Tarea 1 — Análisis integral de corpus en español

Maestría en Cómputo Científico

Objetivo

Aplicar un pipeline completo de Análisis de Lenguaje Natural para caracterizar un corpus etiquetado (5 clases), evaluar leyes empíricas como la de Zipf, extraer estructuras léxicas y gramaticales, seleccionar características, entrenar representaciones y modelos, y explicar los resultados obtenidos con evidencia cuantitativa.

Entregables

- Código en Python (no en Colab), estructurado en módulos.
- Un reporte en formato PDF, con figuras, tablas y discusión de resultados.

Tareas

1. Descripción del corpus

Analiza el corpus y reporta:

- Número de documentos, tokens y vocabulario.
- Hapax legomena y su proporción.
- Porcentaje de *stopwords*.
- Estadísticas por clase (número de documentos, tokens y vocabulario).

2. Ley de Zipf

- Calcula la frecuencia absoluta $f(w)$ de cada palabra w en el corpus y ordénalas de mayor a menor. A cada palabra así ordenada se le asigna un rango r, donde $r = 1$ corresponde a la palabra más frecuente, $r = 2$ a la segunda, y así sucesivamente.
- Representa gráficamente la relación entre **log-rango** y **log-frecuencia**. Es decir, para cada palabra graficar el punto $(\log r, \log f(w))$. La Ley de Zipf predice que los puntos deberían aproximarse a una línea recta decreciente.

- Ajusta una recta mediante regresión lineal sobre los puntos $(\log r, \log f(w))$, de la forma:

$$\log f(r) = \log C - s \cdot \log r,$$

  lo cual equivale al modelo Zipfiano $f(r) \approx \frac{C}{r^s}$.

- En esta formulación:
  - C es una constante de normalización que se aproxima a la frecuencia de la palabra más común ($f(1) \approx C$).
  - s es el exponente de Zipf, que controla la rapidez con que decrecen las frecuencias conforme aumenta el rango. Valores cercanos a $s \approx 1$ son típicos en lenguajes naturales.
- Interpreta el valor del exponente s: si $s > 1$, la frecuencia cae más rápido de lo esperado; si $s < 1$, las palabras raras aparecen relativamente más seguido.
- Discute posibles desviaciones: por ejemplo, la presencia de *stopwords* muy frecuentes, el tamaño limitado del corpus, o palabras raras (hapax legomena) que afectan la cola de la distribución.

3. Palabras importantes por clase

- Elimina palabras vacías y normaliza el texto.
- Identifica las palabras más frecuentes en cada clase.
- Reflexiona si las palabras más repetidas son realmente discriminativas.

4. Patrones gramaticales (POS 4-gramas)

- Etiqueta con POS cada documento.
- Extrae las secuencias gramaticales más frecuentes de longitud 4 en cada clase.
- Discute si estas estructuras difieren entre clases y explica por qué.

5. Representaciones BoW

- Construye representaciones BoW con TF y con TF-IDF.
- Aplica alguna medida estadística (chi-cuadrado, información mutua o *information gain*).
- Obtén el top 20 de características más importantes en cada representación.
- Analiza diferencias entre ambas representaciones.

6. Bigramas

Repite el ejercicio anterior pero utilizando bigramas de palabras. Compara resultados y discute si los bigramas aportan mayor discriminación semántica.

7. Word2Vec y analogías

- Entrena un modelo Word2Vec sobre el corpus.
- Realiza al menos 5 analogías interesantes y discute los resultados.

8. Embeddings de documento y clusterización

- Calcula embeddings de documentos como el promedio de Word2Vec.
- Aplica K-means con $k = 5$.
- Reporta los 5 textos más cercanos al centroide de cada clúster.
- Discute si los clústeres se alinean con las etiquetas originales.

9. Clasificación con partición 70/30

Realiza cuatro experimentos acumulativos con un clasificador (SVM o regresión logística):

(a) Sin preprocesamiento.

(b) Con minúsculas.

(c) Con minúsculas y stemming/lematización.

(d) Con minúsculas, stemming y filtrando palabras con frecuencia mínima de 10.

Compara métricas (accuracy, F1 macro, matriz de confusión) y discute si el preprocesamiento es importante.

10. LSA con 50 tópicos

- Aplica Latent Semantic Analysis (SVD truncado) con 50 tópicos.
- Muestra los términos más relevantes por tópico.
- Identifica qué tópicos son más informativos según una métrica estadística y analiza su coherencia.

Criterios de evaluación

- Corpus, descriptivos y Zipf: 10 pts.
- Palabras por clase y POS-4-gramas: 10 pts.
- BoW y selección de características: 10 pts.
- Bigramas: 10 pts.
- Word2Vec y analogías: 10 pts.
- Clusterización y análisis: 10 pts.
- Clasificación y matrices de confusión: 20 pts.
- LSA y análisis de tópicos: 10 pts.
- Claridad del reporte: 10 pts.