BAIS:6100 Text Analytics

Basic NLP Techniques and Keyword Analysis & Visualization

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

Homework

Homework 3, which corresponds to Modules 3 & 4, is due by 6:00 PM on Wed, Feb 24, not Thu, Feb25, via ICON Assignments

7 questions, 7 points in total

No delay for Homework 3!

Some Change in Homework 3

Homework 3 and the rest of the homework assignments will be a little more self-directed, not providing detailed instructions on how to complete the questions

Midterm Test

- Thu, Feb 25 at 6 pm (Please do not be late!)
  - Instructions (5 minutes)
  - Test (2.5 hours)
- 10-12 questions, 25 points in total
- Materials covered
  - Modules 1-4
  - Homework 1-3
- Rules
  - Open notes, open Internet
  - No communication with anyone else but the instructor

Midterm Test

- **Process**
  - **Questions will be given via an online document**
  - **You will have a Jupyter notebook for the midterm test on IDAS and complete the questions using that notebook**
  - **At the end of test, submit both of your notebook and HTML files to ICON**
- **Student responsibilities**
  - **Prepare your computer: charged battery, power cord, Internet connection, etc.**
  - **Prepare your research IDAS in case the class IDAS is unavailable during the exam**

Midterm Test

- **All questions should be based on what you have learned during class**
- **The format will be very similar to that of homework assignments**
- **The best practice to prepare for the test is to**
  - **review all the details in the notebooks**
  - **familiarize yourself with the core concepts and skills**
- **Tests are expected to be harder than homework assignments mainly due to the time limit**
- **When grading, some level of partial credit can be considered for each question**

Midterm Test

Let the instructor know by Fri, Feb 19 if you will be unable to be present at the test at the scheduled time or if you will need any special accommodations approved by the university

IDAS Research Use

- **We’ll be using the class IDAS system throughout the semester, but you can also use IDAS for research/personal use**
  - **Note that the "class" IDAS and "research" IDAS are separate, i.e., they do not share storage, especially the *classdata* folder**
  - **While the class IDAS system is expected to be available all the time, you can use the research IDAS when the class IDAS is unavailable**
- **Request the research use of IDAS**

  https://workflow.uiowa.edu/form/idas-account

Line Separator in the Printed Output

Suppose you have three strings “`This is the first sentence.`”, “`This is the second sentence`.”, and “`This is the third sentence.`”

When the three strings are printed separated by no character, it should look like the following:

```
This is the first sentence.This is the second sentence.This is the third sentence.
```

When they’re separated by a single new line, it should look like:

```
This is the first sentence.
This is the second sentence.
This is the third sentence.
```

Line Separator in the Printed Output

When they're separated by two new lines, it should look like:

```
This is the first sentence.

This is the second sentence.

This is the third sentence.
```

Remember that the default value for the end parameter in the print function is a new line ('`\n`'). If you don't set the parameter, it uses a new line as default.

Regular Expressions

"How general/specific should my regular expression be?"

```
['rcbarnes@umich.edu',
 'josh.murray@vanderbilt.edu',
 'tarunbanerjee@pitt.edu',
 'E.M.Heemskerk@uva.nl',
 'oubenal1@gmail.com',
 'info@data-mining-forum.de']
```

```
re.findall("[a-zA-Z0-9.]+@[a-zA-Z0-9.-]+\.[a-zA-Z0-9]+", text)
```

vs.

```
re.findall("[a-zA-Z0-9._]+@[a-zA-Z0-9.-]+\.[a-zA-Z0-9]+", text)
```

Regular Expressions

"How general/specific should my regular expression be?"

- **If your goal is to find all matches potentially existing in the world, your regular expression should be comprehensive enough to cover all possible cases**
- **In this course, however, it doesn’t have to be that comprehensive unless specified. All you need is to find those matches specified in the instructions.**