Dealroom Assignement

Zineb Lahrichi

June 2, 2020

1 Task 1

1.1 Explained approach

This first task is an unsupervised clustering task with a given number of labels. There are five different groups with known characteristics given in the excel file and online documentation. The first thing I did was going through all the features to determine what meaningful information can be extracted and understandable for the model to make a good prediction. I first made simple assumptions to start with a simple model. Some of the rules showed in Table 1. might not be general but make enough sens to be considered and can be refined later depending on the results.

Then, I realized that the attributes are all categorical and the information extracted define Boolean and symbolic rules. Therefore, a simple model suiting this type of features and known to perform well would be a decision tree.

I decided to build my own decision tree, based on the splitting attributes defined in Table 1. One of the main advantage of this method is that it is explainable and interpretable and we can easily choose to make it either simple or very complex by adding simple rules.

You will find a Figure of the final decision tree in the last page.

The idea was to build the tree starting with the most obvious rules and splitting the dataset into pure data subsets. For example, schools are entirely determined by their name and organizations are entirely determined by the domain extension of their website URL. Thus we can start by eliminating these two groups for the rest of the filtering process. I also made the assumption that a Tech company is either a startup or a mature company for one of the final nodes.

More details about the data pre-processing steps are explained in the code

1.2 Possible improvements

Such model can surely be improved by refining the rules and adding more nodes to the tree. To do so, we have to base our model on more precise definitions of

Features	meaningful information
LAUNCH DATE	Firms created before 1990 are not startups
WEBSITE	Domain extensions denote the type of the website and the firm in certain cases : ".org", ".gov"
GROWTH STATE	Firms with a late growth stage generally have a strong market presence already. (Mature Companies) Early and seed stage usually means the company has just launched and is still working on its proof of concept (Startup)
TAGLINE/TAGS/NAME	We can define a list of meaningful vocabularies characterizing the categories. Each vocabulary containing a bag of keywords. TECH = ['tech', 'digital','IT'] MATURE = ['world', 'international', 'leader'] SCHOOL = ['school', 'university']

Table 1: Information to be extracted from the features

the labels and examine more specific cases.

2 Task 2

2.1 Explained approach

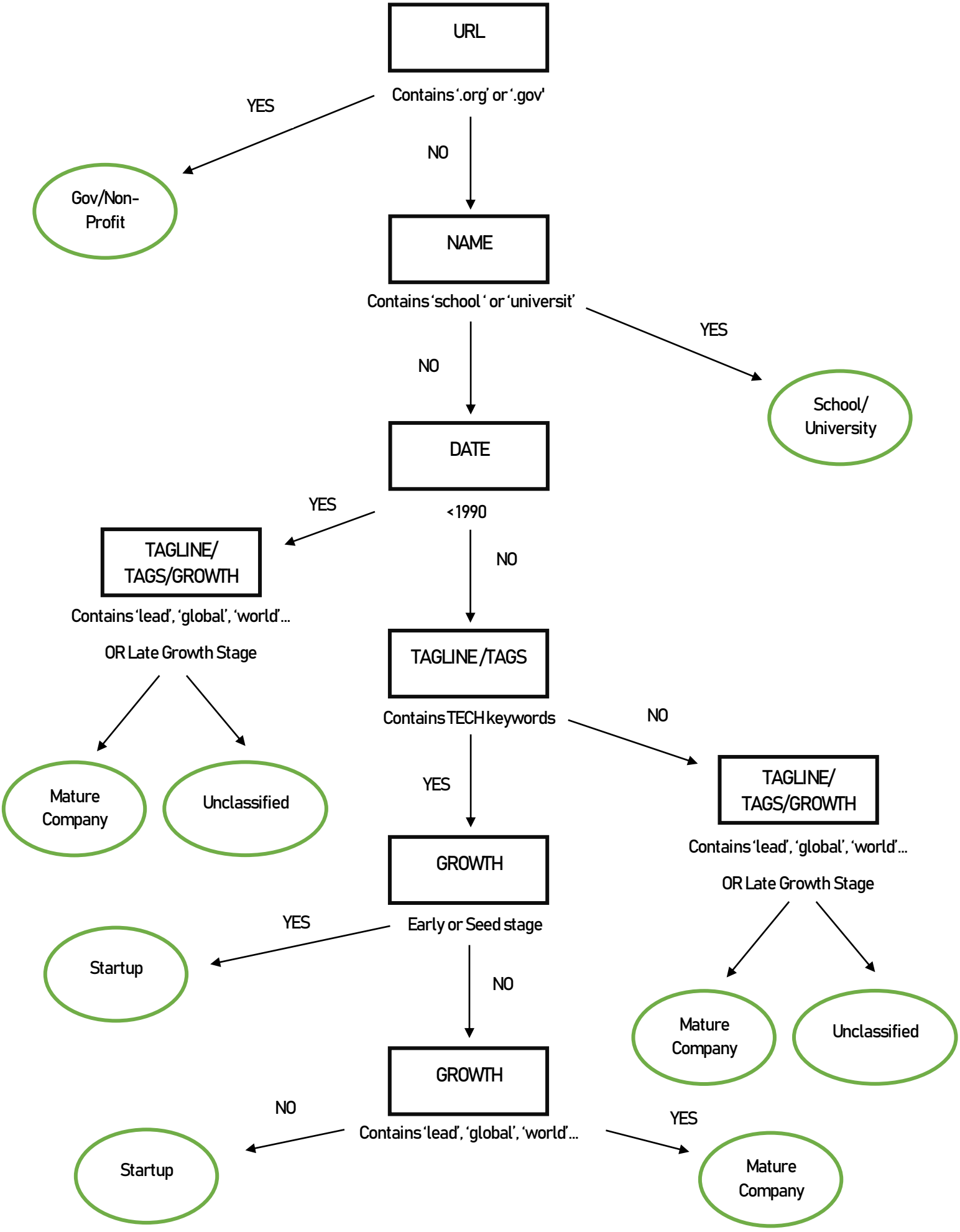
This kind of task was very new for me. I started working with Scrapy but most of my request would return empty outputs.

I searched for more documentation on this topic and found that I needed to find another library able to execute Javascript code.

Using Selenium I managed to extract the list of the companies from the webpage.

2.2 Possible improvements

In the website they classify the companies by field. It could be helpful to extract this information too.


URL
Contains '.org' or '.gov'
YES
Gov/Non-Profit
NO
NAME
Contains 'school' or 'universit'
YES
School/ University
NO
DATE
<1990
YES
TAGLINE/ TAGS/GROWTH
Contains 'lead', 'global', 'world'...
OR Late Growth Stage
Mature Company
Unclassified
NO
TAGLINE /TAGS
Contains TECH keywords
YES
GROWTH
Early or Seed stage
YES
Startup
NO
GROWTH
Contains 'lead', 'global', 'world'...
NO
Startup
YES
Mature Company
NO
TAGLINE/ TAGS/GROWTH
Contains 'lead', 'global', 'world'...
OR Late Growth Stage
Mature Company
Unclassified