Current Topics in Digital Philology

Course Overview

Why This Course? (1)

- Lots of texts
- What do we do with them?
- Some excellent tools:
  - Voyant
  - WebLicht
  - RapidMiner

Why This Course? (2)

BUT what if the tools don't do what you want?

1) Abandon digital methods for this question
2) Change your research question to fit the tool
3) Request that a tool developer add a feature
4) CODE YOUR OWN!

In this course, we will adopt #4

What is this Course?

- “Philology” - We will start with texts (txt files)
- We will learn to clean them up and bring them into an analyzable format (Regex)
- We will learn to represent them as data (vectors and matrices)
- We will learn to manipulate and analyze this data (everything else!)

What will we do?

What will WE DO?

- WE, not I, will be the actors in this class
- And we will DO
  - Each week I will ask you to perform some computational task on some text or corpus
  - The primary grading criteria: DID YOU DO IT?
  - We will also think and reflect
  - But even our thinking and reflecting will be about what we have DONE, i.e., I will ask you to think, reflect, and write about your task from that week

What will YOU DO?

- You will come to class (4 hours/week * 15 weeks)
- You will do the weekly assignments (6 hours/week * 15 weeks)
- You will do a (preferably group) project (6 hours/week * 25 weeks)
- = 300 hours total

Grading

- 10% seminar attendance and participation
- 30% weekly problem sets
- 60% course project
- 100%

How will we do?

- We need a common language
  - Spoken and written: English
  - Coding: Python
- I will NOT grade your English/Python ability
- You must be able to:
  - get your point across in English
    - if I understand, you will pass
  - solve problems in Python

What will we cover?

- Installing Python
  - Pyzo distribution
- Github
  - Open-access and collaborative
  - Will host our course repository
  - You will all put your code and your other assignments there
- Python Crash Course!

Regular Expressions = RegEx

- The most important thing in this class
- Find and display textual patterns
- Textual transformations
  - E)N P E)N
    A)RXH=| N1 DSF A)RXH/
    E)POI/HSEN VAI AAI3S POIE/W

  - <w id="Gen.1.1.1" ana="P" lem="ἐν">ἐν</w>
    <w id="Gen.1.1.2" ana="N1 DSF" lem="ἀρχή">ἀρχῇ</w>
    <w id="Gen.1.1.3" ana="VAI AAI3S" lem="ποιέω">ἐποίησεν</w>

Web Scraping

- Where do we get our txts?
- Download? Great!
- Some texts are embedded in HTML on the web
- BeautifulSoup will help us do this

HTTP Requests and API Calls

- Many websites expose data and tools through Application Programming Interfaces (APIs)
- requests package to use the Perseus API to:
  - get texts
  - automatically lemmatize these texts

Text as Data: Vectors and Matrices

- What a vector is
- How they are useful for text analysis
- How to put them together into matrices
- How to manipulate these matrices to perform data analysis - Linear Algebra
- numpy

Machine Learning

- Allow our computers to learn patterns from our data for:
  - Classification: supervised methods
  - Clustering: unsupervised methods
- Machine Learning methods will be applicable to everything we do this semester
- scikit-learn

Semantic Information Extraction

- Most humanists want to get “meaning” from texts
- Distributional semantics to extract relational meaning
- Semantic drift between and among corpora

Parallel Text Alignment

- Translations exist for many texts we work with
- Explore methods to automatically align these texts
- Consider what we can do with these aligned texts

Dependency Trees the Dynamic Lexicon

- What is a dependency tree?
- What can we do with this information?
- One example: the Perseus Dynamic Lexicon

Visualizing Your Results

- Results in text or tabular form are often difficult to understand
- A picture is worth a thousand words
- Plotting results with matplotlib
- Visualization of networks with NetworkX

And finally...

You will present your projects!

Homework for Tuesday

- Short presentation (2 minutes) about your sources and the questions you would like to ask these sources
- 1 paragraph about this that we can put online

And now for something completely different...

- Let’s install Python
- http://www.pyzo.org/downloads.html

```
print(' '.join(['Hello', 'World!']))
```