“TOPIC MODELLING y BIG DATA”

Bruno Stampete

Tutor: Estrella Pulido Cañabate

INTRODUCCIÓN:

Es un hecho que, con el impulso de las nuevas tecnologías, la cantidad de datos producidos por el ser humano ha ido creciendo de forma constante a lo largo de los últimos años. Desde que el ser humano tiene capacidad de producir información, se estima que se hayan generado unos 5.000 millones de Gigabytes de datos, mientras que en 2011 esta misma cantidad de información se había generado cada dos días y en 2013 cada 10 minutos.
La tasa de generación de datos cada año crece de forma exponencial y si con el adviento de los ordenadores se había dado un primer empuje a la producción masiva de datos, con la propagación masiva de los terminales móviles, el crecimiento de datos es irrefrenable.

Para poder procesar toda esta información se necesitan determinadas herramientas que puedan realizar análisis en tiempos aceptables con la capacidad de cómputo de la que disponemos actualmente.

Los ámbitos en los que se pueden aplicar técnicas de Big Data son muy variados y algunos ejemplos podrían ser:

- Datos de cajas negras de aviones, barcos, trenes y cualquier medio de transporte, empleados para la monitorización de las prestaciones del servicio, para detectar posibles fallos y anomalías.
- Datos procedente de redes sociales, normalmente empleado para análisis de tendencias, opinión, gustos, estudios de marketing y de lanzamientos de productos.
- Datos procedentes de los índices bursátiles para poder analizar las tendencias del mercado y poder predecir futuros comportamientos del mismo.

- Datos procedentes de la red eléctrica y de otras fuentes de energía para poder crear patrones de consumo energético y adelantarse a las necesidades de los clientes y evitar cortes en el suministro.
- Buscadores de internet, para poder buscar con más facilidad entre todos los datos de los que disponen.
- Datos procedentes de los sistemas de transporte público y de análisis de tráfico con los que se pretende analizar los patrones de transporte y las conductas de los usuarios, para poder reforzar los servicios en determinadas horas y evitar atascos o incongruencias en la red.
- Datos procedentes de las compras con tarjetas de crédito para poder analizar el patrón y comportamiento de los consumidores y poder ofrecer unas ofertas personalizadas en función de los hábitos de consumo.

Hay muchos más usos posibles que se le puede dar al análisis intensivo de datos y sobre todo el gran reto es poder cruzar diferentes tipos de datos para obtener patrones de conducta y perfiles de usuario cada vez más completos y profundos, que tengan en cuenta distintos ámbitos de análisis.

Dentro de este último reto, se puede englobar el análisis automático del contenido de textos (Topic Modeling) [5] empleando algoritmos capaces de extraer palabras claves o etiquetas que resumen el texto mismo. Esto es posible gracias a modelos generativos como el Latent Dirichlet Allocation [3] que asume que cada documento (o grupo de documentos) se pueda dividir en N categorías y partiendo de este premisa, coloca cada palabra, en función de las observaciones, en una de las categorías.

OBJETIVOS:

El Objetivo principal de este Proyecto de Fin de Carrera es poder comparar diferentes herramientas que hacen uso, en su mayoría, del algoritmo LDA, para tratar grandes cantidades de información y poder obtener de forma automática un resumen del contenido a través de una serie de etiquetas representativas, encontradas en el texto.

Un primer análisis sería comparar los resultados obtenidos por herramientas que no hacen uso de sistemas distribuidos tales como Mallet [2] con otras que sí hacen uso de sistemas distribuidos de nodos como Hadoop.

En una segunda instancia se analizarán los resultados obtenidos en sistemas distribuidos de un único nodo y se plantea poder evaluar el rendimiento en sistemas distribuidos de múltiples unidades de proceso (multi-node cluster).

Se plantea añadir etapas de pre-procesado del corpus, intentando aprovechar la capacidad de Hadoop de operar sobre grandes cantidades de datos.

MÉTODO Y FASES DEL TRABAJO:

Para la realización de este proyecto se procederá a realizar las siguientes tareas:

- Estudio bibliografico relativo al funcionamiento del Top Modelling y del algoritmo LDA anteriormente mencionados.
- Instalación de la herramienta Mallet, para poder hacer unas primeras pruebas de Topic Modeling.
- Realizar pruebas con Mallet y medir rendimientos.
- Instalación de las herramientas necesarias para el desarrollo del análisis sobre grandes cantidades de datos, tales como Apache Hadoop y del entorno software necesario para su funcionamiento.
- Evaluación del tipo de algoritmos a emplear para desarrollar análisis sobre gran volumen de datos aplicando algoritmos del tipo LDA.
- Obtención de Resultados y de rendimientos en tiempos.
- Prueba con diferentes tipos de corpus de la eficacia del análisis y de posibles interferencias con el resultado.
- Comparación de resultados entre las dos distintas herramientas y obtención de conclusiones.

MEDIOS MATERIALES DISPONIBLES:

Para la realización del proyecto se cuenta con los siguientes medios:

- Fondos bibliográficos de la Escuela Politécnica Superior.
- Corpus de entrenamiento de distintas fuentes tales como: Wikipedia, Twitter, facebook.
- Manuales de Ayuda de la organización Apache y en concreto de la herramienta Hadoop.
- Medios técnicos informáticos: Sistema operativo Linux, Hadoop Apache, Java, Bash, Mallet.

REFERENCIAS:

[1] Hadoop, Apache Software Foundation, https://hadoop.apache.org/

[2] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[3] David M. Blei, Andrew Y. Ng and Michael I. Jordan. "Latent Dirichlet Allocation". Journal of Machine Learning Research 3 (2003) 993-1022.

[4] Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad Alkhouja. "Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce". World Wild Web Conference Lyon April 2012.

[5] Davide Blei. "Probabilistic Topic Models". Communications of the Association for Computing Machinery, April 2012.