Программа для извлечения метаописаний
из неструктурированных текстов
METADATA EXTRACTOR

Руководство администратора

Версия от 2018 г.

Введение

Программа METADATA EXTRACTOR предназначена для извлечения метаописаний из неструктурированных текстов формата pdf. Она помогает частично автоматизировать процесс подготовки метаописаний для системы GeoNetwork.

METADATA EXTRACTOR извлекает данные следующих видов:

1. Заголовок документа
2. Год публикации
3. Имена
4. Локации
5. Названия организация
6. Разное
7. Библиографические ссылки
8. Ключевые слова и словосочетания

METADATA EXTRACTOR работает под управлением операционной системы Windows и Linux.

Note: тестировалось на ОС Windows 10 и GNU/Linux Manjaro 17.1.10

Установка среды python

METADATA EXTRACTOR использует язык программирования Python 3.5.

Установка Python 3.5:

Windows

1. Качаем нужную версию с сайта: https://www.python.org/downloads/windows/
2. Запускаем скаченный .exe файл. Ставим галочку «Add Python to PATH», чтобы появилась возможность запускать интерпретатор без указания полного пути до исполняемого файла при работе в командной строке. Нажимаем «Install Now», Python установится в папку по указанному пути. Помимо самого интерпретатора будет установлен IDLE (интегрированная среда разработки), pip (пакетный менеджер) и документация, а также будут созданы соответствующие ярлыки и установлены связи файлов, имеющие расширение .py с интерпретатором Python. (Рис. 1)

«Рис. 1 Установка Python»

3. После успешной установки вас ждет следующее сообщение (Рис. 2).

«Рис. 2 Конец установки Python»

GNU/LINUX

Arch Linux

sudo yaoutr –S python35 - данный пакет расположен в репазитории AUR

Debian/Ubuntu

sudo apt install python3.6

Установка зависимостей

Установка пакетов для Python производится с помощью пакетного менеджера pip.

Установка Tensorflow

pip3 install --upgrade tensorflow

или

pip3 install --upgrade tensorflow-gpu

для tensorflow-gpu необходимо иметь видеокарту Nvidia с поддержкой технологии CUDA. Так же должны быть установлены: CUDA Tiilkit, cuDNN SDK, GPU driver, CUDA command line tools.

Note: Для более подробной информации смотри: https://www.tensorflow.org/install/

Note: Если tensorflow не работает, то надо просто поставить другую версию, так как библиотека зависит от аппаратной части системы.

Установка numpy

pip3 install numpy

Установка nltk

pip install nltk

далее запускаеи Python

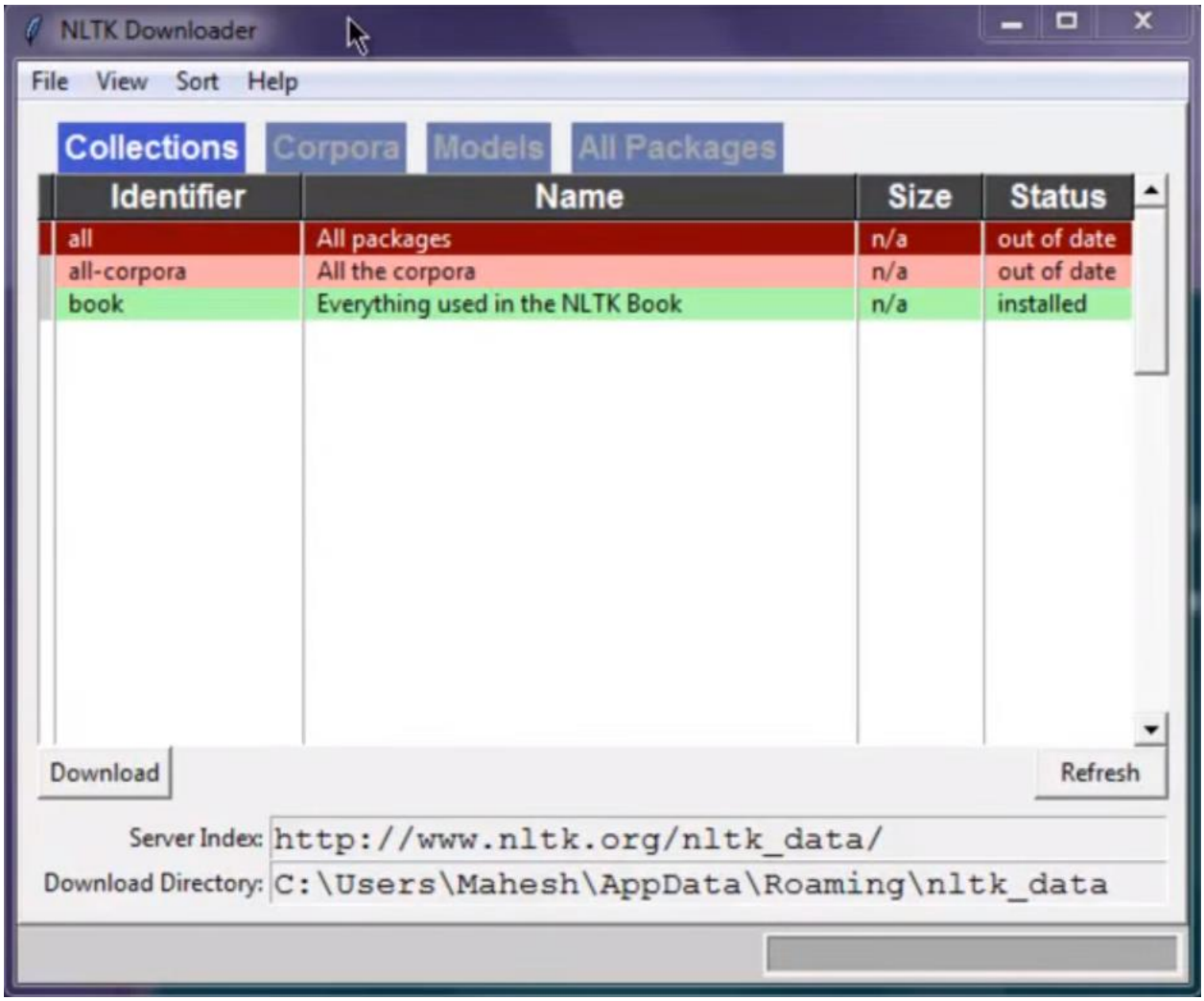
python

запустится среда python

>>> import nltk

>>> nltk.download()

Откроется окно рис. 3:



«Рис. 3 Установка nltk»

Нажимаем кнопку *Download*. После окончания загрузки nltk будет полностью установлен.

Установка pyqt5
pip3 install pyqt5

Установка requests
pip install requests

Установка argparse
pip install argparse

Установка pdfminer

pip install pdfminer.six

Установка PyPDF2

pip install PyPDF2

Теперь все необходимые компоненты установлены.