**UNIVERSIDAD EAFIT**

**MAESTRÍA EN CIENCIA DE DATOS Y ANALÍTICA**

**ST1800 ALMACENAMIENTO Y RECUPERACIÓN DE INFORMACIÓN**

**(draft)**

Proyecto Materia:

ST1800 Almacenamiento y Recuperación de Información.

Objetivo:

Aplicar los conocimientos y conceptos vistos durante la materia en un proceso de Indexación, Búsqueda, Recuperación y Análisis de información textual no estructurada basada en el framework META, SparkML y Amazon AWS

Fuentes de datos:

* Datos semi-estructurados + no-estructurados.
* Archivos de noticias:
  + link para descarga: <https://drive.google.com/drive/folders/13pCI0nBs_ZLzlsIKxvwQiBwB7DGQzmTN?usp=sharing>
    - archivo: news.csv.zip

Actividades:

1. Realizar el entendimiento de los datos
2. Especificar los retos o casos de negocio sobre los datos
   * 1. Realizar indexación (solo en META)
     2. Realizar búsqueda y recuperación (Solo en META)
     3. Modelado de tópicos (en SparkML o AWS comprehend o META)
     4. Análisis de sentimientos (en SparkML o AWS comprehend)
     5. Agrupamiento (en SparkML o AWS comprehend)

<<<<

1. Ingeniería de Datos:
   1. Adecuar el ambiente en nube AWS, activar cuenta aws educate.
   2. Especificar, definir y aprovisionar los servicios AWS necesarios para desarrollar el proyecto.
   3. Ingesta y almacenamiento de datos
      1. realizar la carga de datos en Amazon S3
      2. determinar el esquema más adecuado de almacenamiento de los datos en Amazon S3 para facilitar el procesamiento de texto.
   4. crear los servicios de infraestructura y plataforma necesarios para desarrollar el proyecto:
      1. ej: máquinas virtuales en EC2, clusters EMR, almacenamiento S3, Servicios SageMaker y/o Comprehend, etc.
   5. Configurar Framework META para python:
      1. <https://github.com/meta-toolkit/metapy> basado en: <https://meta-toolkit.org/>
   6. Configurar Framework Spark y SparkML
   7. Configurar Frameworks SageMaker o Comprehend, jupyter o zeppelin, python.
   8. Se utilizará AWS Educate, lo que le permite a cada alumno de la materia tener 50 USD de crédito para desarrollar el proyecto.
2. Desarrollo de los retos o casos de negocio (especificados en el numeral 2)
   1. Sobre la infraestructura AWS
   2. Indexación, Búsqueda y Recuperación (usar framework META-python, sobre AWS)
   3. Minería de texto (desarrollar los casos en Spark/SparkML o los servicios cognitivos AWS (SageMaker o Comprehend)

Equipos de trabajo:

* Máximo 5 alumnos.
* Deben incluir fortalezas combinadas entre Computación, Matemáticas y Estadística.
* Las actividades de Ingeniería de Datos, deberán ser muy soportadas por los Ingenieros de Sistemas o de profesiones similares, o que tengan experiencia en Computación.
* Todos deben estar en capacidad de desarrollo y entender los modelos realizados en python accediendo a los diferentes frameworks ML o NLP, sin embargo dadas las actividades de Ingeniería de datos las cuales van a estar más lideradas por los Ingenieros de Sistemas. Se recomienda división del trabajo en el equipo en esos dos frentes: Ingeniería de Datos y Ciencia de datos (desarrollo de los retos o casos de negocio).
  + Los Ingenieros de datos deben responder por las actividades técnicas en la nube AWS, aprovisionamiento de máquinas, clusters y servicios, y deben ENTENDER el desarrollo de los modelos realizados.
  + Los científicos de datos deben responder por las actividades de desarrollo de los modelos de analítica, y deben ENTENDER los procesos técnicos realizados en AWS.
  + Cada miembro del equipo, debe especificar qué rol va a desempeñar (puede haber un integrante que quiera estar en ambos)
  + La lista de los miembros del equipo, junto con sus roles, debe ser enviado a más tardar el 3 de mayo, por el correo de la materia ST1800 de Eafit Interactiva.
  + A todos se les asignará una cuenta AWS Educate, la cual será más o menos utilizada en función y rol dentro del equipo, sin embargo se espera tener una comunicación e interacción técnica más constante entre el profesor y los Ingenieros de Datos para apoyarlos en O todos los procesos técnicos.

Despliegue en Computación en Nube: Amazon AWS.

Fechas importantes:

* Enunciado: Mayo 1
* Informe/demo de avance: Mayo 28 (fecha de la 3a sesión de clase)
* Entrega final: junio 8