Introducción al procesamiento De lenguaje natural

Parte 1

Agenda

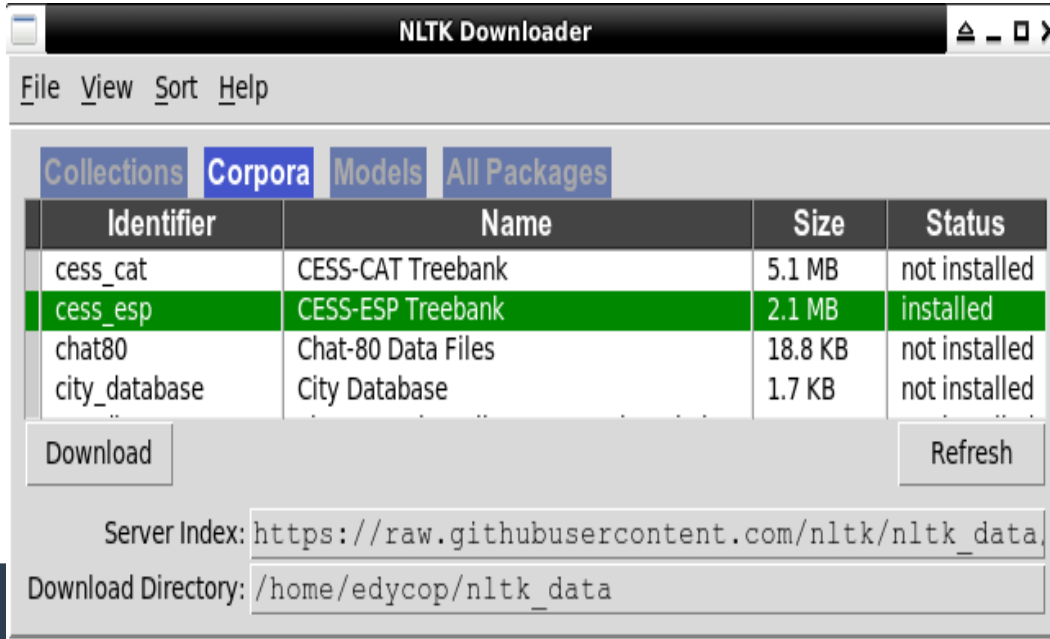
- **Técnicas**
  - Corpus de datos y Wordnet
  - Limpieza de datos
  - Pre-procesamiento
    - Tokenización
      - Palabras claves
    - Stemming - Frecuencia de términos
    - Lemmatización - Frecuencia del documento (IDF)
- **Enfoques**
  - Estadística
  - Redes Neuronales

Corpus de datos

- **Conjunto de entrenamiento**
  - 80%
- **Conjunto de prueba**
  - 20%

Corpus de datos

- **NLKT: (http://www.nltk.org/book_1ed/)**
  - kit de herramientas de lenguaje natural en Python
  - nltk==3.3 en un VirtualEnv
    - pip install nltk
    - >>> import nltk
    - >>> nltk.download()



Corpus de datos

- **Primer inconveniente**
  - Insuficientes corpus morfosintácticos y semánticos en español para investigación y si los hay ... son de pago!
  - Enfoque manual.
    - Web scrapping
    - Procesar archivos PDF
    - Procesar archivos .txt

Corpus de datos

- **Conociendo NLTK**
  - http://www.gutenberg.org/catalog/ (Spanish)
  - Ejemplo de archivo en texto plano: http://www.gutenberg.org/cache/epub/57303/pg57303.txt
    - f = open(file_name, 'r')
    - all_file = f.read()
    - print('=> [Caracteres]: {}'.format(len(all_file)))
    - print('=> [Titulo]: {}'.format(all_file[:69]))

Corpus de datos - Tokenización

- **“Tokenización”**
  - Separa signos de puntuación y palabras
  - Elimina caracteres en blanco y saltos de línea
- **>>> import nltk**
- **>>> nltk.download('punkt')**

- **from nltk import word_tokenize**
  - tokens = word_tokenize(all_file) [144946]
  - tokens = word_tokenize(all_file, '**spanish**') [144952]

Etiquetado PoS

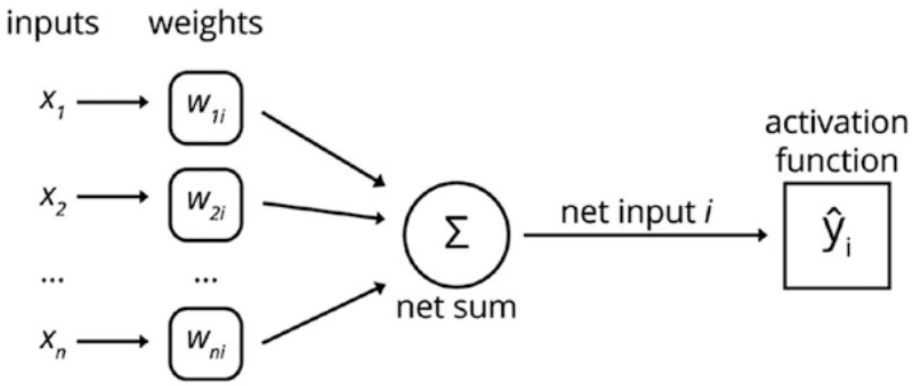
- **PoS = Part of Speech**
  - Asignar a cada palabra de un texto, la parte de la oración que le corresponde:
    - **Sustantivo** (nombre): algo que no cambia
    - **Pronombre**: es relativo a quien habla
    - **Adjetivo**: califica al sustantivo, expresa una característica o propiedad
    - **Articulo**: sustantivo de abstracto a concreto, “un libro” => “el libro”
    - **Verbo**
    - **Adverbio**: Ella trabaja *hoy*
    - **Interjección**: !ay!
    - **Proposiciones**: a, ante, bajo, cabe, con, contra, de, desde, durante, en, entre, hacia, hasta, mediante, para, por, según, sin, so, sobre, tras, versus, vía.
    - **Conjunción**: y, o u

Etiquetado PoS

- **Segundo gran inconveniente**
  - NLTK sólo tiene tagger (PoS tagger) para Inglés
  - https://nlp.stanford.edu/software/tagger.shtml
  -

Modelos

- **Modelos tradicionales de Redes Neuronales**
  - Perceptrón de una sola capa (SLPs)
  - Perceptrón multicapa (MLPs)



Referencias

- **https://pmoracho.github.io/blog/2017/01/04/NLTK-mi-tutorial/**
- **http://www.corpus.unam.mx/cursopln/plnPython/clase08.pdf**
-

Referencias

-
- **https://github.com/PythonNorte/intro_redes_neuronales_acuriale**
-

Continuara ...