WunderSchild  
Back End Developer Test Task

As a test task, we ask you to write a Python parser to extract structured and unstructured data from several websites.

Your solution should include everything needed to run the system (source code, requirements list, database settings, etc.) You can send your solution as a zip archive or you can give us a link to a GitHub repository where your code is uploaded. It is even better to have a Docker container with everything ready to run the parsing in one command.

Please **do not be afraid to ask us any questions**!

**Code style and architectural decisions are important.**

Please write clean production-like code, but keep track of the time. Partial solutions are welcome too, but full solutions are better :)

# Task

You are asked to parse business data from a global business data registry . We are interested in extracting all information available for a big number of companies.

Please store company and personal information to a database of your choice (Neo4j or MongoDB.) It is very important to store the connections between companies and people.

It is advised to use methods of concurrent, parallel, asynchronous programming for parsing.

## Company Information

As a reference, let’s look at a Company page available at <https://opencorporates.com/companies/de/V1109_HRB347>

We want to extract:

1. **Company Number**: V1109\_HRB347
2. **Native Company Number**: Saarbrücken HRB 347
3. **Status:** Currently Registered
4. **Jurisdiction**: Germany
5. **Registered Address** : Industriestraße, 66280 Sulzbach. Germany
6. **Previous Names** :"H Y D A C Technology GmbH"
7. **Directors / Officers :**
   * Andreas Börgmann, prokurist, 6 Sep 2016-
   * Edgar Doctor Weishaupt, prokurist, 11 Dec 2018-
   * Frank Doctor Bauer, prokurist, 2 Feb 2012-
   * Mark Cezanne, prokurist, 29 May 2017-
   * Reiner Grub, prokurist, 25 Jan 2007-
8. **Inactive Directors / Officers**
   1. Klaus Doctor Goebbels, prokurist, - 9 Jan 2008
   2. Renate Huber, prokurist, -24 Feb 2014
   3. Wilfried Klein, prokurist, 29 Jan 2008- 8 Aug 2018

We would need also:

1. Latest Events ( information below this section )
2. Similarly named companies ( information below this section with links )

## Director Information

As a reference, let’s look at a Director page available at <https://opencorporates.com/officers/313483609>

We want to extract:

1. **Company**: "H Y D A C Technology GmbH"
2. **Name**: Andreas Börgmann
3. **Address**
4. **Position**: prokurist
5. **Start Date**: 2016-09-06
6. **Similarly named officers** (information from this section, with link and names of the company form each link)

## Iteration :

In order to extract all info about companies in a given country (in our case Germany) we need to extract the list of valid companies IDs (i.e. V1109\_HRB347).

For example other valid IDs are : T2304V\_HRB12140, R3306\_HRB88180, etc…

The list will allow to perform iteration and to extract all the relevant information about companies in that country.

The candidate should find an algorithm that generates partially or fully the list of valid companies ID.