del sistema

En este apartado se explican los pasos realizados para la implementación del sistema presentado.

6.1 Servidor

La Universitat Oberta de Catalunya ha cedido un servidor para la realización de esta trabajo con la siguiente configuración:

- Número de procesadores: 8 CPU.
- Memoria RAM disponible: 24GB.
- Disco: 100GB de espacio en disco.
- Sistema operativo: Ubuntu 16.04.2.

Todo el sistema Big Data se ha instalado sobre este servidor.

Puesto que las herramientas utilizadas pueden ejecutarse sobre un sistema distribuido, en el caso que fuera necesario, sería posible extender el sistema implementado al conjunto de servidores deseado de forma sencilla.

Finalmente, indicar que debido a las limitaciones de recursos, se han tenido que ajustar los parámetros de configuración de Apache Kafka y Apache Spark para evitar errores de falta de memoria durante la ejecución de la carga y procesamiento de los tweets recibidos.

6.2 Instalación de herramientas

Se ha procedido a la instalación y configuración de todas las herramientas indicadas en los apartados anteriores de forma manual sobre el servidor especificado. A saber:

- Apache Kafka, versión 0.10.2.1.
- Apache Spark, versión 2.1.1.
- MongoDB, versión 3.2.13. En este caso, MongoDB se ha instalado mediante el servicio de apt-get del propio sistema operativo.
- Python, versión 3.5.2.
- Jupyter, versión 4.3.0.

6.3 Integración con Twitter API

Twitter ofrece básicamente dos métodos de acceso a los tweets publicados:

- Mediante un API de acceso a los tweets en tiempo real, es decir, a medida que se van publicando.

- Y por otra parte, es posible realizar búsquedas entre los tweets publicados durante las últimas semanas.

Para la implementación de este trabajo se ha seleccionado el API que da acceso a los tweets en tiempo real por los siguientes motivos:

- En el espíritu de este trabajo está implementar un sistema que sea capaz de mostrar el estado de felicidad de un país de forma diaria, en contraposición a la información publicada por las instituciones internacionales, que se realiza con una frecuencia anual.

- El API de Twitter en tiempo real no tiene ninguna limitación de descarga de tweets. En cambio, el API de búsqueda de tweets tiene un límite que impide descargar tweets a partir de una configuración determinada.

Ahora bien, el principal inconveniente del API de recogida de tweets en tiempo real reside en las capacidades de filtrado de los tweets, porque son mucho más limitadas que el API de búsqueda.

Entre otras limitaciones, no es posible seleccionar los tweets según el campo de país asociado al tweet.

Es por ello que para seleccionar los tweets de los países en cuestión se ha preparado un listado de coordenadas geográficas de todos los países. Para realizar este proceso se ha utilizado el servicio de BoundingBox (Klokan Technologies, 2017), que posibilita cargar las coordenadas geográficas del país, como por ejemplo:

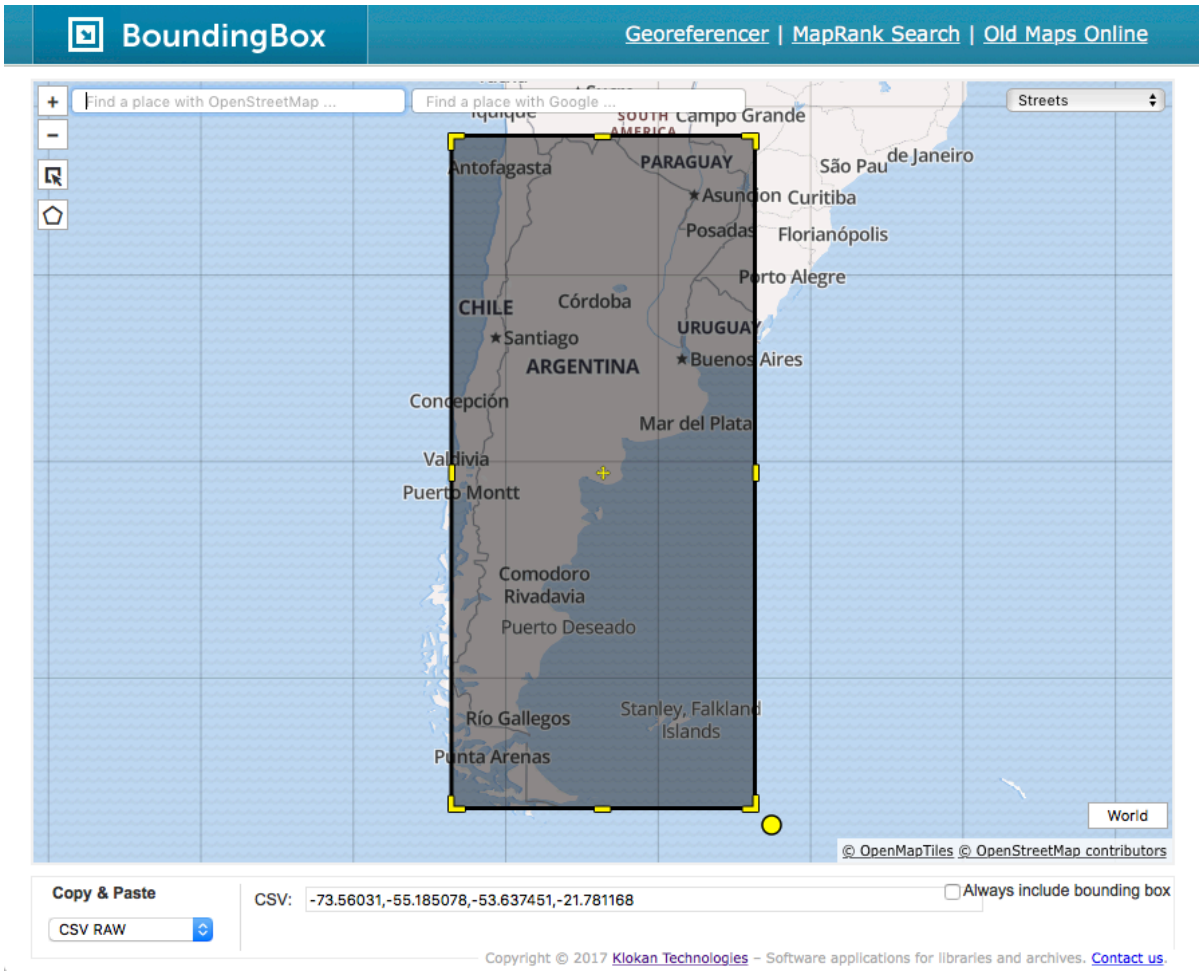

Figura 3. Muestra del funcionamiento de BoundingBox

Las coordenadas utilizadas para cada país han sido las siguientes:

País	Coordenadas asociadas
Costa Rica	-85.9351,8.2876,-82.6392,11.2323
Argentina	-73.56,-55.19,-53.64,-21.78
España	-6.5,36.28,3.3,43.68
Ecuador	-81.156,-5.0159,-75.1925,1.8822
Bolivia	-70.05,-22.98,-57.19,-9.47
Paraguay	-62.64,-27.61,-54.26,-19.29
Venezuela	-73.35,0.65,-59.54,15.92

Honduras	-88.9124,12.9939,-83.1555,16.2252

Tabla 3. Coordenadas utilizadas para todos los países estudiados

Para la implementación de la integración se ha utilizado la librería tweepy (Roesslein, 2017), porque permite acceder al API de Twitter sencillamente desde Python.

6.4 Tratamiento de datos masivos

Debido a las limitaciones del hardware empleado, se han tenido que ajustar las configuraciones de todas las herramientas utilizadas, según se describe a continuación:

- En Apache Kafka se ha activado la compresión de datos para evitar agotar el espacio disponible en el servidor.
- Por su parte, en Apache Spark se ha configurado un volumen máximo de mensajes a aceptar en streaming, de tal forma que se evitar saturar Apache Spark a partir del envío de mensajes desde Kafka. A partir de diversas pruebas, se ha ajustado a 100 mensajes máximo por envío y así evitar errores por falta de memoria en Apache Spark.
- También, en Apache Spark se ha limitado el espacio máximo ocupado por el resultado a 4GB.
- Asimismo, también en Apache Spark se han limitado el espacio de memoria RAM máximo ocupado por el driver y los executors a 8GB.
- Finalmente, también en Apache Spark se han configurado 2 cores como máximo.

7 Carga de datos en el sistema

En este apartado se explica el proceso seguido para la carga de los datos en el sistema.

7.1 Limpieza de datos

El proceso de limpieza de datos incluye la ejecución de las siguientes acciones:

- Para evitar capturar tweets de cuentas de usuario que puedan ser spam, se han descartado cuenta de usuarios con muy pocos seguidores comparado con las cuentas a las que siguen. Se ha seguido la siguiente fórmula, a partir de lo sugerido en (Verma, Divya, & Sofat, 2014):

  ratio = followers_count / friends_count

  Si este ratio es menor que 0,25, entonces el usuario se ha considerado un spammer y sus tweets se han ignorado. El motivo es que habitualmente los usuarios que realmente son spammers siguen a muchos usuarios (friends_count o también following), pero muy pocos usuarios los siguen a ellos (followers_count o followers). En caso contrario, se considera un usuario válido.

  No es el propósito de este trabajo de fin de máster la implementación de un sistema de detección de spammers más sofisticado, por lo que este comprobación se considera suficiente para filtrar las cuentas de posibles spammers.

  Además, cabe destacar también que en ocasiones Twitter puede devolver 0 en ambos campos en condiciones de alto estrés en la generación de tweets. Por este motivo, los tweets con valores 0 en estos dos campos se han considerado como válidos al no poder determinar si encajan dentro de la condición de posible spam.

- Además, tal como se recoge en los requisitos del trabajo, se han descartado los retweets porque se entiende que pueden no expresar un sentimiento personal como un tweet redactado directamente por una persona.

- Asimismo, aquellos tweets que no pertenecen a los países considerados en el análisis se han descartado. En este caso,

Posteriormente, se procede a procesar el mensaje de texto contenido en los tweets recibidos.

7.2 Procesamiento del mensaje de los tweets

Por su parte, el texto del tweet se procesa del siguiente modo:

- Inicialmente, el texto se convierte a minúsculas.
- Después, se cambian las vocales con tildes por vocales sin tildes.
- Posteriormente, las palabras se transforman en tokens a partir de los espacios en blanco y signos de puntuación.
- Como siguiente paso, se descartan los tokens que están considerados como palabras sin significado o stopwords.

Como resultado del proceso anterior se obtiene una lista de palabras que podrán utilizarse para analizar el sentimiento de positividad o negatividad contenido en el tweet.

En el siguiente paso se explica como se ha realizado el análisis de sentimiento de los tweets.

7.3 Análisis del sentimiento de los tweets

En primer lugar, debe destacarse que actualmente resulta muy complejo utilizar un sistema de análisis del sentimiento de los tweets que pueda catalogar los tweets en función del sentimiento de felicidad o tristeza contenido en el tweet en idioma español.

El principal motivo reside en que no existen buenos diccionarios a disposición del público que puedan utilizarse libremente.

Por este motivo, se ha tomado la decisión de utilizar un diccionario de detección de sentimientos positivos y negativos en un texto y asumir que en todo caso un sentimiento positivo se asocia a una sensación de felicidad y un sentimiento negativo por su parte a una sensación de tristeza.

En este trabajo el diccionario utilizado ha sido ElhPolar_esV1.lex (Saralegi, Vicente, & Foundation, 2013) (Saralegi & San Vicente, 2013), por encajar con los requisitos indicados.

Este diccionario incluye 5.211 términos catalogados según el sentimiento positivo o negativo asociado en idioma español, por lo que representa una herramienta adecuada para realizar la catalogación de los tweets en base al sentimiento asociado.

El proceso seguido para determinar el sentimiento de cada tweet ha sido el siguiente:

- En primer lugar, se ha cargado el diccionario de términos. En este punto se han descartado posibles palabras repetidas contenidas en el diccionario.

- Posteriormente, a partir de la lista de palabras procesadas, tal como se ha explicado en el punto anterior, se ha dado una valoración numérica a cada una de las palabras:

    - Si la palabra se encuentra en la lista de términos positivos del diccionario, se ha añadido a una lista de términos positivos asociados al tweet.

    - En cambio, si la palabra se encuentra en la lista de términos negativo del diccionario, se ha añadido a la lista de términos negativos asociados al tweet.

    - Si por el contrario la palabra no se encuentra en el diccionario, se ha descartado.

- Después, se ha realizado el cálculo siguiente: se ha restado el total de palabras negativas respecto al total de palabras positivas. De este modo, se ha obtenido un número entero, que debe interpretarse del siguiente modo:

    - Si el valor obtenido es menor que 0, entonces el tweet tiene un sentimiento negativo.

    - Si el valor es igual a 0, entonces podemos considerar el tweet como neutro.

    - Y si el valor es superior a 0, en este caso debemos concluir que el mensaje del tweet contiene un sentimiento positivo.

8 Análisis de datos y resultados

En este apartado se incluye el proceso de análisis de los datos realizado, así como también los resultados obtenidos.

Para realizar el análisis de datos, se ha utilizado Jupyter y como fuente de datos todos los tweets almacenados en MongoDB.

8.1 Códigos de países

Para poder interpretar correctamente los gráficos que se muestran en este apartado debería tenerse en cuenta la siguiente codificación por país:

Código	País
AR	Argentina
BO	Bolivia
CR	Costa Rica
EC	Ecuador
ES	España
HN	Honduras
PY	Paraguay
VE	Venezuela

Tabla 4. Códigos de países

8.2 Volumen de datos recogidos

El proceso de captura de tweets ha estado en ejecución desde el día 4 de junio de 2017 en adelante. La información contenida en esta memoria incluye todos los datos hasta el momento de la entrega realizada para su corrección.

Durante este tiempo:

- Apache Kafka ha recogido casi 11 millones de tweets. Exactamente, 10.705.991.
- Y por su parte, MongoDB ha almacenado más 5 millones de mensajes procesados. En concreto, 5.258.090.

La diferencia se explica debido al procesamiento de los tweets, en los que se descartan los retweets, los mensajes de posibles spammers y aquellos tweets generados fuera de los países incluidos en el estudio.

Por otro lado, el total de tweets recogidos por cada país se muestra en la siguiente tabla:

País	Número total de tweets
Argentina	2.710.119
Bolivia	13.751
Costa Rica	118.992
Ecuador	177.914
España	1.426.814
Honduras	25.706
Paraguay	133.593
Venezuela	651.201

Tabla 5. Número total de tweets recogidos por país

También, indicar que la variación en el volumen de tweets se debe sobre todo al total de población por país y a las diferencias en la penetración en el uso de las nuevas tecnologías en cada caso.

Además, según el método de captura implementado, se requiere que Twitter haya etiquetado convenientemente cada tweet con el país asociado y en todo caso esto únicamente puede ocurrir cuando el usuario tiene activada la geolocalización. Inevitablemente, en aquellos países con un uso menor de las nuevas tecnologías, también encontraremos previsiblemente menos usuarios con dispositivos que les permitan geolocalizarse.

8.3 Análisis descriptivo general

Los valores estadísticos de los datos recogidos para cada país se muestran a continuación:

country_code	count	mean	std	min	25%	50%	75%	max
AR	2710119.0	0.293061	1.250276	-28.0	0.0	0.0	1.0	26.0
BO	13751.0	0.299833	1.300707	-9.0	0.0	0.0	1.0	10.0
CR	118992.0	0.304701	1.119779	-14.0	0.0	0.0	1.0	13.0
EC	177914.0	0.402031	1.264173	-10.0	0.0	0.0	1.0	18.0
ES	1426814.0	0.335085	1.308550	-20.0	0.0	0.0	1.0	19.0
HN	25706.0	0.405275	1.220996	-8.0	0.0	0.0	1.0	11.0
PY	133593.0	0.297403	1.160998	-12.0	0.0	0.0	1.0	32.0
VE	651201.0	0.230629	1.329253	-15.0	0.0	0.0	1.0	16.0

Figura 4. Valores estadísticos de las distribuciones con outliers

En el proceso de análisis de los tweets se han descartado aquellos valores extremos o *outliers* de los sentimientos calculados en función de su distancia respecto a la media. Concretamente, aquellos valores que están situados 4 veces la desviación estándar de la media de cada país. Este valor se ha seleccionado después de realizar visualizaciones de los histogramas de cada país.

Después de aplicar la limpieza de *outliers*, los valores obtenidos para cada distribución han sido los siguientes:

country_code	count	mean	std	min	25%	50%	75%	max
AR	2697645.0	0.285926	1.174204	-4.0	0.0	0.0	1.0	5.0
BO	13667.0	0.293115	1.204769	-4.0	0.0	0.0	1.0	5.0
CR	118325.0	0.282087	1.041069	-4.0	0.0	0.0	1.0	4.0
EC	176997.0	0.388272	1.178957	-4.0	0.0	0.0	1.0	5.0
ES	1418185.0	0.329124	1.212033	-4.0	0.0	0.0	1.0	5.0
HN	25605.0	0.394064	1.156648	-4.0	0.0	0.0	1.0	5.0
PY	132697.0	0.277904	1.064666	-4.0	0.0	0.0	1.0	4.0
VE	648283.0	0.218961	1.250910	-5.0	0.0	0.0	1.0	5.0

Figura 5. Valores estadísticos de las distribuciones sin outliers

A partir de la totalidad de los datos obtenidos, la media de sentimiento de felicidad por país se puede comparar en el siguiente gráfico:

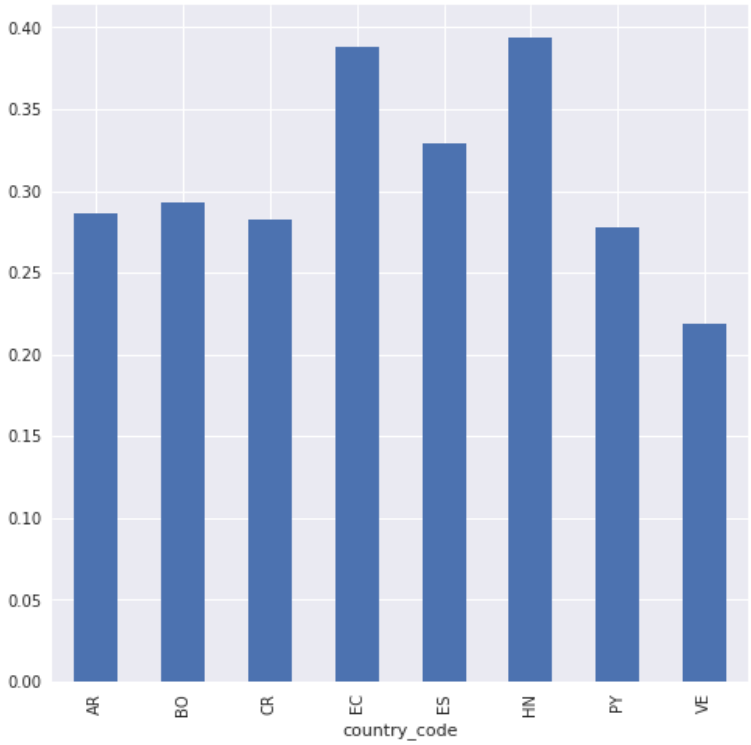

Figura 6. Media aritmética de sentimiento por país

Como puede observarse, estos datos muestran lo siguiente:

- La mediana en todos los casos es 0. Esto es debido a que el proceso de detección del sentimiento en los tweets tiende a valorar con sentimiento neutro muchos tweets recogidos.
- Las medias y desviaciones estándares son similares entre todos los países.

Además, también se han calculado los intervalos de confianza para las medias de todos los países sobre todos los datos recogidos. Los valores obtenidos con nivel de significación 0,05 son los siguientes:

País	Media	Límite inferior	Límite superior
Argentina	0.285926	0.284524	0.287327
Bolivia	0.293114	0.272914	0.313314
Costa Rica	0.282087	0.276155	0.288019
Ecuador	0.388272	0.382779	0.393764

España	0.329124	0.327129	0.331118
Honduras	0.394063	0.379895	0.408231
Paraguay	0.277903	0.272175	0.283632
Venezuela	0.218961	0.215916	0.222006

Tabla 6. Medias e intervalos de confianza por país

Es importante recalcar que sería necesario obtener datos de forma diaria durante más tiempo para que puedan considerarse los resultados de esta memoria representativos desde un punto de vista estadístico. Además, el hecho de recoger únicamente información desde una red social aumenta el sesgo de los cálculos realizados.

8.4 Comparación de la posición de felicidad en Twitter con Word Happiness Report (WHR)

Teniendo en cuenta todo lo explicado en el punto anterior, según los datos recogidos y las limitaciones indicadas, podríamos concluir que el país más feliz sería Honduras y el más triste Venezuela.

En la siguiente tabla puede observarse cuál sería la posición de cada país a partir de las medias calculadas y según el informe World Happiness Report de 2017:

País	Posición Twitter	Posición WHR (ONU)
Honduras	1	8
Ecuador	2	4
España	3	3
Bolivia	4	5
Argentina	5	2
Costa Rica	6	1
Paraguay	7	6
Venezuela	8	7

Tabla 7. Posición de felicidad según los datos recogidos y el informe WHR

Las variaciones que más sorprenden son las asociadas a Honduras y Argentina. Asimismo, cabe destacar que países como España, Bolivia, Paraguay o Venezuela ocupan la misma posición o una muy similar respecto al informe de WHR de 2017.

Finalmente, indicar que en el último informe de WHR se incluye que Venezuela es el país con la evolución peor respecto a su felicidad de todo el mundo. Según los datos recogidos, se confirmaría esta conclusión del informe de 2017 de la ONE puesto que el análisis de la felicidad realizado muestra una diferencia bastante notable entre Venezuela y el resto de países hispanohablantes considerados en este trabajo de fin de máster.

8.5 Evolución de la felicidad de forma diaria

En este apartado se muestran gráficos del comportamiento diario de los países analizados. Los gráficos muestran en valor en media del sentimiento de los tweets publicados en cada país.

En el siguiente gráfico podemos ver la evolución del coeficiente de felicidad calculado por días y semanas desde el momento de la activación de la captura de tweets hasta el momento de la entrega de esta memoria:

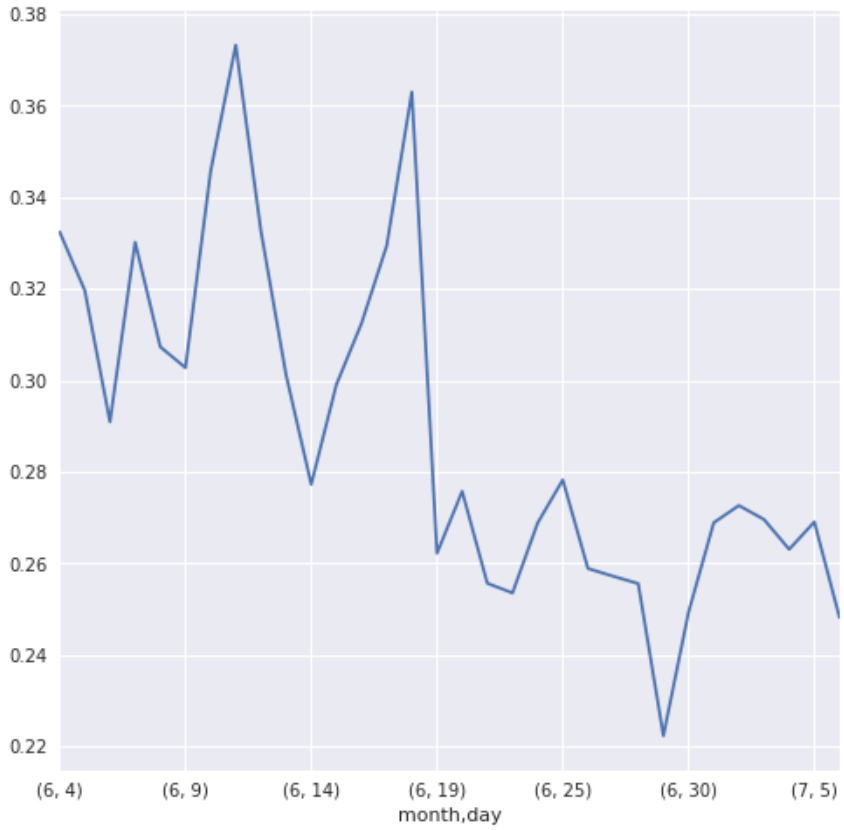

Figura 7. Evolución diaria del sentimiento de todos los países

Desde el punto de vista de cada país, la evolución de la felicidad de forma diaria y semanal se muestra en los siguientes gráficos:

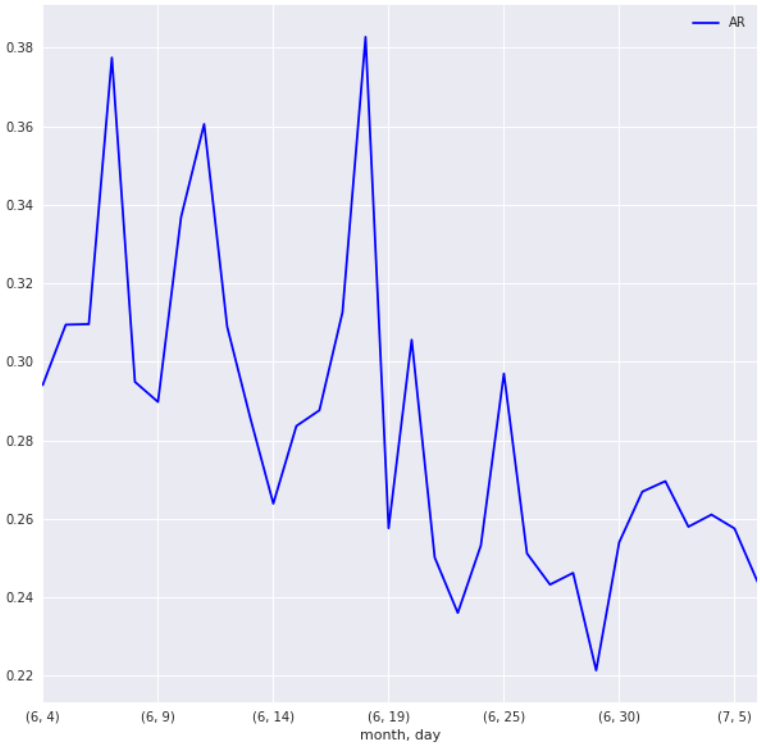

Figura 8. Evolución diaria del sentimiento de Argentina

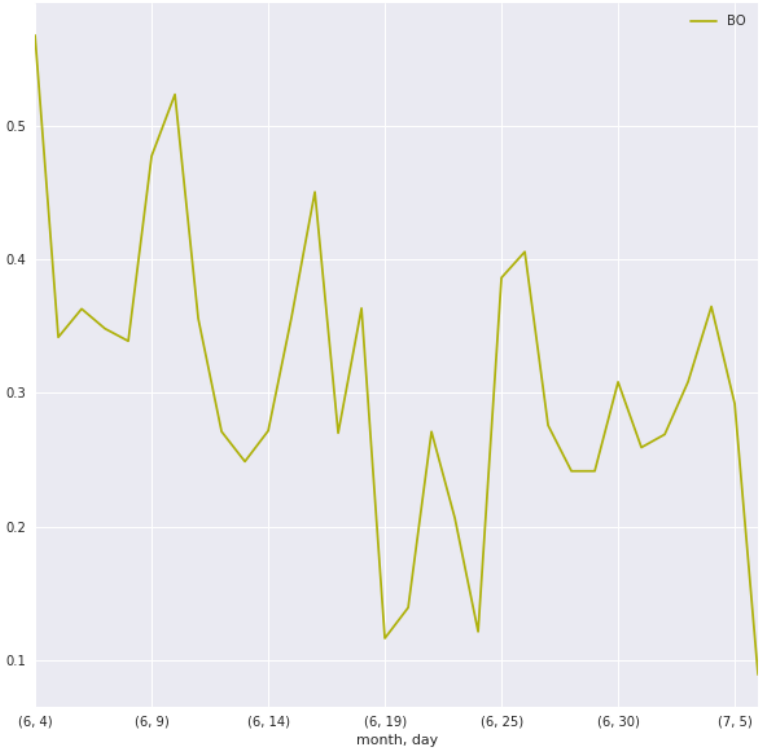

Figura 9. Evolución diaria del sentimiento de Bolivia

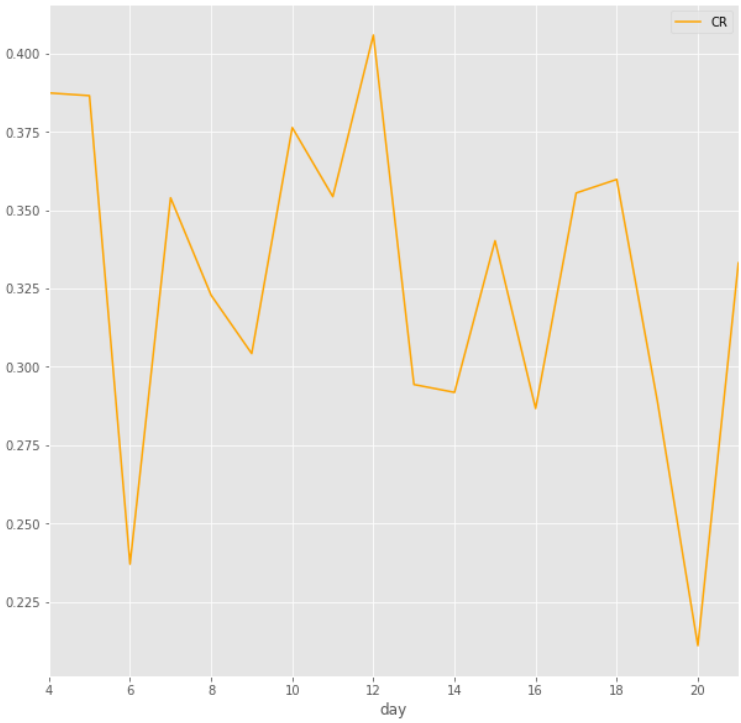

Figura 10. Evolución diaria del sentimiento de Costa Rica

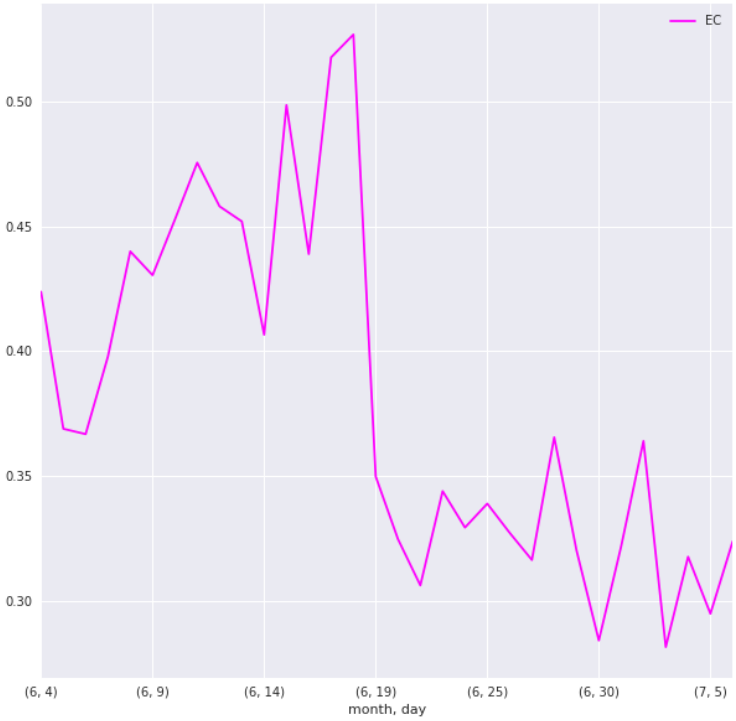

Figura 11. Evolución diaria del sentimiento de Ecuador

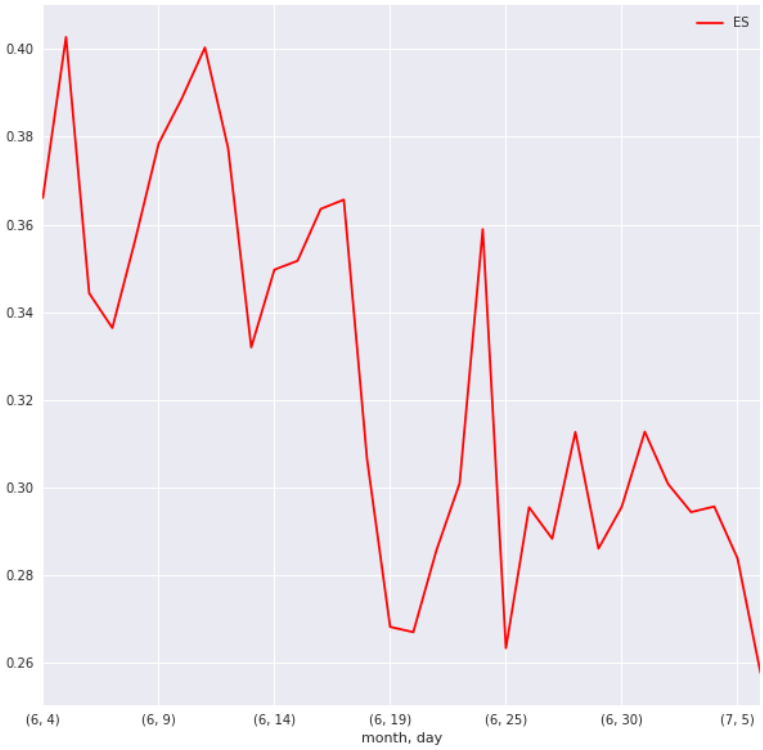

Figura 12. Evolución diaria del sentimiento de España

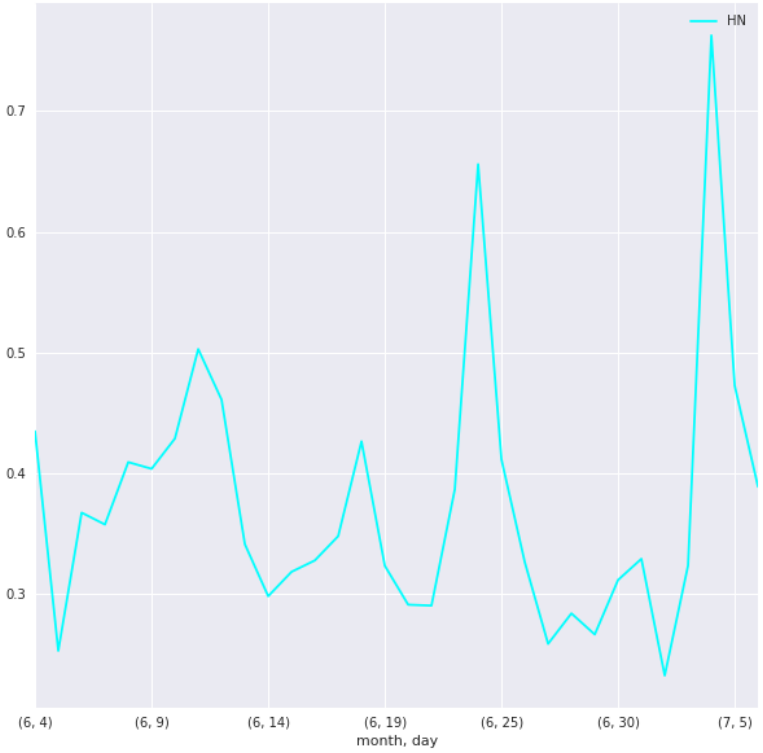

Figura 13. Evolución diaria del sentimiento de Honduras

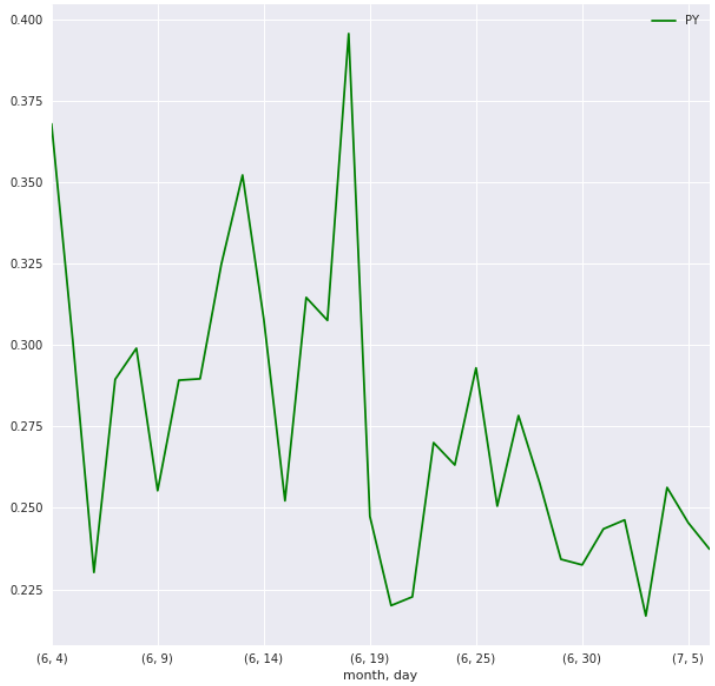

Figura 14. Evolución diaria del sentimiento de Paraguay

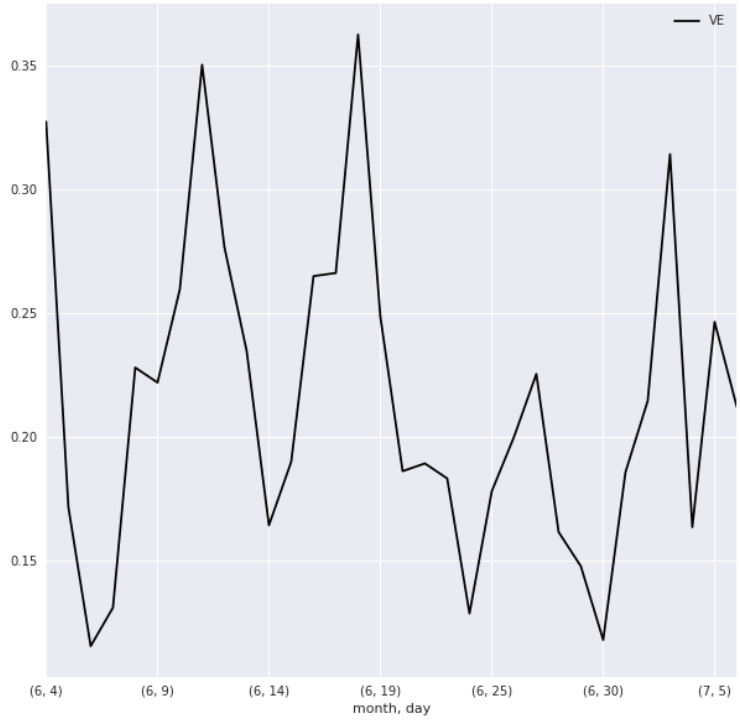

Figura 15. Evolución diaria del sentimiento de Venezuela

De los gráficos presentados se puede comentar lo siguiente:

- En general, los días más felices coinciden con el fin de semana.
- Puesto que Argentina y España tienen un volumen de tweets mucho mayor que el resto de los países, el gráfico de la media general es muy similar al comportamiento de ambos países.
- La tendencia de todos los países durante los días estudiados se dirige a la reducción de la felicidad, excepto en Costa Rica y Honduras.
- Quedaría como trabajo futuro estudiar las subidas y bajadas de felicidad en función de los acontecimientos diarios de cada país.

9 Conclusiones

En este apartado se incluyen las conclusiones alcanzadas mediante la realización de este trabajo de fin de máster.

9.1 Valoración de objetivos alcanzados

El trabajo realizado ha permitido alcanzar todos los objetivos planteados y definidos en esta memoria, con las siguientes excepciones que se enumeran a continuación:

- Debido a las limitaciones de los recursos disponibles no ha sido posible ejecutar el proyecto en un entorno distribuido. Por este motivo, no se ha podido comprobar con seguridad que el proceso implementado podría ejecutarse en un entorno de estas características sin errores.

- Como se ha visto, existe una gran diferencia entre los países respecto al volumen de tweets capturados. Esta situación provoca que la comparación de sentimientos entre países no sea del todo realista, porque para algunos de ellos contamos con muchos menos tweets diarios y en definitiva estamos considerando la opinión y el sentimiento de menos ciudadanos.

- Y relacionado con el punto anterior, la inclusión de un grupo de países con economías tan diferentes y desarrollo tecnológico también variado en realidad provoca que las comparaciones entre los países también sean poco realistas. En algunos países en los que la tecnología llega a pocos ciudadanos, en cierto modo estaríamos creando un sesgo entre aquellos ciudadanos con mejores medios y quizá con un tipo de vida poco habitual en el país, a los que estaríamos dando entrada en nuestros análisis y de este modo llegando a conclusiones incorrectas.

- Tal como se ha comentado, el proyecto permitiría introducir un análisis de sentimientos para nuevos idiomas, además del español. Ahora bien, en este caso debería tenerse en cuenta un problema derivado de los algoritmos de análisis de sentimientos: un diccionario de términos o corpus en inglés podría ser más positivo o más negativo que un corpus en español. De este modo, al comparar dos países con idiomas diferentes, podríamos suponer grados de felicidad diferentes debido a los diccionarios utilizados y no al sentimiento de los ciudadanos. Por lo tanto, este proceso de adición de un idioma sería mucho más complejo que lo planteado inicialmente en el proyecto.

9.2 Seguimiento de la planificación y metodología

Respecto a la planificación y metodología, se puede concluir que la planificación inicial ha sido la adecuada. Ha permitido realizar todas las

entregas acordadas con el Director del Proyecto sin problemas y el avance del proyecto ha sido el adecuado.

Aún así, la carga de datos debería haberse iniciado el 26 de mayo pero finalmente se inició el 4 de junio debido a que la configuración del consumo de mensajes de Apache Kafka desde Apache Spark fue mucho más compleja de lo inicialmente esperado.

El principal motivo se debió a que Apache Kafka no permitía devolver los mensajes capturados desde el principio y siempre enviaba a Apache Spark los últimos mensajes capturados. En esta situación, ante cualquier reinicio de Apache Spark, el sistema acaba perdiendo tweets que se quedaban sin procesar.

Finalmente, la situación quedó corregida a través de la configuración del cliente en Apache Spark, con la indicación de la recepción de los mensajes más antiguos posibles.

Por otro lado, indicar que de manera frecuente el Director del Proyecto y el autor del trabajo de fin de máster han estado en contacto tanto mediante reuniones remotas vía Skype y por correo electrónico, lo que sin duda ha facilitado el éxito de este proyecto.

9.3 Trabajos futuros

En este apartado se incluyen posibles trabajos futuros que permitirían ampliar y mejorar el proyecto presentado:

1. En primer lugar, desde el punto de vista práctico, podrían destacarse las siguientes aplicaciones derivadas del trabajo de fin de máster desarrollado:

    a. Las instituciones internacionales que anualmente preparan y publican informes relativos al nivel de felicidad de los ciudadanos de diferentes países del mundo podrían incorporar los resultados de los análisis de sentimiento de redes sociales para ajustar mucho más sus valoraciones finales. Incluso, se permitiría emitir resultados de felicidad de forma mucho más frecuente.

    b. Desde el punto de vista gubernamental, una solución así podría permitir reconocer las reacciones de los ciudadanos a acontecimientos diversos. De este modo, se podrían aplicar acciones correctivas y planes de actuación ante estos acontecimientos de forma mucho más rápida. Asimismo, permitiría medir el impacto de decisiones tomadas por el gobierno del país respecto a la opinión de los ciudadanos.

2. En segundo lugar, sería muy interesante implementar un sistema de captura de datos económicos en tiempo real para poder comparar la influencia de la economía en la felicidad de un país. De esta forma, se podría contrastar el nivel de felicidad de un país con su evolución económica y obtener conclusiones sobre el efecto en tiempo real que pueden tener noticias como el aumento del paro, caídas en la bolsa, problemas bancarios y otras situaciones similares.

3. También, sería posible analizar zonas diferenciadas dentro de un mismo país, con el objetivo de valorar la felicidad de aquellos lugares que cuentan con una renta per cápita mayor y los que tienen una renta mucho menor. En esta línea, Hedonometer (University of Vermont Complex Systems Center; The MITRE Corporation, 2017) ya valora la felicidad únicamente en EEUU por estados y ciudades.

4. Además, también sería interesante agregar nuevas fuentes de datos de redes sociales adicionales, tales como Facebook. De este modo se podría reducir el sesgo actual asociado a estar analizando datos obtenidos de una única red social.

5. Por otro lado, el proyecto debería ampliarse a nuevos idiomas, como podría ser el inglés, del que es posible disponer de buenos diccionarios para la valoración del sentimiento de los tweets.

6. Asimismo, puesto que el trabajo realizado utiliza herramientas que fácilmente podrían escalar a un volumen mucho mayor de datos, se debéria ampliar el número de países considerados en el análisis, así como evaluar el sistema durante un tiempo de recogida de tweets mucho mayor para tener datos más confiables.

7. Respecto al análisis de los sentimientos, en el futuro podría mejorarse el sistema de análisis de sentimientos para incluir los siguientes requisitos:

    a. En primer lugar, aumentar la precisión del clasificador de tweets y reducir el posible error actual a la hora de valorar la felicidad de los mensajes.

    b. También, debería mejorarse el corpus utilizado con expresiones habituales en Twitter, tales como abreviaturas o hashtags utilizados habitualmente por los usuarios, como también para añadir soporte a smileys.

8. Adicionalmente, con el objetivo de eliminar tweets con mensajes spam, sería interesante ampliar la capacidad de detección de usuarios spammers introduciendo un proceso mucho más sofisticado.

9. También, sería muy interesante estudiar profundamente los cambios en las medias diarias de felicidad de los países, para ver si determinados acontecimientos provocan o reducen la felicidad de los ciudadanos.

10. Finalmente, el análisis de los tweets podría llegar a ampliarse para detectar sentimientos de la ciudadanía hacia temas en concreto. Por ejemplo, medir el grado de felicidad asociado a un evento deportivo o a la publicación de los datos del paro en función de los términos que aparecen en los tweets felices o infelices.

9.4 Conclusiones finales del trabajo realizado

El trabajo realizado ha permitido llegar a las siguientes conclusiones finales:

- Las herramientas de Big Data Open Source actuales están perfectamente preparadas para la implementación de proyectos de estas características. Hoy en día el estado de estas herramientas es lo suficiente maduro como para que a nivel empresarial puedan desarrollarse grandes proyectos de análisis de datos masivos sin tener que recurrir a herramientas no Open Source.

- Como se ha comentado en esta memoria, la integración entre las herramientas Big Data se ha realizado mediante Python. En la actualidad, la mayoría de estas herramientas están implementadas internamente mediante Java o Scala, aunque todas ellas ofrecen interfaces para acceder desde Python. Ahora bien, el proceso de integración podría haber sido mucho más simple con el uso de estos lenguajes, porque las APIs de Python en algún caso son experimentales y no ofrecen los resultados esperados. Además, existe poca documentación on-line y muy poca bibliografía a la hora realizar la integración entre Apache Kafka y Apache Spark mediante Python.

- En la actualidad, no existen buenos corpus en idioma español disponibles libremente para poder utilizar en un proyecto de estas características, al menos, en comparación con el idioma inglés. Esta situación está provocando que no existan muchas iniciativas de proyectos de análisis de datos textuales en idioma español.

- Teniendo en cuenta todas las limitaciones del proyecto y el alcance de los datos recogidos, podríamos concluir que existe una cierta similitud entre las posiciones de ciertos países en comparación con el informe anual de felicidad de los países de la ONU, en el que el 50% de los países mantienen el mismo orden o muy similar en la su posición respecto a la felicidad en los datos obtenidos y en el informe de 2017 de la ONU.

10 Bibliografía

Bakkum, P., Banker, K., Verch, S., Garrett, D., & Hawkins, T. (2016). *MongoDB in Action, Second Edition: Covers MongoDB version 3.0.* New York: Manning Publications.

Bhattacharyya, G. K., & Johnson, R. A. (2014). *Statistics: Principles and Methods, 7th Edition.* John Wiley & Sons.

Bruce, A., & Bruce, P. (2017). *Practical Statistics for Data Scientists.* O'Reilly Media, Inc.

Ecma International. (2013). Obtenido de The JSON Data Interchange Format: http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf

Gallup. (2017). Obtenido de http://www.gallup.com

Klokan Technologies. (2017). *BoundingBox*. Obtenido de http://boundingbox.klokantech.com

Liu, B. (2015). *Sentiment Analysis.* Chicago: Cambridge University Press.

Liu, B., Messina, E., Fersini, E., & Pozzi, F. A. (2016). *Sentiment Analysis in Social Networks.* Cambridge: Morgan Kaufmann.

Macy, M. W., Weber, I., & Mejova, Y. (2015). *Twitter: A Digital Socioscope.* New York: Cambridge University Press.

MongoDB, Inc. (s.f.). *MongoDB*. Obtenido de MongoDB: https://www.mongodb.com

Nandi, A. (2015). *Spark for Python Developers.* Birmingham: Packt Publishing.

OECD. (2017). Obtenido de OECD Better Life Index: http://www.oecdbetterlifeindex.org

OECD.Stat. (2016). *Better Life Index - Edition 2016*. Obtenido de OECD.Stat: http://stats.oecd.org/Index.aspx?DataSetCode=BLI

Project Jupyter. (2017). Obtenido de Jupyter: http://jupyter.org

Roesslein, J. (2017). Obtenido de Tweepy Documentation: http://tweepy.readthedocs.io/en/v3.5.0/

Saralegi, X., & San Vicente, I. (2013). In Proceedings of "XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural". Workshop on Sentiment Analysis at SEPLN (TASS2013). *Elhuyar at TASS 2013* , 143-150.

Saralegi, X., Vicente, I. S., & Foundation, E. (2013). Recuperado el 2017, de ElhPolar dictionary, version: 1.0, LGPLLR license: http://komunitatea.elhuyar.eus/ig/files/2013/10/ElhPolar_esV1.lex

SDSN - ONU. (2017). Obtenido de World Happiness Report: http://worldhappiness.report

Shapira, G., Narkhede, N., & Palino, T. (2017). *Kafka: The Definitive Guide, 1st Edition.* New York: O'Reilly Media, Inc.

Shreedharan, H. (2014). *Using Flume.* New York: O'Reilly Media, Inc.

The Apache Software Foundation, Flume. (s.f.). *Apache Flume*. Obtenido de Apache Flume: https://flume.apache.org

The Apache Software Foundation, Kafka. (s.f.). *Apache Kafka*. Obtenido de Apache Kafka: https://kafka.apache.org

The Apache Software Foundation, Spark. (s.f.). *Apache Spark - Lightning-Fast Cluster Computing*. Obtenido de Apache Spark: https://spark.apache.org

Twitter, Inc. (2017). *Twitter Developer Documentation*. Obtenido de https://dev.twitter.com/overview/api

University of Vermont Complex Systems Center; The MITRE Corporation. (2017). Obtenido de Hedonometer: http://hedonometer.org

Verma, M., Divya, & Sofat, S. (2014). *Techniques to Detect Spammers in Twitter- A Survey.* Obtenido de http://research.ijcaonline.org/volume85/number10/pxc3893279.pdf

Zečević, P., & Bonaći, M. (2016). *Spark in Action.* New York: Manning Publications.

11 Anexos

11.1 Código fuente

Todo el código fuente del proyecto se ha publicado en GitHub en la siguiente dirección:

https://github.com/jcontesti/uoc-mib-project

El código fuente publicado contiene lo siguiente:

- kafka-producer: proceso productor de Apache Kafka, encargado de capturar los tweets de los países incluidos en el estudio.
- spark-consumer: gestión de la conexión con Apache Kafka y sistema de procesamiento y almacenaje de los tweets en MongoDB a ejecutar en Apache Spark.
- data-analysis: notebook para el análisis de los resultados almacenados en MongoDB. Se incluyen visualizaciones adicionales que no se han incluido en esta memoria.