Pasos para hacer una búsqueda en lenguaje natural a partir de una query.

**Query = “Tienes refrescos coca cola?”**

**Frases = [    “Hay en existencia, de  leche para tomar?”,**

**“Venden refrescos de cola y, con gas?”,**

**“Tienen Jugos naturales embotellados?”   ]**

Respuestas = [     “Los lacteos son nutritivos”,

                              “Las gaseosas hace daño”,

                              “la fruta es mejor”   ]

1.- LIMPIAR EL TEXTO DE CARACTERES TIPOGRÁFICOS, Y SE PASA TODO A UN FORMATO COMÚN (minúsculas y sin acento, sin puntos, sin comas, sin signos de admiración, ni exclamación). Quitar palabras no alfabéticas

**Query = tienes refrescos coca cola**

**Frases = [    “hay en existencia de  leche para tomar”,**

**“venden refrescos de cola y con gas”,**

**“tienen Jugos naturales embotellados”   ]**

2.- ELIMINAR STOP WORDS  (es decir, quitar palabras que no aportan significado, tales como: a***, ante, con,contra, de,desde, e, el,ella***,…)

**Query = “tienes refrescos coca cola”**

**Frases = [     “hay en existencia de  leche para tomar”,**

**“venden refrescos de cola y con gas”,**

**“tienen jugos naturales embotellados”   ]**

**Query = “refrescos coca cola”**

**Frases = [    “existencia  leche  tomar”,**

**“refrescos  cola  gas”,**

**“jugos naturales embotellados”   ]**

***quedan sólamente keywords***

3.- LEMATIZACION Y STEMMING (consiste en agrupar términos y palabras comunes entre ellas en una sólo, por ejemplo: computación, computadora, computador, computandos 🡺computacion, eliminar plurales hacia singular) –**no se va a programar el punto 3**

**Query = “refrescos coca cola”**

**Frases = [    “existencia  leche  tomar”,**

**“refrescos  cola  gas”,**

**“jugos naturales embotellados”   ]**

**Query = “refresco coca cola”**

**Frases = [    “existencia  leche  tomar”,**

**“refresco  cola  gas”,**

**“jugo natural embotellado”   ]**

4**.- CREAR DICCIONARIO DE TEXTO** (una lista de todas las palabras –**keywords**—que aparecen tanto en la query como en el conjunto de frases). El diccionario es un conjunto NO HAY PALABRAS REPETIDAS

diccionario = {**refresco, coca, cola, existencia, leche, tomar,  gas, jugo, natural, embotellado**}

5.- CREAR VECTORES, UN VECTOR POR CADA FRASE.

**Query = “refresco coca cola coca”**

V(Query)    **= [1,2,1,0,0,0,0,0,0,0]**    el “refresco” aparece una vez en Query, “coca” aparece 1 vez en query,…

**“existencia  leche  tomar”,**

V(Frases\_1]= **[0,0,0,1,1,1,0,0,0,0]**

¿CÓMO CALCULO LA SIMILITUD ENTRE **V(Query) con V(Frases\_1)**

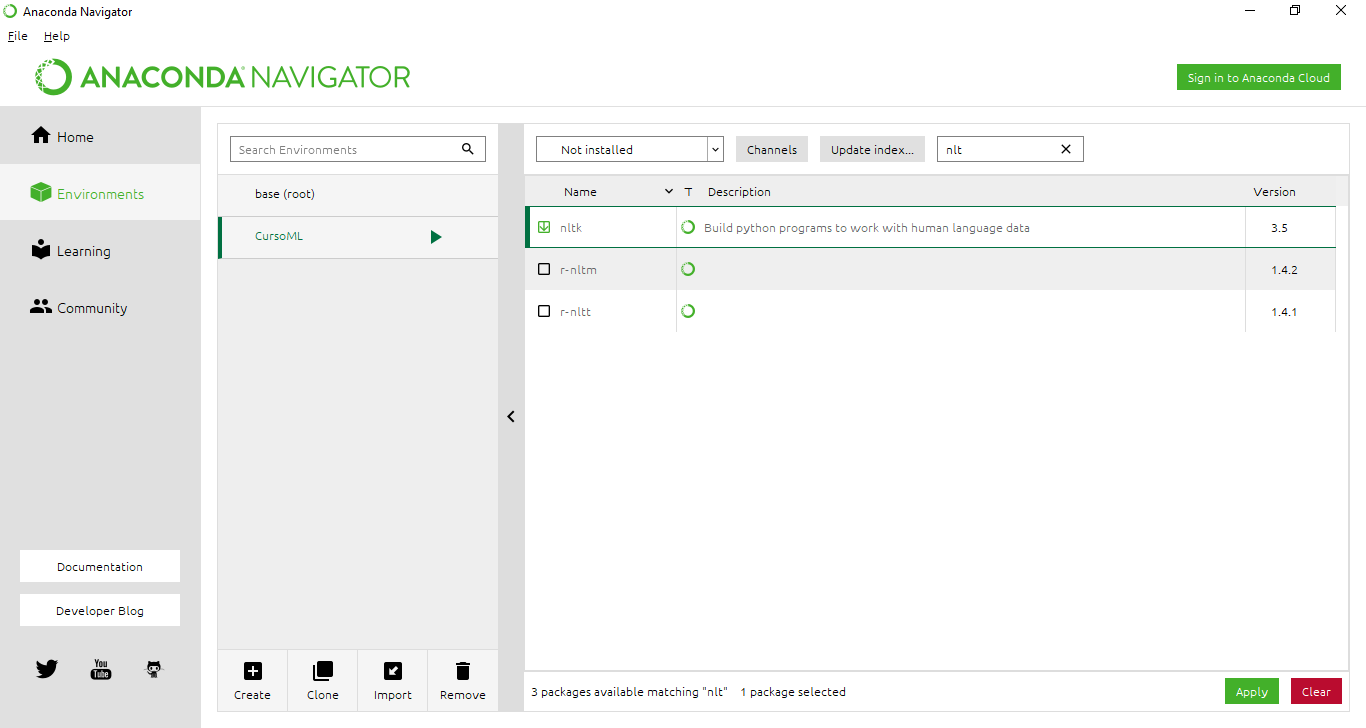
**“refresco  cola  gas”,**

V(Frases\_2] **=[1,0,1,0,0,0,1,0,0,0]**

¿CÓMO CALCULO LA SIMILITUD ENTRE **V(Query) con V(Frases\_2)**

Similitud(V(Query),V(Frase1)) = 1\*0 + 2\*0 + 1\*0 +

Instalar nltk



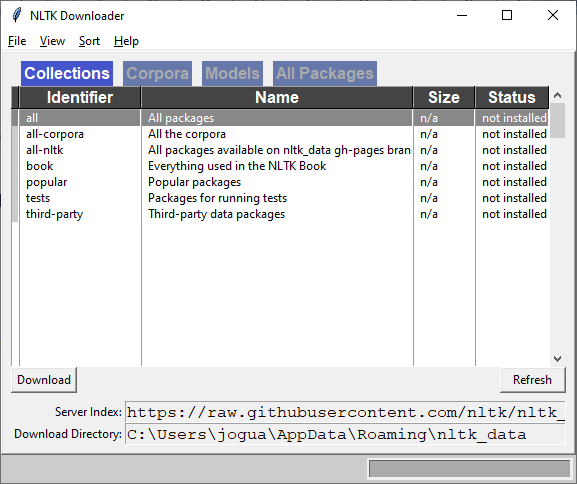
Avisará que va a instalar en total 6 paquetes por dependencias

**En Jupyter notebook teclear:**

import nltk

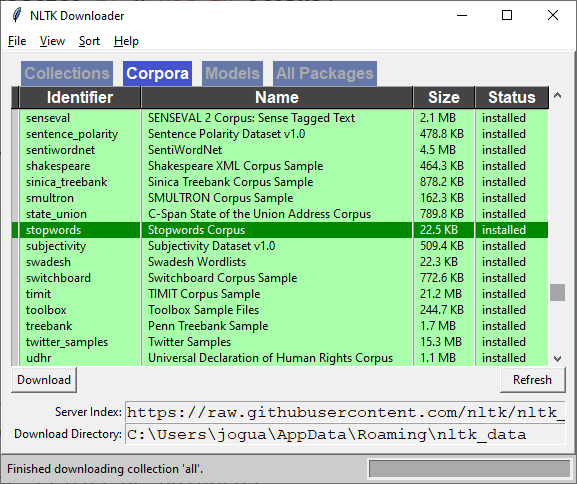
nltk.download()

y ejecutar, eso generará la siguiente ventana:



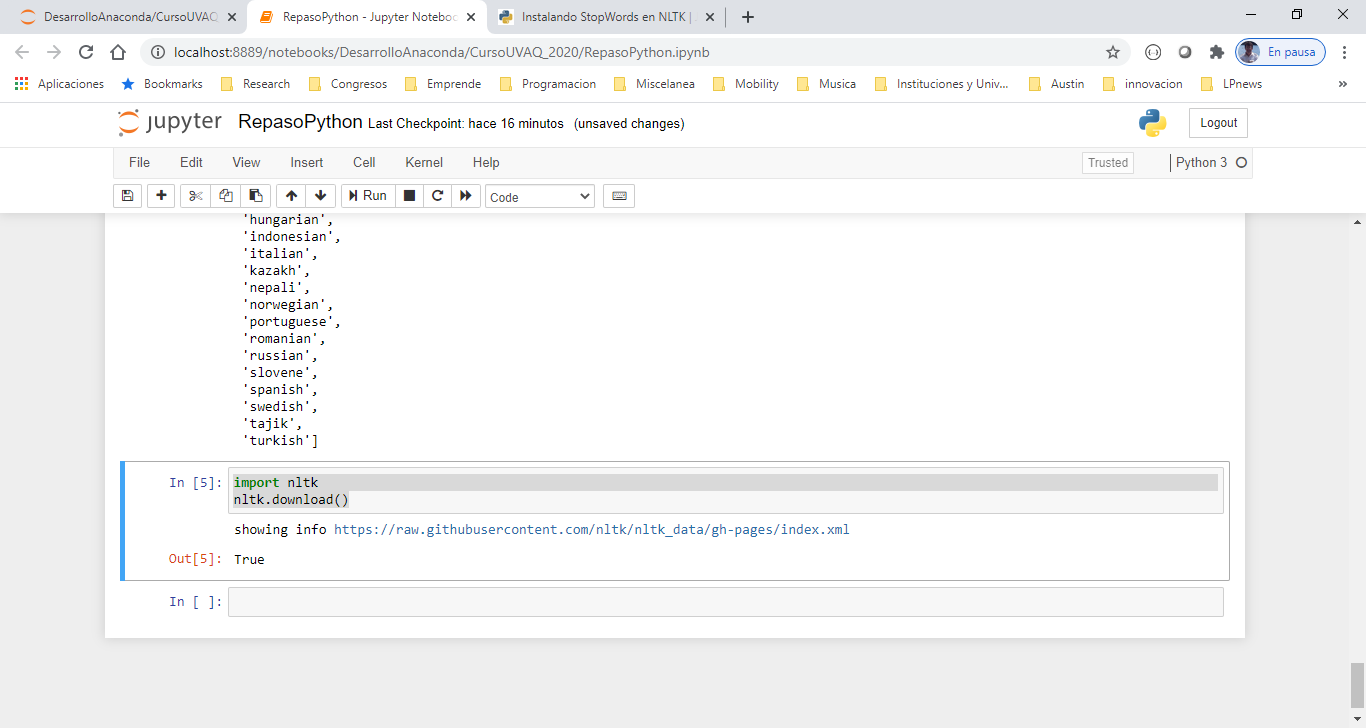
Seleccionar todo (all) y hacer click en Download

Para bajar exclusivamente el corpus stopwords



Y dar download

Al terminar la instalación notebook se ve así:



Escribir código parcial  
  
import nltk

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

frases = ["Hay en existencia leche para tomar?", "Venden refrescos de cola y con gas .,.,.,.?",

"Tienen Jugos naturales embotellados?"]

respuestas = ["Claro que tenemos leche lala","No pura pepsi","Jumex y Vigor nos proveen"]

## 1.- función para limpiar frases

def getCleanPhrase(listaPalabrasDeUnaFrase):

textoAlfabetico = [ w.lower() for w in listaPalabrasDeUnaFrase if w.isalpha() ]

stopWords = stopwords.words("spanish")

textoLimpio = [ w for w in textoAlfabetico if w not in stopWords]

return textoLimpio

## Función para crear vectores de una frase en forma de lista de palabras

def createVector(wordList, diccionario):

vector= [wordList.count(w) for w in diccionario ]

return vector

frasesCrudas\_a\_ListaDePalabras = [ word\_tokenize(f) for f in frases ]

print("\n Lista de listas de palabras: ",frasesCrudas\_a\_ListaDePalabras )

frasesLimpias = [ getCleanPhrase(f) for f in frasesCrudas\_a\_ListaDePalabras ]

print("\n Frases limpias en lista de palabras:",frasesLimpias)

query = input("\n Order >>>>")

queryLimpia = getCleanPhrase( word\_tokenize(query) )

print("\n query limpia: ",queryLimpia)

listaDeLasFrases = []

for l in frasesLimpias:

listaDeLasFrases = listaDeLasFrases + l

listaDeTodasLasPalabras = listaDeLasFrases + queryLimpia

diccionario = set(listaDeTodasLasPalabras)

print("\nDiccionario:",diccionario)

## Crear VECTORES

vectoresDeFrases = [createVector(l,diccionario) for l in frasesLimpias]

print("\n Lista de vectores:",vectoresDeFrases)

vectorQuery = createVector(queryLimpia,diccionario)

print("\n Vector del query:",vectorQuery)