Natural Language Processing NLP

- و هي من أساسيات الـ ML & DL
- تختص بشكل أساسي في التعامل مع النصوص و ما يسمي تحليل الإنطباع Sentiment Analysis
- يتم استخدامها عبر اثنين من الموديلوز في sklearn وهي :

preprocessing.LabelEncoder	لتحويل النصوص في الفيتشرز الي ارقام
preprocessing.OneHotEncoder	لصناعة مصفوفة الواحد من النصوص
feature_extraction.text.CountVectorizer	لقراءة النصوص الطويلة و معالجتها

LabelEncoder

من اجل تحويل البيانات الغير رقمية (سترنج او بوليان) الي ارقام , من اجل عمل التوقع او التصنيف , يتم استيراد هذين المكتبتين من سكليرن
تسمية متغير باسم الدالة انكودر , مع مراعاة انها ستقوم باختيار كل قيمة مرة واحدة دون تكرار و يكون الهدف هو منع ادخال اي string للخوارزم لانه سيعجز عن فهمها, ولكن فقط ارقام يستطيع حسابها
, و يتم تطبيقها علي العمود المطلوب (الذي يحتوي علي الاسماء) فتتحول البيانات من نصوص الي ارقام

خطوات تنفيذها :

- تكوين الـ df
- استدعائها
- عمل الكائن
- تطبيقها علي العمود المحدد في الـ df
- عمل التحويل transform
- إضافة او تعديل العمود في الـ df بالارقام الجديدة
- يمكننا رؤية قائمة النصوص
- و يمكننا عمل التحويل العكسي

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

le.fit(df['score'])

le.transform(df['score'])

df['score'] = le.transform(df['score'])

list(le.classes_)

list(le.inverse_transform([2, 2, 1]))
```

OneHotEncoder

و هي تقوم بتحويل العمود الذي يحتوي علي نصوص الي عدد من الأعمدة الجديدة , يساوي عدد الكلمات المختلفة , **بحيث كل عمود يكون فيه اصفار و قيمة 1 فقط عندما تتواجد القيمة**

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	nan	Yes
France	35	58000	Yes
Spain	nan	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

	0	1	2	3	4
0	1	0	0	44	72000
1	0	0	1	27	48000
2	0	1	0	30	54000
3	0	0	1	38	61000
4	0	1	0	40	63777.8
5	1	0	0	35	58000
6	0	0	1	38.7778	52000
7	1	0	0	48	79000
8	0	1	0	50	83000
9	1	0	0	37	67000

خطوات تنفيذها :

- تكوين الـ df
- استدعائها
- عمل الكائن
- تطبيقها علي المصفوفة المطلوبة بعد تحويلها
- عمل التحويل transform و صياغته في مصفوفة جديدة , ثم عمل مقلوب لها
- إضافة او تعديل العمود في الـ df بالارقام الجديدة

```
from sklearn.preprocessing import OneHotEncoder

ohe = OneHotEncoder()

ohe.fit(data_ value)

ohe.transform(data_value).toarray()

df['Female'] = newmatrix[0]
```

CountVectorizer

و هي تقوم بقراءة النصوص الأطول , و حذف الكلمات المألوفة , ثم عمل وظيفة مشابهة لوظيفة LabelEncoder , و بعدها يمكن استخدام اي خورازم معين لعمل التصنيف او التوقع

و يتم استدعائها بالكود

```
from sklearn.feature_extraction.text import CountVectorizer
```