Course Description:

Digitization efforts over the last two decades have made vast corpora available and in recent years humanities scholars have increasingly tried to use computational techniques to explore these archives. This course will provide an introduction to the use of computational natural language processing techniques for the humanities. Beginning with an introduction to the Python programming language, we will cover the core concepts required for working with text corpora from the basics of acquiring data, string manipulation, regular expressions, to the use of programming libraries for text clustering and classification.

Course Information:

Instructor: Anupam Basu (Department of English, Washington University in Saint Louis)

Time and Location: TF 11:10-12:35, Olson 207

Email: abasu22@wustl.edu

Office Hours: W 1:00-2:00 or by appointment

Recommended Text:

- [Natural Language Processing with Python](#)

We will use this book as a reference. Make sure you get the edition that uses Python 3.

Requirements and Evaluation:

- Class participation: 20%
- Class assignments: 20%
- Midterm assignment: 20%
- Final project: 40%

Any class notes will be uploaded on the course website in the form of Jupyter notebooks. Sample corpora to be used in class will be in the files section. All assignments also need to be turned in on as Jupyter notebooks.
Course Schedule:

Week 1: Python Language Basics
  - Introduction. Install Python and associated libraries
  - Python language components: variables, operators, expressions
  - Numeric data-types: booleans, integers, floats
  - Loops and conditionals
  - Sequence data-types: lists, tuples, and strings
  - Hashing data-types: dictionaries

Week 2: Working with Text.
  - More about strings: regular expressions and files
  - Functions, modules, and packages
  - Frequency distributions
  - Zipf's Law
  - Basic plotting in python
  - Type-token ratios

Week 3: Basic NLP
  - The Python Natural Language Toolkit Library (NLTK)
  - Tokenization
  - Stemming
  - Lemmatization
  - Significant words: Tf-idf scores.

Week 4: Machine Learning for NLP
  - Libraries for machine-learning: Scikit-Learn, Gensim, Numpy
  - Vectors and vector spaces.
  - Authorship attribution
  - Topic modeling
  - Feature sets: Frequency, tf-idf, topic models as features
  - Plotting high dimensional spaces with PCA
  - Clustering / Classification