Syllabus for DATA 531, Fall 2021,
Text Analytics with Information Retrieval
by Prof. Roger Bilisoly

Text (Free):
Python Tutorial, Release 3.7.0, by Guido van Rossum and the Python Development Team, 2018, https://bugs.python.org/file47781/Tutorial_EDIT.pdf. Chapters 3 through 7 cover similar material as learning module PY1. In addition, there are both lecture and homework PowerPoint slides for each learning module available as PDF files from the class web page.

Required Software (Free):
Python 3, IDLE (which comes with Python 3), and Anaconda (which has Spyder, Jupyter notebooks and PyCharm), all of which are free to download and use. Python modules are packages that, in general, need to be both installed and imported. Anaconda comes with many of these preinstalled, which is one reason why it’s convenient to use.

Prerequisites and Requirements:
The prerequisite for this course is DATA 511 or permission of the instructor. This course requires extensive use of the programming language Python, the string searching methodology of regular expressions (also called regexes, which are implanted in the Python module, ‘re’), and text analysis techniques. Programming takes time to learn, so doing well in this class will require (1) regular practice with Python to improve your programming skills, and (2) trying your hand at analyzing text documents.

If you plan to take both DATA 531 (offered each fall semester) and DATA 532 (offered each spring), these can be taken in either order. See “How do DATA 531 and 532 Work?” below for further information.

If You Want to do a Thesis in Text Analytics.
To start a thesis, you must take both DATA 531 and 532 before officially doing so. The specializations in the Data Science M.S. require three classes (total of 12 credit hours), but for text analytics, the third class is an elective. Finally, it is never too early to think about your thesis, so keep the following in mind.

- If you have not taken DATA 532 yet, then let’s talk sometime during weeks 8-10 and to plan to make your final project potentially useful towards a thesis.
- If you have taken DATA 532, then let’s talk early in the semester and use this class as a springboard for a thesis proposal and, hopefully, you’ll be able to finish a chapter by the end of this semester.

How do DATA 531 and 532 work?
Both DATA 531 and DATA 532 are based on working through learning modules, which are described below. Each module has two PDFs: lecture slides and homework exercises (which come with answers.) Together these prepare you for the quiz in that module. All the quizzes combined are worth 50% of your course grade. The final project, which is described below, is worth the other 50%.

The learning modules in DATA 531 and 532 overlap: both have PY1, PY2, PY3 (on Python) and RE1, RE2 (on regular expressions). Which modules you do this semester depends on (1) whether you have a background in programming, (2) whether you have experience with Python, or (3) whether you have taken DATA 532. For each of these cases, see the respective table below with suggestions on which modules to do when.
Suggestions for which learning modules you should do and when to do them.

- X means do that module in the time frame given.
- ? means attempt that module if you have time to do it.
- Weeks 11-15 are for working on your final project. As indicated by the question mark, you might consider working on a module and doing your final project, or just focus on the latter.

### New to Programming.

<table>
<thead>
<tr>
<th>Learning Module</th>
<th>Weeks 1-5</th>
<th>Weeks 6-10</th>
<th>Weeks 11-15</th>
</tr>
</thead>
<tbody>
<tr>
<td>PY1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PY2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE1</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>RE2</td>
<td></td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>PY3</td>
<td></td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>Advanced</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Have Programming Experience but New to Python.

<table>
<thead>
<tr>
<th>Learning Module</th>
<th>Weeks 1-5</th>
<th>Weeks 6-10</th>
<th>Weeks 11-15</th>
</tr>
</thead>
<tbody>
<tr>
<td>PY1</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PY2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE1</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>RE2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PY3</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced</td>
<td></td>
<td></td>
<td>?</td>
</tr>
</tbody>
</table>

### Have taken DATA 532 already.

<table>
<thead>
<tr>
<th>Learning Module</th>
<th>Weeks 1-5</th>
<th>Weeks 6-10</th>
<th>Weeks 11-15</th>
</tr>
</thead>
<tbody>
<tr>
<td>PY1</td>
<td>Review as needed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PY2</td>
<td>Review as needed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE1</td>
<td>Review as needed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RE2</td>
<td>Review or Finish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PY3</td>
<td>Review or Finish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thesis Preparation</td>
<td></td>
<td></td>
<td>Start writing a chapter for your thesis!</td>
</tr>
</tbody>
</table>

Lecture and homework slides for PY1, PY2, PY3, RE1, and RE2 are available on the class web page. As of the start of this semester (fall, 2021), SKL is still being written.
Core Learning Modules for DATA 531 and 532:

PY1 Learning Module, Introduction to Python Programming.
This module focuses on learning the basics of Python programming by writing short pieces of code. It covers data types, collections, arithmetic and logical operators, Boolean values, strings and string functions, branching and looping, how to import modules (such as ‘collections’), and reading files.

PY2 Learning Module, More Python Programming.
This module builds on PY1. It covers additional topics such as comprehensions, nested lists, text formatting, writing functions, reading and writing files. The goal is to start writing longer, more complicated programs and analyze longer texts such as Charles Dickens’ *A Christmas Carol* by looking at its words, lines, and paragraphs.

RE1 Learning Module, Introduction to Regular Expressions.
Regular expressions (regexes) are a string pattern matching language that is supported by many programming languages (including Python, of course.) This learning module covers basic syntax including topics such as defining character ranges, character classes, quantifiers, assertions, flags, capturing matches, greedy vs. non-greedy matches, re.compile(), re.sub(), re.findall(), as well as regex and match objects. These techniques will be applied to text analysis tasks.

RE2 Learning Module, Advanced Regular Expressions.
Building on the RE1 learning module, this looks at more sophisticated applications of regexes. Some addition topics are covered such as lookahead and lookbehind (both negative and positive forms). However, the emphasis is putting regexes to use by making a concordance for Charles Dickens’ *A Christmas Carol*, looking at sentence segmentation, and how to make your program download texts directly from the web.

PY3 Learning Module, Advanced Python Programming.
This module builds on PY1, PY2 and RE1. Its focus is on writing programs that read in a text, clean it, then analyze it. Some additional programming ideas such as functional programming are introduced. In addition, the Vector Space Model (VSM), cosine similarity, and Levenshtein edit distance are introduced to analyze text, which will be applied to the collection of William Shakespeare’s 154 sonnets.

Suggested order of doing the learning modules.
PY1, PY2, RE1, RE2, PY3 is the order I wrote these and is what I suggest. However, your background in Python and/or regular expressions may allow other sequences.
Advanced Learning Modules for DATA 531 and 532 (In Development):

SKL Learning Module, Scikit-learn.
Once texts have been transformed into a table of numeric values (as done in PY3), then this can be analyzed with Machine Learning (ML) techniques. This learning module focuses on the Scikit-Learn (sklearn) Python module, which has a large number of ML functions, https://scikit-learn.org/stable/. For example, tasks such as clustering texts (for instance, with k-means) or classifying new texts (for instance, with k-nearest neighbors) will be studied.

NLTK Learning Module, Natural Language Processing.
Natural Language Processing (NLP) is a collection of algorithms that analyze one or more text(s) to reveal the linguistic structure of these. This learning module focuses on the NLTK Python module, https://www.nltk.org/, which implements NLP. This has been touched on in RE2, which introduced sent_tokenize() from NLTK. This module explains how to use more sophisticated tools of NLP such as how to use a POS-tagger (POS = part of speech such as nouns, verbs, adjectives, etc.).

Possible Future Advanced Topic Modules:


The Final Project: First, Pick the Appropriate One!

For your final project you need (1) one or more text(s) to work with and (2) an initial idea of how to analyze them. To clarify, texts can mean books, letters, emails, tweets from Twitter, Twitch chat, Instagram posts, Discord posts, and so forth. Your goal is to try to discover something that a non-technical reader or user of these texts would find interesting.

Please post on the Discussion Board’s “Final Project Proposals” which text(s) you plan to analyze and an idea for text analysis. I’ll either approve or give suggestions. The deadlines for this are on the class web page.

You also need to pick which of the three Final Projects on the next page fits your situation. In brief, if you have not taken DATA 532, then do Final Project 1. If you have taken DATA 532 already, then do Final Project 2 unless you want to do a thesis in text analytics, then do Final Project 3. All three projects focus on analyzing a collection of texts, but if you’re thinking of doing a thesis, then your project should contribute to that goal. For example, it could end up being a chapter of your thesis.
• **Final Project 1:** This requires finding one or more text(s) from the web, which are to be analyzed with at least one text mining technique using Python. Search for texts that are interesting to you and think about how these might be analyzed. If you need ideas on what text mining tools you might use, please contact me and we can discuss it. Finally, post to the “Final Project Proposals” thread under “Discussion Threads” and include two pieces of information. First, post a link to your texts: each student must analyze different texts, so this is a chance to claim priority. Second, in a short paragraph, describe at least one text mining technique that you want to use along with what you hope to achieve with your analysis. Note that several students can use the same technique. Once you get my acceptance, you have until the last week of classes to finish writing it up.

• **Final Project 2:** This builds on Final Project 1 by adding the requirement that you combine some Natural Language Processing (DATA 532) techniques with what you learned in this class and apply this to your collection of texts. Otherwise, the instructions are the same as Final Project 1.

• **Final Project 3:** This builds on Final Projects 1 and 2. Now is a chance to try out an idea that could be a chapter of your thesis. So, if you’re reading this, let’s talk about your potential thesis topic. If you convince me that you have a good idea for a thesis, then let’s talk about creating a thesis proposal. That is, let’s use this semester to plan how you can graduate sooner with a M.S. in data science.

**Guidelines for the Final Project**

Once your proposal is approved, start it with an executive summary stating briefly what your project is about and explain your main results. Point out the most interesting conclusions from your analyses. In general, executive summaries give the reader an overview of your work. Make it between one and two pages long and don't use any graphics.

The main body of the report addresses (1) how you prepared your texts, (2) how you analyzed them, and (3) what results you obtained from your analyses. Make sure to address the following points.

• First, how did you parse the text to get individual words (or phrases, lines, sentences, paragraphs) using Python? What problems arose? What challenges did you face? Were all the texts written with the same conventions? Did you use regular expressions?
• Second, give a detailed description of your analysis and its results. How does it compare to what you thought might be discovered?
• Third, give your conclusions and discuss what you would like to do If you had more time. Think seriously about the latter since if you take DATA 532, you could do some of these ideas, or use these additional ideas for a thesis.
• Finally, submit all your code in one separate .py file. Make sure the code is commented so that the programming logic is clear. For example, how do the blocks of code work together? What files are inputted and outputted?
How the Final Project Will be Graded

Your final project must be double spaced and from 20 to 30 pages long (including graphics but excluding your complete Python code.) Email me (1) your final report as a PDF and (2) all Python code combined in a single .txt file via the class Blackboard Vista email system. You might be thinking: why not send your Python code as a .py file? Microsoft Outlook does not allow .py files to be sent.

Please name your files Last_Name_Final_Report.docx and Last_Name.txt because this helps me grade your projects. For example, I’ll put all the PDFs for this class in one folder, which requires each PDF to have a distinct name. Using your surname in the file name should make this happen.

Your final report will be graded using the following weights.

- 20% on the executive summary.
- 10% on organization. That is, do you present your results in a logical and easy-to-understand fashion?
- 10% on visual layout. That is, are the tables, figures, plots, etc., integrated into the text of the report in a logical fashion?
- 10% on readability. That is, is your report well written?
- 30% on how well you address the relevant issues to your problem.
- 20% on your conclusion.

Once everybody’s reports have been graded, I will email you your final project PDF with my remarks added using an Apple pencil in iPad.

Where to find texts to analyze?

Answer: on the web, of course. For example, most books published before 1926 are out-of-copyright in the U.S., and many of these have been transcribed. However, many other types of texts are available online such as articles and reports, and of course various types of social media. Here’s some ideas on where to start.

- Science articles at the Z-Library Project, https://booksc.org/
- This is new to me but looks promising, PDF Drive, https://www.pdfdrive.com/.

If you find a good source of texts, please tell the class by posting on the “Sources of text online” forum in the Discussion Board. And if you are struggling with picking a final project topic, let’s brainstorm together by setting up a meeting.
CCSU Policies

University’s Academic Integrity Policy: It is your responsibility to know the Student Code of Conduct (see [http://web.ccsu.edu/studentconduct/codeofconduct.asp](http://web.ccsu.edu/studentconduct/codeofconduct.asp)) and to read the Academic Integrity Website (see [http://www.ccsu.edu/academicintegrity/](http://www.ccsu.edu/academicintegrity/)). Some examples of misconduct are: plagiarism, having another person do your work for you; and posting offensive messages in a discussion thread. However, for a thorough discussion, use these two links.

Withdrawing from Class: The last day to easily withdraw (receive a grade of W) from any CCSU course is **November 17, 2021**. Please consult with your academic advisor prior to deciding to withdraw. Cessation of attendance, notice only to the instructor, or telephone calls to the Enrollment Center does not count as an official notice of your intention to drop the course. After the above date, withdrawals require approval of both me and the department chair.

Disability Statement: Please contact me privately to discuss your specific needs if you believe you need course accommodations based on the impact of a disability, medical condition, or if you have emergency medical information to share. I must have a copy of the accommodation letter from Student Disability Services in order to arrange your class accommodations. Contact Student Disability Services, if you are not already registered with them. They maintain confidential documentation of your disability and will assist you in coordinating reasonable accommodations for this class.

Statement on Discrimination and Harassment: Central Connecticut State University strives to maintain our campus as a place of work and study for faculty, staff, and students that is free of all forms of prohibited discrimination and harassment based upon age; ancestry, color; gender identity and expression; intellectual disability; learning disability; mental disorder; physical disability; marital status, national origin; race; religious creed; sex, (including pregnancy, transgender status, sexual harassment and sexual assault); sexual orientation; or any other status protected by federal or state laws. Any student who has concerns about this should contact the Office of Equity and Inclusion at 860-832-1652 or [https://www.ccsu.edu/diversity/](https://www.ccsu.edu/diversity/), Student Affairs at 860-832-1601, or his/her/their faculty member.

Sexual Misconduct, Intimate Partner Violence and Stalking: Central Connecticut State University (CCSU) will not tolerate sexual misconduct against students, staff, faculty, or visitors in any form, including but not limited to the following: sexual assault, sexual exploitation, sexual harassment or stalking, as defined in CCSU policies. For additional information, please consult the CCSU policies at [http://www.ccsu.edu/diversity/policies/index.html](http://www.ccsu.edu/diversity/policies/index.html). **To file a report,** contact: Equity and Inclusion (860-832-1652); Student Affairs (860-832-1601); Student Rights and Responsibilities (860-832-1667) or the University Police (860-832-2375).

For support and advocacy, contact: The Victim Advocate, 860-832-3795; Student Wellness Services, 860-832-1926, the Women’s Center, 860-832-1655, the local YWCA’s Sexual Assault Crisis Services Hotline, 860-223-1787 (confidential) and Prudence Crandall Center for Domestic Violence (confidential), 888-774-2900 (24-hour hotline). [https://web.ccsu.edu/studentaffairs/resources/files/HowtoHelpBooklet2019.pdf](https://web.ccsu.edu/studentaffairs/resources/files/HowtoHelpBooklet2019.pdf) has more information.