Task 1

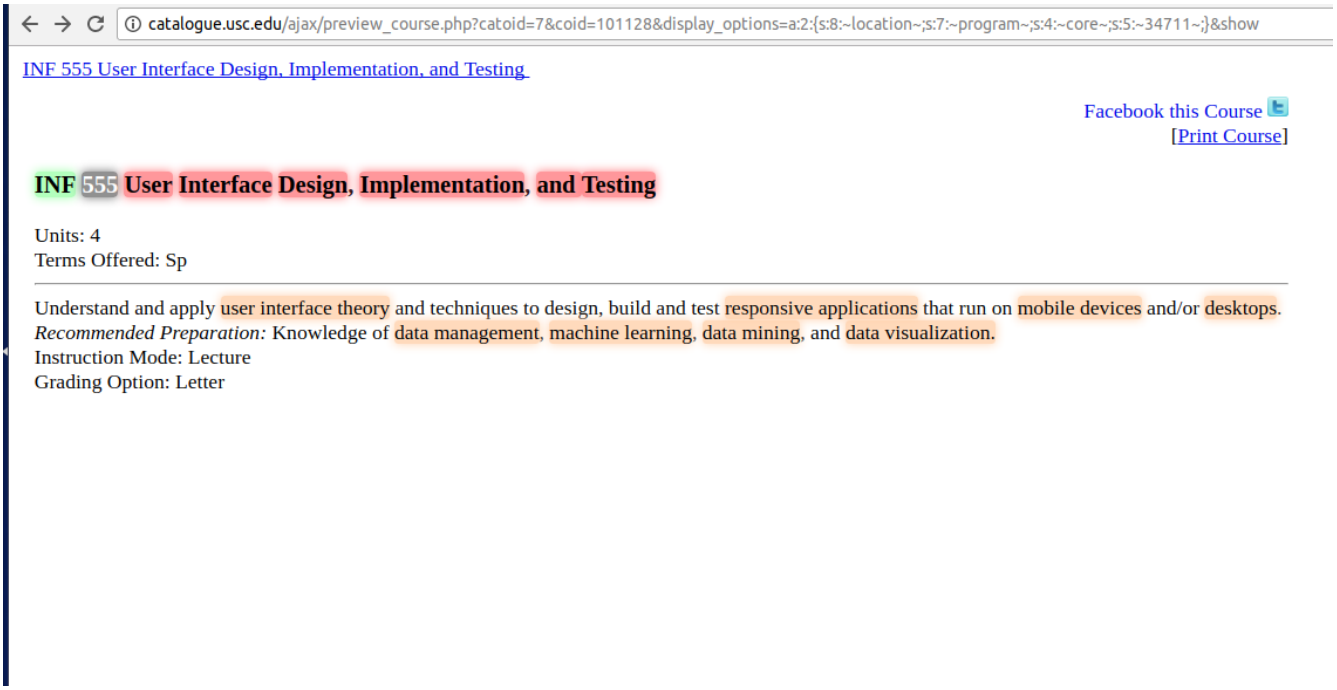
- Target website: Programs, Minors and Certificates on USC Catalogue

- Link: http://catalogue.usc.edu/content.php?catoid=7&navoid=2081

- Would first go into a program's description page, then extract and label text information of each course from the AJAX request page. (Example page shown below.)

Task 2

- Example page:



- I use **python-crfsuite** to train the CRF model.

- Tags defined for the target page:

- **D** for **Department**
- **C** for (course) **Code**
- **T** for **Topic**
- **S** for **Subject**
- **I** for **Irrelevant**

- 50 labeled records in **/training** folder / 20 labeled records in **/testing** folder

- Features:

- **word.lower()** # lower case of the word
- **word[-3:]** # last three characters of the word
- **word[-2:]** # last two characters of the word
- **word.isupper()** # if the word is in upper case
- **word.istitle()** # if the word is title-cased
- **word.isdigit()** # if the word is composed by numbers
- **postag** # POS tag created by nltk library

- Performance

	precision	recall	f1-score	support
D	0.96	1.00	0.98	67
C	1.00	0.99	0.99	67
T	0.90	0.93	0.91	160
S	0.89	0.82	0.85	343
I	0.92	0.94	0.93	742
avg / total	0.91	0.91	0.91	1381

Task 3

- I am interesting in finding important words of a course which describe course contents and help users understand more about what they may learn from it. The corresponding tag label is **Subject**. Also, I an interesting in linking similar courses and analyzing relationship between courses, so I apply **Department** for department code (such as INF), **Code** for course code (such as 558), and **Title** for words that appear in the course name. Example page shown above in page 1.

- I apply features regrading format (if upper case/ title-cased, suffix...etc) since it helps us reveal useful facts such as proper nouns, which have high potential as a **Subject** tag. Key words usually appear in certain position of a sentence, so POS tag is also applied to help analyzing the semantics.

- The classifier seems to work well. I think it is conceivable because the tags I choose mostly have significant characteristics. **Department** must be uppercased, about 2-4 characters word; **Code** is numerical and must follow **Department**; **Topic** must be title cased word following **Department** and **Code**. As for **Subject**, features on suffix (like -ion, -ent, -ing in the example page) and the POS tag may have helped a lot.