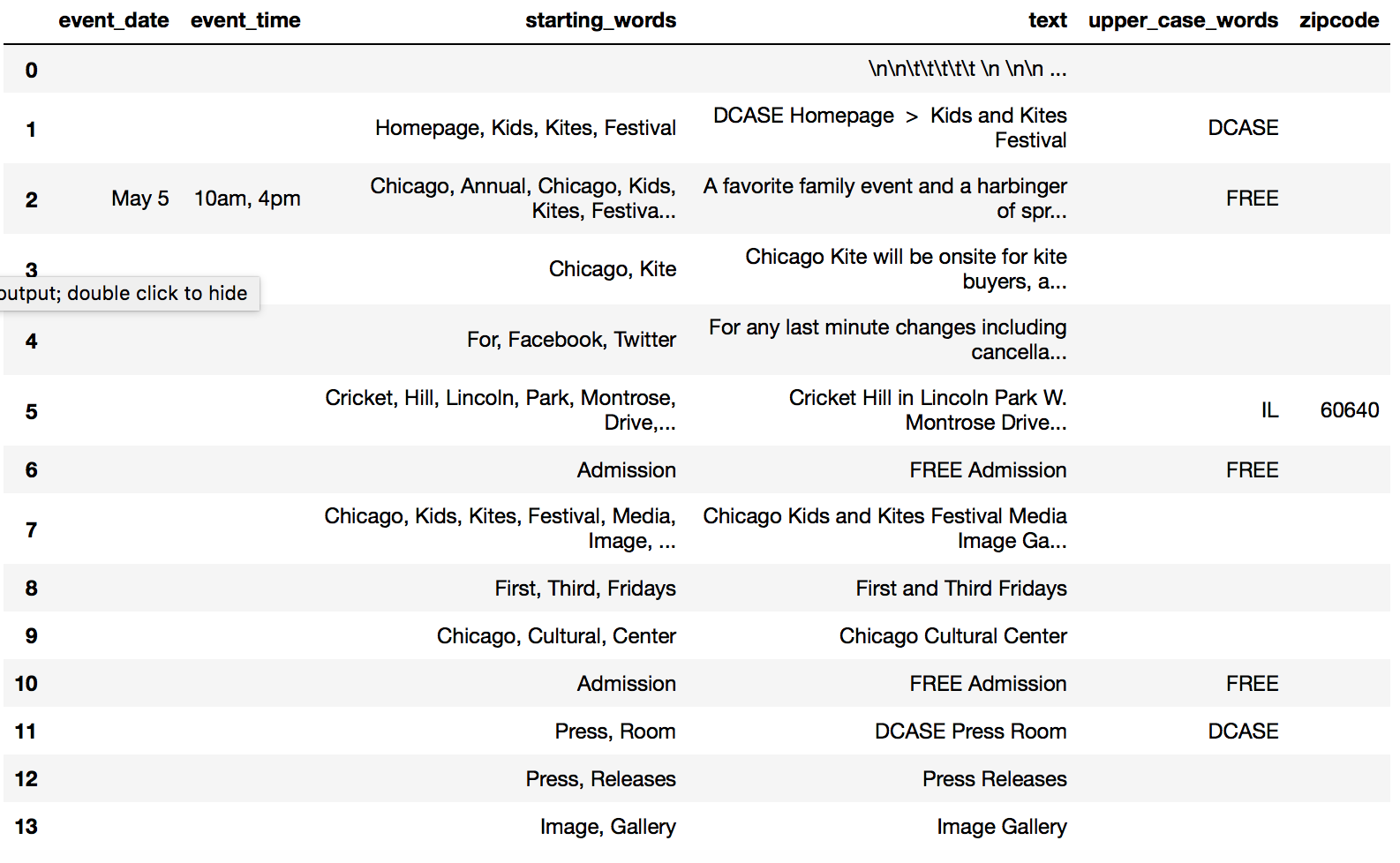
Computational Content Analysis: Memo 1

Xingyun Wu

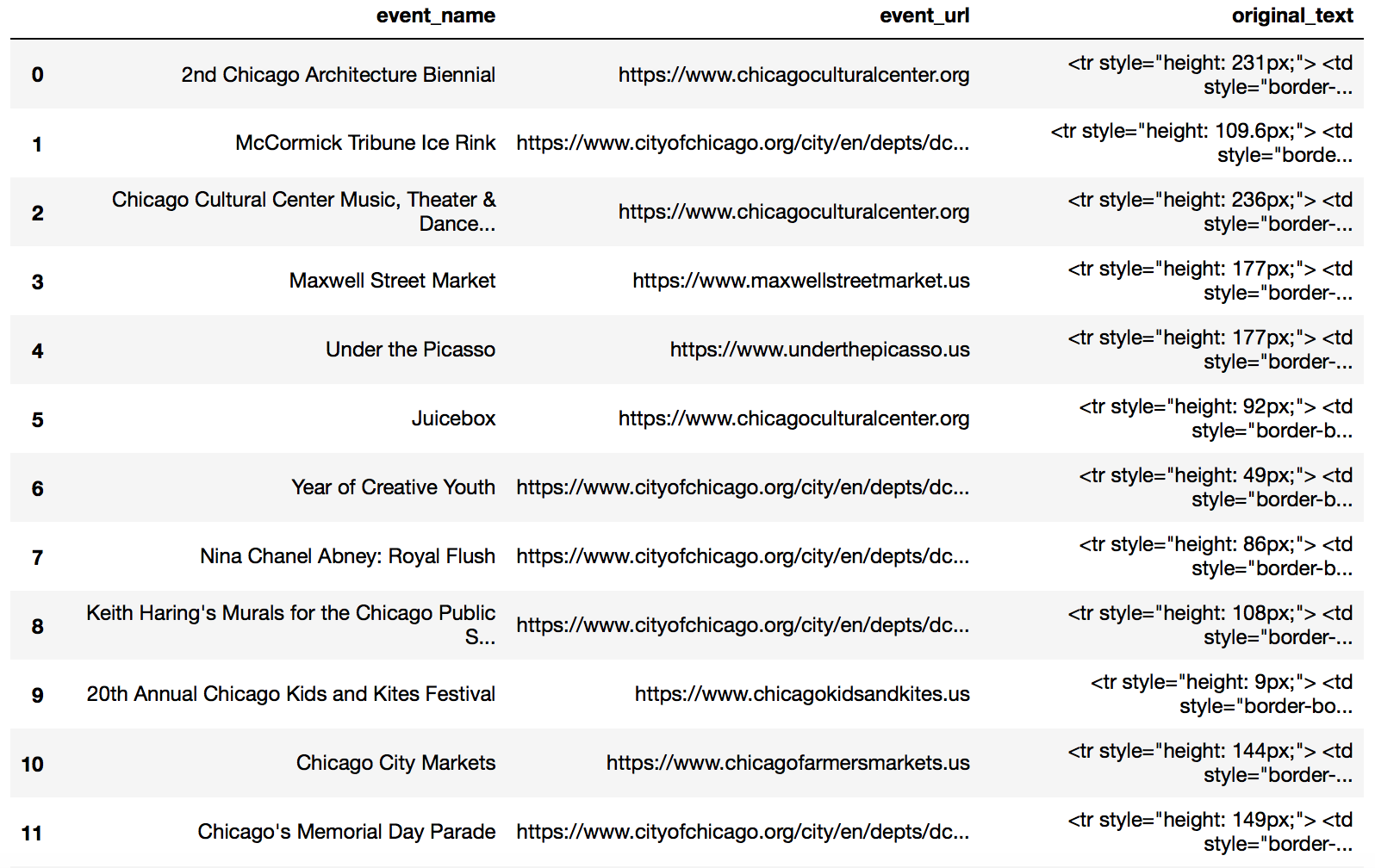
1/10/2018

1. Summarize results from preliminary analysis

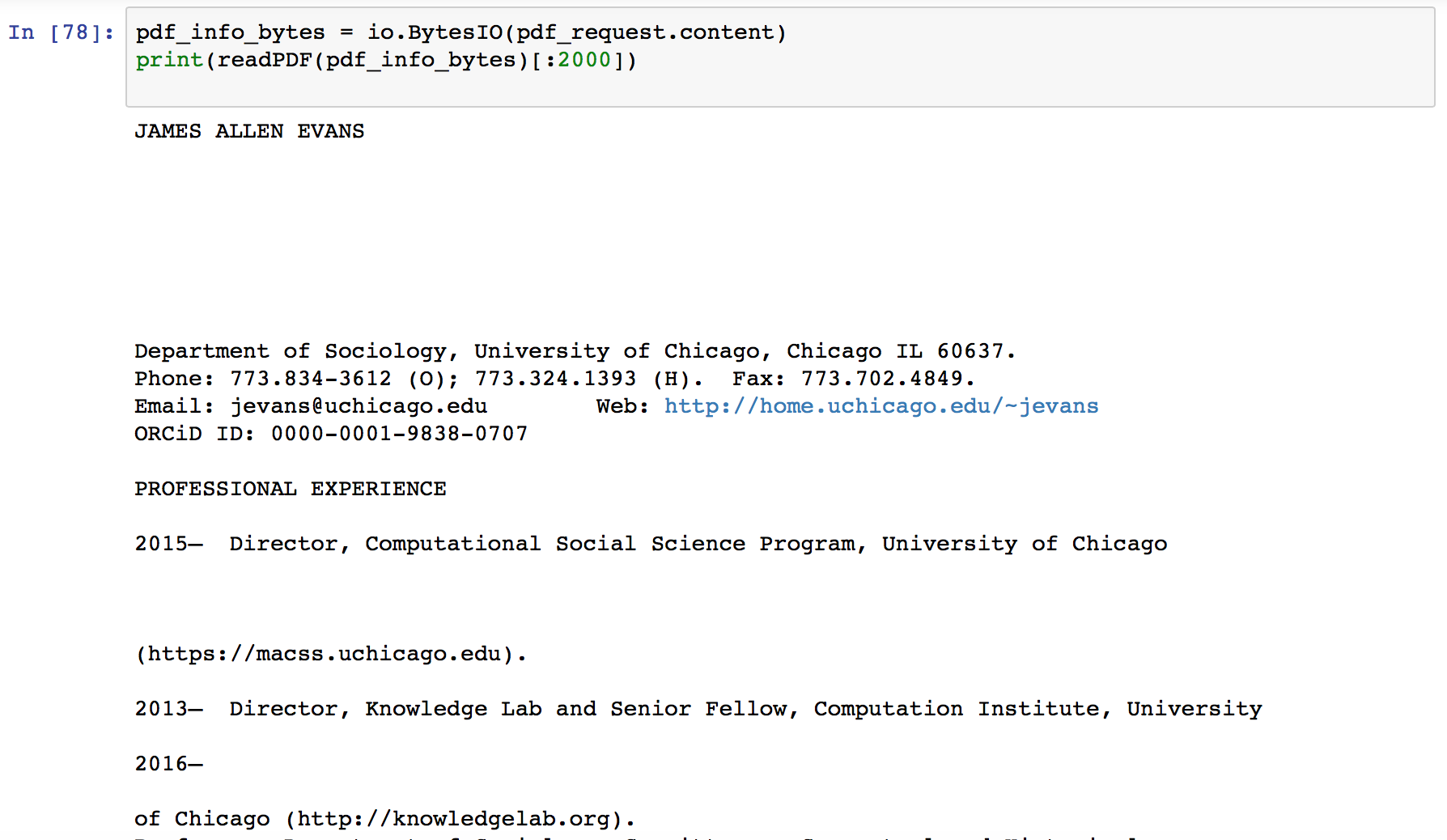
In this homework, I try to scrape the webpage of the 20th Chicago Kids and Kites Festival, from the City of Chicago’s official website. I first extract all the texts from that webpage using BeautifulSoup with a Regular Expression. Then I use other 5 Regular Expressions to extract from each paragraphs for detailed information of: starting words, time, date, zip code, and upper case words.

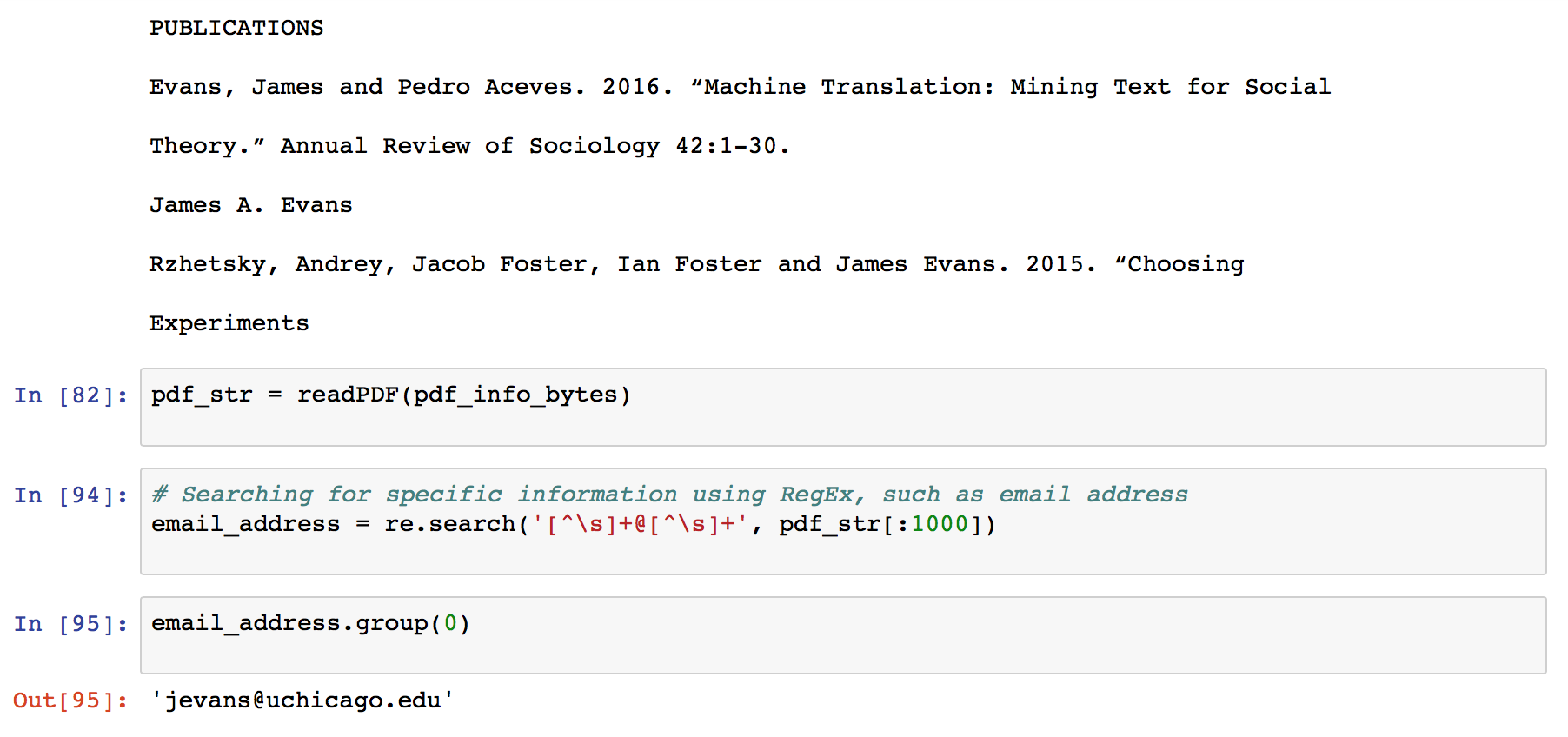


Then in task 2, I use the spidering technique to get urls of all the events listed on the official website of the City of Chicago’s official website.



In task 3, I first download the pdf file of the CV of Prof. James Evans, from UChicago’s website, directly into memory. Then I extract its content as strings. And I tried to use RegEx to extract his email address, to check whether I could get detailed information.





1. Identifies and interprets textual examples that facilitate qualitative validation of the patterns summarized

This homework does not require us to look for patterns. Generally speaking, the results look good.

1. Critically evaluates the method’s drawbacks and scope conditions for its beneficial development

The methods introduced in this Jupyter notebook could only deal with resources that have clean format. The instructions use Wikipedia websites, which is very clean. However, when I tried to scrape the website of the lists of cultural events, I met many problems. It is very hard to construct a recursive function to go deep into lower levels of webpages, since the webpages for cultural events just do not have lower level pages.

Another problem is that there are many websites that does not welcome people to scrape. Some just abandon scrapers, which needs to be fixed to use “sleep”. Others just occasionally change the structure of their source code, which is rare but still happened. So the built scrapers could not always work.