Tarea 3: Clasificación de sentimientos base de datos IMBD

Luis Maximiliano López Ramírez

Septiembre 27, 2024

1 Instrucciones

1. **Descargar el siguiente notebook:** `https://github.com/azunre/transfer-learning-for-nlp/blob/master/Ch2-3/tlfornlp-chapters2-3-imdb-traditional.ipynb`
2. **Revisión del Código:** El código tiene un pequeño error a la hora de cargar los datos. ¡Hay que encontrar cuál es el error! También se pueden buscar alternativas para leer y cargar la base de datos correctamente.
3. **Entrenamiento con Features BOW:** Entrenar con las características BOW (Bag-of-Words) los modelos indicados en el notebook. Comparar el desempeño con las siguientes configuraciones:
   (a) **Configuración 1:** `maxtokens = 50, maxtokenlen = 20`
   (b) **Configuración 2:** `maxtokens = 100, maxtokenlen = 100`
   (c) **Configuración 3:** `maxtokens = 200, maxtokenlen = 100`
4. **Entrenamiento con TF-IDF:** Repetir el proceso anterior quitando las características BOW y utilizando ahora el método `TF-IDF` mediante la función `TFidVectorizer()` de `scikit-learn`.

2 Tablas Comparativas

- `Configuraciones`
  - `Inciso a: {"maxtokens": 50, "maxtokenlen": 20}`
  - `Inciso b: {"maxtokens": 100, "maxtokenlen": 100}`
  - `Inciso c: {"maxtokens": 200, "maxtokenlen": 100}`
-

Se usaron features BOW y features TF-IDF (TFidVectorizer())

Table 1: BOW - Tabla 1

	Inciso a	Inciso b	Inciso c
Logistic Regression Time	0.352871	0.692264	1.103024
Logistic Regression Accuracy	0.680000	0.735000	0.761667
SVC Time	5.318035	8.017462	11.882643
SVC Accuracy	0.665000	0.715000	0.746667
Random Forest Time	2.084996	2.652875	3.456817
Random Forest Accuracy	0.658333	0.721667	0.763333
Gradient Boosting Time	18.737749	39.877150	56.840677
Gradient Boosting Accuracy	0.650000	0.663333	0.748333

Table 2: TF-IDF (TFidVectorizer) - Tabla 2

	Inciso a	Inciso b	Inciso c
Logistic Regression Time	0.241724	0.333917	0.777298
Logistic Regression Accuracy	0.703333	0.776667	0.805000
SVC Time	11.379516	17.148620	45.520317
SVC Accuracy	0.720000	0.763333	0.808333
Random Forest Time	4.171154	5.544183	12.661570
Random Forest Accuracy	0.705000	0.736667	0.751667
Gradient Boosting Time	50.776217	91.489628	321.756221
Gradient Boosting Accuracy	0.663333	0.738333	0.766667

3 Observaciones y Conclusiones

Conclusiones Basadas en las Tablas de Resultados

Las dos tablas presentan el desempeño de diferentes modelos de clasificación (Regresión Logística, SVC, Random Forest y Gradient Boosting) en términos de tiempo de entrenamiento y precisión (*Accuracy*) bajo dos enfoques distintos para la representación de las características: **BOW** (Bag-of-Words) y **TF-IDF** (*TFidVectorizer* de scikit-learn). Cada tabla contiene los resultados para tres configuraciones de parámetros (**Inciso a, b y c**), que varían los valores de `maxtokens` y `maxtokenlen`.

1. Comparación de Desempeño entre BOW y TF-IDF

- El método **TF-IDF** muestra consistentemente mejores resultados en términos de precisión (*Accuracy*) que **BOW** para la mayoría de los modelos y configuraciones.
- En *Logistic Regression*, la precisión mejora significativamente al utilizar **TF-IDF**, pasando de 0.761667 (BOW, Inciso c) a 0.805000 (TF-IDF, Inciso c).
- Para *SVC*, el salto de precisión entre **BOW** y **TF-IDF** también es notable, alcanzando un valor máximo de 0.808333 en la configuración **TF-IDF** frente a 0.746667 en la mejor configuración de **BOW**.
- El modelo de *Gradient Boosting* es el más beneficiado por el uso de **TF-IDF**, con un incremento significativo en precisión de 0.748333 (BOW, Inciso c) a 0.766667 (TF-IDF, Inciso c).

2. Impacto de las Configuraciones (Incisos a, b y c)

- A medida que se incrementan los valores de `maxtokens` y `maxtokenlen` (de **Inciso a** a **Inciso c**), todos los modelos tienden a mejorar su precisión. Esto es más evidente en el modelo *SVC* y *Gradient Boosting*.
- En **TF-IDF**, los tiempos de entrenamiento para *Gradient Boosting* incrementan exponencialmente al aumentar los valores de `maxtokens`, pasando de 50.776217 segundos (Inciso a) a 321.756221 segundos (Inciso c), lo que sugiere que configuraciones más grandes son muy costosas en términos computacionales.

3. Comparación entre Modelos

- **Regresión Logística:** Aunque es el modelo más simple, muestra una buena precisión en **TF-IDF** (0.805000 en Inciso c) y es competitivo con *Random Forest*.
- **SVC:** Tiene una mejora significativa en **TF-IDF** frente a **BOW**, alcanzando su mayor precisión (0.808333) en **TF-IDF** con **Inciso c**, pero con tiempos de entrenamiento altos.
- **Random Forest:** Muestra un desempeño robusto y consistente, manteniendo un buen balance entre precisión y tiempo en ambas representaciones.

- **Gradient Boosting:** Aunque logra las mejores precisiones en la mayoría de las configuraciones, el costo en tiempo es considerablemente más alto, especialmente en **TF-IDF** con configuraciones elevadas.