Week 2
LEC Support Session

CAB431

Professor Yuefeng Li
School of Computer Science
Queensland University of Technology

Outline

1. Week 2 Lecture Review
2. Week 2 Review Questions
3. First Workshop – week 2 workshop
4. CAB431 – Questions & Feedback
   - Where can we find the zoom ID and password for the online workshop?
   - RESPONSE:
     - in Canvas, the page **Contact your teaching team**

Week 2 lecture review

1. Processing Text
   - Text Statistics
   - Tokenizing
   - Stopping and stemming
   - Phrases and N-grams
2. Information Extraction
   - Named Entity
   - HMM - *Hidden Markov Model*
3. Document Structure and Markup
   - HTML tags
   - Hyperlinks

Week 2 Review Questions

- **Processing Text**
  - From Words to Terms and Text Statistics
    - find useful index 'terms' or text features from words.
    - Text processing is significant to the results of text analysis.
  - **Document Parsing**
  - **Questions**
    - **Question 1** - open an XML file, find all terms and their frequencies and represent it as a dictionary. At last, plot the distribution of the top-10 terms.
    - **Question 2** – multiple choice + short answer
    - **Question 3 (N-grams) -** design a python program to print bigrams and trigrams .

Week 2 Review Questions cont.

- **Information Extraction**
  - Named entity
    - The process of recognizing them and tagging them in text is sometimes called semantic annotation.
    - Two main approaches have been used to build named entity recognizers: rule based and statistical.
  - Hidden Markov Model
    - Markov property - e.g., The context of a word can be described by modeling the generation of the sequence of words.
  - **Questions**
    - **Question 4 (Markov chain)** - (1) and (2), to understand transition matrix; (3) optional – how to make prediction using transition matrix.
    - **Question 5.** (This question is optional) - design a python Viterbi function to find a sequence of states (X) in an HMM, for a given Y, a sequence of observation.

Week 2 Reivew Questions

- **Document Structure and Markup**
  - To recognize document structure and make it available for indexing.
  - **Question 6 -** Design a python program to extract all hyperlinks (or destination links) in a html file.

Week 2 Workshop

- Friday 9-10:30am (S503), zoom ID: 870 0969 2757 Password: 625042
- This week's workshop is about basic I/O operations and pre-processing text data using Python.
- Getting familiar with Python for doing practical questions.
- Know how to use basic Python for simple text pre-processing.

CAB431 – Questions & Feedback

- Where can we find the zoom ID and password for the online workshop?
- RESPONSE:
  - in Canvas, the page **Contact your teaching team**