BAIS:6100 Text Analytics

Python Basics for Text Processing Part 1

Kang-Pyo Lee

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

2

Class Survey Results

Homework

Homework 1, which corresponds to Module 1, is due by **6:00 PM** on **Thu, Feb 11** via ICON Assignments

You will receive a **20%** deduction for each day that it is late, including the first/same day

16 questions, 7 points in total

Homework

Homework 1 is expected to take a student with a good understanding of the topics 1-2 hours to complete it

There will be a question at the end of the notebook asking you to share how much time it actually took you to complete the homework

Homework

35 pts assigned to 5 homework assignments

- HW 1: 7 pts
- HW 2: 7 pts
- HW 3: 7 pts
- HW 4: 7 pts
- HW 5: 7 pts

Homework

How to complete a homework assignment:

1. Go to Terminal on IDAS
2. Copy the Homework1.ipynb file in the *classdata* folder to your home directory (You do not have to change the file name)
3. Open the notebook and complete the questions
4. When you are done, click Kernel > Restart & Run All to create a clean version of the notebook and then save the notebook
5. Click File > Download as > Notebook (.ipynb) to download the notebook from IDAS to your computer (You should have downloaded the Homework1.ipynb file)
6. Click File > Download as > HTML (.html) to download the HTML version of the notebook to your computer (You should have downloaded the Homework1.html file)
7. Go to the Assignments on ICON and submit both the two files (*.ipynb and *.html)
8. Validate your submission by opening the submitted HTML file on ICON to check if you've submitted the right file

Homework Instructions & Rules

- **The format of homework assignments is the same as that of in-class exercises; the only exception is that you may need to save the final outcome from a question in a specified variable**
- **Pay attention to every single condition in the question and make sure to reflect them all in your answer → Missing anything may lead to losing points**
- **You can feel free to use any approach you would want to each problem as long as it meets all the conditions specified in the question**
- **Every question should be based on what you have learned in class, so try reviewing the class materials first before you try Googling**

Homework Instructions & Rules

- Use only one cell per question, the cell starting with *# Your answer here*
- Each code cell must be run with no errors → Make sure to Restart & Run All before submission to create a clean version
- Do not contain any unnecessary code in your answer (It may affect your answer)
- All problems must be solved in a programmable way
- You can discuss with each other, but do not share code under any circumstance
- You are responsible for any wrong submission → Make sure to submit the right ones and validate the submitted HTML version of the notebook

Homework Grading

- Each assignment is manually graded by the instructor
- Official solution to each homework will be shared after grading (Official solution does not always mean the only solution)
- If you lose any points from a question, that should be because you fail to meet any of the conditions in the question, not because your solution is not identical to the official solution
- When losing points, there can be partial credit depending on the question
- You can always ask the instructor about your grades

Remember...

- **Do your best to pay attention to every little detail in homework assignments! (Those details will make a difference in your grades and your data science career)**
- **Do not share code under any circumstance!**
- **Validate your submission on ICON!**