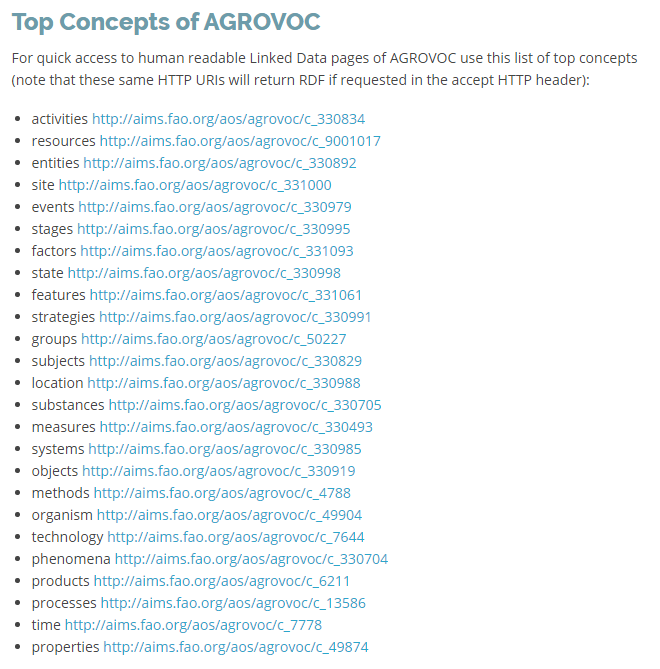
爬虫功能说明

**1. 背景：**

由于项目需要，现在需要官方关于农业实体的分类，那么所有农业名词分类应该构成一棵树，根节点就是root，第一层节点为所有农业品种的大类，即为top concepts。下一层又是每个top concept的小类。（如：小麦的下一层节点可能包括春小麦，冬小麦；春小麦的孩子可能包括一些小麦的品种名称）。

**2. 需求概述：**

现在需要爬取一些这样的信息，http://aims.fao.org/standards/agrovoc/linked-data页面的中间部分给出的是除了根节点的一些top concepts，如下图所示：



即根节点下面就包括上面这些节点，每个节点下面都有它的孩子节点，比如可以打开第一个activities后面的链接，可以得到如下界面：



左上角就是该节点的信息，包括地址和该节点的名称。我们需要的信息就是这个节点的名称，和该节点的孩子信息，而孩子节点的信息就存储到narrower这一块。可以注意到narrower这一块含有很多链接地址，每一个地址就代表一个孩子。比如我们可以打开第一个孩子的链接<http://aims.fao.org/aos/agrovoc/c_130> ，可以得到如下界面：



其实所有界面都是一样的，左上角是该节点的一些信息，administration即为该节点的名称，narrower这一块就是该节点的孩子信息。我们需要的也就是这两部分信息，即节点名称和narrower这一块的信息。

再示范一次，我们打开这个节点的第一个narrower，即地址为<http://aims.fao.org/aos/agrovoc/c_28938>，打开后可以得到如下界面：



Public administration即为该节点的名称，narrower块为该节点的孩子地址链接，可以看到该节点只有一个孩子，链接地址为<http://aims.fao.org/aos/agrovoc/c_37882>，打开后可以看到一个名称为governance的节点，它也只有一个孩子，地址为<http://aims.fao.org/aos/agrovoc/c_aca7ac6d>，打开后可以看到一个名称为land governance的节点，如下所示：



我们可以注意到，该节点没有narrower这一块，这就意味着该节点是叶子节点，没有孩子节点，那么我们此次深度遍历就到此结束，下面就返回上一层遍历其他没有访问到的孩子。总的来说，爬取信息的过程就类似一个深度遍历的过程，遍历到的每个节点要爬取其名称以及其孩子节点。

**3，结果处理：**

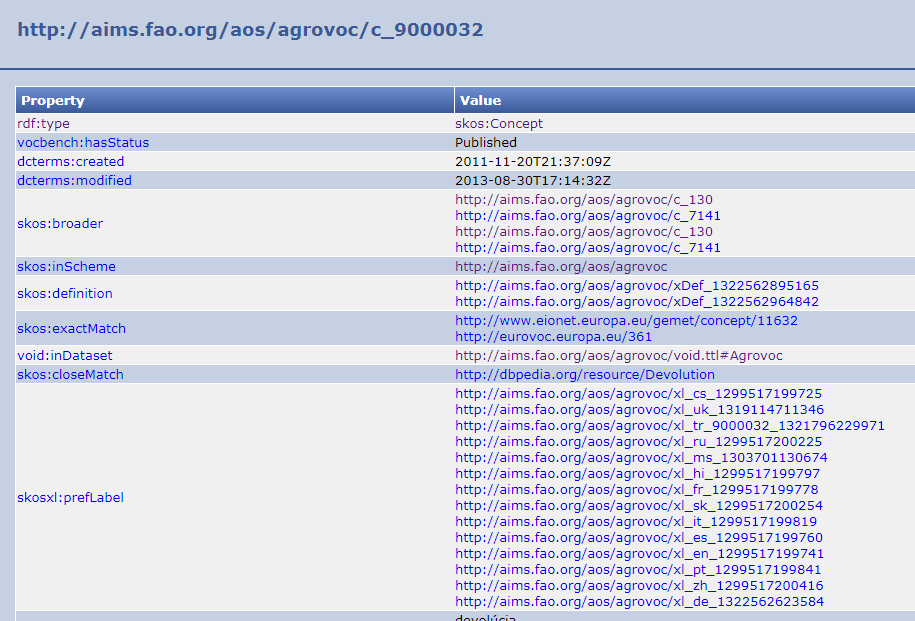
当通过深度优先遍历把所有节点的名称爬取下来后，可以得到一个森林的结构（因为没有根节点），那么我需要的结果就是人为添加一个根节点，将其生成一棵树，以json的格式输出，得到一个json文件。

同时，还需要一个包含所有节点名称的txt文档，文档结构应当按如下方式存🡪节点：子节点1，孩子节点2，孩子节点3……（

如：activities：administration，maintenance，surveying，insulation……

如：administration：public administration，land administration

注意：在打开各个节点过程中可能会存在一些节点只有链接地址，没有名称的现象，此时我们可以直接存储其节点编号，具体情况如下：



该节点没有节点名称信息，那么我们要存储其节点编号：c\_9000032，即链接地址的最后部分。