Tarea: limpieza de datos y texto con Terminal

Jesús Alejando Pérez Granados

A01253993

Materia: Análisis de ciencia de datos

Maestra: Mauricio González Soto

Grupo 101

Tecnológico de Monterrey Campus Monterrey

13 de febrero de 2026

Comandos utilizados (pipeline completo)

```
        # Get into bash
bash
        # Get into the file
cd Documents/TC2004B/TC2004B-S1-2026-ClassDirectory/limpieza_datos
        # Download the book
curl -o moby.txt https://www.gutenberg.org/cache/epub/2701/pg2701.txt
        # Saving a copy of only the book
sed -n '/\*\*\* START OF/,/\*\*\* END OF/p' moby.txt | sed '1d;$d' > moby_limpio.txt
        # Total word count
awk '{total += NF} END {print total}' moby_limpio.txt
        # Average words per line
awk '{sum += NF; n++} END {print sum/n}' moby_limpio.txt
        # Prints lines of over 20 words
awk 'NF > 20 {print NR, NF, $0}' moby_limpio.txt
        # Average letters per word
awk '{for(i=1;i<=NF;i++) sum+=length($i); n+=NF} END {print sum/n}' moby_limpio.txt
        # Prints maximum and minimum line lengths
awk 'NR==1 {min=max=NF} { if(NF<min) min=NF; if(NF>max) max=NF; sum+=NF; n++
} END {print "min",min,"max",max,"avg",sum/n}' moby_limpio.txt
        # Prints the top 20 most used words with how many times they’re used
$ awk '{
   for(i=1; i<=NF; i++) {
      gsub(/[^a-zA-Z]/, "", $i)
      if(length($i) > 0) {
         word = tolower($i)
         count[word]++
      }
   }
}
END {
   for(w in count) print count[w], w
}' moby_limpio.txt | sort -rn | head -20
```

Extra

```
        # Prints all instances of the word “thereof” with three lines before and after
grep -i -A 3 -B 3 "thereof" moby_limpio.txt
```

Top 20 palabras más frecuentes

14423 the
6379 and
6593 of
4639 a

4579 to
4155 in
2940 that
2522 his
2368 it
1943 i
1776 but
1749 he
1721 with
1721 as
1706 is
1636 was
1599 for
1476 all
1364 this
1311 at

Estadísticas del texto (palabras totales, promedio por línea, etc.)

Word count: 212796
Promedio de letras por palabra: 4.78263
Palabras por línea
Promedio: 9.70077
Mínimo: 0
Máximo: 19

Reflexión crítica

El uso de la terminal es bueno para manejar simplemente archivos. Esto es ventajoso en muchas ocasiones porque es de las maneras más rápidas de procesar y no es necesario cargar los archivos. También permite organizar y crear archivos de manera rápida. Sin embargo, claramente tiene limitaciones, pues no es capaz de analizar semánticamente y, aunque es rápido, es porque también es lo más básico. Del mismo modo, se complica más al intentar analizar grandes volúmenes de archivos distintos de manera conjunta.