Articles relating to industry sectors

Elasticsearch Pilot project

28 Nov 2018

The purpose of this document is to provide the scope, objectives and deliverables for this Proof of Concept pilot.

**Objective:**

To identify which articles on NewsAPI relate to which industry sector using Elasticsearch and the descriptions of industry codes (NACE code taxonomy).

**Outcome:**

Store the articles in a PostgreSQL database along with the set of industry codes which best categorise the article. Each article which refers to one or more industry codes should be stored in the database. For each article, please capture meta-data including:

* Article title
* Article source (in our example all articles will come from NewsAPI)
* Date and time stamp
* Each of the NACE industry codes matching the articles (many articles will have multiple NACE codes)
  + Matching NACE codes should be ranked from the most relevant (most strongly correlated to the subject of the article) to the least relevant

**Method:**

What is most important is to test how quickly, accurately and cheaply we can use Elasticsearch to identify the industry sectors covered in the article by using a taxonomy.

NOTE: It is possible that the above analysis can be done using an open-source tool or service like OpenCalais by REUTERS ([demo here](http://www.opencalais.com/opencalais-demo/)). We may decide to test the accuracy of our model versus OpenCalais.

However, please focus on finding the best way to identify taxonomy terms and their relevance in a stream of articles and other unstructured data.

**Resources:**

[NewsAPI link](https://newsapi.org/docs/get-started)

NewsAPI allows free access to news articles published on a given day (i.e. it is free to take news articles published in the last 24 hours but it costs money to find articles in their archives)

Your API key is: 6f756ed31b104120b842b037a51e58c1

Getting [started guide](https://newsapi.org/docs/get-started) is here.

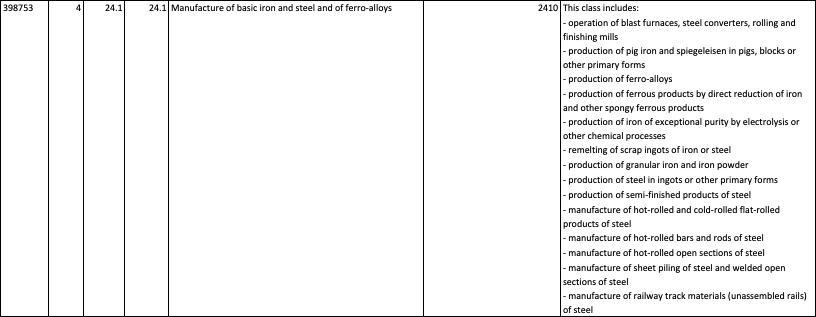
**Taxonomy:**

In the NACE taxonomy, there are 997 different codes. Each code has language which describes that industry sector.

In the example below, the first column is the Order followed by Level, Code, Parent and a description.

We need to create Elasticsearch queries to find all articles which contain the words in the “Description” and the column “This Item Includes”.





To find articles relating to this industry sector, our Boolean query would search for:

[“manufacture” AND “iron” OR “steel” OR “ferro-alloy” ]

AND

[“blast furnace” OR “Steel Converter” OR “rolling mill” OR “finishing mill” OR “pig iron” OR “spiegeleisen” OR “ferrous product” OR “iron ingots” OR “steel ingots” OR “semi-finished steel” OR “hot-rolled” OR “flat-rolled” OR “cold-rolled” OR “sheet piling” OR “steel section” OR “sheet piling” OR “railway track”]

For any article returned by this query, we need to:

* Assign the code 2410
* Tag all the keywords caught in the article from the query
* Save meta data on the article (source, date and time, etc.)
* Save a link to the article
* Store the code, keywords, meta data and link in the PostgreSQL database
* Rank the relevance of each NACE code so we end up with the most relevant NACE code down to the least relevant NACE code.

**GitHub:**

[GitHub link](https://github.com/Risclarion/ElasticsearchPilot)

ElasticsearchPilot is the project name in Github