data&analytics
INNOVACIÓN Y TECNOLOGÍA

Uso de NLP y Machine Learning para Clasificación de Textos

data&analytics
INNOVACIÓN Y TECNOLOGÍA

Hola!

María Isabel Limaylla

Estudiante de Doctorado en Computación

maria.limaylla@gmail.com
www.linkedin.com/in/mariaisabellimayllalunarejo

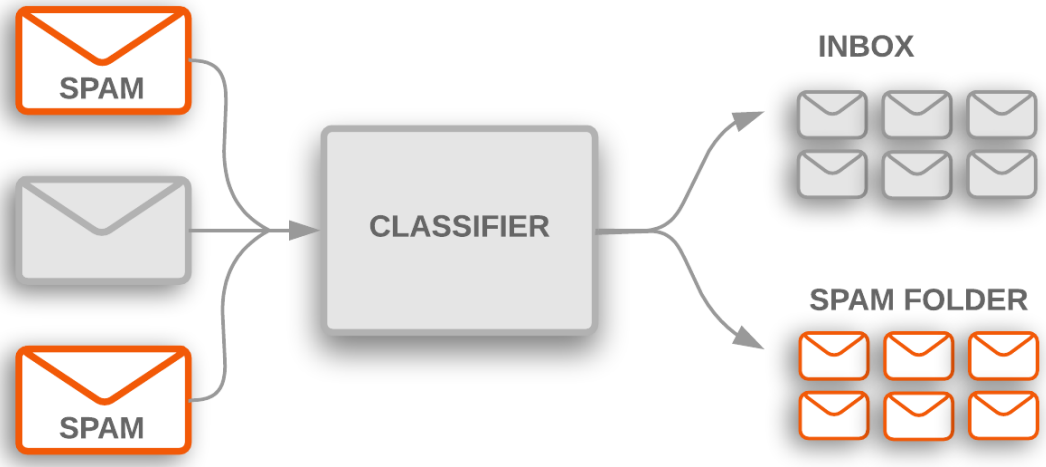
UNIVERSIDAD NACIONAL DE INGENIERÍA

UNIVERSIDADE FEDERAL DO PIAUÍ

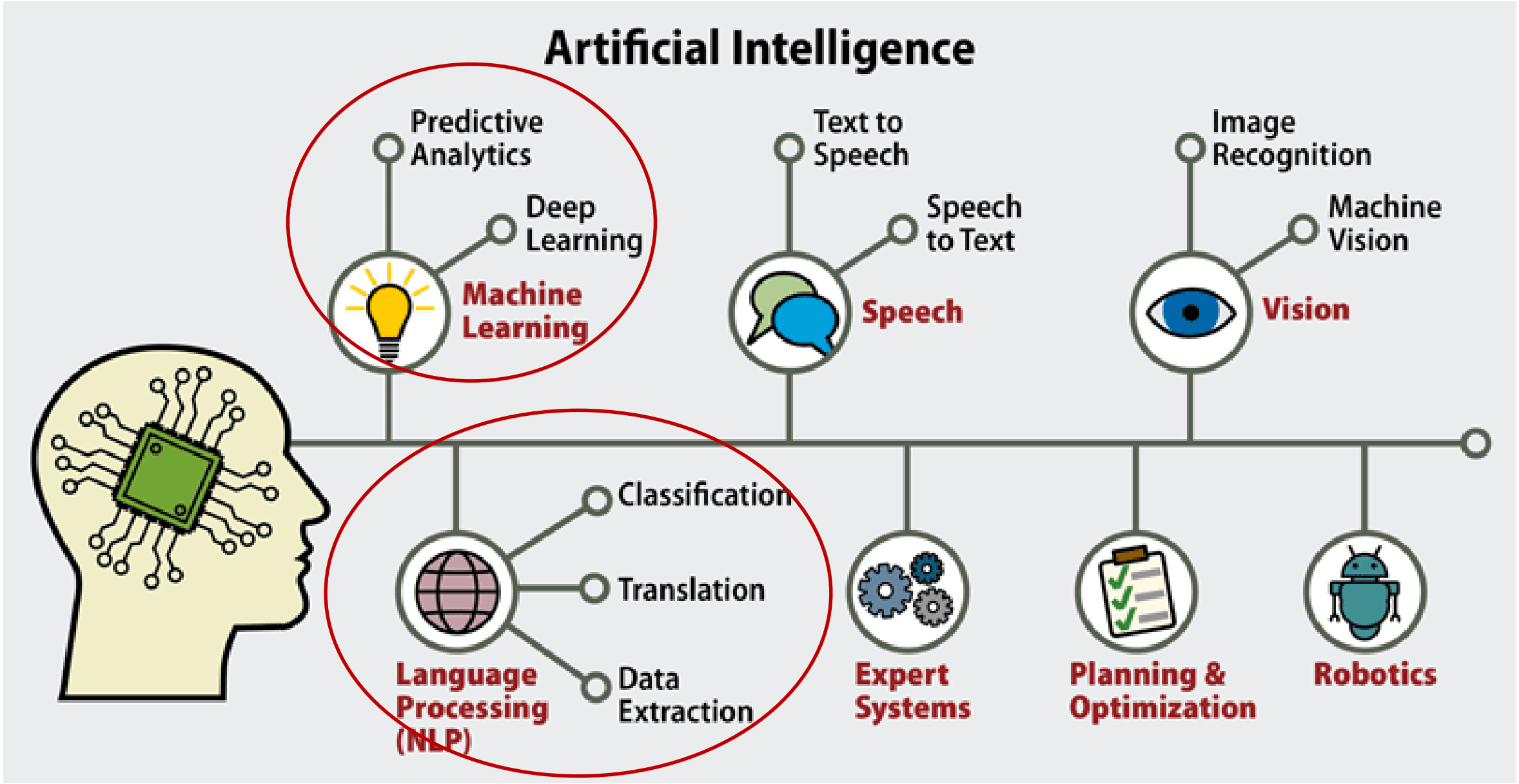
UNIVERSIDADE DA CORUÑA

Clasificación de Texto

- Clasificación de un texto en base a su contenido.
- Los clasificadores de texto se pueden utilizar para organizar y categorizar cualquier tipo de texto, desde documentos, estudios médicos y archivos.



https://developers.google.com/machine-learning/guides/text-classification



https://www.datamation.com/artificial-intelligence/what-is-artificial-intelligence.html

Python – Google Colab

data&analytics
INNOVACIÓN Y TECNOLOGÍA

Natural Language Analysis
with Python NLTK

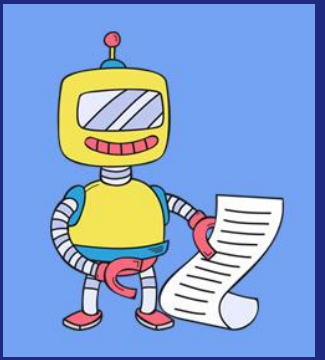
Google colab

data&analytics
INNOVACIÓN Y TECNOLOGÍA

1.Definición

Qué es Procesamiento de Lenguaje Natural?

- Lenguaje Natural: Comunicación por seres humanos
- Ejemplos: Inglés, Español, Portugués
- Procesamiento de Lenguaje Natural es cualquier manipulación computacional del lenguaje natural



Aplicaciones de NLP

data&analytics
INNOVACIÓN Y TECNOLOGÍA

Voice assistants (Alexa, Siri, Google Home etc)

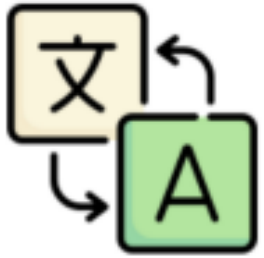
Question answering systems

Text summarization algorithms

Chatbots

Sentiment Analysis

Machine Translation



Técnicas de NLP

data&analytics
INNOVACIÓN Y TECNOLOGÍA

- Segmentación de Palabras (Tokenización)
- Limpieza de datos (stopwords)
- Análisis Lexicográfico (stemming, lematizing)
- Etiquetado Gramatical (POS)
- Vectorización de Texto

NLTK

data&analytics
INNOVACIÓN Y TECNOLOGÍA

Instalación:

```
import nltk
nltk.download('book')
```

Carga de Textos:

```
from nltk.book import *
text9
```

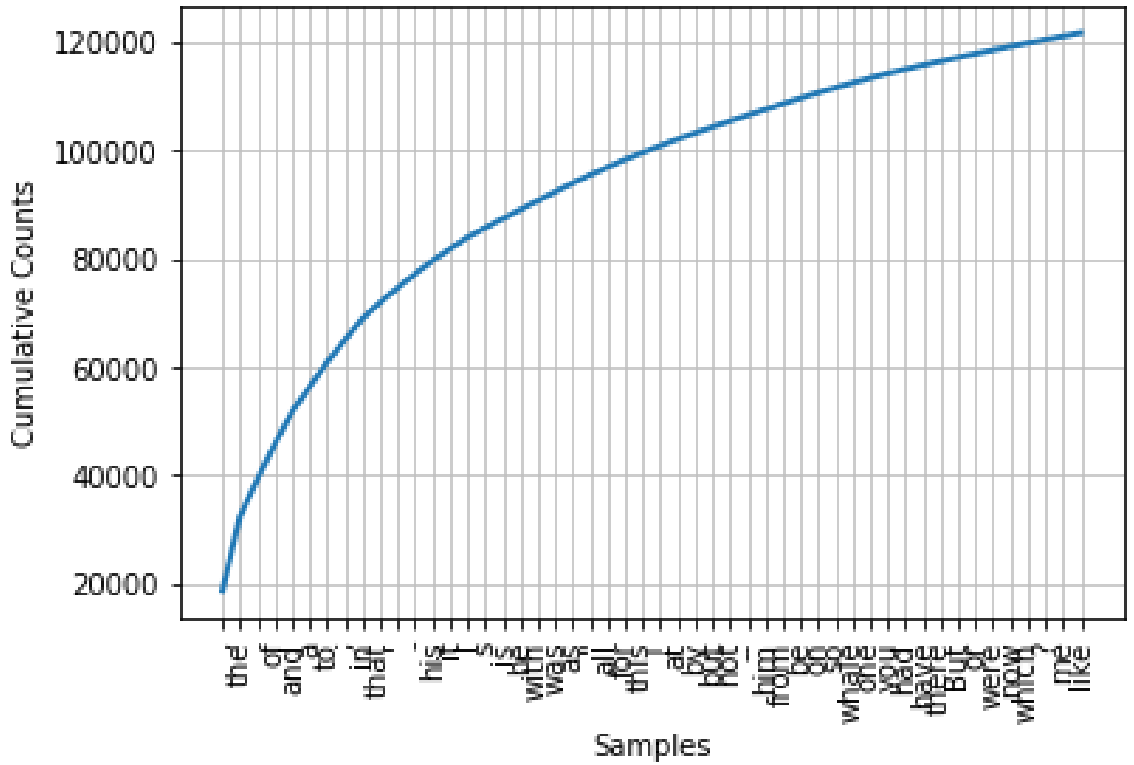
```
<Text: The Man Who Was Thursday by G . K . Chesterton 1908>
```

https://www.nltk.org/book/ch01.html

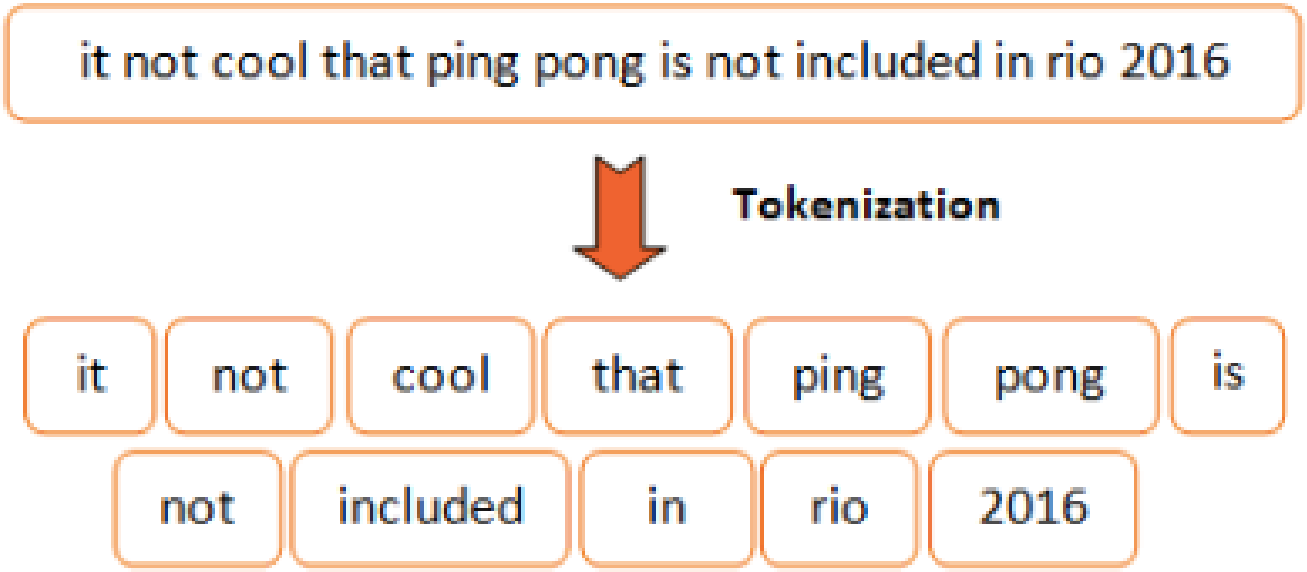
Distribución de Frecuencia

data&analytics
INNOVACIÓN Y TECNOLOGÍA

```
fdist = FreqDist(text1)
fdist.most_common(50)
fdist.plot(50, cumulative=True)
```



Tokenización




```
text = word_tokenize("And now for something completely different")
Tokens: ['And', 'now', 'for', 'something', 'completely', 'different']
```

Stopwords

data&analytics
INNOVACIÓN Y TECNOLOGÍA

Stop Words: remover palabras comunes pero que no proveen utilidad al descubrimiento del contexto (el, la, de, los, y, etc...)

```
from nltk.corpus import stopwords

def removeStopword(texto):
    stopwordSpanish = stopwords.words('spanish')
    textNew= [w.lower() for w in texto if w not in stopwordSpanish]
    print(textNew)
```

Stemming - Lematization

data&analytics
INNOVACIÓN Y TECNOLOGÍA

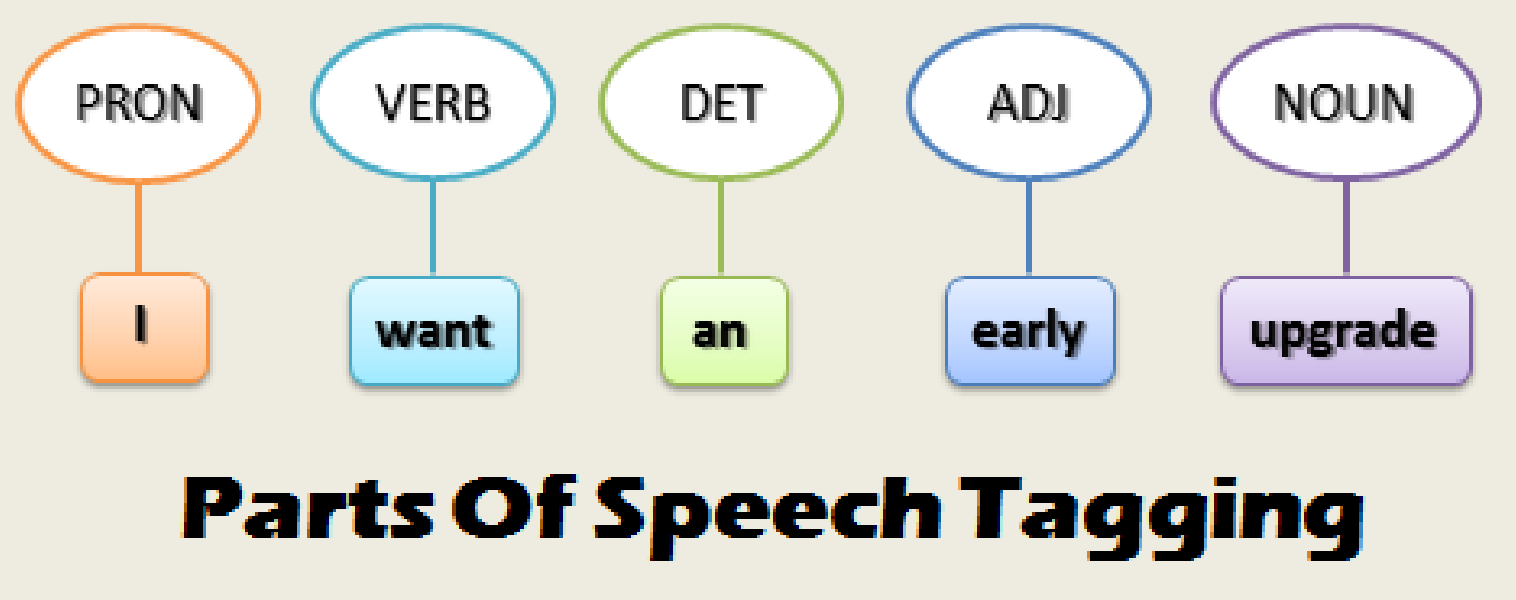
campo --> camp
casita --> cas
vendedor --> vend
panadería --> pan

comiendo --> comer
limones --> limón
corruptas --> corruptos
nueces --> nuez

```
tokens = word_tokenize(raw)
porter = nltk.PorterStemmer()
lancaster = nltk.LancasterStemmer()
print('PorterStemmer',[porter.stem(t) for t in tokens])
print('LancasterStemmer',[lancaster.stem(t) for t in tokens])
wnl = nltk.WordNetLemmatizer()
print('WordNetLematizer',[wnl.lemmatize(t) for t in tokens])
```

Etiquetado Gramatical POS

data&analytics
INNOVACIÓN Y TECNOLOGÍA



Etiquetado Gramatical POS

data&analytics
INNOVACIÓN Y TECNOLOGÍA

```
text = word_tokenize("And now for something completely different")
nltk.pos_tag(text)
```

```
[('And', 'CC'),
 ('now', 'RB'),
 ('for', 'IN'),
 ('something', 'NN'),
 ('completely', 'RB'),
 ('different', 'JJ')]
```

```
nltk.corpus.brown.tagged_words(tagset='universal')
```

```
[('The', 'DET'), ('Fulton', 'NOUN'), ...]
```

Vectorización de Texto

data&analytics
INNOVACIÓN Y TECNOLOGÍA

- One-hot Encoding
- Bag of Words (BOW)
- TF-IDF
- Incrustaciones de palabras (Word embeddings)

One-Hot Encoding

data&analytics
INNOVACIÓN Y TECNOLOGÍA

	a	cat	dog	is	it	my	old
"It is a dog."	1	0	1	1	1	0	0
"my cat is old"	0	1	0	1	0	1	1

Bag Of Words (BOW)

data&analytics
INNOVACIÓN Y TECNOLOGÍA

Simple example using **word counts**:

	a	cat	dog	happy	is	it	my	not	old	wolf
"It is a dog."	1	0	1	0	1	1	0	0	0	0
"my cat is old"	0	1	0	0	1	0	1	0	1	0
"It is not a dog, it a is wolf."	2	0	1	0	2	2	0	1	0	1

TF (Term Frequency)

data&analytics
INNOVACIÓN Y TECNOLOGÍA

Term frequency (TF): Increases the weight for **common** words in a document.

$$tf(term, doc) = \frac{number\ of\ times\ the\ term\ occurs\ in\ the\ doc}{total\ number\ of\ terms\ in\ the\ doc}$$

	a	cat	dog	is	it	my	not	old	wolf
"It is a dog."	**0.25**	0	**0.25**	**0.25**	**0.25**	0	0	0	0
"my cat is old"	0	**0.25**	0	**0.25**	0	**0.25**	0	**0.25**	0
"It is not a dog, it a is wolf."	**0.22**	0	**0.11**	**0.22**	**0.22**	0	**0.11**	0	**0.11**

IDF (Inverse document frequency)

data&analytics
INNOVACIÓN Y TECNOLOGÍA

term	idf
a	log(3/3)+1=**1**
cat	log(3/2)+1=**1.18**
dog	log(3/3)+1=**1**
is	log(3/4)+1=**0.87**
it	log(3/3)+1=**1**
my	log(3/2)+1=**1.18**
not	log(3/2)+1=**1.18**
old	log(3/2)+1=**1.18**
wolf	log(3/2)+1=**1.18**

Inverse document frequency (IDF): Decreases the weights for **commonly** used words and **increases** weights for **rare** words in the vocabulary.

$$idf(term) = log\left(\frac{n_{documents}}{n_{documents\ containing\ the\ term} + 1}\right) + 1$$

$$e.g.\ idf("cat") = 1.18$$

TF-IDF

data&analytics
INNOVACIÓN Y TECNOLOGÍA

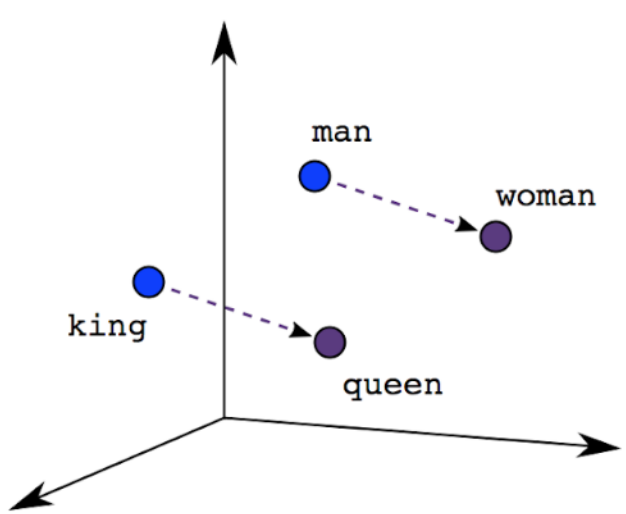
Term Freq. Inverse Doc. Freq (TF-IDF): Combines term frequency and inverse document frequency.

$$tf_{iaf}(term, doc) = tf(term, doc) * idf(term)$$

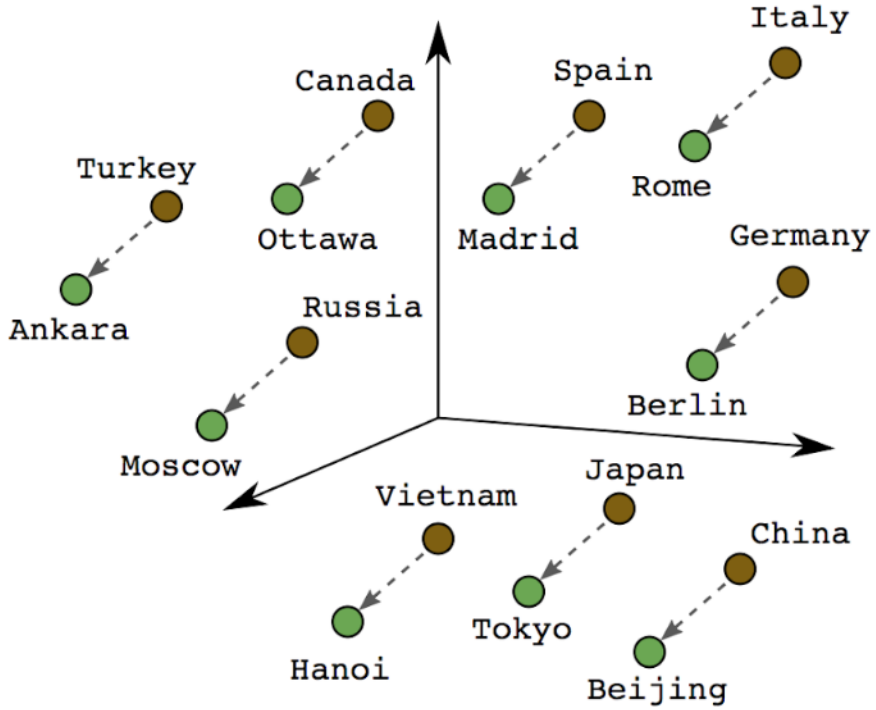
	a	cat	dog	is	it	my	not	old	wolf
"It is a dog."	**0.25**	0	**0.25**	**0.22**	**0.25**	0	0	0	0
"my cat is old"	0	**0.3**	0	**0.22**	0	**0.3**	0	**0.3**	0
"It is not a dog, it a is wolf."	**0.22**	0	**0.11**	**0.19**	**0.22**	0	**0.13**	0	**0.13**

Word Embedding

data&analytics
INNOVACIÓN Y TECNOLOGÍA



Male-Female



Country-Capital

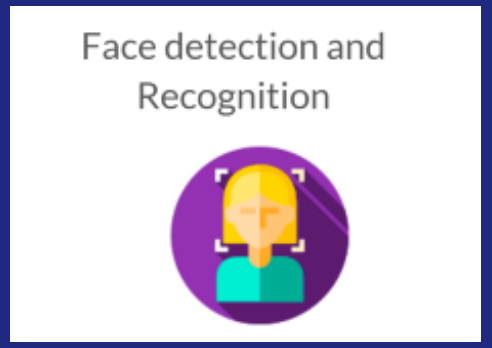
https://developers.google.com/machine-learning/guides/text-classification/step-3

data&analytics
INNOVACIÓN Y TECNOLOGÍA

2.Definición

Qué es Machine Learning?

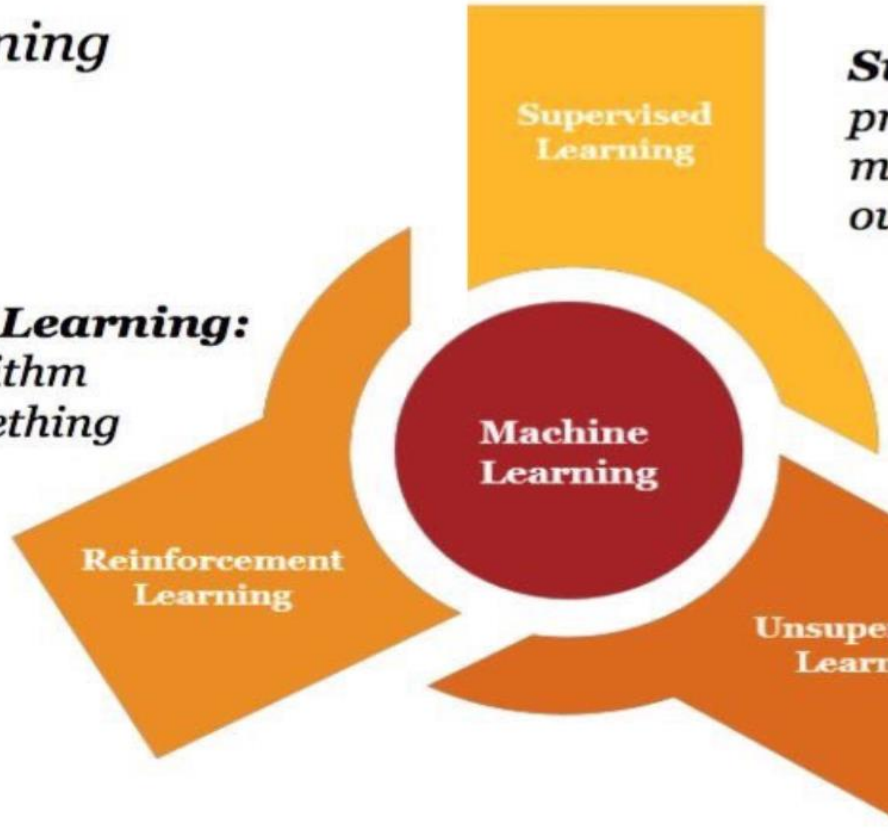
- El aprendizaje automático es un subcampo de la inteligencia artificial que utiliza algoritmos para aprender automáticamente cómo realizar una tarea determinada sin estar programado explícitamente con reglas.



Machine Learning : 3 Types of Learning

Background

Types of machine learning



Supervised Learning:
pre-labelled data trains a model to predict new outcomes

Example: Sorting LEGO blocks by matching them with the colour of the bags

Classification & Regression Algorithms

Reinforcement Learning:
feedback to algorithm when it does something right or wrong

Rewards & Recommendations algorithms

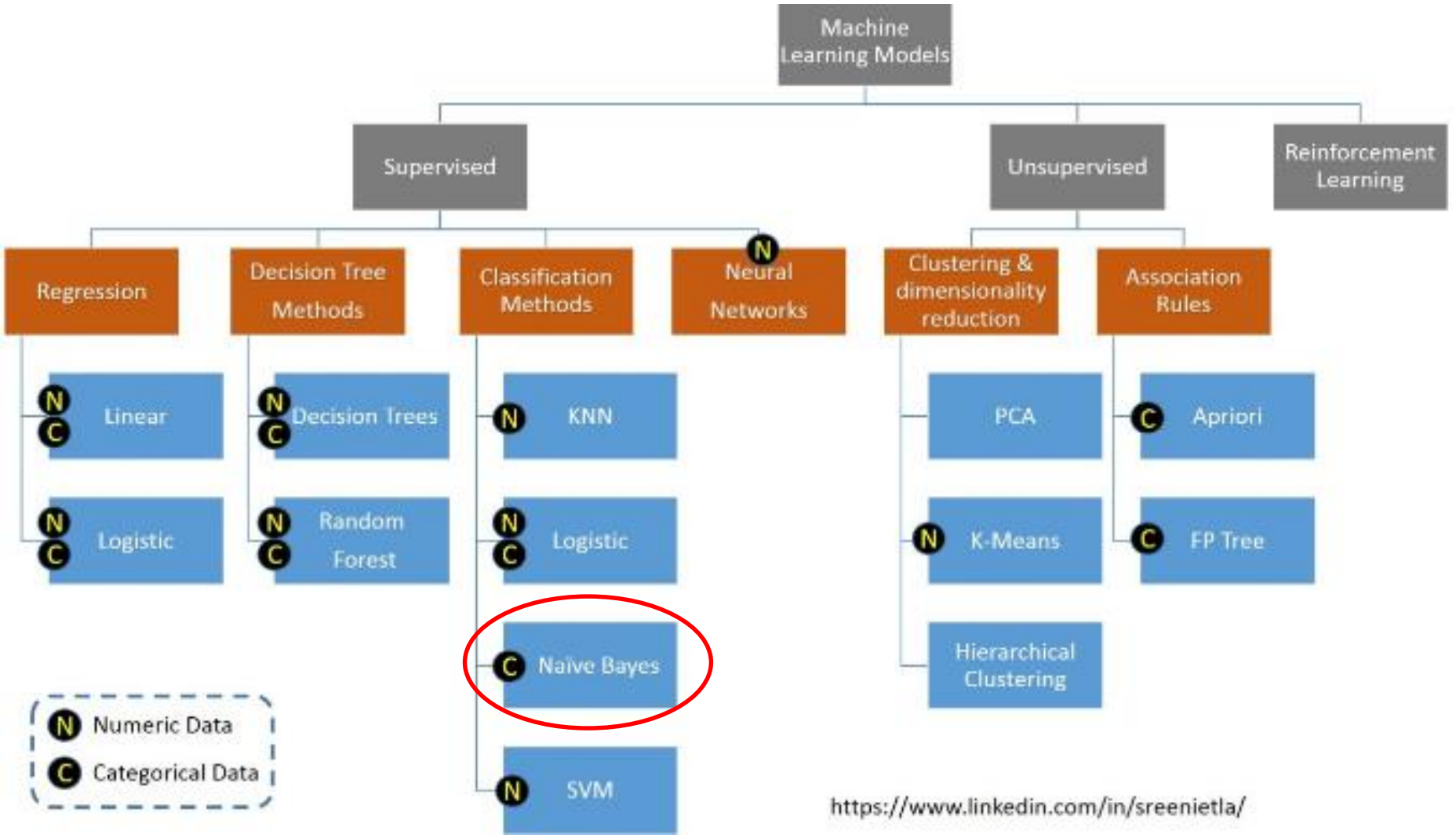
Example:
Child gets feedback 'on the job' when it does something right or wrong

Unsupervised Learning:
Non-labelled data self organises to predict new outcomes (e.g. clustering)

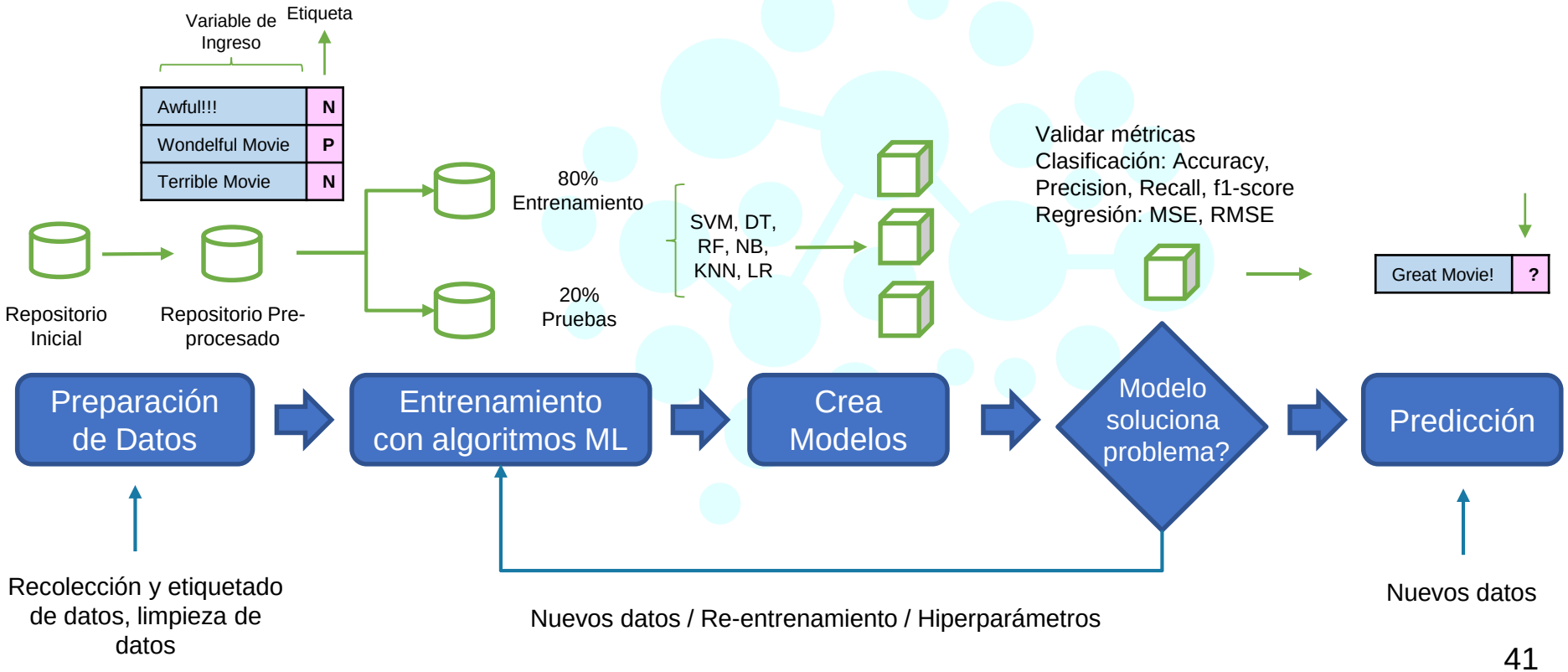
Clustering Algorithms

PwC

Source: PwC @mikequindazzi


Machine Learning Models
Supervised
Unsupervised
Reinforcement Learning
Regression
Decision Tree Methods
Classification Methods
Neural Networks
Clustering & dimensionality reduction
Association Rules
Linear
Logistic
Decision Trees
Random Forest
KNN
Logistic
Naive Bayes
SVM
PCA
K-Means
Hierarchical Clustering
Apriori
FP Tree
Numeric Data
Categorical Data
https://www.linkedin.com/in/sreenietla/

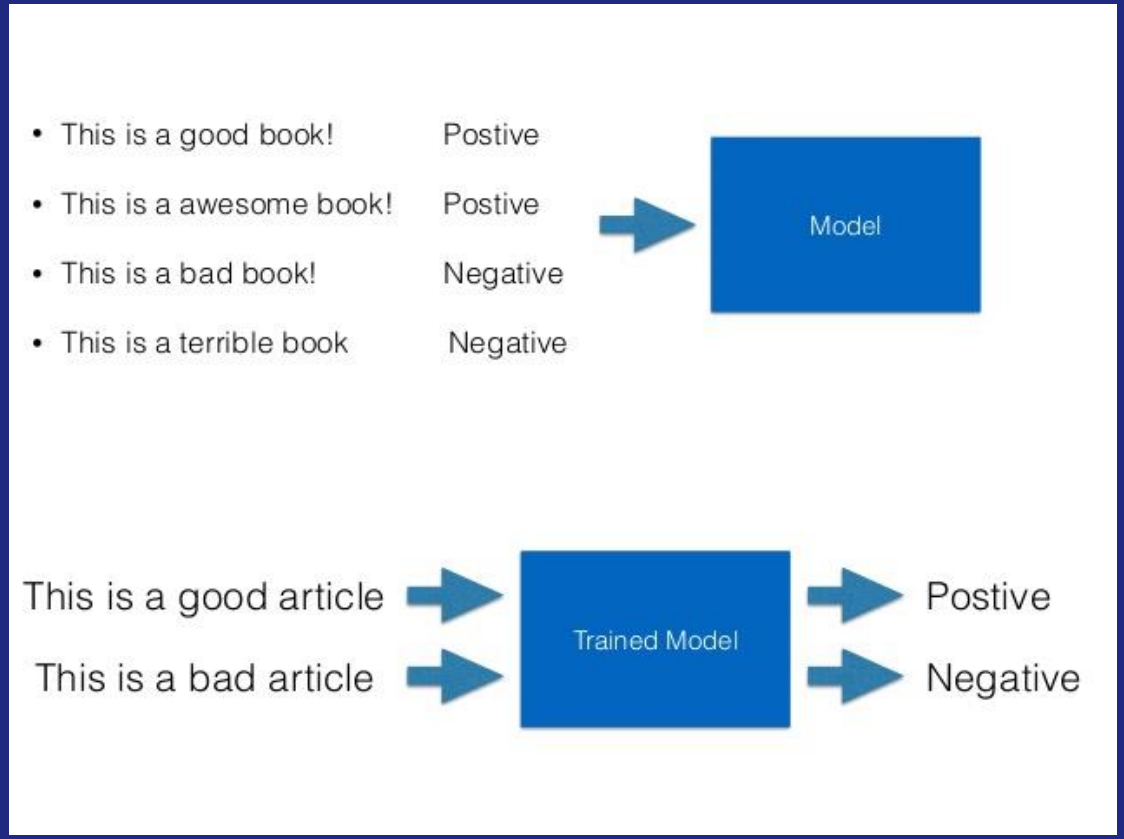
Ciclo de Vida Experimentos Machine Learning



3.Clasificación Supervisada

Y su uso para el análisis de sentimientos

Uso de procesamiento de lenguaje natural, análisis de texto y lingüística computacional para clasificar masivamente documentos de manera automática, en función de la connotación positiva o negativa del lenguaje ocupado en el documento.



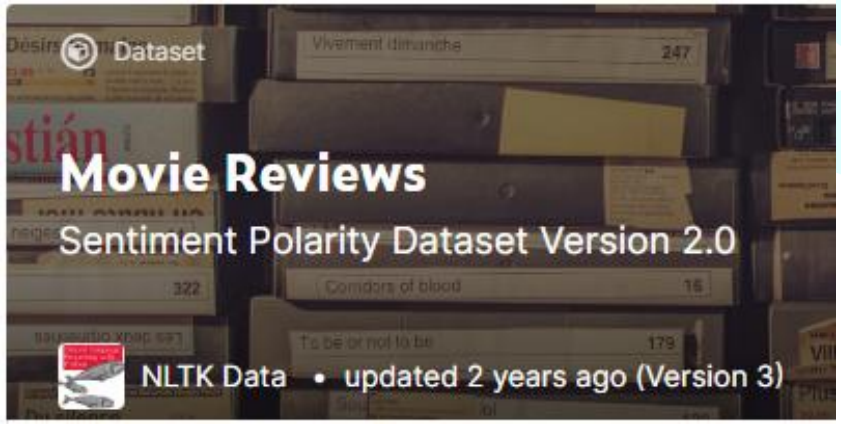
https://es.wikipedia.org/wiki/An%C3%A1lisis_de_sentimiento

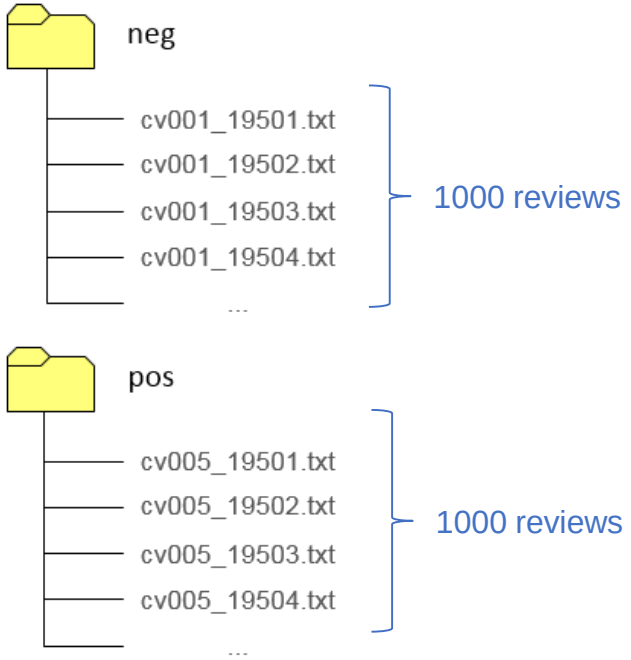
Para qué sirve?

- Conocimiento sobre el cliente (qué les gusta, qué no les gusta, problemas, etc.)
- Mejora en la gestión de la reputación en redes sociales.
- Analizar el impacto en usuarios de ciertos acontecimientos sociales, culturales, etc.

Movie Reviews

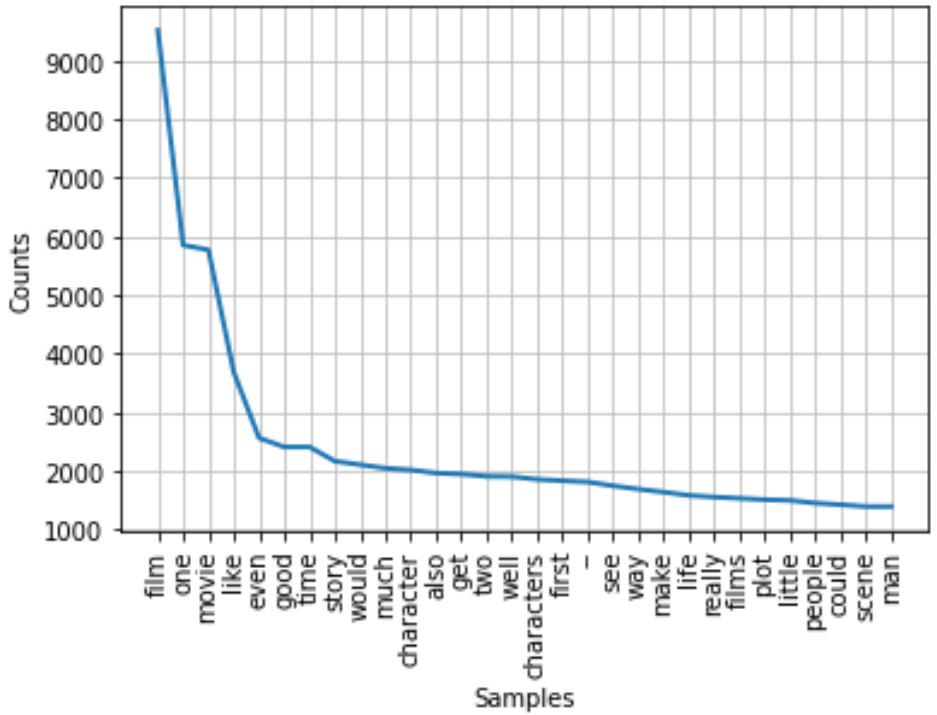
- 1000 reviews etiquetados como positivos y otros 1000 etiquetados como negativos.





Movie Reviews

```
total positivos 1000
total negativos 1000
1583820
710578
10 palabras más frecuentes [('film', 9517), ('one', 5852),
cantidad de veces que se repite la palabra happy 215
```



BOW

```
def tokenize(words):
    words_clean = []
    for word in word_tokenize(words):
        word = word.lower()
        if word not in string.punctuation and word not in stopword_2:
            words_clean.append(stemmer.stem(word))
    return words_clean

def bag_of_words_CountVec(X1):
    matrix_vectorizer = CountVectorizer(tokenizer=tokenize,analyzer='word',
                                        max_features=1000)
    X = matrix_vectorizer.fit_transform(X1).toarray()
    return X,matrix_vectorizer
```

BOW

data&analytics
INNOVACIÓN Y TECNOLOGÍA

```
def DatosBOW():
    df = pd.DataFrame()
    X1 = [movie_reviews.raw(fileid) for fileid in movie_reviews.fileids()]
    y = [movie_reviews.categories(fileid)[0] for fileid in movie_reviews.fileids()]
    print(len(X1))

    (X,count_vectorizer) = bag_of_words_CountVec(X1)

    return X,y,count_vectorizer
```

Naive Bayes Clasificador

data&analytics
INNOVACIÓN Y TECNOLOGÍA

```
def clasificadorBOW():
    (X,y,count_vectorizer) =  DatosBOW()
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,random_state=0

    mnb = MultinomialNB()
    classifier = mnb.fit(X_train, y_train)
    y_pred = classifier.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    print(accuracy)

    lista_texto = []
    lista_texto.append('This is an ugly movie.')
    lista_texto.append('This is an excelent movie.')
    lista_texto.append('This is an different movie.')
    prediccion = classifier.predict(bag_of_words_CountVec2(lista_texto,count_vectorizer))
    print(prediccion)
```

Matriz de Confusión

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$ (sensitivity)

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$F\text{-}score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)}$$

$$Specificity = \frac{TN}{TN + FP}$$

https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/

4.Información Adicional

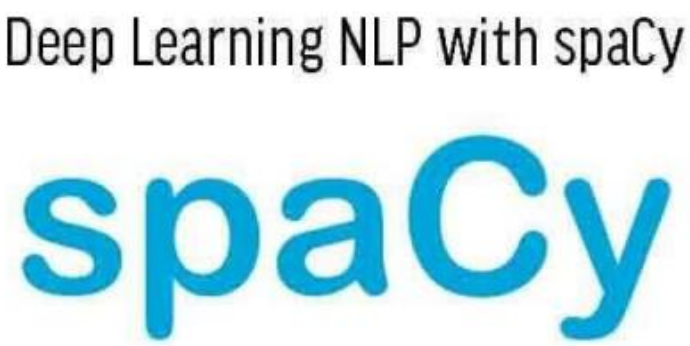
Datasets

- UCI Machine Learning Repositorios (http://archive.ics.uci.edu/ml/index.php)
- Kaggle Datasets (https://www.kaggle.com/datasets)
- Google Datasets Search (https://datasetsearch.research.google.com)
- Airbnb (http://insideairbnb.com/get-the-data.html)
- Lists of DataSets (https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research)
- En español (https://lionbridge.ai/datasets/22-best-spanish-language-datasets-for-machine-learning/)

https://es.wikipedia.org/wiki/An%C3%A1lisis_de_sentimiento

Herramientas para NLP

data&analytics
INNOVACIÓN Y TECNOLOGÍA



Herramientas para NLP

data&analytics
INNOVACIÓN Y TECNOLOGÍA

Herramientas para NLP – Java/R

data&analytics
INNOVACIÓN Y TECNOLOGÍA

Stanford Core NLP

Stanford University Natural Language Processing

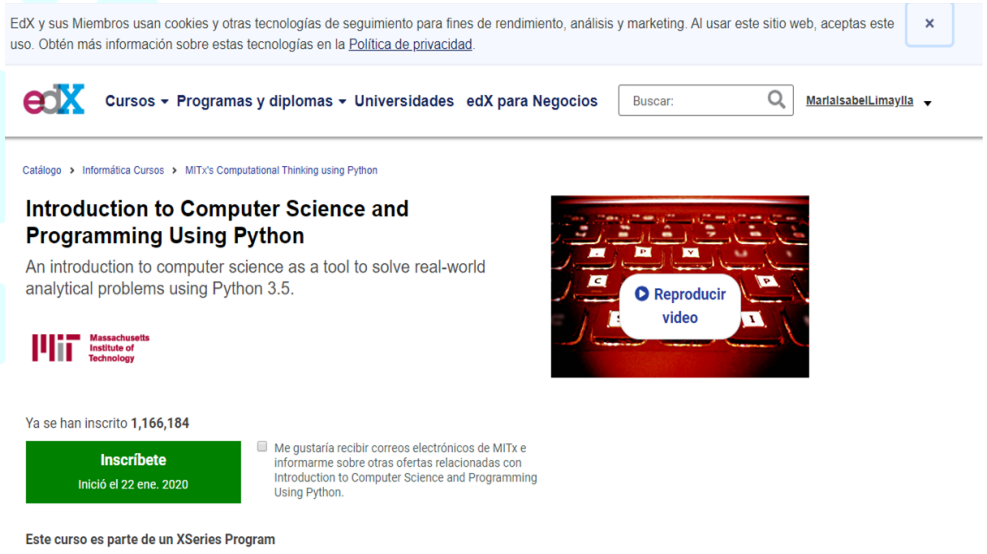
OPEN NLP™

Machine Learning con R y caret

mlr

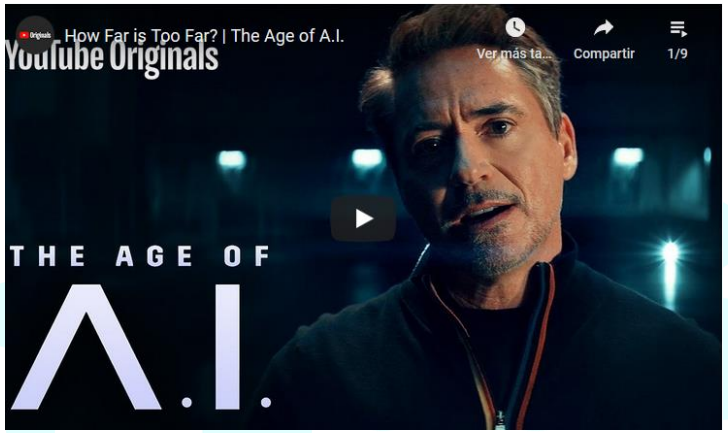
MOOCs

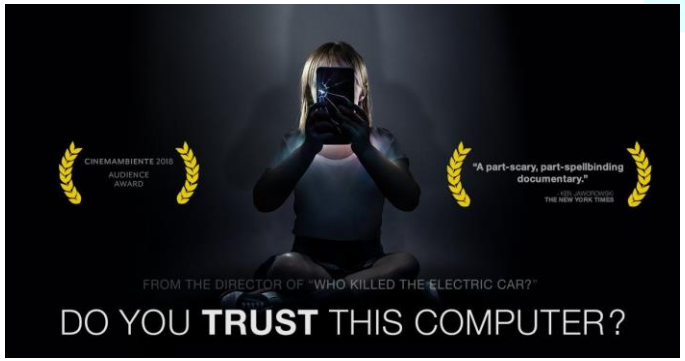
- Udemy – Curso Básico de Machine Learning
- EDX - The Analytics Edge
- EDX - Introduction to Computer Science and Programming Using Python
- Coursera - Machine Learning
- AWS Machine Learning Accelerator - Natural Language Processing

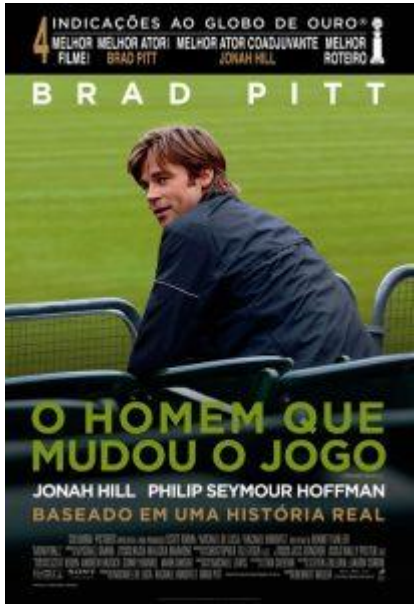


Videos

data&analytics
INNOVACIÓN Y TECNOLOGÍA







Gracias!!!

data&analytics
INNOVACIÓN Y TECNOLOGÍA

Alguna pregunta?

maria.limaylla@gmail.com
www.linkedin.com/in/mariaisabellimayllalunarejo