BAIS:6100 Text Analytics

Text Similarity
Keyword Network Analysis

Kang-Pyo Lee

Course Schedule

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 24)
6	Mar 4	Modules 5 & 6. Text Data Collection 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 31)
11	Apr 8	Group Project Presentations and Course Wrap-Up	Project Deliverables

Homework

Homework 5, which corresponds to Modules 7, 8, 9, and 11, is due at 6:00 PM on Wed, Mar 31, not Thu, Apr 1, via ICON Assignments

10 questions, 7 points in total

No delay for Homework 5!

Class IDAS

Class IDAS will be down for an hour some time over the weekend, so the system admin can deploy the software for shared group folders

Final Test

- Thu, Apr 1 at 6 pm (Please do not be late!)
  - Instructions (5 minutes)
  - Test (2.5 hours)
- 6-7 questions, 25 points in total
- Materials covered
  - Modules 6-11
  - Homework 4-5
- Rules
  - Open notes, open Internet
  - No communication with anyone else but the instructor

Final 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**

Final 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

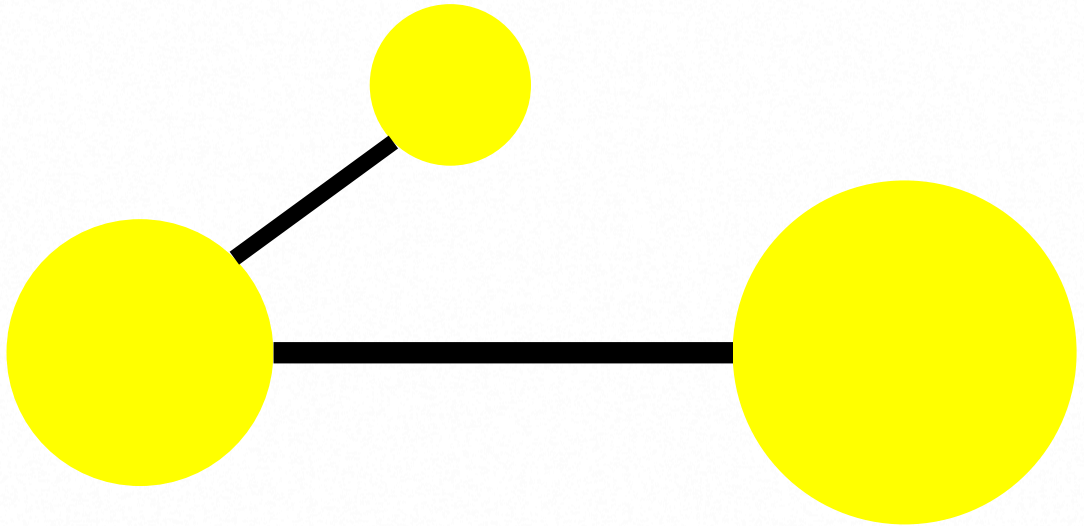
Final Test

Let the instructor know by Fri, Mar 26 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

Network Analytics

Network analytics aims to understand relationships among entities

A network is defined as a graph with nodes (vertices), edges (links), and, optionally, their weights



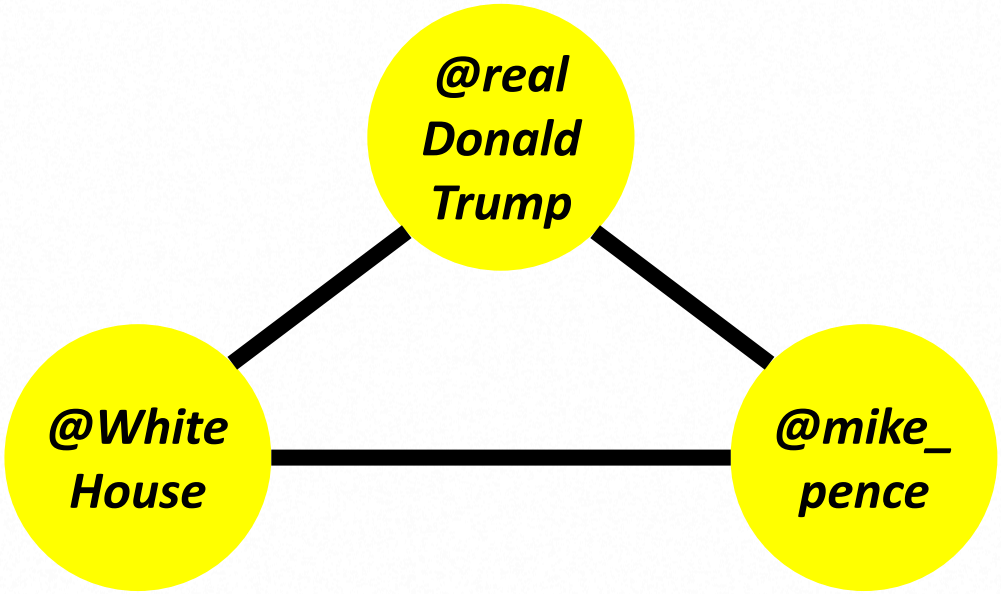
The larger a node weight is, the bigger the node becomes

The larger an edge weight is, the thicker the edge becomes

Network Analytics

Social Network

- Node – person
- Edge – (follow) relationship



Keyword Network

- Node – keyword
- Edge – (co-occurrence) relationship