Prueba técnica ML Engineer
W & J

Aspirante: Jairo Andrés Hernández Mosquera

Contexto

HUFFPOST

HuffPost es un sitio web estadounidense sobre opiniones y noticias como política, cultura, bienestar, etc. Fue fundado en 2005 por Andrew Breitbart, Arianna Huffington, Kenneth Lerer y Jonah Peretti. El conjunto de datos abordado se compone de alrededor de 200.000 titulares con de noticias del año 2012 al 2018 obtenidos de HuffPost

Imágenes recuperadas de:

https://revistas.uca.es/index.php/eureka/article/view/5018

https://www.prisa.com/es/noticias/noticias-1/asi-es-el-nuevo-huffpost

El conjunto de datos

Imágenes recuperadas de:

https://thumbs.dreamstime.com/b/peri%C3%B3dico-con-los-titulares-92098359.jpg

Características

- ➔ **Titular**
- ➔ **Párrafo descriptivo corto de la noticia**
- ➔ **Autor**
- ➔ **Fecha**

Retos

Imágenes recuperadas de:

https://www.weblite.com.my/blog/use-words-that-sell-to-convert-readers-into-customers

1. ¿Se pueden catalogar las noticias con la descripción y los titulares? Compara tu clasificación con las categorías incluidas en el set de datos.
2. ¿Existen estilos de escritura asociados a cada categoría?
3. ¿Qué se puede decir de los autores a partir de los datos?
4. Ahora, utilizando técnicas de aprendizaje no supervisado, trata de identificar temas, “protagonistas” u otras entidades de las noticias.
5. Basándote en el texto de la descripción corta, caracteriza este dataset.
6. ¿Qué otra información útil se puede extraer de los datos?

Imágenes recuperadas de:

http://www.dcerca.es/

1. Clasificación

¿Es posible Catalogar una noticia nueva de forma automática?

2. Análisis Estilometrico

- Cantidades caracteres especiales
- Puntuación
- Longitud de las oraciones
- Medidas de contenido en el texto
- ….

- Diagramas
- Modelos de aprendizaje automáticos

→

Caracterización de una categoría (Entretenimiento, Crimen..)

3. Estilo del autor

Análisis Estilometrico

- Cantidades caracteres especiales
- Puntuación
- Longitud de las oraciones
- Medidas de contenido en el texto
- ….

- Diagramas
- Modelos de aprendizaje automáticos

→

Caracterización de un autor (Entretenimiento, Crimen..)

Análisis Descriptivo →

“El autor que siempre escribe de los mismos temas”

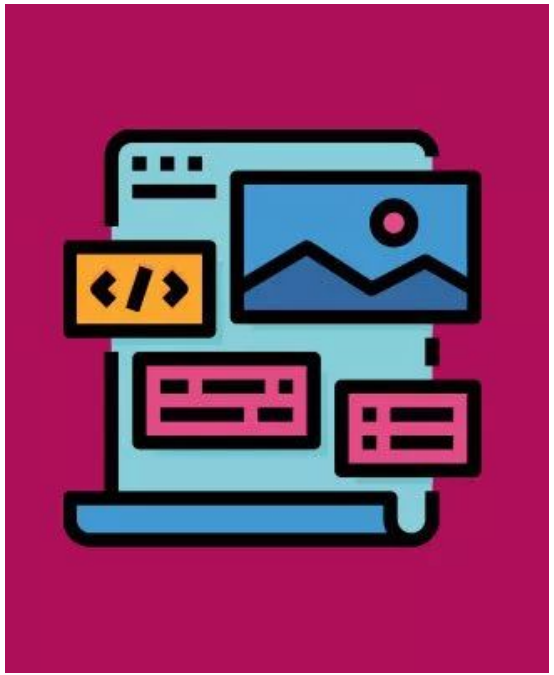
alta relación entre variables →

Diagramas, Modelos de aprendizaje automáticos

Imágenes recuperadas de:

hhttps://www.tutopress.co/blog/los-mejores-temas-para-usar-con-elementor/

4. Temas a tratar



¿De qué temas trata cada autor?

Imágenes recuperadas de:

hhttps://www.medium.com

5. Otra información a tener en cuenta

- La incidencia del tiempo en el contenido