Universitat Oberta de Catalunya

uoc.edu

Plataforma Big Data para el análisis de métricas sociales mundiales

Jordi Contestí Llull

Máster en Inteligencia de Negocio y Big Data

Trabajo de Fin de Máster

Julio de 2017

Índice

Introducción y contexto del trabajo

- Actualmente, la ONU y otras instituciones publican informes con la clasificación mundial de la felicidad por países
- Desventajas más importantes: información agregada y frecuencia anual
- Las redes sociales se han convertido en un termómetro social
- ¿Es posible utilizar la redes sociales para valorar la felicidad de todo un país?
- Objetivo del proyecto: implementar un sistema Big Data que permita estudiar el nivel de felicidad de diferentes países a partir de los datos publicados en las redes sociales
- Muestra de 8 países de habla hispana: Argentina, Bolivia, Costa Rica, Ecuador, España, Honduras, Paraguay y Venezuela
- Red social: Twitter

Fases del trabajo

1. Planificación de tareas e hitos
2. Redacción de requisitos funcionales y técnicos
3. Estudio de las herramientas
4. Diseño del sistema
5. Implementación y carga de datos
6. Análisis de datos y resultados

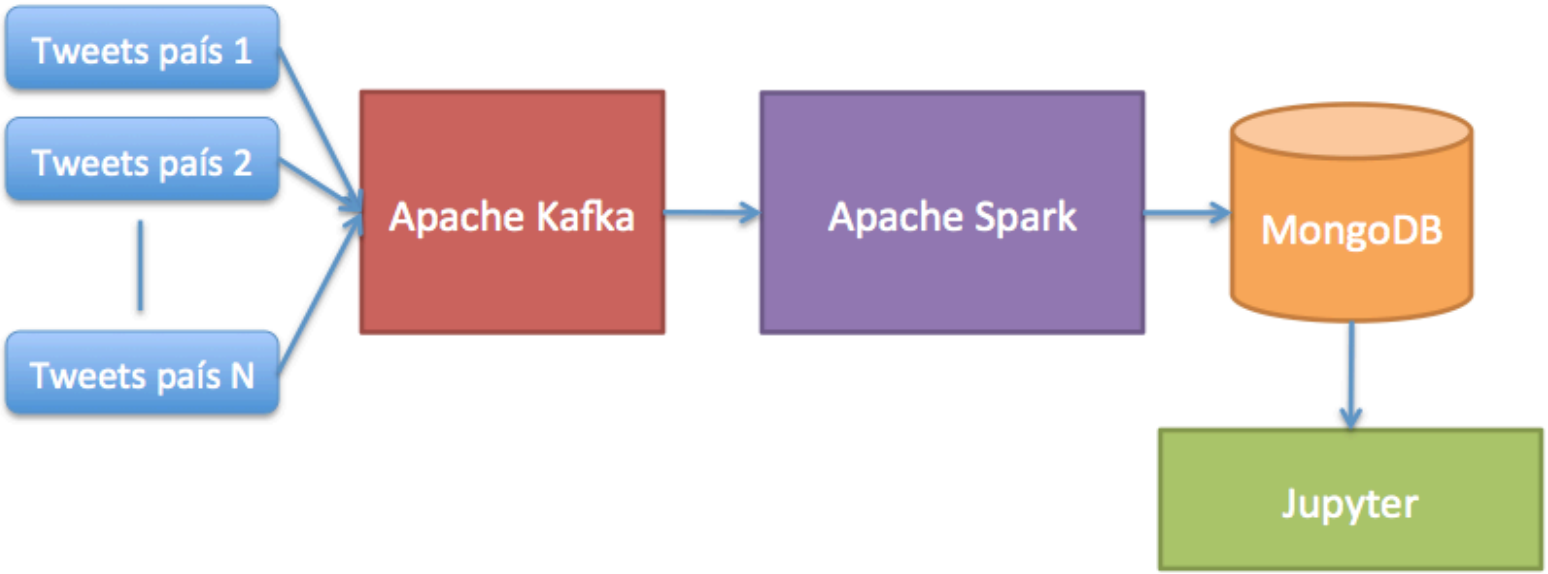
Requisitos y alcance del trabajo

- Captura de todos los tweets de los países estudiados de forma diaria durante un tiempo determinado
- Eliminación de spam, retweets y tweets no situados en los países deseados
- Cálculo del sentimiento felicidad de cada uno de los tweets. Utilizado diccionario de sentimientos positivos y negativos
- Almacenamiento de los tweets y cálculo del sentimiento
- Análisis de los datos obtenidos
- Comparación con el informe de felicidad anual de la ONU (World Happiness Report)
- Sistema escalable: mayor volumen de países y tweets
- Herramientas Open Source

Estudio y selección de herramientas

- Captura de los tweets: Apache Kafka
- Procesamiento de los tweets: Apache Spark
- Almacenamiento: MongoDB
- Análisis y visualización de datos: Jupyter y librerías Python
- Lenguaje para la integración entre herramientas: Python

Diseño del sistema



Análisis de datos y resultados

- Recogidos más de 10 millones de tweets entre el 4 de junio y 6 de julio 2017
- Gran diferencia en volumen de tweets entre países
- Resultados no pueden considerarse suficientemente representativos
- Posiciones obtenidas y comparación con el informe de la ONU:

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

Trabajos futuros

- Aumento del volumen de datos para ganar confiabilidad en los resultados
- Integración de más redes sociales
- Mejora del algoritmo de análisis de felicidad
- Adición de nuevos idiomas y países
- Incorporación y cruce de datos adicionales como información económica
- Ajuste de los datos de informes internacionales con los datos recogidos desde las redes sociales
- Contraste de las alteraciones de felicidad con sucesos acontecidos en los países

Conclusiones

- Las herramientas Big Data Open Source son válidas para este tipo de proyectos
- La integración entre las herramientas mediante Python es compleja
- No existen diccionarios públicos adecuados para el análisis la felicidad en idioma español
- Los resultados obtenidos muestran que las redes sociales podrían permitir la medición de la felicidad de los países en tiempo real

Universitat Oberta
de Catalunya

Gracias por su atención

Jordi Contestí Llull
Máster en Inteligencia de Negocio y Big Data
Trabajo de Fin de Máster
Julio de 2017

UOC.universitat
@UOCuniversitat
UOCuniversitat