LING 369: MINING LANGUAGE DATA WITH PYTHON

DAN SIMONSON
DES62@GEORGETOWN.EDU
OFFICE HOURS: AFTER CLASS, MIDNIGHT MUG

1. Course Description

During the last 25 years, natural language processing has advanced from using hand-curated "toy systems" to employing robust and sophisticated statistical techniques. While the improvements have been substantial, applying those improvements to real world data still requires linguistic intuition and analysis. This course will discuss techniques for handling, manipulating, and making discoveries from large and small quantities of unstructured (and often unruly) language data. Topics will include: data acquisition, text encoding and data formats, information extraction, domain adaptation, and evaluation measures. A large focus will be placed on implementation; students will build systems with these components, exploiting unstructured linguistic data for use and measuring their success quantitatively.

2. Objectives

By the end of the class...

- How do you get data?
- What are the best practices for storing / handling data?
- What is the structure of language data (and data in general) at the most basic levels?
- What angles of attack are there for dealing with language data? What's best to start with?
- What NLP tools are best? What's worst? In what conditions do these apply?
- Once NLP pre-processing has been applied, what do you do with it?

3. Classroom Policies

You're expected to be on time, and participate meaningfully in class. Bodily presence is not enough: please attend class both in body and in mind.

Cell phone use is not permitted. In the event of an emergency, please step outside the classroom.

Laptop use is discouraged when not absolutely necessary for instruction. If it is absolutely necessary based on your needs please let me know. If you must use a laptop in class, please keep the Wifi turned off to refrain from Internet use, unless necessary for instruction.

4. Assignment Submission

Failure to follow assignment submission policies may result in a penalty at the instructor's discretion.

Assignments must be on time. Failure to submit assignments on time will result in a penalty, at the instructors discretion. The penalty will vary between assignments.

Hand-written assignments should be legible; black pen is preferred, but pencil is fine. *Do not submit assignments in blue or purple ink*, as I grade in blue, and being colorblind, purple looks the same to me.

4.1. **Electronic Submission**. In this course, assignments must be submitted electronically. Electronic submissions must meet the following criteria:

- Submit one file.
- If it is a document, it must be a PDF.
- Document filenames should be named `LastNameFM_x.pdf`, where `LastName` is your last name, `F` is your first initial, `M` is your middle initial, and `x` is the assignment number.
- You *may not* submit photos of written work. You must find a scanner[1] and scan it. Photos will be rejected without further review.

5. Evaluation

Assignments in this class will each be assigned a point value. The total point value for all assignments will be

Table 1. Breakdown of Grade Constituents

Participation	10%
Workshop Participation	30%
Homeworks	30%
Final Project	30%

Table 2. Grade Scale for the Class, g is the student's final grade.

A	$92.5000 \leq g$
A-	$89.5000 \leq g \leq 92.4999$
B+	$87.5000 \leq g \leq 89.4999$
B	$82.5000 \leq g \leq 87.4999$
B-	$79.5000 \leq g \leq 82.4999$
C+	$77.5000 \leq g \leq 79.4999$
C	$69.5000 \leq g \leq 77.4999$
D	$59.5000 \leq g \leq 69.4999$
F	$g < 59.5$

5.1. **Participation**. (10 points total) You need to show up to class on time and make meaningful contributions, following all classroom policies.

5.2. **Workshop Participation**. (10 points per workshop) Students are prepared as required to present in workshop classes and make meaningful contributions to the discussions of others presentations.

5.3. **Homeworks**. (10 points per homework) Each homework may be composed of preliminary portions of the final project and exercises to help round out students' practical skills. For full credit, students complete homeworks accurately, completely, and on time.

5.4. **Final Project**. (30 points) Students will implement what they've learned in this class to complete a final project, a start-to-finish application where they take language data, extract meaningful features from that data using NLP techniques, and make some determination about that data.

[1]There's a scanner in the Gelardin Media Center.

6. Instructional Continuity

I bike 3.5 miles into campus. Whoever's responsible for canceling classes must live in the ICC or under a rock. Either way, there may be circumstances where it's too treacherous for me to make it into campus, in which case I'm required to have "instructional continuity" plans.

I'll likely set up some kind of Google Hangout for us to use and broadcast class from my kitchen, in the event of a cancellation.

7. Students with Disabilities

If you have a disability that will affect your performance in this class please contact the Academic Resource Center (arc@georgetown.edu). The Academic Resource Center is the office on campus responsible for ensuring compliance with the Americans with Disabilities Act (ADA). They will help work out a plan for you to thrive best in this class.

8. Honor Code

Students are expected to respect and adhere to the honor code.

Unless otherwise specified, students may discuss the content of homeworks and exercises, and this is encouraged. *However, they do need to provide a write-up in their own words.*

Students may not discuss the content of midterms and finals. Violations of this will be prosecuted to the greatest extent possible.

9. Course Outline

9.1. **Breakdown by Date**. See Table (3). The course as a whole is divided into a few overarching topics:

- Basics: Creating simple NLP-driven applications.
- Getting Data: More sophisticated techniques for getting data
- Using Data: More sophisticated techniques for using corpora and NLP data.
- Wrap-Up: Finishing up the class.

9.2. **Deadlines**.

- 2016-6-13: Homework 1 and Workshop 1
- 2016-6-22: Homework 2 and Workshop 2
- 2016-6-30: Workshop 3
- 2016-7-5: Homework 3
- 2016-7-7: Project Presentations
- 2016-7-7?: Projects Due

Table 3. Topics covered in course by day.

	Part 1: **Basics**
2016-6-6	Overview: The Typical Pipeline, For Fun and Profit
2016-6-7	Work Environment: Bash, Editors, and Python Fundamentals
2016-6-8	Deploying a Basic Pipeline
2016-6-9	Interpreting Output
2016-6-13	Workshop
	Part 2: **Getting Data**
2016-6-14	Data Formats, Loading Data, Writing Data, Unicode, Encoding
2016-6-15	Scraping Tweets
2016-6-16	Scraping Websites, etc., and Cleaning Data
2016-6-20	Saving Corpora with No Regrets
2016-6-21	Shell Scripting, CoreNLP, Using Remote Servers
2016-6-22	Workshop
	Part 3: **Using Data**
2016-6-23	Writing a Corpus Reader
2016-6-27	Getting Stuff from NLP Output
2016-6-28	Association Measures
2016-6-29	Visualizing the Language
2016-6-30	Workshop
2016-7-4	**Holiday**: No Class
	Part 4: **Wrap-Up**
2016-7-5	Basic Data Mining
2016-7-6	Conclusions
2016-7-7	Workshop: Project Presentations