NLP

Para que sirve:

-Clasificacion: Binaria (1 o 0).

-Entity recognition: reconocer en el texto entidades (personas, empresas, fechas, lugares) y taggearlas.

-Traduccion

-Topic modeling: entender cuales son los temas que estan en un texto y extraerlos.

-Resumenes automaticos.

# BAG OF WORDS

Extraer un vocabulario de nuestro texto.

Extrar todos los terminos que intervienen en dicho texto y vamos a crear por cada texto un vector binario donde va a aparecer un 0 o un 1 segun aparece o no la palabra en el texto.

Tokenizacion: de que forma separamos nuestro texto en palabras.

Stopwords: que hacemos con los mismos

Rarewords: palabras que no aparecen en 5 o mas documentos podemos querer filtrarlas.

Stemming o lematizacion: Convertir palabras en una raiz comun.

Category y categorical la convertimos en "categ"

# TF-IDF

Tf: Term frequency.

IDF: Inverse doc frequency (log n/ft)

Generamos un vector de TF\*IDF y luego lo podemos utilizar en un modelo de ML como clasificador o red neuronal para predecir el label.

# N-Gramas

vector de ngramas, eligiendo cuales son los que usaremos, excluyendo lo mas raro y loas mas populares

Podemos usar alguna variante de FEATURE HASHING fijando la cantidad de columnas que voy a tener.

Los ngramas los voy a hashear entre un numero de 0 y k-1.

En un vector de tamaño fijo puedo representar ngramas ocupando menos espacio.

# EMBEDDINGS

un vector que representa a una palabra.

Esos vectores son densos, de tamaño 100-300 dimensiones aprox.

(ejemplos word2vec, glove, starspace)

Se entrengan de 2 formas:

-Por cada palabra tomo cuales son sus vecinos y trato de predecir cual es la palabra (continuos BOW).

-En base a la palabra tratar de predecir los vecinos (skipgrams).

Vectorizacion (transformar n embeddings en uno solo).

-Promedio

-Doc2Vec

-Sentence2Vec

# REDES NEURONALES CONVOLUCIONALES EN 1 DIMENSION (CONV 1-D)

Cada palabra tiene asociado un embedding.

La CONV 1-D va a usar filtros de un cierto tamaño, definimos n filtros de k palabras cada uno.

La red le va a aplicar este filtro a las n palabras y nos va a dar un numero.

Y asi lo va a aplicando a todas las palabras.

Con los resultados se le aplica un maxpooling y se queda con uno solo.

La red de 2 va a aprender a detectar "bigramas".

Si hay n filtros voy a tener n resultados.

A los resultados le aplico otra convolucion mas.

Llega a una softmax que nos dice a que clase pertenece cierto texto.

# OTRAS OPCIONES

CONV 1-D

LSTMS

BD-LSTMS

ATTENTION MODELS

RNNS

CHAR-RNNS

CHAR-CONV 1-D (idem anterior pero a cada caracter)