How to build a De-jargonizer in new language?

1. Download and install Python 2.7 and Scrapy 1.3. these will be used for creating 2 crawlers, One for getting the URLs to create the vocabulary matrix, and the second for reading the URLs.
2. Copy the contents of our Crawler folder (<https://github.com/NoamAndRoy/JargonProject/tree/master/Crawler>).
   1. seed.txt is the starting URL for the crawler – change this file so it will fit your desired initial URL
   2. "scienceCategoris.txt" contains prefixes of URLs that will be categorized as "Science".
   3. BBCSpiders.py contains the rules for the crawler for moving through the pages and the rules for parse pages, this file should be changed based on the site you want to crawl (it's important to remember to change the "name" variable based on what you called the crawler).
3. Run the first crawler.
4. There are 2 outputs: "GeneralUrls.txt", "ScienceUrls.txt" during the software development we used these two as separate files, but at the end we combined them together.
5. After you finished with the first crawler, copy the contents of the Crawler(<https://github.com/NoamAndRoy/JargonProject/tree/master/GeneralSpider>) folder.
   1. BBCSpider.py contains the rules that tells the crawler how to parse pages and extract the words from them, this file you will need to change based on the site you want to crawl.
   2. Copy the file generated by the first crawler "GeneralUrls.txt" to the crawler folder.
6. Run the second crawler.
7. You need to create an interface to let the users use your De-jargonizer, the TextGrading class (<https://github.com/NoamAndRoy/JargonProject/blob/master/Website/Jargon/Handlers/TextGrading.cs>) is used to analyze a text based on the matrix you created.

You will need to change it based on the language you use.

* 1. public static string CleanWord(string i\_Word) removes extra characters from a string.
  2. private static List<string> getCleanedWords(string i\_Text) return a list of words from text.