In the first stage of linking entities, we get the raw records from the WARC file. The aim is that extract a row text of every valid record which is consisted of an HTML file. At the very beginning, the WARC file is read in binary and analyzed by ArchiveIterator from the Warcio library. Then, we select records whose rec\_type shows the response in all records and get their record\_id as key from the header. Next, we can obtain the whole HTML file in stream as the content. By now, we have the record id and its corresponding content in HTML. To gather the text of a web, we import Beautiful Soup library to parse the HTML file by LXML parsing analyzer and access the text from it. After extracting the raw text, for better efficiency and performance of entities detection, we try to clean the text using the regular expression. The useless emojis and stickers will disturb the following work. Therefore, we concentrate more on characters and basic punctuations from the main language like English, Chinese, Japanese, and so on. Plus, to save memory and time of processing, the cleaned text is assembled to prevent repeating and saved as a set. In the last of this stage, we store the key of record id and the value of its text into one dictionary dumped in a JSON file for later stages.