Natural Language Processing for Company Business Descriptions: Preprocessing and Data Cleaning Step

Wei Ding, Wusheng Liu, Kwok Ping Ng
Technical University Munich
wei.ding@tum.de, kwokping.ng@tum.de, wusheng.liu@tum.de

1 Files Cleaning

1.1 Json Files

Each website is stored in a json file. Each page/tab under the website with its web content is stored as a key-value entry in this json file. Page/tab names are keys and texts from each tab are contents. Through a brief view of part of the data, we find that some json files are empty, some scrawled tabs and contents are invalid. So it is necessary to do the files cleaning.

1.2 Workflow

The workflow of file cleaning is shown in Figure 1. Firstly we removed the empty json files and tabs with empty content. We found that recorded tab keys have the pattern: Date_Different level of tabs_Webpage filename extension (for example *20161203_ipsc_about-us_executive-leadership_.html* as in Figure 2). So we used the regular expression to extract tabs which do not match this pattern.
Then we checked the content of these unmatched tabs. A lot of them has this error message as content. (*Error Messages: "The Wayback Machine has not archived that URL.\nHelp make the Wayback Machine more complete!\Save this url in the Wayback Machine..."*) So we only remove these kind of unmatched tabs. We also check whether the matched tabs also contain this error message and removed them. We also removed the empty files after cleaning these tabs.

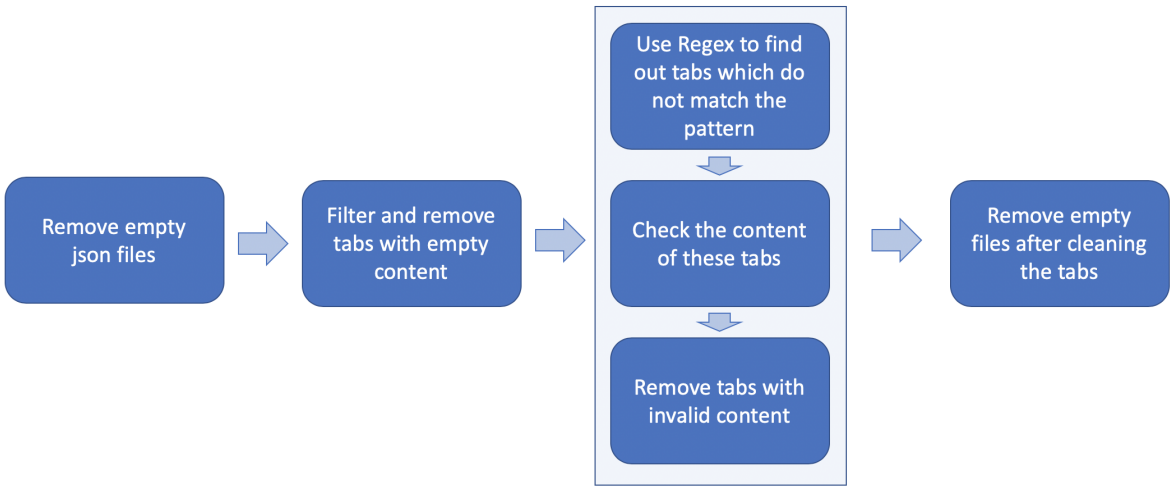

Figure 1: The workflow of file cleaning

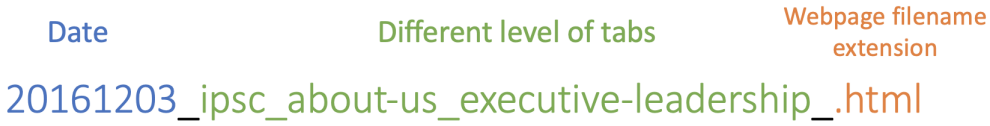

Figure 2: Name structure of each tabs

1.3 Files Processing Results

Table 1 shows the number of valid files after each processing steps. We found that the number of final valid files from the *Sum* folder is less than half of the total files in that folder, which means that we need to focus on the files from *Sum_all* for the further classification.
Table 2 is the statistics of number of tabs (for example *20161203_ipsc_about-us_executive-leadership_.html* as a tab) in the whole json files. These tabs may indicate the content of the webpages. From the numbers, we see each valid json file in *Sum* folder has average **3.64** valid tabs, while each valid file in *Sum_all* folder has around **27** valid tabs. Valid tabs in *Sum* folder has higher rate than that in *Sum_all* folder because files in *Sum* have been filtered by some business logic. Similarly, valid tabs per file in Sum are fewer than in *Sum_all*. We think that for classification purpose, we need to investigate the business logic to filter the tabs in *Sum_all* folder to achieve better classification results.

No.	File Type	No. in "Sum" folder	No. in "Sum_all" folder
a	Total Files	9,910	9,910
b	Empty Files	4,862	29
c	Empty Files after removing empty tabs and removing "The Wayback Machine" tabs*	5,036	2,057
d	Valid Files(d = a-c)	4,874	7,853
e	Valid Files %(e = d/a)	49.2%	79.2%

Table 1: Files Results in Cleaning * The tab contains error message "The Wayback Machine" is considered as an irrelevant tab

No.	Tab Type	No. in "Sum" folder	No. in "Sum_all" folder
A	Total Tabs	18,528	235,713
B	Empty Tabs	753	21,913
C	Empty Files after removing empty tabs	5,031	1914
D	Tabs with name not start with date and not end with (html$\|\|htm\|\|php\|\|org\|\|asp\|\|aspx$) **	20	2,465
E	Tabs with content "The Wayback Machine"*	6	419
F	Valid Tabs(F = A - B -E)	17,769	213,381
G	Valid Tabs %(G = F / A)	95.9%	90.5%
H	No of Valid Tabs per File(H = F/d)	3.64	27.17

Table 2: Tabs Results in Cleaning * The tab contains error message "The Wayback Machine" is considered as an irrelevant tab ** The valiad tab always has format as stated above and other file type like htm, php, org, asp and aspx. For those tab name does not fulfill this format will be analyzed independently

2 Tabs Names Preprocessing

2.1 Working Flow

Each tab has different level of sub-tabs. For example as in Figure 2, *executive-leadership* is the sub-tab of *about-us*, which is the sub-tab of company *ipsc*. So we split the tabs to get single sub-tabs according to "_". We then counted the occurrence of each sub-tabs. If a sub-tab appears several times in a website, it is still considered as once. After that we analyzed the occurrences and highlighted some important sub-tabs manually. Figure 3 illustrates the tab processing in this step.

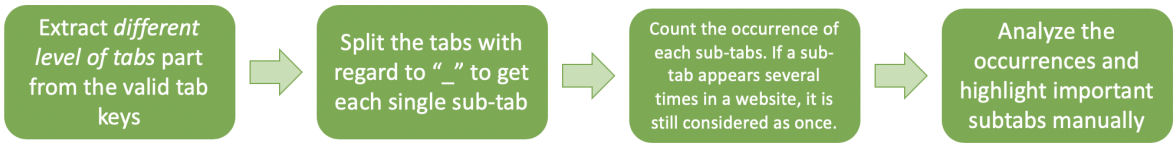

Figure 3: The workflow of tab processing

2.2 Tab Analysing

In this preprocessing step, we are trying to analyze the name of the tabs of all the company website. There are two main purposes for this step. The first goal is to identify the tabs with high occurrence rate across the company websites. Meanwhile, the second goal is to identify the tabs with high correlation with the tasks we are going to perform at later steps, such as classifying company categories, extracting founders information, etc.

For the first goal, Figure 4 is the distribution of the company tabs with occurrence rate of over 200 for the data in the *Sum_all* folder and over 100 for the data in the *Sum* folder.

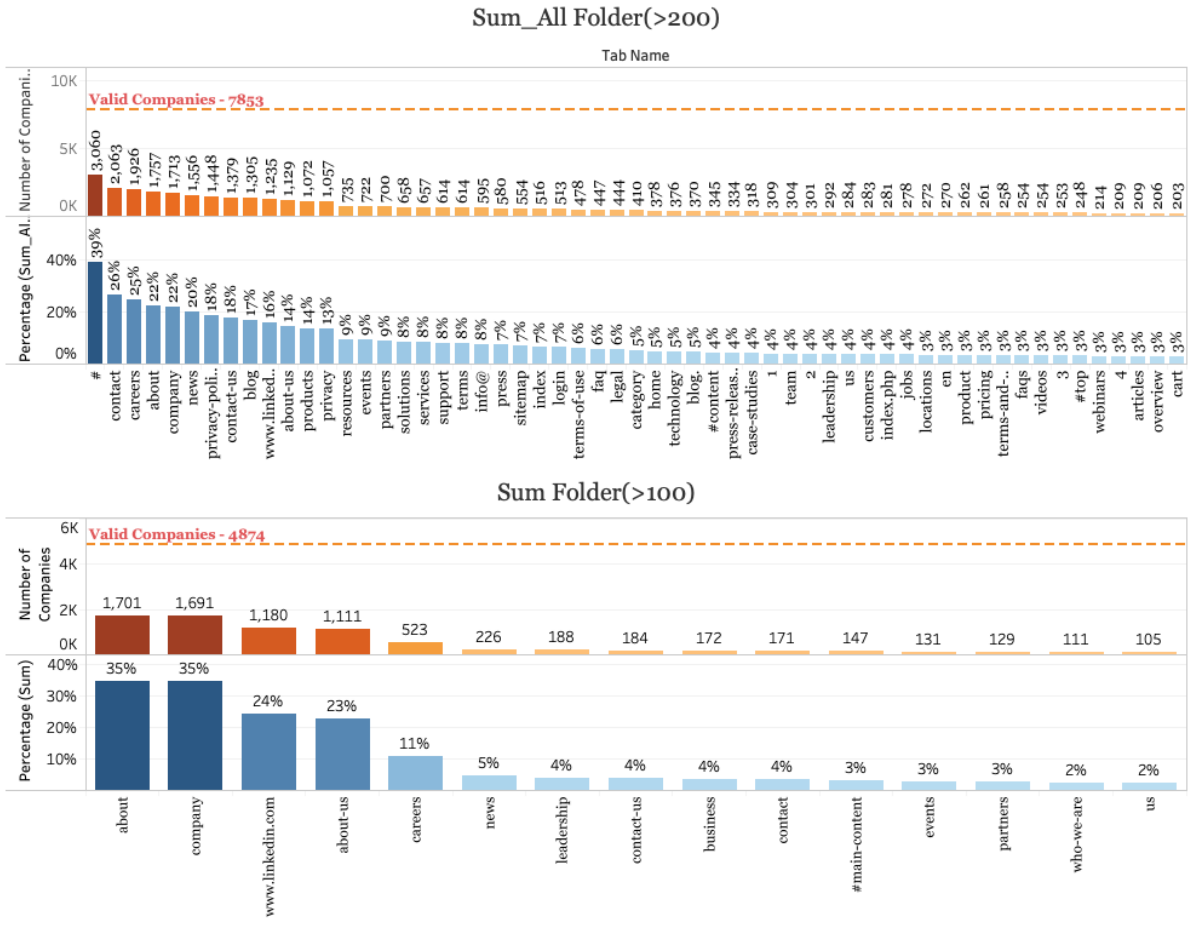

Figure 4: Tabs which have high occurrences. The upper figure shows these high frequent figures in *Sum_all* and the lower figure is for *Sum* folder. Orange color shows the occurrence number and blue color shows the percentage of the valid files. # stands for the main page of some companies.

The sub-tab names in *Sum* folder are more concentrated in fewer keywords than that in *Sum_all* folder files. There are only 5 sub-tabs which appear in over 10% percentage of the company in *Sum* folder.

But there are 13 sub-tabs in *Sum_all* folder which appear in over 10% percentage of the company. This indicates we can focus less in *Sum* folder to conduct analysis.

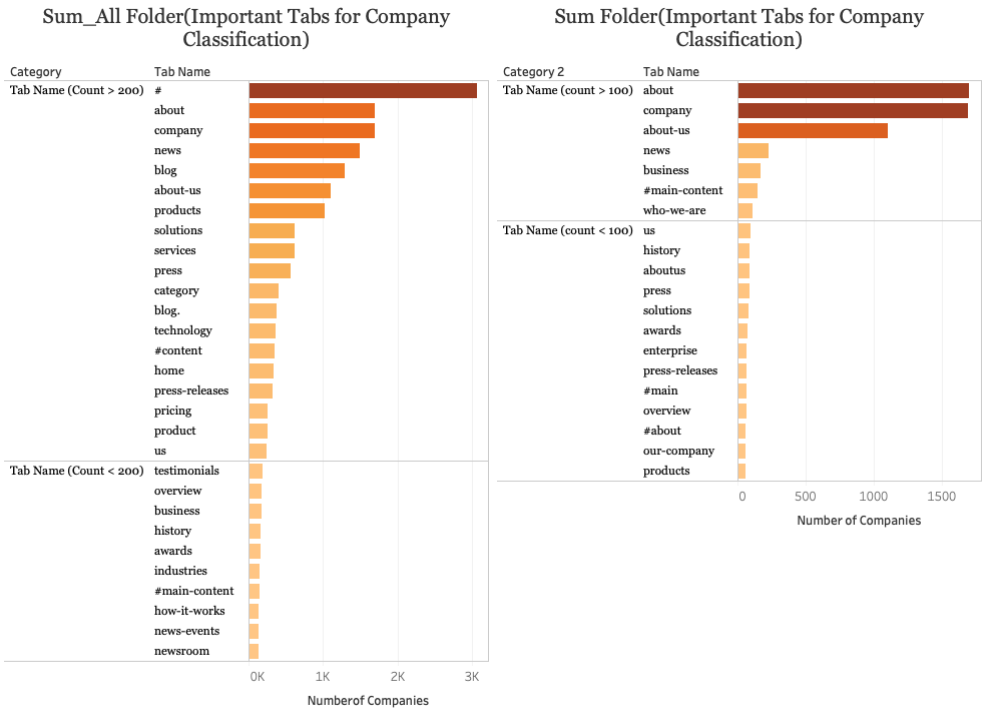

Figure 5: Important sub-tabs we have selected for the company classification. Left column shows the sub-tabs appeared in *Sum_all* folder while right column shows those appeared in *Sum* folder.

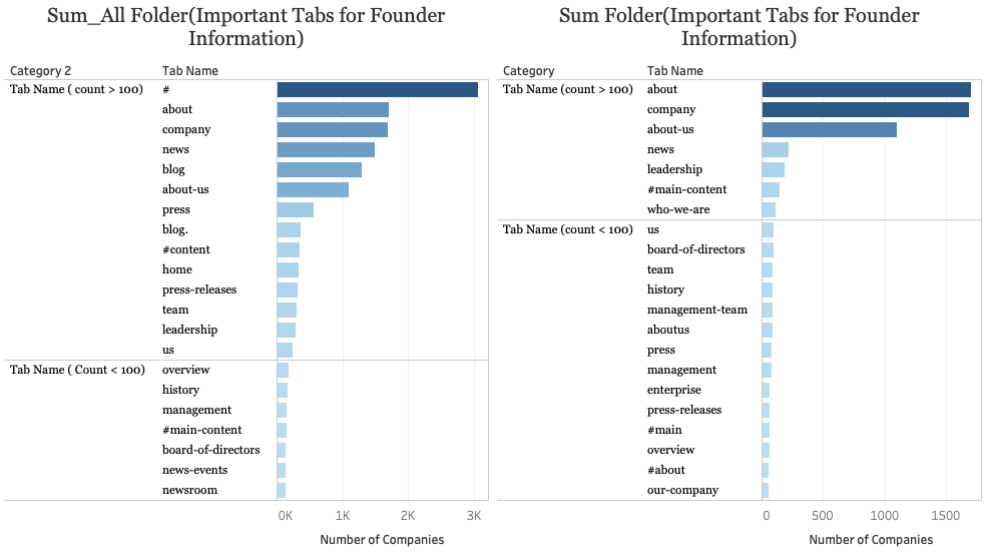

Figure 6: Important sub-tabs we have selected for selecting founder information. Left column shows the sub-tabs appeared in *Sum_all* folder while right column shows those appeared in *Sum* folder.

To fulfill the second goal, the tabs names in Figure 5 and in Figure 6 are are identified manually as having high correlation with the tasks we are going to perform in the future. We consider each tab name

with occurrence over 50 for *Sum* folder and over 100 for *Sum_all* folder to decide whether it contains key information for **company classification** or **founder information**. Note that some tab names may contain both.
We agree that some important tabs names that obviously contains company classification such "about-us", "company" or "news". For the important sub-tabs from *Sum_all* folder, most of them appear in more than 200 web pages, while most of the helpful sub-tabs in *Sum* folder appear less than 100 times.
For the founder information, tab names like "leadership", "about-us" are more likely to have such information.

3 Next Steps

For the next step, we will firstly focus on giving the labels for each tab. Since 16% of the companies in folder *Sum_all* have a webpage for LinkedIn. We assume we can get the categories of the company from LinkedIn. So we will get the labels for each companies mainly based on scrawling the LinkedIn. Besides, we also need to remove the duplicated sentences in a same json file but from different tabs.