COM5507 Social Media Data Acquisition and Processing
Week 7. Mining the social web: Web data formats and web services

Lectured by: Dr. Xinzhi ZHANG
Research Assistant Professor, Department of Journalism
Hong Kong Baptist University
24 Oct 2018 @ CityU M4003

Agenda

- Diving into the social web
- Web data formats
  - JSON
  - XML
- Web services
  - API
    - An example of New York Times
  - Cloud Computing Services
    - An example of Amazon Web Services (AWS)

SNS (Social networking sites), social media

- boyd & Ellison (2007) listed three major functions that SNS provided for users:
  - establish a public or semi-public profile;
  - show a list of those who sharing a connection, and
  - view and traverse this list within this bounded system.
- *The history of social media in 90 seconds [Video]*

The social media's evolution timeline. URL: http://www.drapersonline.com/pictures/636xAny/4/3/7/1304437_social-media-timeline.jpg

Mining the social media data

- To acquire and process social media data, we need to understand more data formats and applications on the (social) web.

Data & communication on the web

- With the HTTP Request/Response well understood and well supported, there was a natural move toward exchanging data between *programs* using these protocols.
- The web has evolved from web 1.0 to web 2.0.
- There is an agreed way to *represent* data communicating between applications
- Two commonly used formats: XML and JSON

XML: eXtensible Markup Language

- Primary purpose is to help information systems share structured data
- It started as a simplified subset of the Standard Generalized Markup Language (SGML), and is designed to be relatively human-legible

http://en.wikipedia.org/wiki/XML

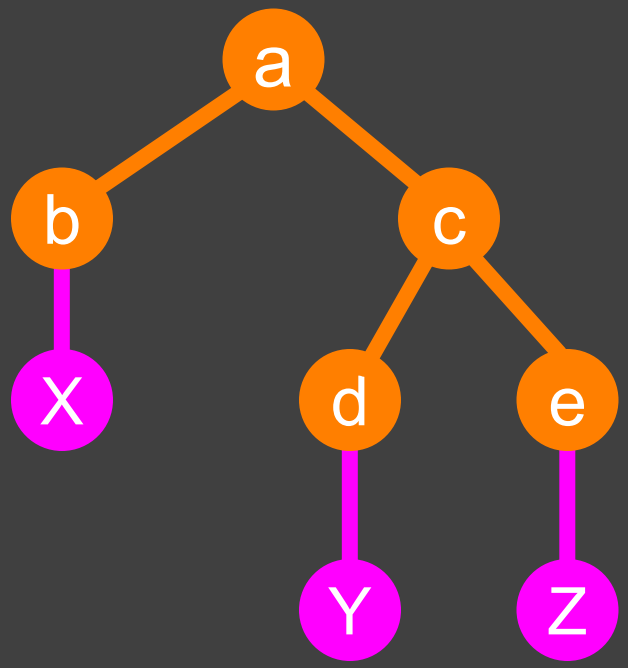
```xml
<?xml version="1.0"?>
<data>
    <country name="Liechtenstein">
        <rank>1</rank>
        <year>2008</year>
        <gdppc>141100</gdppc>
        <neighbor name="Austria" direction="E"/>
        <neighbor name="Switzerland" direction="W"/>
    </country>
    <country name="Singapore">
        <rank>4</rank>
        <year>2011</year>
        <gdppc>59900</gdppc>
        <neighbor name="Malaysia" direction="N"/>
    </country>
    <country name="Panama">
        <rank>68</rank>
        <year>2011</year>
        <gdppc>13600</gdppc>
        <neighbor name="Costa Rica" direction="W"/>
        <neighbor name="Colombia" direction="E"/>
    </country>
</data>
```

https://docs.python.org/3.7/library/xml.etree.elementtree.html

XML as a Tree

```
<a>
  <b>X</b>
  <c>
    <d>Y</d>
    <e>Z</e>
  </c>
</a>
```

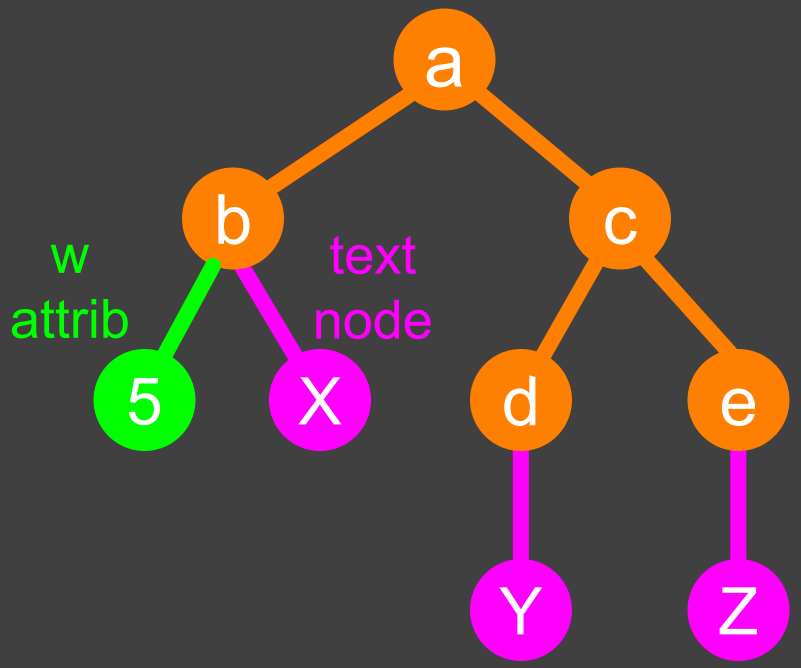
Elements Text



XML Text and Attributes

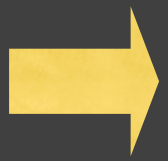
```
<a>
 <b w="5">X</b>
 <c>
  <d>Y</d>
  <e>Z</e>
 </c>
</a>
```

Elements Text

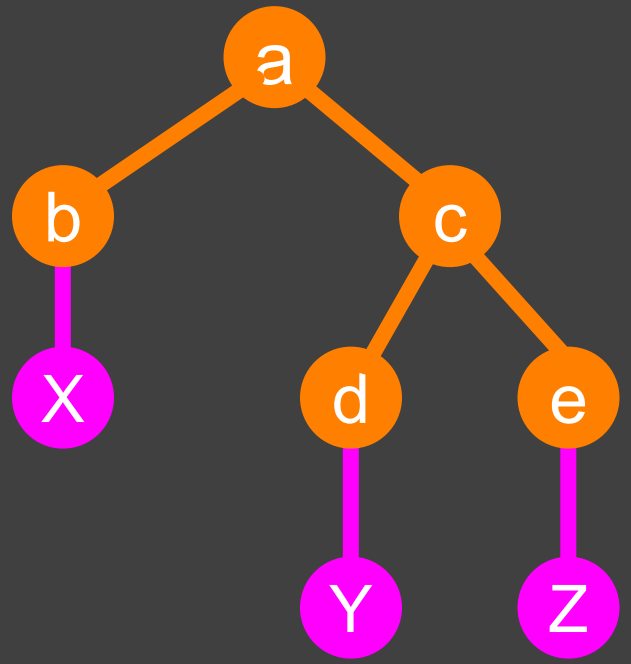


XML as Paths

```
<a>
  <b>X</b>
  <c>
    <d>Y</d>
    <e>Z</e>
  </c>
</a>
```



/a/b	X
/a/c/d	Y
/a/c/e	Z



Elements Text

XML Schema

- Describing a “contract” as to what is acceptable XML
- Description of the legal format of an XML document
- Expressed in terms of constraints on the structure and content of documents
- Often used to specify a “contract” between systems - “My system will only accept XML that conforms to this particular Schema.”
- If a particular piece of XML meets the specification of the Schema - it is said to “validate”

http://en.wikipedia.org/wiki/Xml_schema

http://en.wikibooks.org/wiki/XML_Schema

Many XML Schema Languages

- Document Type Definition (DTD)
  - - http://en.wikipedia.org/wiki/Document_Type_Definition
- Standard Generalized Markup Language (ISO 8879:1986 SGML)
  - - http://en.wikipedia.org/wiki/SGML
- XML Schema from W3C - (XSD)
  - - http://en.wikipedia.org/wiki/XML_Schema_(W3C)

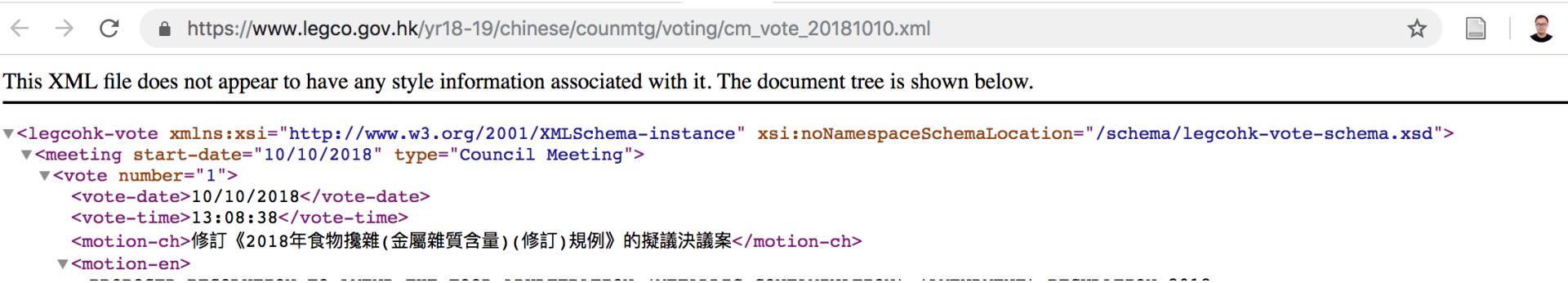
http://en.wikipedia.org/wiki/Xml_schema

XSD XML Schema (W3C spec)

- We will focus on the World Wide Web Consortium (W3C) version
- It is often called “W3C Schema” because “Schema” is considered generic
- More commonly it is called XSD because the file names end in .xsd

http://www.w3.org/XML/Schema

http://en.wikipedia.org/wiki/XML_Schema_(W3C)



https://www.legco.gov.hk/yr18-19/chinese/counmtg/voting/cm_vote_20181010.xml

Working with XML

- XML basic
- Hong Kong LegCo voting records
- In-class demonstration

- *socialweb_01_xml.ipynb*

JavaScript Object Notation

- JSON: JavaScript Object Notation
- Douglas Crockford - “Discovered” JSON
- Object literal notation in JavaScript

Working with JSON

- In-class demonstration
- Twitter data

“Half-baked” open datasets – in XML and JSON

- Hong Kong LegCo voting record (https://www.legco.gov.hk/general/english/counmtg/yr16-20/mtg_1718.htm#toptbl)
- Open government in mainland China
  - Shenzhen (http://opendata.sz.gov.cn/)
  - Shanghai (http://www.datashanghai.gov.cn/home!toHomePage.action)
  - Changsha (http://www.changsha.gov.cn/xxgk/szfxxgkml/)

Social media data collection

- Social media data can be acquired by scraping – if the data can be fetched via public approach.
- A more reliable but more restricted method to acquire social media data (especially within-boundary user data) is via the application programming interface (API).

Service Oriented Approach

- Services publish the “rules” applications must follow to make use of the service (API)

Application Program Interface

- An Application Program Interface (API) is a contract for interaction
- An API specifies an interface and controls the behavior of the objects specified in that interface. The software that provides the functionality described by an API is said to be an "implementation" of the API.
- An API is typically defined in terms of the programming language used to build an application.

http://en.wikipedia.org/wiki/API

Working with API

- In-class demonstration
- The New York Times API

More APIs

- Twitter API: https://developer.twitter.com/en/docs.html
- Guardian API: https://open-platform.theguardian.com/documentation/

Case 1: "#ddj" in the Twittersphere

- Data harvest: Application Programing Interface (API) + "*twitteR*" package in R.
- Search terms: Coddington (2015).
  - (a) data journalism (including "data-driven journalism"),
  - (b) computational journalism, and
  - (c) computer-assisted reporting.
- All tweets containing these keywords and hashtags were included (i.e., "#datajournalism" for the hash-tag search and "data journalism" for the term search).
- N = 6,951, from 25 Nov 2016 – 28 Dec 2016 (a four-week span).

Search more areas: investigative reporting and data visualization

- Application Programing Interface (API) + "*twitteR*" package in R.
- Search terms:
  - "investigative reporting", "investigative report", "investigative journalism", "investigative news"
  - "data visualization", "data visualisation", "dataviz"
- All tweets containing these keywords and hashtags were included
- N = 14,794 for investigative reporting
- N = 19,350 for data visualization
- All from 25 Nov 2016 – 28 Dec 2016 (the same four-week span as reported in study 1).

Table 1 *Background information of top posting users (screen names) on Twitter (updated till 8 May 2017)*

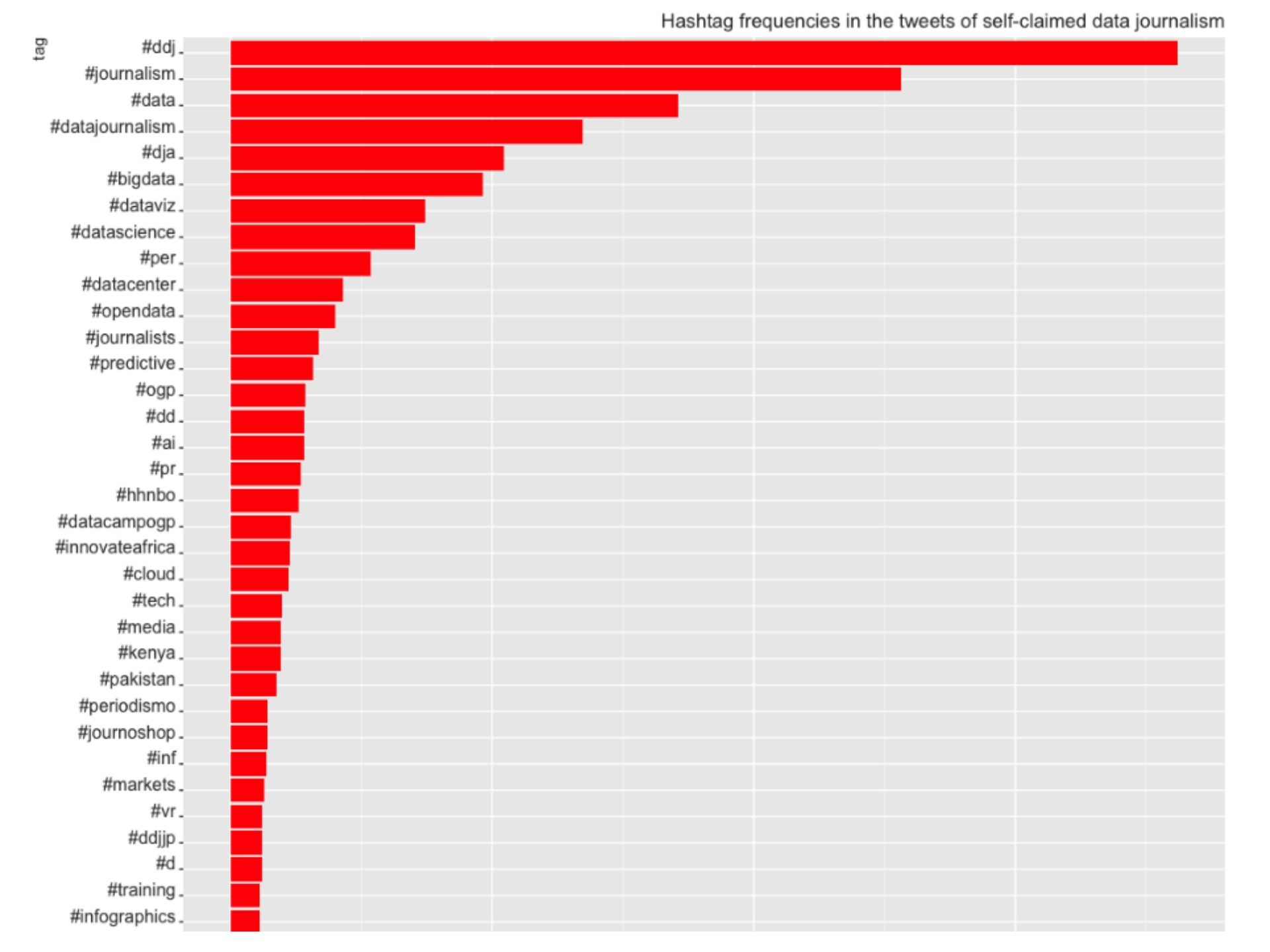
Screen name	Brief description	No. of posts in the sample	% within the sample	Total Tweets	Following	Followers
Tech_Journalism	Rich Miller, the founder and editor at large of Data Center Knowledge.	149	2.14%	12.9K	898	3,067
gijn	The Global Investigative Journalism Network, an association of 145 nonprofits in 62 countries.	109	1.57%	12.5K	3,861	18.2K
VisualOfData	A professional account on data visualization and infographics	65	0.94%	4,260	5,624	5,531
flpires	Francisco Lavrador Pires, an Engineer and Business Learning Analyst.	64	0.92%	24.6K	931	445
MegaDataMama	A professional account on data analytic.	41	0.59%	77.4K	79	1,510
ujigis	The GIS Association of Japan.	41	0.59%	19.7K	1,420	1,168
newslinn	A professional account helping citizens, protesters and activists connect directly with local journalists.	39	0.56%	8,933	2,195	2,333
postoditacco	A professional account on Webaholic. Fact-checker. Storyteller; also working on BI and ADV for an international publisher	32	0.46%	30.1K	2,044	2,518
Bahareh360	Bahareh R. Heravi (Verified), Scientist & data journalism enthusiast, Assistant Professor of University College Dublin	31	0.45%	4,400	2,090	1,756
dmedialab	Disrupt MediaLab, a professional account.	24	0.35%	57.6K	77	10.3K
journalism	An aggregation of journalism and news publishing-related blogs and sites.	22	0.32%	29.6K	151	3,776
newsnerdnews	A professional site for digital journalists.	22	0.32%	21.3K	276	212
paulbradshaw	Professor Paul Bradshaw, an online journalist and blogger, who leads the MA in Multiplatform and Mobile Journalism at Birmingham City University.	19	0.27%	46.1K	10.4K	25.4K

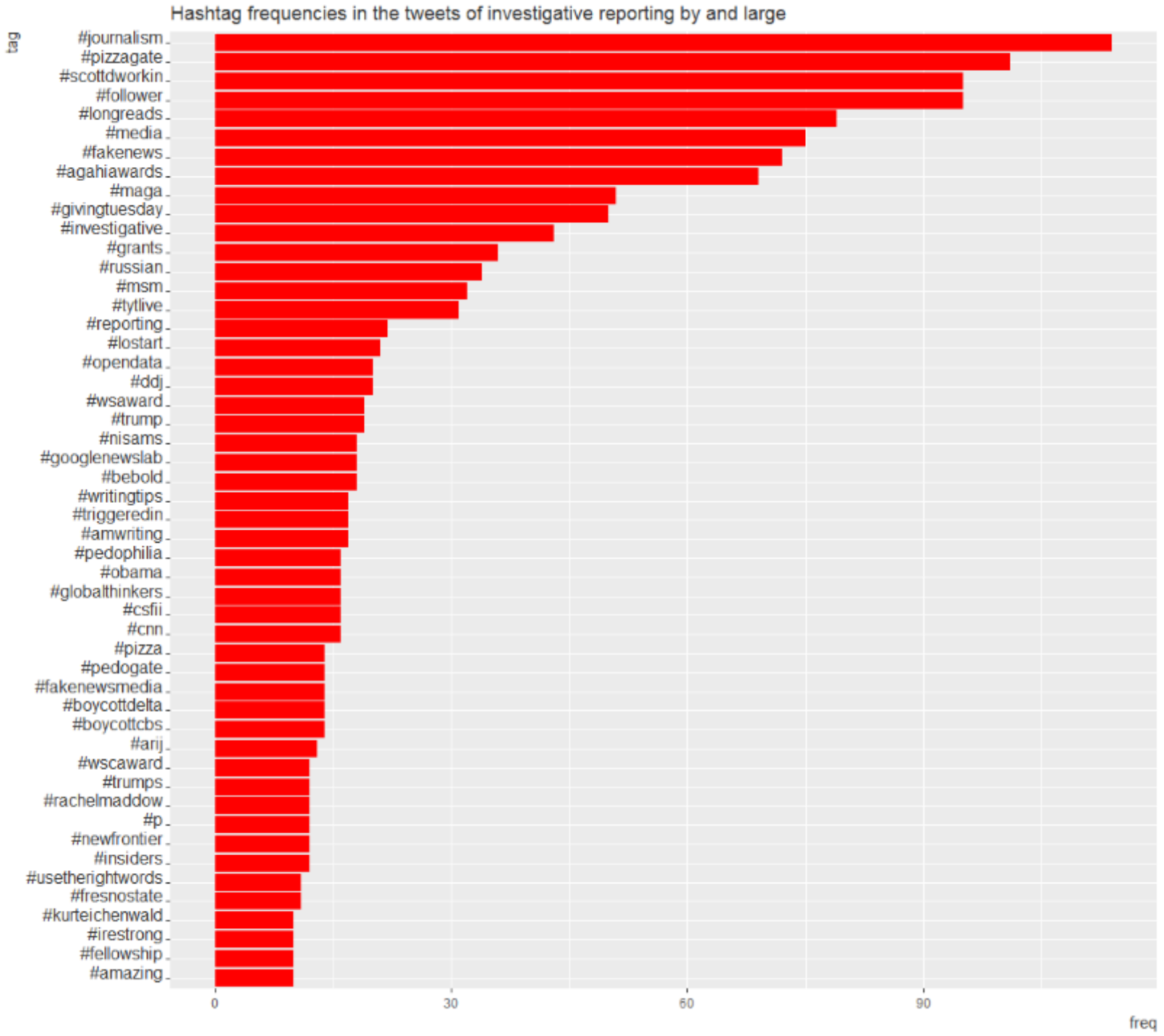
Background information of top posting users (screen names) on Twitter discussing the three types of topics

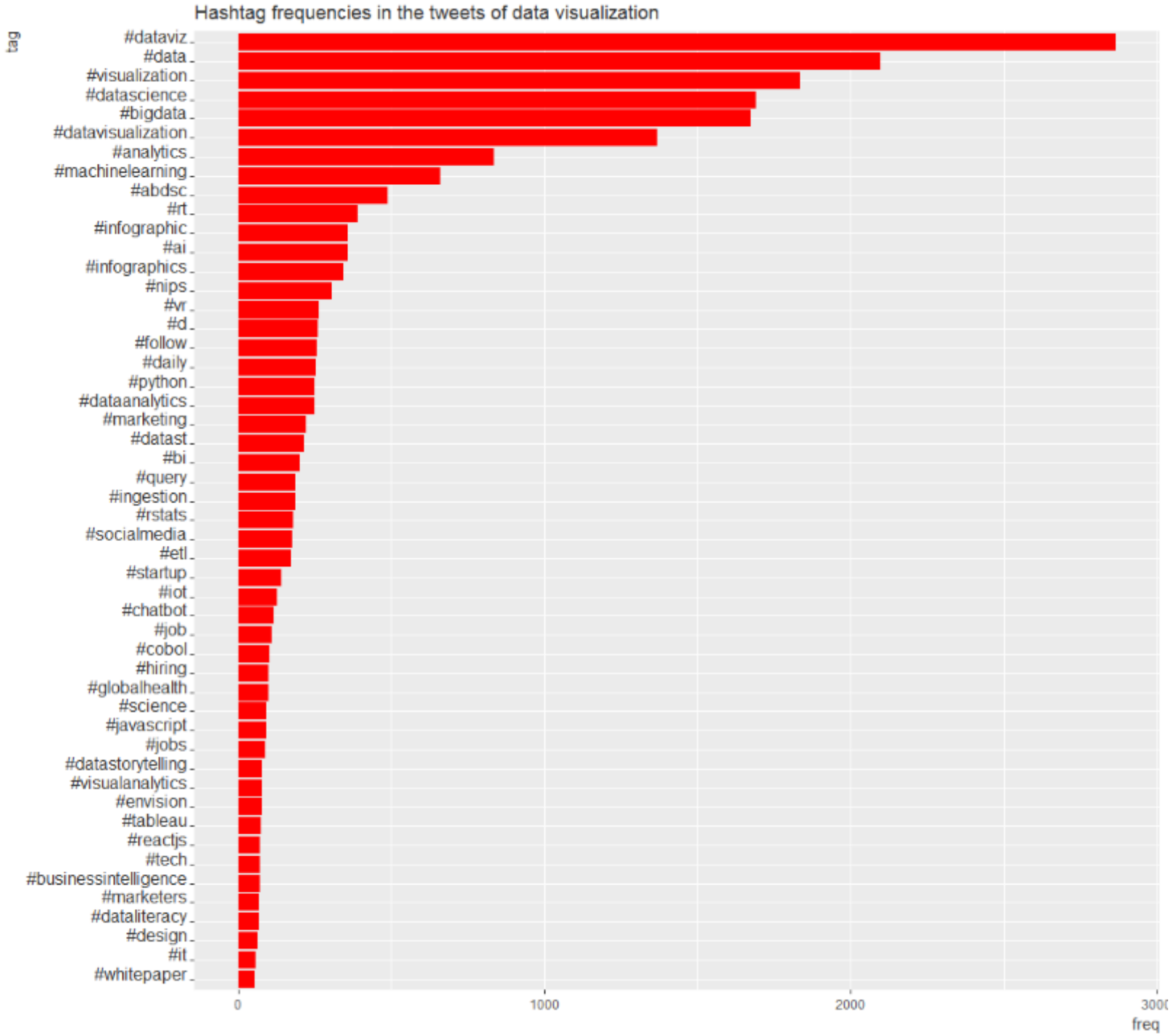
ddj			Inv. J			Dataviz		
Screen name	No. of posts	% within the sample	Screen name	No. of posts	% within the sample	Screen name	No. of posts	% within the sample
Tech_Journalism	149	2.14%	ShorensteinCtr	24	0.16%	topiclydataviz	528	2.70%
gijn	109	1.57%	kataripsias	18	0.12%	InfographicsIts	274	1.40%
VisualOfData	65	0.94%	mcgrawcenter	18	0.12%	ITProjectBoard	123	0.63%
flpires	64	0.92%	janesasseen	16	0.11%	MegaDataMama	117	0.60%
MegaDataMama	41	0.59%	reveal	16	0.11%	micoyuk	94	0.48%
ujigis	41	0.59%	gijn	15	0.10%	TT_DataVisual	70	0.36%
newslinn	39	0.56%	ucbsoj	14	0.09%	alevergara78	65	0.33%
postoditacco	32	0.46%	WSoyinkaCentre	14	0.09%	jha_jhapk	64	0.33%
Bahareh360	31	0.45%	gijnAfrica	13	0.09%	mannitan	63	0.32%
dmedialab	24	0.35%	newslinn	13	0.09%	SecurityTube	53	0.27%
journalism	22	0.32%	Lukey627	12	0.08%	OttLegalRebels	51	0.26%
newsnerdnews	22	0.32%	miamihigh09	12	0.08%	STraqr	48	0.25%
paulbradshaw	19	0.27%	AmericanRN1027	11	0.07%	AlexSecanove	41	0.21%

#Hashtag

- Hashtag: normally starting with “#” and followed by a sequence of terms, is like a label or keyword adopted by users on social media.
- It enables users to search messages with a specific theme or topic, and it represents how users categorize the key themes of the tweets, thereby reflecting how they attempt to make sense of the tweets.
  - It indicates a domain of knowledge;
  - It refers to the active users’ behavior (information searching, tagging, categorizing);
  - It captures the topic and themes;


Hashtag frequencies in the tweets of self-claimed data journalism
tag
#ddj
#journalism
#data
#datajournalism
#dja
#bigdata
#dataviz
#datascience
#per
#datacenter
#opendata
#journalists
#predictive
#ogp
#dd
#ai
#pr
#hhnbo
#datacampogp
#innovateafrica
#cloud
#tech
#media
#kenya
#pakistan
#periodismo
#journoshop
#inf
#markets
#vr
#ddjjp
#d
#training
#infographics


Hashtag frequencies in the tweets of investigative reporting by and large
tag
#journalism
#pizzagate
#scottdworkin
#follower
#longreads
#media
#fakenews
#agahiawards
#maga
#givingtuesday
#investigative
#grants
#russian
#msm
#tytlive
#reporting
#lostart
#opendata
#ddj
#wsaward
#trump
#nisams
#googlenewslab
#bebold
#writingtips
#triggeredin
#amwriting
#pedophilia
#obama
#globalthinkers
#csfii
#cnn
#pizza
#pedogate
#fakenewsmedia
#boycottdelta
#boycottcbs
#arij
#wscaward
#trumps
#rachelmaddow
#p
#newfrontier
#insiders
#usetherightwords
#fresnostate
#kurteichenwald
#irestrong
#fellowship
#amazing
0
30
60
90
freq


Hashtag frequencies in the tweets of data visualization
tag
#dataviz
#data
#visualization
#datascience
#bigdata
#datavisualization
#analytics
#machinelearning
#abdsc
#rt
#infographic
#ai
#infographics
#nips
#vr
#d
#follow
#daily
#python
#dataanalytics
#marketing
#datast
#bi
#query
#ingestion
#rstats
#socialmedia
#etl
#startup
#iot
#chatbot
#job
#cobol
#hiring
#globalhealth
#science
#javascript
#jobs
#datastorytelling
#visualanalytics
#envision
#tableau
#reactjs
#tech
#businessintelligence
#marketers
#dataliteracy
#design
#it
#whitepaper
0
1000
2000
freq

Cloud computing service

- Cloud computing is the on-demand delivery of compute power, database storage, applications, and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing.
- Amazon AWS is a commonly used service of this kind.

https://aws.amazon.com/what-is-cloud-computing/

Working with AWS

- In-class demonstration
- Tutorial preview: https://aws.amazon.com/what-is-cloud-computing/

Exercise 3

- Apply an API account from a media or social media institution
  - Answer all the questions in a transparent and reproducible manner;
- Use “Twitter Search API” to harvest any keywords or hashtags of your interest
- Store the output in the local space

Exercise 4

- Launch an Amazon AWS EC2 instance
- Create a .txt file containing the text: “this is the file for Amazon AWS EC2,” and upload it to your volume
- Create a ”print-hello-world” .py file, and execute it on your server (hint: you may need to work with CLI)
- Bonus (10 marks): execute your individual assignment #01 on your server, and store the outputs in the server as well. Download the file to the local space.

Acknowledgements / Contributions

Some of the contents in this week's slide are referred from Charles R. Severance (www.dr-chuck.com) of the University of Michigan School of Information and open.umich.edu and made available under a Creative Commons Attribution 4.0 License. Please maintain this last slide in all copies of the document to comply with the attribution requirements of the license. If you make a change, feel free to add your name and organization to the list of contributors on this page as you republish the materials.

Initial Development: Charles Severance, University of Michigan School of Information
- largely modified and extended by Xinzhi Zhang, Hong Kong Baptist University in Oct 2018, for CityU COM5507.