INF 558 Information Graph – Assignment 3

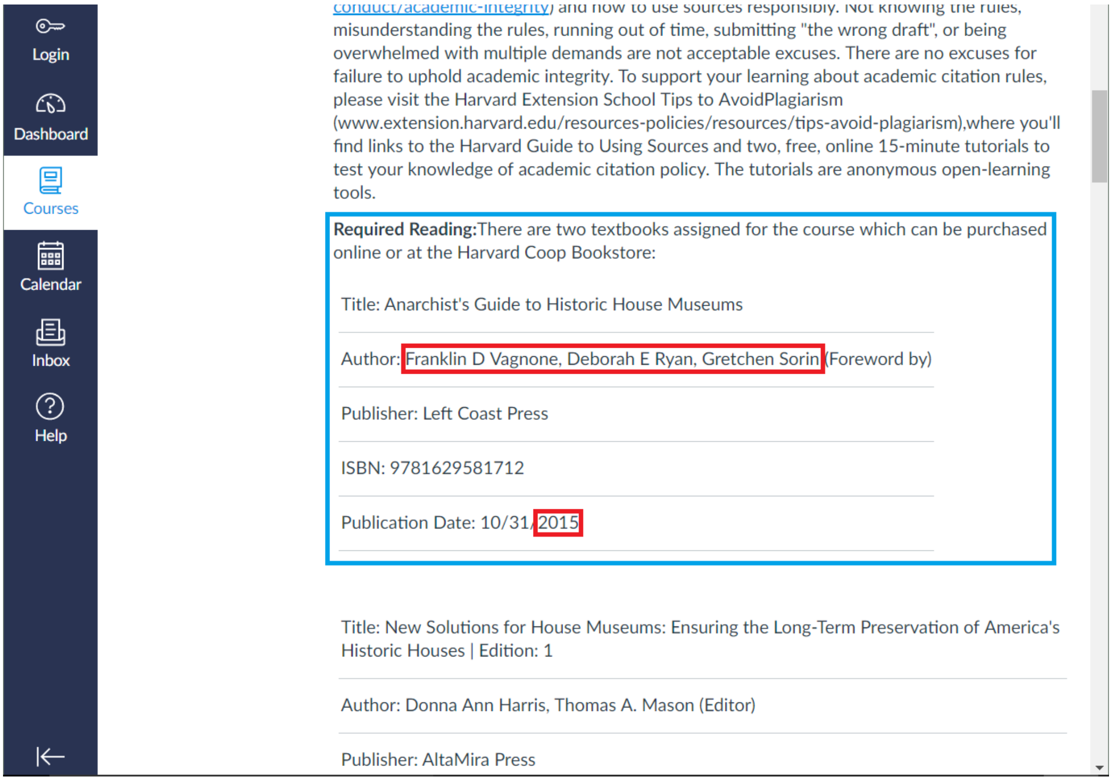
Student: Cheng-Lin Li USC ID: 9799716705

Task 1:

1. Identify the source of unstructured text and the key information.

- Objectives:
    - Target website: Harvard Extension School
    - Target object: textbooks
    - Target information:
        - Authors/ name
        - publish year.
- Challenges:
    - Every syllabus has different format.
        - Please refer figure 1 and 2.
    - Textbooks format are different.
        - Please refer figure 1 and 2.
    - The order of author name is different.
        - First name Last name
        - Last name, First name
        - Last name
    - The location of author name is different.
        - Before book name
        - After book name
    - Different number of authors
        - 1 author
        - Multiple authors.

INF 558 Information Graph – Assignment 3



conduct/academic-integrity) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting "the wrong draft", or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity. To support your learning about academic citation rules, please visit the Harvard Extension School Tips to AvoidPlagiarism (www.extension.harvard.edu/resources-policies/resources/tips-avoid-plagiarism),where you'll find links to the Harvard Guide to Using Sources and two, free, online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.

Required Reading: There are two textbooks assigned for the course which can be purchased online or at the Harvard Coop Bookstore:

Title: Anarchist's Guide to Historic House Museums

Author: Franklin D Vagnone, Deborah E Ryan, Gretchen Sorin (Foreword by)

Publisher: Left Coast Press

ISBN: 9781629581712

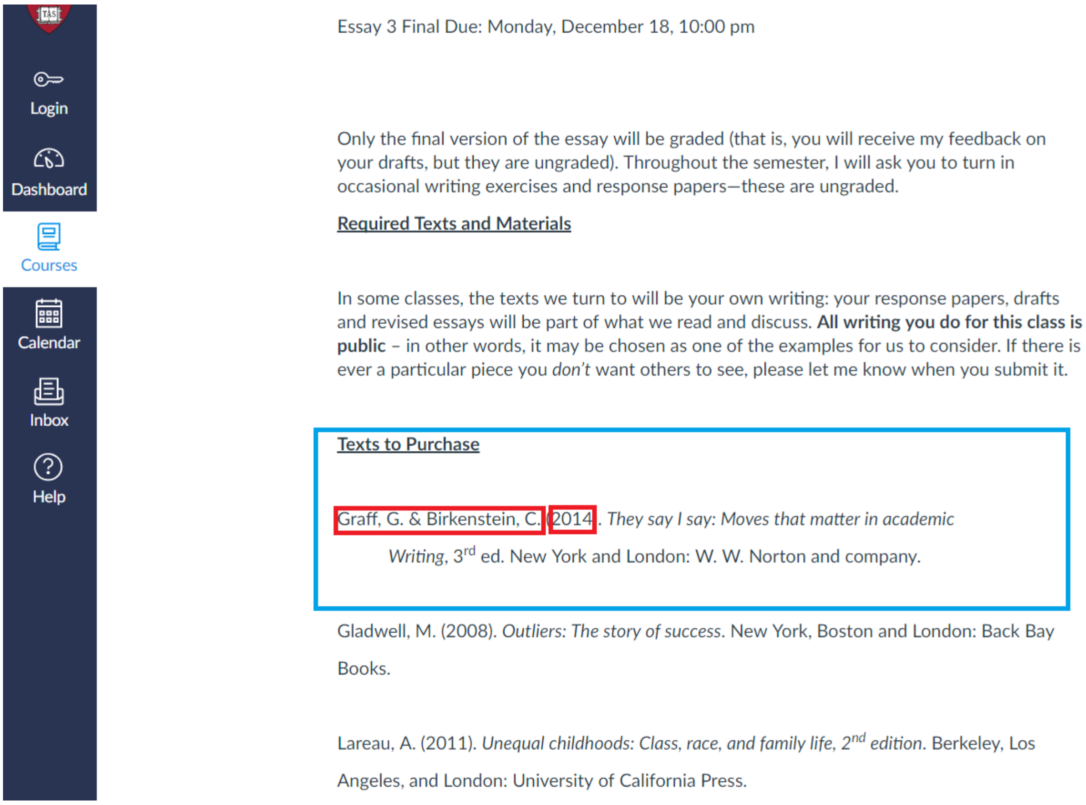
Publication Date: 10/31/2015

Title: New Solutions for House Museums: Ensuring the Long-Term Preservation of America's Historic Houses | Edition: 1

Author: Donna Ann Harris, Thomas A. Mason (Editor)

Publisher: AltaMira Press

Figure 1. Sample syllabus



Essay 3 Final Due: Monday, December 18, 10:00 pm

Only the final version of the essay will be graded (that is, you will receive my feedback on your drafts, but they are ungraded). Throughout the semester, I will ask you to turn in occasional writing exercises and response papers—these are ungraded.

Required Texts and Materials

In some classes, the texts we turn to will be your own writing: your response papers, drafts and revised essays will be part of what we read and discuss. **All writing you do for this class is public** – in other words, it may be chosen as one of the examples for us to consider. If there is ever a particular piece you *don't* want others to see, please let me know when you submit it.

Texts to Purchase

Graff, G. & Birkenstein, C. (2014). *They say I say: Moves that matter in academic Writing*, 3rd ed. New York and London: W. W. Norton and company.

Gladwell, M. (2008). *Outliers: The story of success*. New York, Boston and London: Back Bay Books.

Lareau, A. (2011). *Unequal childhoods: Class, race, and family life, 2nd edition*. Berkeley, Los Angeles, and London: University of California Press.

Figure 2. Sample syllabus

INF 558 Information Graph – Assignment 3

Task 2: Training the CRF

1. CRF packages: https://python-crfsuite.readthedocs.io/en/latest/
2. Label / Tag definitions:
    a. **F = First Name**
    b. **M = Middle Name**
    c. **L = Last Name**
    d. **A = Abbreviation = ‘.’**
    e. **D = Dash = ‘-‘**
    f. **Y = Year of publish**
    g. **N = Irrelevant**
3. Features:
    a. Word features:
        i. Lower case word.
        ii. Length of the word.
        iii. Is the word all upper case?
        iv. Is this a title word?
        v. Is this a digital number?
    b. Previous word and next word has the same features:
        i. Lower case word.
        ii. Length of the word.
        iii. Is the word all upper case?
        iv. Is this a title word?
        v. Is this a digital number?
        vi. Is the word a dot?
        vii. Is the word a dash?
        viii. Is the word a comma?
4. Programs:

- 4-1. Information Extraction and Tokenize:

```
Usage: python extract-textbook.py <input html files path>
</output/path> [delimiter for token] [irrelevant label]
      The program requires Python 3.6 to execute.
      - input_path = input file path of html files
      - output_path = Output file path of csv file
      - delimiter for token = Optional, if provide,
         Output sentences will convert to token + delimiter
      - irrelevant label = Optional, if provide, the label/tag will be
         appended after the delimiter, default=’N’
       Output format:
```

```
                  One line per sentence:
               Or, if delimiter='|' and label='N'
                  One|N
                  word|N
                  per|N
                  row|N
```

- 4-2. CRF training program

```
Usage: python crf.py <training data input file> <testing data input file>
<model file name>
      The program requires Python 3.6 to execute.
      - training data input file : Tokenized training data input file
        location.
      - testing data input file = Tokenized testing data input file
        location.
      - model file name = Trained CRF model file path and name
```

5. Dataset:

- Draft data:
  - ./source/html
    - raw html files
  - ./source/tokenized_draft_extract.csv
    - textbook information extract from html and tokenized by delimiter ='|' and irrelevant label='N'
  - ./source/ final_label_data.txt
    - Manual label and adjust data file.
- Training Dataset:
  - ./training/training-author.txt
    - 50 records for training set.
- Testing Dataset:
  - ./testing/testing-autohr.txt
    - 20 records for testing set.
- Trained model file:
  - author_year.crfsuite
    - Trained model to tag auther name and publish year.

Task 3: Report:

1. What was the information you were looking for? Describe the labels you chose and why. Also include at least one screenshot of the webpage you are using and show where is the information you are looking for.
   - As task 1 presented, author name and publish year of textbooks are the target information of this assignment. Figure 1, and Figure 2 are reference webpages which marked these targets.
2. What kind of tags did you tag your data with? Explain your choices.
   - Please refer task 2-2.
3. Report your classifier's precision, recall and F-1 measure. Why did your classifier perform well (or not satisfactorily)?
   - Precision, Recall, and F-1:

Tag	precision	recall	f1-score	support
A	1.00	0.86	0.92	7
F	0.62	0.77	0.69	13
L	0.60	0.36	0.45	25
M	0.60	1.00	0.75	3
Y	0.92	1.00	0.96	11
Avg. / Total	0.72	0.66	0.67	59

   - Tag Y and A are perform well.
     - Y is a four digits number, these features can be specifically identified.
     - A is a '.', but not every '.' is abbreviation. It is relatively hard to identified.
   - Tag M, L and F are not satisfied.
     - M are between L and F, but L and F are hard to identify.
     - L and F are title word, but corresponding location are always changing. Comma feature cannot help on this.
     - Some of books only marked single Last name as author name, which causes the recall of L is extremely lower than other tags.
   - Smaller C1 – coefficient for L1 penalty can help recall improve on tag L but hurt precision of tag F.

References

1. T. Peng, M. Korobov, “python-crfsuite — python-crfsuite 0.9.5 documentation", *Python-crfsuite.readthedocs.io*, 2017. [Online]. Available: https://python-crfsuite.readthedocs.io/en/latest/index.html.
2. M. Korobov, “scrapinghub/python-crfsuite", GitHub, 2017. [Online]. Available: https://github.com/scrapinghub/python-crfsuite/blob/master/examples/CoNLL%202002.ipynb.
3. N. Okazaki, "CRFsuite - A fast implementation of Conditional Random Fields (CRFs)", Chokkan.org, 2017. [Online]. Available: http://www.chokkan.org/software/crfsuite/.