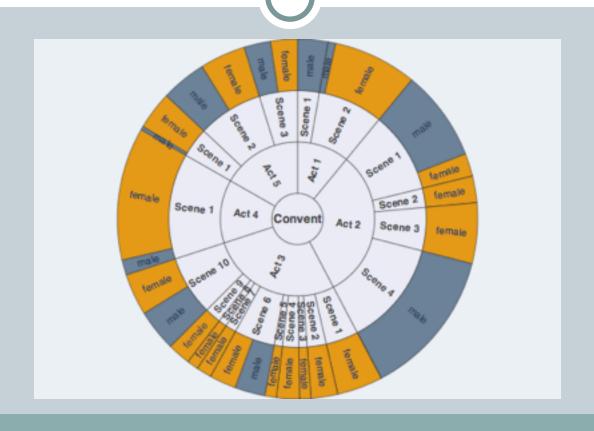
Workshop Overview



Goals

- This course will introduce students to cutting edge ways of structuring and analyzing digitized text-as-data, and will do so by exploring questions fundamental to the humanities.
- Topics Covered
 - Principles of Natural Language Processing
 - Introduction to Python for NLP
 - Discriminating Words
 - Dictionary Methods
 - Textual Classification
 - Topic Modeling
- Give time for you to apply what your learn to your own corpus.

Two Goals

- Intuition about what computational text analysis entails
 - What types of questions can you answer?
 - What types of evidence does it produce?
 - Range of techniques available
- Practical tools for implementing techniques
 - Understanding Python
 - Working scripts for specific techniques
 - How to learn on your own

Introduction: Text Formats

To analyze text, we need to move away from word processing and pdfs. What does this mean?

Text Formats

The way we read text: easy on the eyes



Text Formats

• The way computers read text: everything must be uniquely identified (e.g. html).

```
| Col. | Fig. 18 | Fig. | High | Col. | Col. | Improve Cases" | 1 | And the Cases | 1 
This has not been approximated to the second control of the second
                   (interpret, etg)_case, 'earlyce' setg_(12), (1), (1) 'comment(1) of conjust
violet engi-'exploration tools for 'early 'interpretable. Introduction 'fraction. Introduction. Interpretable. Interpretable. Interpretable. Introduction. Interpretable. Inte
                   ments id-"spointeriormen, name., spoints inspire, content, postent, ".
                   - dies rejugiterense, entingende in, entundier im, o. dies sejugiterense, von des sejugiterense, enting terminale men ent jerningster im, o. dies sejugiterense, enting terminale men ent jerningster im, o. dies sejugiterense, enting terminale men ent jerningster im, o. dies sejugiterense, enting terminale men entire men 
                      deci ammo "mercialin" document "The New York Time: Find Dressing news, multimetia, reviews a spinion on washington, business, sports, movies, travel, books, jobs, education, real estate, cars a more a myritem com." (>
morite heart "O' content" heart property ()
                   checks name "FF" partients "manager" //-
               -deriz dams "Respitization-text" contents "name descrip at the ord-delign Arguery mythes, coarrespondence particles and the coarrespondence pa
design property of the content of th
      ** The Control of 
                      rowth name"addrage" content:"homepage.sytimes.com/index.html" />
```

Text Formats

- This week we'll cover two text formats:
 - Raw text, or .txt
 - Delimiter-separated files, or .csv

Word (Processed) Text

A Highbrow Essay on Woman:

A Dissertation on the Economic Function of Woman with the part played therein by scientific bulletins and deep thinkers.

By Eugene Wood

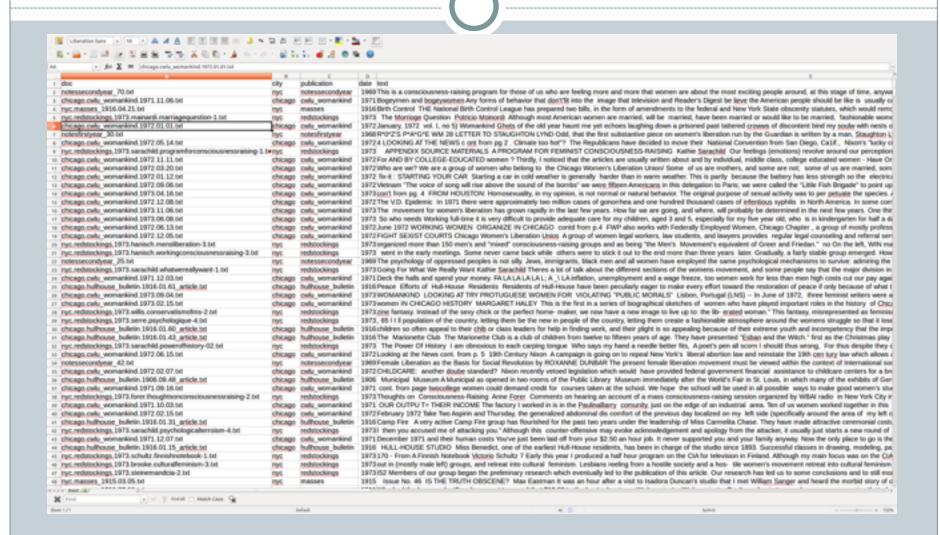
If there is any one thing in the reading line that I dote upon more than another, it is a bulletin, a real scientific bulletin, whether it be on the Stomach Contents of Arctomys Miurus or The Method of Procedure in Making Salt-rising Bread. Those fellows go at it so thoroughly. Right up to the handle. They don't have to worry whether the editor will like it or not. They don't care whether it will hit the public or not. If anything, they'd a little rather it didn't. It can't be very scientific if people read it and enjoy it. They aren't like literary folks, who when they take hold of a subject must not do more than pull out a few of the prettiest tail-feathers.

They pluck the subject as bare as a teacup. And then they take the hide off it. And then they cut it open and have a look at its insides, and dissect away every muscle from every bone, so that when they get all through, and washed up, that subject hasn't one secret left. They know it backwards and forwards, lengthwise and crosswise, up and down, and outside and inside.

Raw Text (.txt)

A Highbrow Essay on Woman:\n A Dissertation on the Economic Function of Woman with the part played therein by scientific bulletins and deep thinkers. \nIf there is any one thing in the reading line that I dote upon more than another, it is a bulletin, a real scientific bulletin, whether it be on the Stomach Contents of Arctomys Miurus or The Method of Procedure in Making Salt-rising Bread. Those fellows go at it so thoroughly. Right up to the handle. They don't have to worry whether the editor will like it or not. They don't care whether it will hit the public or not. If anything, they'd a little rather it didn't. It can't be very scientific if people read it and enjoy it. They aren't like literary folks, who when they take hold of a subject must not do more than pull out a few of the prettiest tail-feathers.\n\tThey pluck the subject as bare as a teacup. And then they take the hide off it. And then they cut it open and have a look at its insides, and dissect away every muscle from every bone, so that when they get all through, and washed up, that subject hasn't one secret left. They know it backwards and forwards, lengthwise and crosswise, up and down, and outside and inside.\nSo when I received a few days ago a Teachers' College Bulletin on "The Economic Function of Woman," by Edward T. Devine, PhD, Professor of Social Economy of Columbia University, I just knocked off work on that hurry job I had, part of the pay for which is going to reward the insurance company for my not dying this year, and settled myself to a really enjoyable intellectual sozzle. Here was something that nobody else could ever read clear through unless he was paid for it or had to read it in order to get a term-standing. And I'm interested in Woman. Most men are, if you'll notice. More or less. It is a subject that is brought to the male attention so often, so very often when you consider the whole period from the cradle to the grave. And then, again, this seemed a particularly promising viewpoint from which to consider Woman, what, if any account, is she?\nThere is not an extended piece of writing, however foolish it may seem, from which it is entirely impossible to get one good idea. And I will say for Dr. Devine that he sets forth some very sound and sensible things. I am sure of this because they're exactly what I think. When he says that students of the economic processes haven't paid as much attention to Consuming as they have to Producing, I think he's quite right. (I want the printer and the editor to let these capital letters stand as they are because I want to give the impression that I am a Deep Thinker. Nobody can be a Deep Thinker without capital letters sticking up through his copy like bristles on a cucumber. If I can't have any other symptoms of a Deep Thinker than Capital Letters, I must have them.)

Excel (Processed) File



.CSV

doc city publication date text word count org identifier wave notessecondyear_70.txt nyc notessecondyear 1969 "This is a consciousness-raising program for those of us who are feeling more and more that women are about the most exciting people around, at this stage of time, anyway, and that the seeds of a new and beautiful world' society lie buried in the consciousness of this very class which has been abused and oppressed since the beginning of human history. It is a program planned on the assumption that a mass liberation movement will develop as more and more women begin to perceive their situation correctly and that, therefore, our primary task right now is to awaken '.class'" consciousness in ourselves and others on a mass scale. The following outline is just one hunch of what a theory of mass consciousness-raising would look like in skeleton form. I. The "bitch session" cell group A. Ongoing consciousness expansion 1. Personal recognition and testimony a. Recalling and sharing our bitter experiences b. Expressing our feelings about our experiences both at the time they occurred and at present c. Expressing our feelings about ourselves, men, other women d. Evaluating our feelings 2. Personal testimony - methods of group practice a. Going around the room with key questions on key topics b. Speaking our experience - at random c. Cross examination 3. Relating and generalizing individual testimony a. Finding the common root when different women have opposite feelings and experiences b. Examining the negative and positive aspects of each woman's feelings and her way of dealing with her situation as a woman B. Classic forms of resisting consciousness, or: How to avoid facing the awful truth 1. Anti-womanism 2. Glorification of the oppressor 3. Excusing the oppressor (and feeling sorry for him) 4. False identification with the oppressor and other socially privileged groups 5. Shunning identification with the oppressor and other socially privileged groups 5. with one's own oppressed group and other oppressed groups 6. d groups 6. Romantic fantasies, utopian thinking and other forms of confusing present reality Thinking one has power in the traditional role-can "get what one wants,"" has with what one wishes reality to be etc. 8. Belief that one has found an adequate personal solution or will be able to Self-cultivation, rugged individualism, seclusion, and other forms of go-it-alonism rs?? C. Recognizing the survival reasons for resisting consciousness D. power behind the throne, etc. find one without large social changes 10. Self-blame!! ""Starting to Stop"" - overcoming Ultra-militancy; and others?? repressions and delusions Daring to see, or: Taking off the rose-colored glasses a.. Reasons for repressing one's own consciousness 1) Fear of feeling the full weight of one's painful situation 2) Fear of feeling one's past wasted and meaningless (plus wanting others to go 3) Fear of despair for the future Analyzing which fears are valid and which invalid through the same obstacles) Examining the objective conditions in one's own past and in the lives of most women throughout History 2) Examining objective conditions for thepresent redstockings 1

Other formats (we won't cover)

html, json, xml, pickel

Warning!

When creating your own corpus:

Do not save text in .doc, .docx, .rtf, .xls, or pdf formats





Provided Corpuses (aka: Corpora)

• Literary:

- txtLab 450: a novel per year for 150 years in 3 languages
- Underwood/Sellers: 750 volumes of C19 poetry, canon/archive

• Music Reviews:

- From metacritic.com, 1990-2015
- First paragraph from professional reviews (versus user reviews)
- Random sample of 5000 reviews from ~171,000 reviews
- Includes metadata on release date, artist, album, reviewer, and genre
- Stored in .csv format
- If you have one, we hope you can try these lessons on your own corpus!

- Monday Afternoon: Jump Right In with an introduction to NLTK
 - Introduction to the way computers "read" text
 - Text tokens and counting words (bag of words)
 - Part-of-Speech tagging and counting tagged words
 - What you'll learn:
 - Basic things you can do with Natural Language Processing
 - Starting thinking in a computational/ distance reading framework
 - Illustration: Can you identify a novel from only the most frequent nouns and verbs?
 - Corpus: Two mystery novels

- Tuesday: Back to Basics Working with Texts in Python
 - Morning: Python basics lists, strings, and all the important things
 - Afternoon: The Pandas Dataframe
 - What you'll learn:
 - How to use python to do what you already do, but faster
 - ➤ Understand the Python version of Excel (called Pandas), and learn why it's 1000x better than Excel
 - Illustration: graphing the character space using Pandas
 - Corpus: Literary novels

- Wednesday: Document Term Matrix (DTM) and Word Scores
 - Morning: scikit-learn and word scores
 - What you'll learn:
 - ➤ What is a Document Term Matrix and why do we <3 it
 - ➤ How to identify important or interesting words, not just frequent words
 - ➤ Illustration: What words distinguishe reviews from different genres
 - Corpus: Music Reviews
 - Afternoon: Sandbox!
 - × You do you

- Thursday: Thematic Frequencies, or the Frequency of Themes
 - Morning: Dictionary Method
 - Counting groups of words to measure presence of themes
 - Corpus: Music Reviews
 - Afternoon: Text Classification
 - Classify text into different categories to measure presence of themes
 - Corpus: Literary texts
 - × Sandbox!
 - You do you
 - What you'll learn:
 - ➤ Two different methods to use computers to "code" or "categorize" text into pre-determined themes. Can it replace human coding?

- Friday: Computational Inductive Analysis
 - Morning:
 - Topic Modeling (users, beware!)
 - Unsupervised machine learning
 - Corpus: Literary texts
 - × Wrap-Up
 - What you'll learn:
 - How to use computational methods to inductively identify themes in your text
 - A look back at how far you've come in a week
 - A look toward the future
 - Afternoon:
 - Lightening Talks (all workshops together)

Cheesy (but genuine) Thoughts

- You can learn to program!
 - Programming is a *creative* endeavor
 - There is no one right solution, no one way to code
 - Requires problems solving skills, artistic thinking
- You will get frustrated, you will hate your computer, and you will want to give up
 - Nothing will work the first time you try. Doesn't matter, try again
 - No matter what problem you have, 100 other people have had the same problem, 100 others have the answer (check StackOverflow, Google)
 - Look to each other (and StackOverflow) for help
- Once it works, it is beautiful

Cheesy (but genuine) Thoughts

- This Workshop only works if it is active, not passive
 - We know the scripts work on our computer. Get them working on your computer
 - Modify them, tie them up with a bow, and take them home with you
- This is a new and developing field, so now more than ever: **question everything and push boundaries**. You are the future of this field.

Our Goals

Theoretical

- General understanding of the past, present, future of text analysis
- Overview of computational text-analysis, how it fits into traditional, or close reading
- Intuition about programming/scripting and how computers read text

Technical

- Basic technical understanding of programming/scripting in Python
- Deeper understanding of some key techniques
- Knowledge of how to learn on your own
 - First reproduce, then modify
 - Use Stack Overflow

Practical

- How to translate these techniques to your own corpus/research
 - ➤ Go home with completed scripts you can modify/use in your own research
- Ideas and a plan of action for how to advance your own research, contribute to this field