**TERTULIA MATUTINA**

IDEA GENERAL:

Se trata de recopilar las noticias a diario de una serie de periódicos. De ahí sacar los topics más relevantes del día. Cruzarlos con la repercusión que están teniendo en Twitter y finalmente obtener alguna presentación donde se pueda ver cuáles son los temas más tratados en la prensa.

PIPELINE DEL PROYECTO:

1. Definir Data Set:

El dataset lo formamos nosotros scrapeando a diario diferentes periódicos de manera que recopilamos todas las noticias de estos periódicos.

En posteriores pasos será cruzado con información obtenida de Twitter.

1. Arquitectura y validación de los datos:

1ª opción hacerlo desde la nube. Si no fuera posible servidor de google y collaboratory.

En la opción de Google Cloud la idea es lanzar funciones a diario para recopilar las noticias. Con esa información construir nuestra base de datos y que sirva como dataset de entrada a nuestra red neuronal. A partir de ahí obtenemos una serie de topics (los más destacados) que alimentan una nueva consulta en Twitter y nos aporta más información en relación a esos topics destacados. Finalmente toda esta información se usa como entrada en una herramienta de visualización, Tableau, permitiéndonos representar aquellos aspectos que nos interesen.

1. Análisis Exploratorio:

Usamos tanto Tableau como el notebook de la API de Twitter para hacer visualizaciones de los datos que vamos obteniendo.

1. Preprocesamiento:

En el proceso de scraping, la noticia obtenida requiere de un tratamiento previo para poder disponer del texto real de la noticia. Hay que eliminar etiquetas de formato, links a otras noticias, signos de puntuación, etc…

A la hora de trabajar con la red será necesario otro pre procesado propio de NLP (normalizado del texto, conversión a minúsculas, eliminación de signos de puntuación, lematización, etc…)

1. Modelado:

Red neuronal que usando técnicas de NLP nos permite obtener los topics diarios.

1. Informe:

Presentación final con los temas más destacados del día y su repercusión en Twitter.