Manual de Instrucciones

Tokenizer WordCount Using Hadoop (Java Version)

Enlaces y Archivos Necesarios

- **Link Dataset Original**: Consulta el archivo linkDataset.txt para obtener el enlace de descarga del dataset original.
- **Nota importante**: Todo el desarrollo debe hacerse usando datasetCompleto.txt. Si solo tienes datasetCompleto.csv, convierte el archivo manualmente o solicita una copia en .txt.

Preparación del Entorno

1. Preparación de Archivos

- Asegúrate de tener el archivo datasetCompleto.txt.
- Colócalo en el directorio WorkingFiles.

2. Diccionario de Palabras

- El archivo Dictionary.txt también debe estar en la carpeta workingFiles.
- Este archivo contiene las palabras a eliminar del análisis (stopwords, malas palabras, etc.).

Preprocesamiento del Dataset

Compilación

En la terminal, navega al directorio del archivo Preprocesamiento.java y compílalo:

```
javac Preprocesamiento.java
```

Ejecución

Ejecuta la clase compilada:

```
java Preprocesamiento
```

- Esto generará datasetProcesado.txt dentro del directorio WorkingFiles.
- Este archivo es el que se usará en el análisis de frecuencia.

Nota: Si prefieres no correr el preprocesamiento, puedes usar una versión ya procesada del archivo (ver linkDataset.txt).

Análisis de Frecuencia con Hadoop

Configuración de Hadoop

1. Asegúrate de que Hadoop esté instalado y funcionando:

```
hadoop version
```

2. Crear Carpeta en HDFS

```
hdfs dfs -mkdir /user/hadoop/input
```

3. Subir el Dataset Procesado

```
hdfs dfs -put workingFiles/datasetProcesado.txt /user/hadoop/input
```

Análisis de Frecuencia de Una Palabra

1. Navegar al directorio del JAR

2. Ejecutar el JAR

```
hadoop jar wordcount.jar WC /user/hadoop/input /user/hadoop/output1
```

3. Descargar Resultados

```
hdfs dfs -get /user/hadoop/output1/part-r-00000 Frequency1/
```

4. Procesar Resultados (Java)

Compila y ejecuta ResultProcessor1.java:

```
javac FrequencyAnalysis.java
```

```
java FrequencyAnalysis
```

Esto generará freq-results-sorted.txt con las palabras más frecuentes ordenadas.

Análisis de Frecuencia de Dos Palabras (Bigramas)

1. Navegar al directorio del JAR

2. Ejecutar el JAR

```
hadoop jar wordcount2.jar WordCount /user/hadoop/input /user/hadoop/output2
```

3. Descargar Resultados

```
hdfs dfs -get /user/hadoop/output2/part-r-00000 Frequency2/
```

4. Procesar Resultados (Java)

Compila y ejecuta ResultProcessor2.java:

```
javac FrequencyAnalysis2.java
```

```
java FrequencyAnalysis2
```

Esto generará freq-results-sorted.txt con los bigramas más comunes ordenados.