Machine Learning con Embeddings

Modelado de Tópicos

Catherine L. Scott

Arian Milanes

Modelo 1: BoW/TF-IDF

BoW: creamos un gran vocabulario con todas las palabras únicas del corpus. Cada documento se convierte en un vector numérico donde contamos cuántas veces aparece cada palabra del vocabulario.

TF-IDF: Asigna un "peso" o importancia a cada palabra.

- **TF (Frecuencia de Término)**: Le da más importancia a las palabras que aparecen muchas veces en un mismo documento.
- **IDF (Frecuencia Inversa de Documento)**: Le quita importancia a las palabras que aparecen en muchos documentos.

Modelo 2: Word2Vec

Word2vec toma como entrada un gran corpus de texto y produce un espacio vectorial, asignando cada palabra única en el corpus a un vector correspondiente en el espacio.

Modelo 3: Doc2Vec

Doc2vec aunque es similar a Word2Vec, en este caso los vectores representarán la información de un documento, no de una palabra.

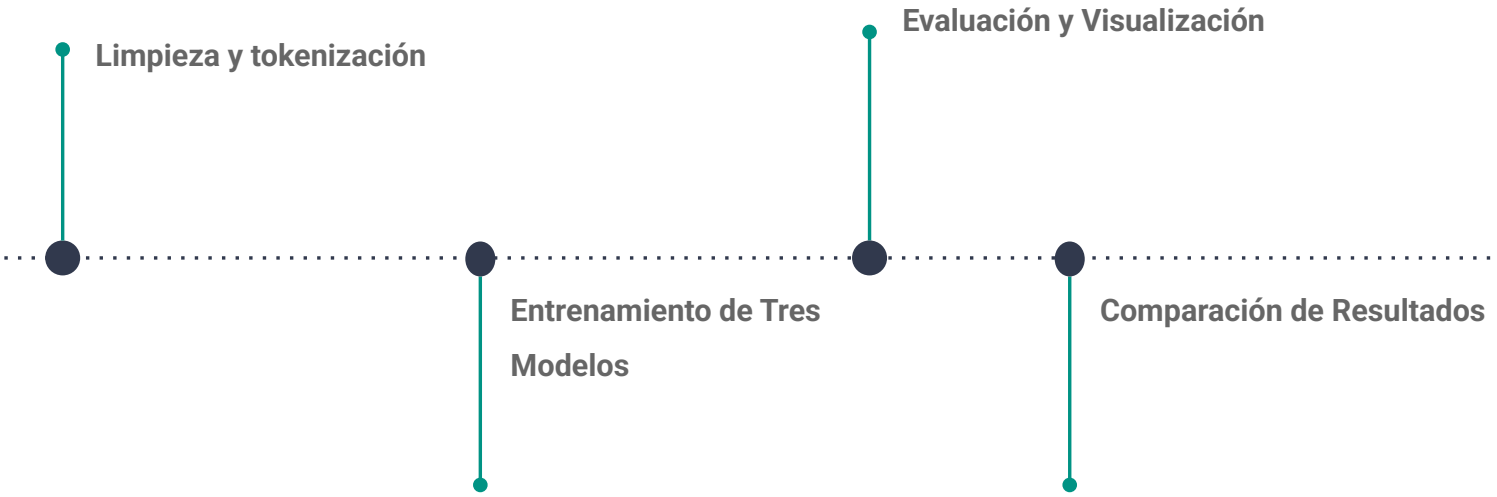
Métrica de Evaluación: Puntuación de Silueta (Silhouette Score)

La Puntuación de Silueta evalúa qué tan denso y bien separado es un cluster. En otras palabras, nos dice si los documentos dentro de un mismo tópico están muy relacionados entre sí y, al mismo tiempo, muy distintos a los documentos de otros tópicos.

¿Cómo se interpreta el puntaje? El valor va de -1 a +1:

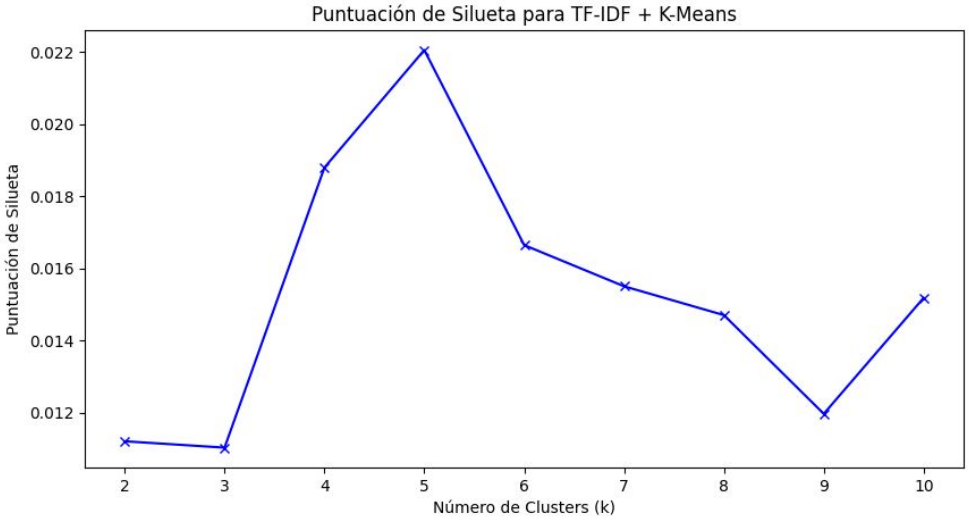
- **+1**: Indica que los clusters son perfectos: densos y muy bien separados.
- **0**: Indica que los clusters se superponen o que los puntos están muy cerca del límite entre dos clusters.
- **-1**: Indica que los puntos probablemente han sido asignados al cluster incorrecto. ¡El peor resultado posible!

Pipeline


Limpieza y tokenización
Entrenamiento de Tres Modelos
Evaluación y Visualización
Comparación de Resultados

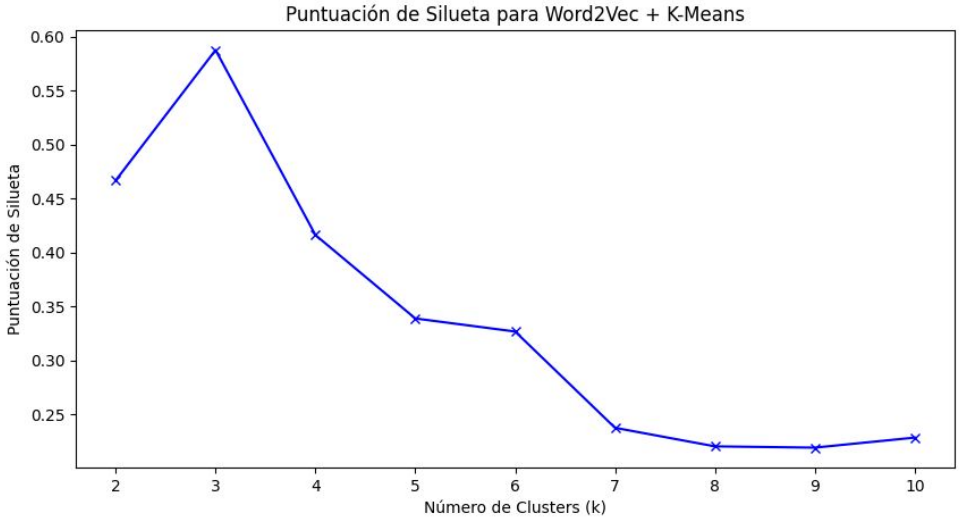
```
--- Modelo 1: TF-IDF + K-Means ---
Puntuación de Silueta para k=2: 0.011203482543527635
Puntuación de Silueta para k=3: 0.01102693039659593
Puntuación de Silueta para k=4: 0.018796974236023486
Puntuación de Silueta para k=5: 0.022058292651149652
Puntuación de Silueta para k=6: 0.01664635655562566
Puntuación de Silueta para k=7: 0.015504018788235508
Puntuación de Silueta para k=8: 0.01470309002682577
Puntuación de Silueta para k=9: 0.011965093947677164
Puntuación de Silueta para k=10: 0.015182630959034353
```

Resultados:
BoW/TF-IDF



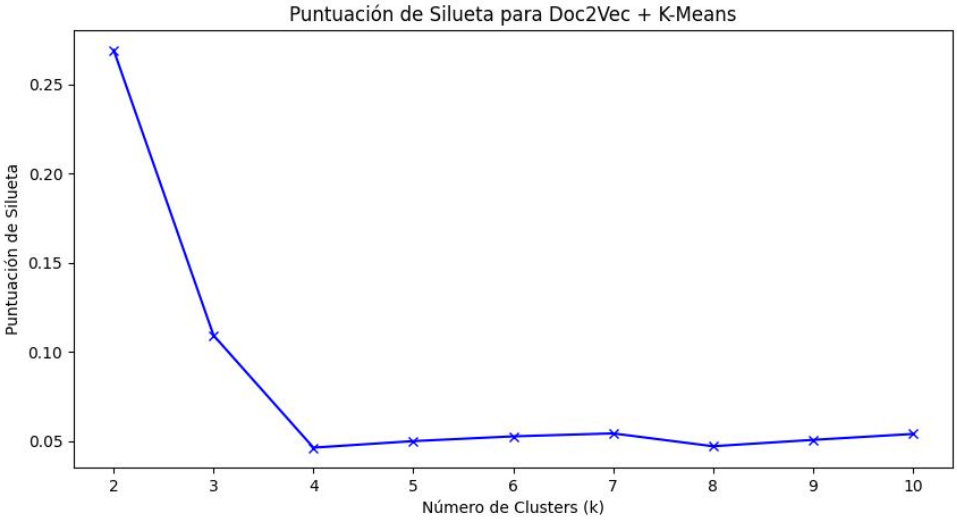
Resultados: Word2Vec

```
--- Modelo 2: Word2Vec + K-Means ---
Puntuación de Silueta para k=2: 0.46679569062314125
Puntuación de Silueta para k=3: 0.5876510919107314
Puntuación de Silueta para k=4: 0.4162096049255658
Puntuación de Silueta para k=5: 0.3388772472527688
Puntuación de Silueta para k=6: 0.3267518473367397
Puntuación de Silueta para k=7: 0.2375456114016033
Puntuación de Silueta para k=8: 0.22034020949644686
Puntuación de Silueta para k=9: 0.2191584926278471
Puntuación de Silueta para k=10: 0.22845657500649597
```



```
--- Modelo 3: Doc2Vec + K-Means ---
Puntuación de Silueta para k=2: 0.26922351121902466
Puntuación de Silueta para k=3: 0.1092555820941925
Puntuación de Silueta para k=4: 0.04630756750702858
Puntuación de Silueta para k=5: 0.04996853694319725
Puntuación de Silueta para k=6: 0.05263279378414154
Puntuación de Silueta para k=7: 0.05429334565997124
Puntuación de Silueta para k=8: 0.04709077998995781
Puntuación de Silueta para k=9: 0.05065475031733513
Puntuación de Silueta para k=10: 0.053970590233802795
```

Resultados: Doc2Vec



```
--- Comparación de las Mejores Puntuaciones de Silueta ---
TF-IDF + K-Means: Mejor k=5, Mejor Puntuación de Silueta=0.022058292651149652
Word2Vec + K-Means: Mejor k=3, Mejor Puntuación de Silueta=0.5876510919107314
Doc2Vec + K-Means: Mejor k=2, Mejor Puntuación de Silueta=0.26922351121902466
```

Comparativa