PRESS RELEASE PROJECT

There is 3TB of data. There are multiple documents which contain around 50-60 xml files on average. The first xml file of each document seems to be a directory for all the other xml files present in the document.

You can track which document each XML file belongs to using the ‘Content-ID’ field which matches for all the xml files with the directory file.

The Identifier tag under “dc:metadata” gives the unique Id of the press release.

The ID identifies each press release file and will help map it back to the document it belongs to.

**PRESS RELEASE DOCUMENT SCHEMA**

Each XML file has a schema which consists of several XML Tags. There are three main nodes in each XML File: ArticleDocHead, nitf:body and Metadata.

To get a detailed view of the full schema of each XML file, refer the document Press Data Schema.docx

Notes:

* We are concerned with tags under “ClassificationGroup” such as Subject, Company, Industry.
* **Nitf:hedline** : This is the Title of the article.

**Key Node/Tag Definitions:**

Metadata: It is one of the nodes of the article and has several sub-elements that contain all the tags that define the content of the article. It is relevant across all the articles.

Classification: To define the Publication type (Ex: NewsWire and Company Filings) and language of the press release.

Classification Group: Used to group together multiple classifications of different schemes under a single (more general) classification scheme. For example, separate classifications using schemes for companies, subjects, people, and law firms can be grouped together under a single scheme.

* Score:  Optional attribute used to identify the frequency or applicability of a given classification within a document. Subject, Industry, Company, Country, etc have scores.
* Company: Name of the company being talked about in the article (usually the title will also mention the company name).
* Legal: The tags under this classification scheme states any sort of legal information about the press release. Ex: Copyright Law, Governments, Computer & Internet Law.
* Subject: Tags which give context on what the press release content is about.
* Industry: The industry to which the news/company is being talked of in the article.

**Analyzing the XML Files:**

The sample size considered here is 10038.

First each document containing multiple XML files was split into individual XML files. The code for this is in the file ReadingFiles.txt

Then we read the files individually by referencing the path of the files in the local computer. You can change the path to the path of where the files are stored in your local drive.

To understand what tags could be of relevance, it is necessary to be able to extract the different elements of the XML file. The code for the same is in the file All\_Elements.txt.

Based on the elements from the code and manual analysis of the XML files, we concluded that the tags Subject, Company, Industry, Title would be of most relevance.

The code to extract Subject tags is in the file Subject\_Code.txt. Similarly, you can get the Company tag values by replacing <element1.attrib["classificationScheme"] == "subject"> to “company” in the code.

[Note: If you want to check for a specific subject in the code then you can add another if statement in the for loop. For example,

for element2 in element1.findall("classificationItem"):

if(element2[1].text == 'ACQUISITIONS')]

To get the Industry Tags from the sample, you can refer to the code in the file Industry.txt

ACQUISITIONS:

Based on the analysis of several tags, the focus is on understanding the press release files related to acquisitions. There are two things we analyzed in the files.

One is to see which files have the Subject “ACQUISITIONS”. You can use the code in file Subject\_Code.txt (refer note) to get this data.

Secondly, we wanted to capture the files which have any acquisition related word in the ‘Title’ and analyze the Companies that are tagged in those files as well as what kind of acquisitions are mentioned in the press releases. The code for this is in the file Acquisitions\_Code.txt

ACQUIRER AND ACQUIREE:

To further understand if we can obtain both the Acquirer and Acquiring company from the Title, we have formulated some NLP rules for Title. We will apply these rules to all the files which have any pattern of the word “Acquisition” in the Title.

TECHNOLOGY INDUSTRY:

To study specifically about the Tech Industry, we have created a sample data specific to Subject Tags that are related to the Technology Industry. The code for the same is in the file Tech\_Code.txt

Note: You can add code to this if you want other tags such as Industry. You can refer to the Pharma\_Code.txt file for reference.

PHARMACEUTICALS INDUSTRY

To study specifically about the Pharma Industry, we have created a sample data specific to Subject Tags that are related to the Pharma Industry. The code for the same is in the file Pharma\_Code.txt.

**Output files to refer are listed below:**

SubjectData(2).xlsx – This has all Subjects present in the sample of 10038 files. The Subjects marked in Red are Tech related subjects and the Subjects marked in Yellow are Pharma related subjects.

CompanyData(2).xlsx - This has all Companies present in the sample of 10038 files.

IndustryData.xlsx – This contains all the Industry names and Industry codes in the sample.

Phama\_Subject.xlsx – This contains all the file details of files with Subject related to Pharmaceuticals industry. Comments around Analysis is mentioned in the file.

Tech\_Subject.xlsx - This contains all the file details of files with Subject related to Pharmaceuticals industry. Comments around Analysis is mentioned in the file.

All\_Acquisitions\_Title.xlsx – It consists of all the files where Title contains the word “acquires”, “acquisition” and “acquire”.

Acquisitions\_Title.xlsx – This file contains the details such as Subject, Company, etc for all the files that have Subject “Acquisitions” and have acquisitions word in Title of the file.