Please make sure you can access the class IDAS system

https://notebooks.hpc.uiowa.edu/bais61000exa

BAIS:6100 Text Analytics Introduction

Kang-Pyo Lee

How to Participate in Our Virtual Class

- Be aware every session is being recorded and the recording will be available after class
- Stay muted while you are just listening and unmute yourself when you need to speak
- Turning on the camera to show your face is recommended, but not required
- When you have a question for the instructor, you can
  - either interrupt the instructor using your microphone
  - or use the chat box indicating this is a question/comment for the instructor
- Avoid sending the instructor private messages on Zoom unless you have to
- You may leave (or rejoin) the Zoom session whenever you want
- Feel free to let the instructor know if there's anything you feel uncomfortable with during virtual class

Outline

- Introductions
- Overview of the Course
- Text Analytics
- Python, Jupyter Notebook, and IDAS
- Module 1: Python Basics for Text Processing, Part 1

Instructor

Name: Kang-Pyo Lee

Motto: "Learn from data!"

Education: Seoul National University, Ph.D. in Computer Science

Previous Work: Data Scientist at Samsung Big Data Center

Current Work: Lecturer at Business Analytics, Tippie College of Business

Data Scientist at Iowa Initiative for Artificial Intelligence (IIAI)

Adjunct Lecturer at Biostatistics, College of Public Health

Research Interests: social media analytics, text analytics, machine learning, big data

Courses and Workshops

Credit courses

- BAIS:6040 Data Programming in Python (Business Analytics)
- BAIS:6100 Text Analytics (Business Analytics)
- BIOS:7600 Big Data Analysis with Python (Biostatics)

Training workshops

- Introduction to Python Data Programming
- Machine Learning with Python
- Web Scraping with Python
- Social Media Analytics with Python

Welcome Message from Advisors & Site Directors

Welcome to the beginning of a great semester!

If you have any questions throughout the semester, don't hesitate to connect with your advisor (based on program/site). Visit our Tippie College of Business website for contact information or to schedule a virtual advising appointment.

Professional MBA & Business Analytics Programs Team:


Lisa Smith
Cedar Rapids


Angela Ross
Des Moines


Francine Bryce
Des Moines


Chelsea Hillman
Quad Cities

Online MBA Program Team:


Michel Pontarelli
Online MBA program


Jan Fasse
Online MBA program


Nicole Vogt
Online MBA program

"Your advisors and Site Directors would like to welcome you to the Spring 2021 session. If you have any questions about degree progress, registration for future classes, etc. you can reach out to them via email. You likely already have their email addresses, but you can also find them on the Tippie website. They wish you a successful winter session!"

Self-Introduction

- **Briefly introduce yourself!**
  - **This is probably the first and the last opportunity to introduce yourself to the whole class**
  - **This may help you find your team members for the group project**
  - **Sharing why you decided to take this Text Analytics course would be appreciated**
  - **Turn on the camera while introducing yourself if you don’t mind**

Goal & Scope of This Course

This course aims to introduce the concepts and techniques of text analytics using the Python programming language

Goal & Scope of This Course

We are going to use Python as the only programming language of this course

Students are not allowed to use R or any other programming languages

Goal & Scope of This Course

The main topics include:

- Python basics for text processing
- Natural Language Processing (NLP) techniques
- Keyword analysis and visualization
- Text data acquisition
- Term-document matrix representation
- Text classification
- Text clustering and topic modeling
- Text similarity
- Keyword network analysis

Course Schedule (Subject to Change)

Week	Date	Topics	Due
1	Jan 28	Introduction to Text Analytics Introduction to Python, Jupyter Notebook, and UI Interactive Data Analytics Service (IDAS)	
2	Feb 4	Module 1. Python Basics for Text Processing, Part 1 : Strings, Collections, Built-in Functions, Flow Control, and User-Defined Functions	
3	Feb 11	Module 2. Python Basics for Text Processing, Part 2 : Files, Dataframes, and Pattern Matching Using Regular Expressions	HW 1
4	Feb 18	Module 3. Basic Natural Language Processing (NLP) Techniques : Tokenization, Part-of-Speech Tagging, Stemming, Lemmatization, N-grams, Noun Phrase Extraction, Language Detection and Translation, and Gender Prediction Module 4. Keyword Analysis and Visualization	HW 2
5	Feb 25	Test 1	HW 3 (Feb 23)
6	Mar 4	Modules 5 & 6. Text Data Acquisition Using Twitter APIs and Web Scraping Group Project Announcement	
7	Mar 11	Module 7. Document-Term Representation Module 8. Text Classification	Hw 4
8	Mar 18	Module 9. Text Clustering and Topic Modeling	Project Proposal
9	Mar 25	Module 10. Text Similarity Module 11. Keyword Network Analysis	
10	Apr 1	Test 2	HW 5 (Mar 30)
11	Apr 8	Group Project Presentations and Course Wrap-Up	Project Deliverables

Course Activities

8 formal and active-learning lecture sessions with in-class hands-on practice

5 individual homework assignments

2 individual tests

1 group project

Coursework

5 homework assignments, 35%

(equally weighted)

2 tests, 50%

(two in-class exams, equally weighted)

1 group project, 15%

Final Letter Grades

A: ≈ 60% of students

B: ≈ 40% of students

C, D, F: as needed

The A and B ranges will be divided into +/- designations

Late Assignments

- **All homework assignments are expected on time**
- **You may turn in an assignment late, but you will receive a 20% deduction for each day that it is late, including the first/same day**

Media/System Requirements

- **Check the ICON course website frequently for announcements, assignments, etc.**
- **Make sure to receive notifications from ICON via email and not to miss any important communications**
- **We are going to use the class IDAS system throughout the semester**

Communications

Feel free to ask the instructor any questions about class

- **If your question is something to be shared with the whole class, you may post it on the ICON course website as a reply on Announcements or as a post on Discussions (Do not share code under any circumstance)**
- **If you don’t want to share your question with others, email the instructor at kangpyo-lee@uiowa.edu or send a message on ICON (Expect to receive a response within 24-48 hours)**

Office Hours

Office hours will be held on Wednesdays from 10 am to 11 am via Zoom by appointment

Attendance

- **No attendance policy for regular lecture sessions**
- **You may miss regular sessions, and do not have to let the instructor know you are going to miss a regular session**
- **Always make sure to check every single announcement made during each session you missed via class recording**

Attendance

- **All students are expected to be present for the two tests and group presentations at the regularly scheduled times**

- **Discuss with the instructor at least one week in advance**
  - **in the event that you must miss any of those non-regular lecture sessions above**
  - **if you have specific accommodations that have been approved by the university (e.g., SDS)**

Class Size

26 enrolled students

with 1 instructor and no TAs

Student Honor Code

Students must adhere to the Tippie Master's Honor Code that emphasizes the importance of honesty and integrity

DO NOT SHARE CODE for homework assignments under any circumstance

Prerequisites

MSCI/BAIS:6060
Data Programming in R

OR

MSCI/BAIS:9060
Data Programming in R

OR

MSCI/BAIS:6040
Data Programming in Python

AND

MSCI/BAIS:6070
Data Science

OR

MSCI/BAIS:9110
Advanced Analytics

Prerequisites

For those who are unfamiliar with Python:

- The first three sessions will cover the basics of Python focused on text analytics, which should help you get used to Python
- You will need to put some extra effort and time to learn Python

For those who have taken Data Programming in Python:

- You will find the first half of the course to be similar to what you learned before, but everything will be focused on text analytics
- The second half will be totally different

Course Syllabus

For more details, refer to the full text of the course syllabus posted on the ICON course website

Datasets

Hashtag Tweets

Twitter Hashtag	Year	Tweets Collected
#ai	2020	207,528
#bitcoin	2020	304,667
#blacklivesmatter	2020	798,902
#bts	2020	2,747,841
#covid19	**2020**	**3,681,594**
#fakenews	2020	190,958
#innovation	2020	51,414
#mentalhealth	2020	76,898
#metoo	2020	89,554
#startup	2020	55,297

Timeline Tweets

Twitter User	Owner	Activity	Followers	Tweets Collected
@justinbieber	Justin Bieber	Musician	#2 (114M)	3,140
@katyperry	Katy Perry	Musician	#3 (109M)	3,191
@Cristiano	Cristiano Ronaldo	Footballer	#5 (90M)	3,149
@TheEllenShow	Ellen DeGeneres	Comedian	#9 (79M)	3,199
@KimKardashian	Kim Kardashian	TV personality and businesswoman	#11 (68M)	3,182
@cnnbrk	CNN Breaking News	News channel	#15 (59M)	3,200
@BillGates	Bill Gates	Businessman and philanthropist	#19 (53 M)	3,200
@nytimes	The New York Times	Newspaper	#25 (48M)	3,200
@NASA	NASA	Space agency	#32 (42M)	3,200
@elonmusk	Elon Musk	Industrial designer and tech entrepreneur	#33 (42M)	3,199

What Is Text Analytics?

Text analytics is the type of data analytics that focuses on the process of automatically deriving information from text

Types of Data

Whether there is
a predefined data model

Structured Data

- **Data in a tabular format with columns and rows**
- **Tables in relational DBs**

Predefined, fixed schema

vs.

Unstructured Data

- **Text data**
- **Multimedia data such as images, audios, and videos**

vs.

Semi-structured Data

- **XML (eXtensible Markup Language)**
- **HTML (HyperText Markup Language)**
- **JSON (JavaScript Object Notation)**

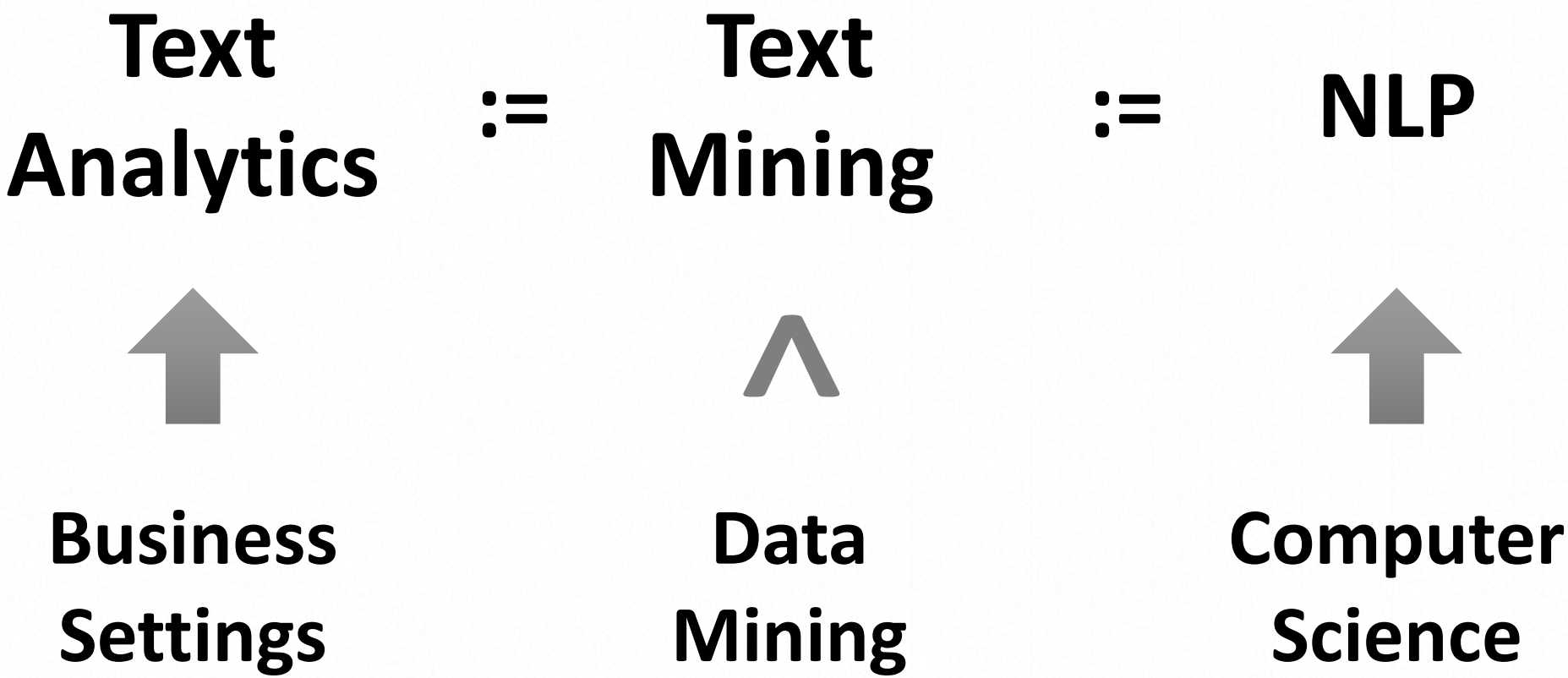
Flexible schema

What Is Text Analytics for?

The overarching goal of text analytics is to turn raw text into data for analysis, via application of natural language processing (NLP) and different types of algorithms and analytical methods

Text mining on Wikipedia: https://en.wikipedia.org/wiki/Text_mining

Terminology



Process of Text Analytics

Text analytics usually involves the process of

1) **structuring the input text (usually by parsing)**
2) **deriving patterns within the structured data (usually by applying statistical methods)**
3) **and finally evaluation and interpretation of the input (usually manually by humans)**

Something to Be Aware of

Text analytics is challenging due to the nature of language

Do not always expect 100% accuracy from text analytics techniques

Text Analytics

100% accurate!

Not 100% accurate!

```
In [1]: s = "Really? Are you nuts? Have you looked at yourself lately?"

In [2]: s[:7]
Out[2]: 'Really?'

In [3]: s.lower()
Out[3]: 'really? are you nuts? have you looked at yourself lately?'

In [4]: from textblob import TextBlob
        tb = TextBlob(s)

In [5]: tb.tags
Out[5]: [('Really', 'RB'),
         ('Are', 'NNP'),
         ('you', 'PRP'),
         ('nuts', 'NNS'),
         ('Have', 'VBP'),
         ('you', 'PRP'),
         ('looked', 'VBN'),
         ('at', 'IN'),
         ('yourself', 'PRP'),
         ('lately', 'RB')]

In [6]: tb.sentiment
Out[6]: Sentiment(polarity=-0.04999999999999999, subjectivity=0.4)
```

Text Analytics

Parsing level vs. Application level

Text manipulation & transformation
Tokenization & n-grams
Stemming
Lemmatization
Part-of-speech (PoS) tagging
Dependency parsing
Pattern matching

Text summarization
Text classification
Text clustering
Topic modeling
Sentiment analysis (Opinion mining)
Text similarity & full-text search
Named Entity Recognition (NER)
Language detection & translation

Text Analytics in Python

Built-in functionality & string methods of Python

+

Text analytics libraries

e.g., pandas, nltk, gensim, scikit-learn, textblob

Python as a Programming Language

python™ is a general-purpose high-level programming language

Python as a Programming Language

Python is a general-purpose high-level programming language

Can be used to build just about anything:

web development
data analysis and artificial intelligence
networking
scientific computing
building productivity tools, games, and desktop applications
etc.

Python as a Programming Language

Python is a general-purpose high-level programming language

Written in a form that is close to our human language, enabling programmers to just focus on the problem being solved

```
a = "I'm learning Python data analytics."
a.replace("Python", "R")
```

Python as a Data Analytics Tool

The nature of Python makes it a perfect-fit for data analytics

Easy to understand and learn
Readable and flexible code
Easy integration with other applications
Open access to an extensive set of libraries
Active community & ecosystem

Comparison with Other Data Science Software

	Proprietary	Open-Source
Traditional	SAS, SPSS, STATA, MATLAB, Minitab	Python, Jupyter, WEKA The University of Waikato, R, RStudio
Latest	RapidMiner, Tableau, Power BI	Open for Innovation KNIME, Orange, H2O.ai

Python Script

A Python script is a text file that contains executable Python program statements

Python Script

A first way to write and run a Python script

1. Install Python on your computer
2. Write a Python script using a text editor
3. Save the script as a file with the file extension .py
4. Open a command line tool (e.g., Command Prompt or PowerShell on Windows and Terminal on Mac) and move to the directory where the script file is saved
5. Type the following command and press enter:

```
python FILE_NAME.py
```

Writing a Python Script

print_text.py

```
import pandas as pd

df = pd.read_csv("classdata/tweets/timeline_UN.csv", sep="\t")

series = df["text"][:10]

for item in series:
    print(item + "\n")
```

Running a Python Script

```
(base) kangplee@jupyter-notebook-research-kangplee:~$ python print_text.py
RT @WFP: Famine feeds on: 🍞#ClimateChange🌎 Chaos💥ConflictCOVID-19 has compounded existing issues and pushed millions of p
eople to t…

RT @UNReliefChief: Our best chance of getting ahead of this virus now is if all wealthy nations, especially those with mult
iple deals with…

🌳Forest &amp;amp; land conservation☀ Renewable energy🌾Climate-friendly farming techniques👷 Green businesses &amp;amp; jobs… http
s://t.co/698QTIsN0S

RT @Refugees: 2020 was a record low for refugee resettlement.We urge States to offer more resettlement places and help save
 lives of refu…

RT @UNICEF: "The very little we know about the impact of the conflict on children in Tigray is deeply troubling. @unicefchi
ef To reach f…

On Wednesday's #HolocaustRemembranceDay we honour the memory of the six million Jews &amp;amp; millions of others who peris… ht
tps://t.co/lxdxEKGDEk

RT @mbachelet: On this day, we are reminded of the horror to which hatred and lies can lead.Words have consequences.We need
 to ensure t…

RT @UNDP: 50 countries. 17 languages. 1.2 million people. The results are in of our #PeoplesClimateVote - the largest surve
y of public op…

Constituents who are asking questions &amp;amp; raising issues are very powerful. -- UN Envoy Mark Carney explains that e… http
s://t.co/dXxEfWvr0y

RT @UNESCO: The Holocaust began with words - and in the era of the internet and social media, the power of propaganda is mo
re devastating t…

(base) kangplee@jupyter-notebook-research-kangplee:~$
```

iPython & Jupyter Notebook

iPython is a Python command shell for interactive computing

Jupyter Notebook (formerly iPython Notebook) is a web-based interactive data analysis environment that supports iPython

Why Jupyter Notebook?

Interactive

Easy to share

Jupyter Notebook

print_text.ipynb

Print Text in a CSV File

- Developed by Kang Lee
- Last updated on January 27, 2020

Import Modules

```
In [1]: import pandas as pd
```

Load the CSV file into a Pandas Dataframe

```
In [2]: df = pd.read_csv("classdata/tweets/timeline_UN.csv", sep="\t")

In [3]: df.shape
Out[3]: (3200, 7)
```

Select Data from the Dataframe

```
In [4]: series = df["text"][:10]

In [5]: len(series)
Out[5]: 10
```

Iterate over the Series Printing the Value

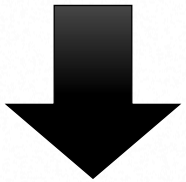
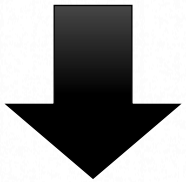
```
In [6]: for item in series:
            print(item + "\n")
```

```
RT @WFP: Famine feeds on:🌪 #ClimateChange🌍 Chaos💥 ConflictCOVID-19 has compounded existing issues and p
eople to t…

RT @UNReliefChief: Our best chance of getting ahead of this virus now is if all wealthy nations, especially
deals with…

🌳 Forest &amp;amp; land conservation☀ Renewable energy🌾 Climate-friendly farming techniques👷 Green business
s://t.co/698QTIsN0S
```

Jupyter Notebook vs. Jupyter Hub

	**Jupyter Notebook**	**Jupyter Hub**
	**RStudio Desktop**	**RStudio Server**
		
	Useful for individual work with smaller data	Useful for team work with larger data and teaching

Interactive Data Analytics Service (IDAS)

IDAS is a campus resource to support large-scale and collaborative data analytics using interactive tools such as Jupyter Notebook for Python and RStudio for R

IDAS

- **Applications**
  - Jupyter Notebook for Python, R, and Julia
  - RStudio for R
- **Use types**
  - Research (genera) use
  - Class use
- **All 4 options available now**

IDAS Links

- User communication channels
  - Homepage: https://hpc.uiowa.edu/interactive-data-analytics-service-idas
  - ITS service page: https://its.uiowa.edu/interactive
  - Wiki Documentation: https://wiki.uiowa.edu/display/hpcdocs/Interactive+Data+Analytics+Service+Documentation
- Requests
  - User account request: https://workflow.uiowa.edu/form/idas-account
  - Class use request: https://workflow.uiowa.edu/form/idas-class-request
  - Software request: coming soon
- Access
  - Jupyter: https://notebooks.hpc.uiowa.edu/
  - RStudio: https://rstudio.hpc.uiowa.edu/

Class IDAS System

Be advised that

- You need an Internet connection to access the system
- You have no root (admin) access
- All the files you see in your Jupyter environment are located in the server, not in your local computer (You will have to back up all your files by downloading them to your local computer once the semester is over)
- The system will be available 24/7 throughout the semester
- There is a shared folder named *classdata*, in which you can find all the notebook files and data files for this course (You only have read access)

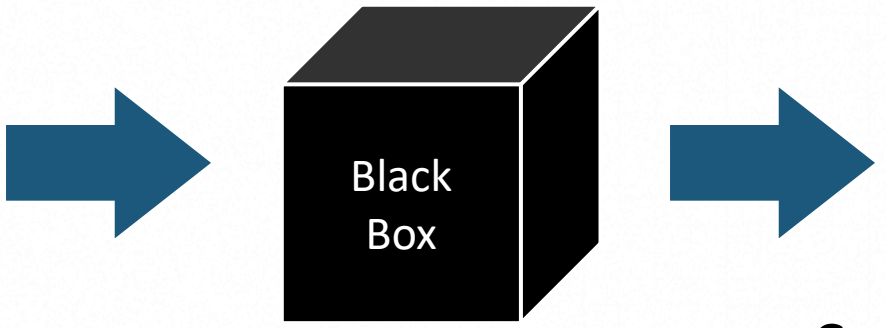
Python Data Analytics Libraries/Packages

Useful to know:

- Each library has its own purpose and usage
- A library takes the form of a package
- Library repository: PyPI
- A library is typically developed, maintained, and upgraded by a team/organization of developers (versioning and dependencies are important!)
- Installing a package is a one-time process, you just load it after installation

Python Data Analytics Libraries/Packages

Python Package



Parameters

Something you expect from the library

You do not have to implement each component yourself!
All you need to care about is to find the right package and use it the right way

PyPI (Python Package Index): a repository of software for the Python programming language

Python Data Analytics Libraries/Packages

Reasons you should use commonly-used Python packages rather than writing the code yourself

Convenient to use

Often well-tested

Possibly faster than your code

Popular Python Data Analytics Libraries/Packages

Package	Usage
numpy, scipy	Numerical & scientific computing
pandas	Data manipulation & aggregation
mlpy, scikit-learn	Machine learning
keras, tensorflow, theano	Deep learning
statsmodels	Statistical analysis
nltk, gensim, textblob	Text processing
networkx	Network analysis
bokeh, ipywidgets, matplotlib, plotly, seaborn	Visualization
beautifulsoup, scrapy, selenium	Web scraping

Data Analytics Settings for This Course

Component	Name
Python version	Python 3 (vs. Python 2)
Data analytics environment	Jupyter Notebook (vs. Wing IDE, PyCharm, PyDev, Spyder)
Data analytics software toolkit	Anaconda (vs. Enthought Canopy)
Data analytics libraries	pandas for data analysis beautifulsoup & selenium for web scraping nltk, genism, scikit-learn & textblob for text processing

Useful Resources for Learning Jupyter Notebook

Jupyter Notebook for Beginners: A Tutorial
https://towardsdatascience.com/jupyter-notebook-for-beginners-a-tutorial-f55b57c23ada

Advanced Jupyter Notebooks: A Tutorial
https://towardsdatascience.com/advanced-jupyter-notebooks-a-tutorial-3569d8153057

Jupyter Notebook for Beginners: A Tutorial
https://www.dataquest.io/blog/jupyter-notebook-tutorial/

28 Jupyter Notebook Tips, Tricks, and Shortcuts
https://www.dataquest.io/blog/jupyter-notebook-tips-tricks-shortcuts/