UNIVERSIDAD EAFIT
INGENIERÍA DE SISTEMAS
STO0263 TÓPICOS ESPECIALES EN TELEMÁTICA, 2019-2
GRUPOS 031 Y 032

PROYECTO 4 – BIG DATA / SPARK
- Grupo máximo 3 alumnos –

Objetivo principal del proyecto 4:

Cada alumno integrante del grupo de proyecto4 debe tener una experiencia de diseño proyectos de analítica e implementación en tecnología big data – apache spark.

Si en la sustentación NO se evidencia ninguna de las experiencias por algún integrante, no se considera ganado el proyecto4 (nota mayor o igual a 3.0).

Productos:

1. Producto ejecutando en ambiente de Producción y Desarrollo
2. Repositorio github que contenga todos los programas desarrollados.
    a. El documento "readme.md" donde se encuentre todo el proceso de diseño, desarrollo, pruebas, instalación y ejecución del proyecto4
    b. En el github, documento 'marcodereferencia.pdf', donde describa globalmente los fundamentos teóricos, conceptuales y matemática de la técnica de analítica de texto empleada. Debe leer, sintetizar y resumir algunos papers fundacionales del modelo de analítica de texto utilizado. Citado en formato APA.

Clúster Apache Spark:

El desarrollo del proyecto4, se podrá realizar y ejecutar en uno de los siguientes ambientes:

- Amazon AWS – EMR + Notebooks, Para Producción (opción 1)
- En un clúster público como el de Databricks Community, Para Producción (opción 2)
- En el cluster Hadoop/Spark del DCA para el curso: (con user/pass de la VPN), para Desarrollo
    - https://hdp1.dis.eafit.edu.co
    - https://zeppelin1.dis.eafit.edu.co
    - https://hdp1shell.dis.eafit.edu.co

Finalmente, el producto debe poder correr en alguno de los anteriores ambientes sin mayores cambios.

Actividades:

Apropiar globalmente una metodología de desarrollo de proyectos de big data analítica, que siga el 'ciclo de vida' de los datos.

a. Entender bien el problema (entendimiento del negocio)
b. Entendimiento de los datos, preparación, limpieza, etc
c. Familiarización con las APIs de analítica ML que tiene SparkML, otras librerías como Pandas de Python, o Librerías especializadas en Text Mining como NLTK, Spicy.
d. Diseño, desarrollo y pruebas de modelo.
e. Ejecución y entrega

Para lo anterior deberá ejecutar correctamente el ciclo de vida:

1. Fuentes
2. Ingesta
3. Almacenamiento
4. Procesamiento
    a. Preparación de datos
    b. Procesamiento analítico
5. Aplicación
    a. Apps web
    b. Visualización
    c. api

Criterios de evalucación:

10% ejecución del proyecto en al menos un ambiente de producción y el ambiente de desarrollo.
20% Funcionalidad de PREPARACIÓN DE DATOS implementada correctamente.
30% Funcionalidad de TEXT MINING seleccionada
10% Repositorio Github con código y documentación sobre el proceso, desarrollo y ejecución

10% Documento de ‘marcodereferencia.pdf’ en el github (con referencias a artículos científicos, formato APA que sustente las citaciones)
20% Sustentación grupal e individual

Fechas de entrega:

20 de noviembre 2019

Sustentaciones: 20,21 y 22 de noviembre de 2019.

Código de Honor:

Cada uno de los autores, debe declarar explícitamente que el trabajo es original y cual fue su aporte en el desarrollo del proyecto4, o si es copiado de algún sitio en internet (porque hay muchas implementaciones de este problema) deberá referenciar o citar el sitio de donde tomo el trabajo y declarar entonces cual fue su aporte con esta copia en el proyecto4.

Esta declaración debe ser colocada en README.md del github del proyecto4 por cada autor.

ANALÍTICA DE TEXTO PARA PROCESAMIENTO DE NOTICIAS

Contexto:

La minería o analítica de texto, son un conjunto de modelos, técnicas, algoritmos y tecnologías que permiten procesar texto de naturaleza NO ESTRUCTURADA.

La minería de texto (text mining) permite transformar el texto en una forma estructurada, de tal forma que facilite una serie de aplicaciones como Búsqueda en texto, relevancia de documentos, entendimiento natural del lenguaje (NLP), traducción automática entre idiomas, análisis de sentimientos, detección de tópicos entre muchas otras aplicaciones.

Quizás el procesamiento más sencillo de todos sea el wordcount, el cual consiste en determinar la frecuencia de la palabra por documento o por todo el dataset.

Problema:

Se tiene un conjunto de noticias en texto libre, sobre el cual se desea realizar:

PRIMERA PARTE: PREPARACIÓN DE DATOS

LAS NOTICIAS DEBEN SER PRE-PROCESADAS PARA PREPARAR LOS DATOS PARA LA ANALITICA, DENTRO DE LAS SUGERENCIAS DE PREPARACIÓN ESTÁN:

1. Remover caracteres especiales (. , % ( ) ‘ “ ….
2. Remover stop-words
3. Remover palabras de longitud 1
4. Stemming / lemmatization

SEGUNDA PARTE: explorar al menos una técnica de analítica de texto NLP de los diferentes modelos y aplicaciones de SparkML-NLP, deberá ser creativos para encontrar dicha navegación, entre las opciones:

- Agrupamiento de textos (algoritmos como k-means)
- Detección de comunidades (algoritmos como Louvain)
- Detección de tópicos y Clasificación de Textos (LDA, Word2vec, Doc2Vec)
- Análisis de sentimientos
- Generación automáticas de keyworks
- Ranking de documentos para búsqueda y recuperación
- Etc.

Datos:

Los datos de trabajo para el proyecto son un conjunto de noticias publicados en kaggle:

https://www.kaggle.com/snapcrack/all-the-news

puede usar otros datasets más adecuados para la problemática concreta de text mining.

Referencias:

- https://en.wikipedia.org/wiki/Search_engine_indexing
- https://en.wikipedia.org/wiki/Inverted_index
- https://nlp.stanford.edu/IR-book/html/htmledition/a-first-take-at-building-an-inverted-index-1.html