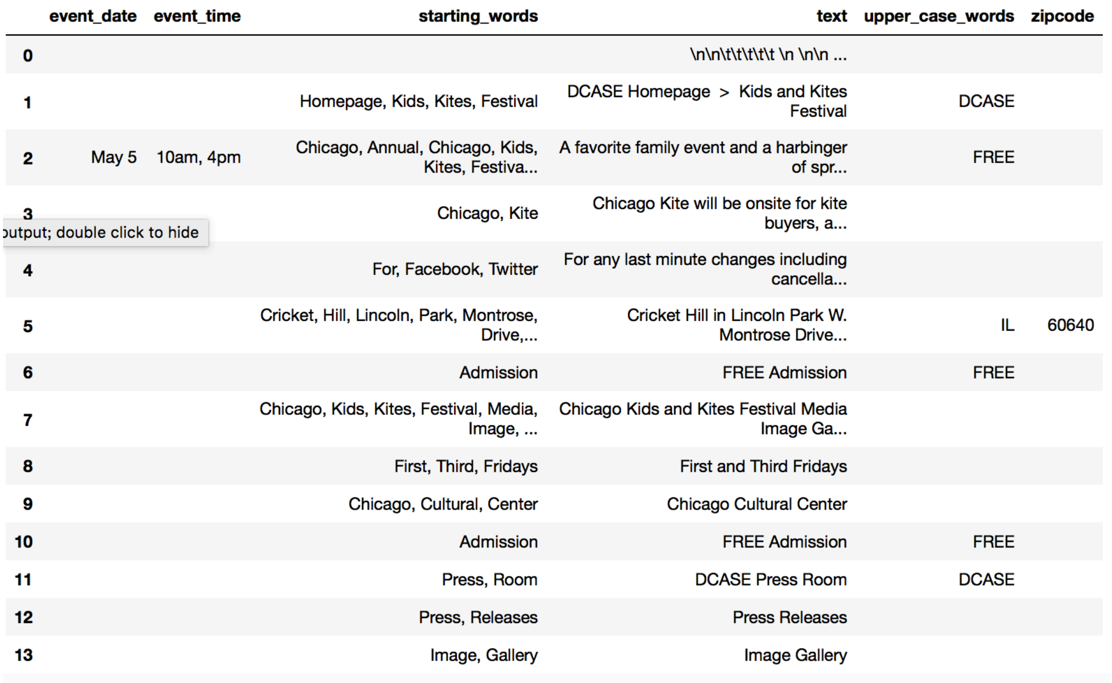
Computational Content Analysis: Memo 1

Xingyun Wu

1/10/2018

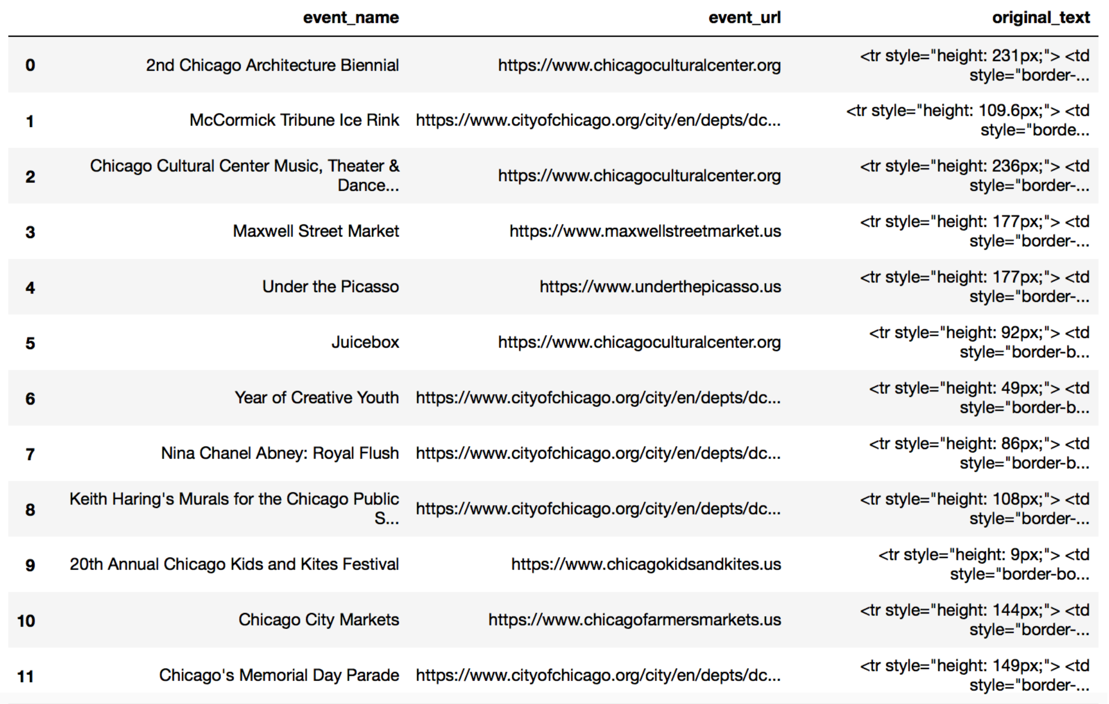
1. Summarize results from preliminary analysis

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.



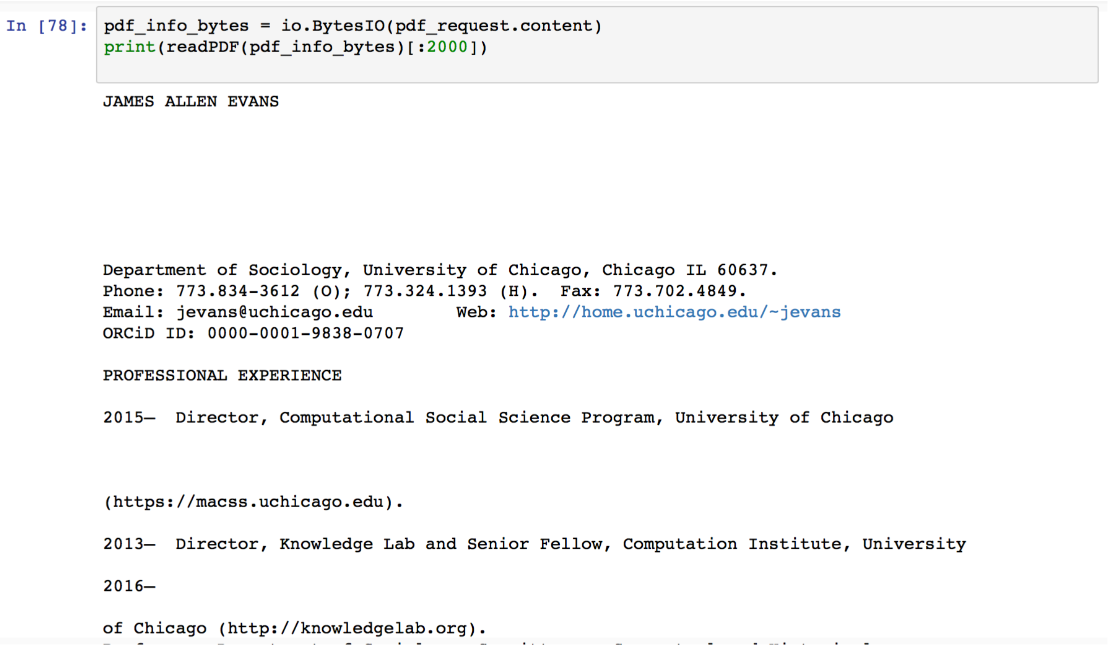
	event_date	event_time	starting_words	text	upper_case_words	zipcode
0				\n\n\t\t\t\t\t \n \n\n ...		
1			Homepage, Kids, Kites, Festival	DCASE Homepage > Kids and Kites Festival	DCASE	
2	May 5	10am, 4pm	Chicago, Annual, Chicago, Kids, Kites, Festiva...	A favorite family event and a harbinger of spr...	FREE	
3			Chicago, Kite	Chicago Kite will be onsite for kite buyers, a...		
4			For, Facebook, Twitter	For any last minute changes including cancella...		
5			Cricket, Hill, Lincoln, Park, Montrose, Drive,...	Cricket Hill in Lincoln Park W. Montrose Drive...	IL	60640
6			Admission	FREE Admission	FREE	
7			Chicago, Kids, Kites, Festival, Media, Image, ...	Chicago Kids and Kites Festival Media Image Ga...		
8			First, Third, Fridays	First and Third Fridays		
9			Chicago, Cultural, Center	Chicago Cultural Center		
10			Admission	FREE Admission	FREE	
11			Press, Room	DCASE Press Room	DCASE	
12			Press, Releases	Press Releases		
13			Image, Gallery	Image Gallery		

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.



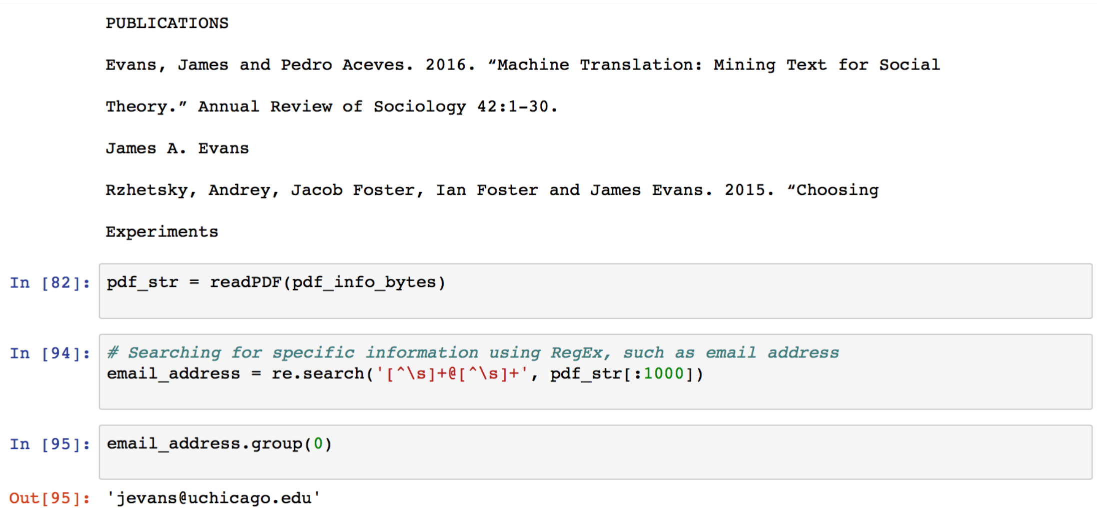
	event_name	event_url	original_text
0	2nd Chicago Architecture Biennial	https://www.chicagoculturalcenter.org	<tr style="height: 231px;"> <td style="border-...
1	McCormick Tribune Ice Rink	https://www.cityofchicago.org/city/en/depts/dc...	<tr style="height: 109.6px;"> <td style="borde...
2	Chicago Cultural Center Music, Theater & Dance...	https://www.chicagoculturalcenter.org	<tr style="height: 236px;"> <td style="border-...
3	Maxwell Street Market	https://www.maxwellstreetmarket.us	<tr style="height: 177px;"> <td style="border-...
4	Under the Picasso	https://www.underthepicasso.us	<tr style="height: 177px;"> <td style="border-...
5	Juicebox	https://www.chicagoculturalcenter.org	<tr style="height: 92px;"> <td style="border-b...
6	Year of Creative Youth	https://www.cityofchicago.org/city/en/depts/dc...	<tr style="height: 49px;"> <td style="border-b...
7	Nina Chanel Abney: Royal Flush	https://www.cityofchicago.org/city/en/depts/dc...	<tr style="height: 86px;"> <td style="border-b...
8	Keith Haring's Murals for the Chicago Public S...	https://www.cityofchicago.org/city/en/depts/dc...	<tr style="height: 108px;"> <td style="border-...
9	20th Annual Chicago Kids and Kites Festival	https://www.chicagokidsandkites.us	<tr style="height: 9px;"> <td style="border-bo...
10	Chicago City Markets	https://www.chicagofarmersmarkets.us	<tr style="height: 144px;"> <td style="border-...
11	Chicago's Memorial Day Parade	https://www.cityofchicago.org/city/en/depts/dc...	<tr style="height: 149px;"> <td style="border-...

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.



```
In [78]: pdf_info_bytes = io.BytesIO(pdf_request.content)
         print(readPDF(pdf_info_bytes)[:2000])
```

```
JAMES ALLEN EVANS







Department of Sociology, University of Chicago, Chicago IL 60637.
Phone: 773.834-3612 (O); 773.324.1393 (H).  Fax: 773.702.4849.
Email: jevans@uchicago.edu        Web: http://home.uchicago.edu/~jevans
ORCiD ID: 0000-0001-9838-0707

PROFESSIONAL EXPERIENCE

2015–  Director, Computational Social Science Program, University of Chicago



(https://macss.uchicago.edu).

2013–  Director, Knowledge Lab and Senior Fellow, Computation Institute, University

2016–

of Chicago (http://knowledgelab.org).
```



```
PUBLICATIONS

Evans, James and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social

Theory." Annual Review of Sociology 42:1-30.

James A. Evans

Rzhetsky, Andrey, Jacob Foster, Ian Foster and James Evans. 2015. "Choosing

Experiments
```

```
In [82]: pdf_str = readPDF(pdf_info_bytes)
```

```
In [94]: # Searching for specific information using RegEx, such as email address
         email_address = re.search('[^\s]+@[^\s]+', pdf_str[:1000])
```

```
In [95]: email_address.group(0)
```

```
Out[95]: 'jevans@uchicago.edu'
```

2. 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.

3. 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.