College of William & Mary
Government 391 / Linguistics 464 / Data Science 440
Fall 2021

Data science for the social sciences & humanities: Text as Data

Prof. Maurits van der Veen

Class	**Office**
TuTh, 2-3:20 pm	355 Chancellor Hall
Location: 114 Chancellor Hall	maurits@wm.edu
Office hours: in person or on Zoom, by appointment	

Course description

Who wrote that anonymous memo? Can we track COVID infections by analyzing Google searches? Do Facebook status updates provide clues about whether U.S. regional dialects are increasingly blending together? Are wars foreshadowed by low-level conflict reported in online media? And how do conscious and unconscious biases show up in newspaper reporting?

Scholars have investigated all of these questions — and many more like them — using computational tools that identify patterns in language and text, taking as inputs the growing volumes of data that are gathered daily from our devices, computers, and smartphones. The combination of these tools and these data has become known as the "big data revolution", and it is transforming our understanding of the world in which we live.

This course covers text mining and language data analysis in an interdisciplinary manner accessible to non-computer science students in the humanities and social sciences. While a basic familiarity with python programming is a prerequisite, a much more important requirement is a lively curiosity about the answers and insights that can be extracted from big data.

Data science techniques are amazingly powerful at helping us find patterns in large quantities of texts, and their results can illuminate important questions in important and exciting new ways. However, the texts a society produces encapsulate prevailing mores, biases, and norms. As a result, text-as-data analyses are both uniquely positioned to identify such biases and, simultaneously, especially vulnerable to unwittingly reproducing such biases or producing biased outcomes. The likely presence of biases in both our source material and our analyses — and the importance of being aware of these pitfalls — will be an important thread throughout the course.

Course objectives

This course has three primary objectives:

1. Develop a hands-on understanding of how computational analyses of texts are used to create knowledge and insights in the humanities and social sciences.
2. Develop familiarity with use of the computational techniques for text mining and data analysis.
3. Foster data-oriented problem solving skills.

College Curriculum (COLL)

This course satisfies the COLL 200 requirement. The course is rooted in the Natural World and Quantitative Reasoning (NQR) knowledge domain, and reaches out to the Arts, Letters, and Values (ALV) and Cultures, Societies, and the Individual (CSI) domains.

Readings

No textbooks need to be purchased for the class. Most of the material will be from individual research articles, book chapters, and news articles. Readings are available online or through the electronic journals feature of the W&M library website. Required readings are listed on the course schedule.

However if you want to dig deeper in any given area, I recommend two textbook resources. The first is a very helpful introductory reference (with code) to many of the key building blocks of computational text analysis; moreover, it can be found online. The second is pitched at a higher level, and may be particularly helpful for those of you with a basic computer science background as you develop your course projects.

- *Natural Language Processing with Python* (also known as the NLTK book) by Bird, Klein, and Loper
- *Applied Text Analysis with Python* by Bengfort, Ojeda, and Bilbro

In addition, if you want to improve your python programming skills, I highly recommend *Learn Python 3 the Hard Way* by Zed Shaw.

COVID-19

A year and a half into the pandemic, there is still a lot of uncertainty about the safest and best ways to provide a good classroom experience. My goal is to do everything I can to contribute to everyone staying healthy and being able to participate fully in the class. In particular:

1. We currently have an indoor mask mandate. But if the College decides to lift this mandate over the course of the semester, I will not stop wearing a mask, and I request that all of you continue masking too. We cannot know everyone's health status and that of those close to them: not everyone can be vaccinated, for instance (as of now no children under 12 can be, among others), and the best we can do to protect everyone's health is to continue to wear masks indoors.

2. If you experience COVID-19 symptoms during the semester, you should make an appointment with the Student Health Center or a private healthcare provider for a clinical assessment and testing if necessary. If you test positive or are identified as close contacts, you must complete the form at *Report COVID* to initiate case management that will assist with isolation requirements and help you navigate classes and study. Moreover, I will work with you to make sure that you will miss as little as possible during your absence from class, and that you will not be penalized if you cannot do an assignment on time.

3. If I experience COVID-19 symptoms or am required to isolate because of close contact with someone who has COVID, or if a campus-wide outbreak forces a return to online course delivery for everyone, we will switch to Zoom, and I will try to minimize any other necessary changes to the course.

Evaluation

You are responsible for your own learning. My role is to engage you in learning more about data science for the social sciences and humanities, and to guide you in expanding your educational and intellectual interests. Your grade will be composed of:

● Participation	25%
● Labs / programming assignments	30% (5% each)
● Final Project	45%
(draft: 15%; presentation: 10%, final version: 25%)	

Grade scale

A	94-100	C+	77-79	D-	60-63
A-	90-93	C	73-76	F	0-59
B+	87-89	C-	70-72		
B	83-86	D+	67-69		
B-	80-82	D	63-66		

Labs / programming assignments
Labs and programming assignments are intended to give you hands-on experience with the principles and concepts discussed in class. Most weeks have a programming assignment. Usually I will introduce that week's topic as well as demonstrate the goal product (or something like it) on Tuesday, and make the assignment instructions available the same day. On Thursday we'll discuss more advanced work using the topic/tools of the week; you'll get much more out of class on Thursday if you have begun that week's programming assignment already.

Final project
You will design your own final project and present your project to the class. Projects must involve some programming component, and can be completed individually or in groups of two.

Course Policies

Attendance and class participation
Success in this course is predicated on regular class attendance. When you miss class, you miss an opportunity to learn something new and to gain a deeper understanding of the course material, as well as the chance to ask questions, learn from your peers, and show me where I may need to slow down, speed up, or retrace particular steps. Moreover, your risk missing out on important details, changes to the syllabus, and discussion of assignments.

If you are unable to attend class, whether due to COVID exposure or any other reason, please let me know as soon as possible, and we'll find the best way for you to catch up. Note: if you miss a class, do not ask me if you missed anything, because the answer will always be 'yes'. It is best to get notes from a classmate, review them, and after that ask me any questions about the material that might arise from doing so.

Readings
It is your responsibility to do the readings ***before*** class. Think critically about the points that are being made in the reading, as well as the methods used in the analysis; try to work out the problems discussed in the reading for yourself, and formulate your own thoughts in response to the readings. Come prepared to discuss.

Assignments
Assignments are to be submitted via Blackboard. You should complete them individually, but discussion with your classmates is allowed. If you discussed the work with another student, please mention that at the top of the assignment. Assignments are due at the start of class. Assignments that are late will lose 10 points (out of 100) for each day they are late. If your homework assignment is more than a week late, it will not be accepted, and you will receive a 0.

Project meetings
After you have chosen your project and begun working on it, I'll schedule individual meetings with each project group to discuss how best to proceed, and how to handle any obstacles you might encounter. These meetings will take place either on Zoom or in in person, outside (to minimize indoor meetings).

Office hours and email
I am available for both Zoom and live meetings by appointment. Please do get in touch if you have any questions regarding the reading material, the assignments, or the course in general. You are also welcome to email me if you have minor questions, but better still is to post a question on our class Slack, where others can see it too.

Accommodations
Any student who needs accommodation is requested to consult with Student Accessibility Services (SAS) as early as possible. I will follow the recommendations from SAS, and all discussions, issues, concerns, and accommodations will remain confidential.

Academic Integrity and the Honor Code
I adhere to the College's official policies regarding academic honesty. More information about the College's Honor Code can be found here. You are expected to be familiar with the College's policies on academic honesty. There will be **zero tolerance** for any cases of plagiarism, fabrication, cheating and facilitation of other forms of academic dishonesty.

Course technology

Communication (outside class)
I have set up a class Slack workspace, DS4SSH-Fall2021, where you can post questions and receive answers from me or from your fellow students. I will send an invitation link after the first class. To make sure everyone has successfully added themselves to the Slack workspace, I would like you to post a message in the #course channel with your current idea of what you think you might want to do for your course project. This can be very broad and I will not hold you to it, but it will help me check that everyone is on Slack, while at the same time giving me some information about your interests which will help me tailor my lectures.

Coding
Programming will be done using Python in Jupyter Notebooks. You have the option of using W&M's JupyterHub: https://jupyterhub.wm.edu or of downloading Anaconda's Jupyter distribution to your own computer (www.anaconda.com). Depending on how much demand there is on your wifi set-up, the latter may be preferable; however, either approach is fine.

Calendar

The calendar that follows is not final, and we may need to change or add readings and assignments here and there, especially if COVID intervenes and causes a change in the College's scheduling. Any such changes will be announced in class and on Blackboard.

I. Introduction to digital humanities / digital social science

Week 1: Introduction to the course

Thursday, September 2: **Introduction**

- Brief overview of the course
- Setting up: Jupyter, Slack

Week 2: Promise and pitfalls

Tuesday, September 7: **Promise**

- Blei, David M., and Padhraic Smyth. 2017. "Science and data science" *PNAS*, 114(33): 8689-8692. (http://www.pnas.org/content/pnas/114/33/8689.full.pdf)
- Dhar, Vasant. 2013. "Data science and prediction." *Communications of the ACM*, 56(12): 64-73. (http://jupiter.math.nctu.edu.tw/~yuhjye/assets/file/reading_list/data_science_and_prediction.pdf)

Thursday, September 9: **Pitfalls**

- Subbaraman, Nidhi. 2017. "Scientists taught a robot language. It immediately turned racist." *BuzzFeed*, April 13 (https://www.buzzfeednews.com/article/nidhisubbaraman/robot-racism-through-language)
- Gerber, Alison, Cristian Norocel Oy, and Francesca Bolla Tripodi. 2021. "Why academics shouldn't move fast and break things. *Medium*, April 30 (https://medium.com/@alison.gerber?p=8f14c9ef567d)
- Hsu, Jeremy. 2018. "The Strava heat map and the end of secrets." *Wired*, January 29 (https://www.wired.com/story/strava-heat-map-military-bases-fitness-trackers-privacy/)

II. Corpus analytics

Week 3: Basic corpus analytics using NLTK

Notes / assignments:

- Assignment 1: Corpus analytics using NLTK (due Sep. 21)
  (for help installing nltk, see: https://www.nltk.org/install.html)

Tuesday, September 14: **NLTK Introduction**

- Overview of the tools and functionality of NLTK

Thursday, September 16: **Corpus analytics using NLTK**

- In-class exercises using NLTK

Week 4: Collocations / word usage

Notes / assignments:

- Assignment 2: Compare word usage across two (sub-)corpora (due Sep. 28)

Tuesday, September 21: **Talking by & about gender**

- Jones, Jennifer J. 2016. "Talk 'like a man': The linguistic styles of Hillary Clinton, 1992-2013" *Perspectives on Politics* 14(3): 625-642.

Recommended

- Atir, Stav, and Melissa J. Ferguson. 2018. "How gender determines the way we talk about professionals." *PNAS* 115(28): 7278-7283.

Thursday, September 23: **Talking about fake news (& Islam & immigrants)**

- Li, Jianing, and Min-Hsin Su (2020) "Real talk about fake news: Identity language and disconnected networks of the US public's 'Fake News' discourse on Twitter." *Social Media + Society,* April 2020: 1-14.

Recommended

- Martin, Patrick, and Sean Phelan. 2002. "Representing Islam in the wake of September 11." *Prometheus* 20(3): 263-269.

- Blinder, Scott, and William L. Allen. 2016. “Constructing immigrants: Portrayals of migrant groups in British national newspapers, 2010–2012”. *International Migration Review*, 50(1): 3-40.

Week 5: Sentiment analysis

Notes / assignments:

- Assignment 3: Sentiment analysis of BLM newspaper coverage (due Oct. 5)
- Prepare for next week: download newspaper coverage using NexisUni (sign up for a particular set of months/papers)
- *No class on Thursday; discussion on Slack*

Tuesday, September 28: **Introduction to sentiment analysis**

- Yoo, Joseph, Jordon Brown, and Arnold Chung. 2018. “Collaborative Touchdown with #Kaepernick and #BLM: Sentiment Analysis of Tweets Expressing Colin Kaepernick’s Refusal to Stand during the National Anthem and Its Association with #BLM.” *Journal of Sports Media* 13(2): 39-60.

Thursday, September 30: **Beyond “just” sentiment: looking at emotions in texts.**

- Rodriques de Andrade, Francisca Marli, et al. 2021. “Twitter in Brazil: Discourses on China in times of coronavirus” *Social Sciences and Humanities Open,* 3(1).

Recommended

- Soroka, Stuart N., Lori Young, and Meital Balmas (2015). “Bad news or mad news”? *The ANNALS of the American Academy of Political and Social Science* 659: 108-121.

Week 6: Topic modeling

Notes / assignments:

- *Deadline to choose your project is this week!*
- Assignment 4: Identify topics in newspaper coverage of inequality (due Oct. 12)

Tuesday, October 5: **Introduction to topic models**

- Jockers, Matthew L., and Mimno, David. 2013. “Significant themes in 19th - century literature” *Poetics* 41(6): 750-769.

Recommended

- Van Galen, Quintus, and Bob Nicholson. 2018. “In search of America.” *Digital Journalism*, 6(9): 1165-1185.

Thursday, October 7: **Applications of topic models**

- Bagozzi, Benjamin E., and Berliner, Daniel. 2016. “The Politics of Scrutiny in Human Rights Monitoring: Evidence from Structural Topic Models of US State Department Human Rights Reports.” *Political Science Research & Methods* 6(4): 661-677.

Recommended

- Mueller, Hannes, and Christopher Rau. 2017. “Reading between the lines: Prediction of political violence using newspaper text.” *American Political Science Review*, 112(2): 358-375.

III. Building a corpus

Week 7: Getting your own data: web scraping

Notes / assignments:

- Assignment 5: Scrape comments on NYT newspaper article(s) & repeat a previous assignment (due Oct. 19)

Tuesday, October 12: **Introduction to web scraping**

Thursday, October 14: **Scholarly work using web-scraped data**

- Breeze, Ruth. 2021. “Claiming credibility in online comments: Popular debate surrounding the COVID-19 vaccine.” *Publications* 9:34.

Week 8: Getting your own data: what’s out there?

Notes / assignments:

- *Draft of project paper will be due Nov. 2! (all except actual analysis)*

Tuesday, October 19: *Fall Break (no class)*

Thursday, October 21: **Brainstorming session: finding a corpus for your project**

IV. Applied NLP

Week 9: Syntactic parsing / semantically annotated corpora

Notes / assignments:

- *Draft of project paper due Nov. 2!*

Tuesday, October 26: **Syntactic parsing and sentence analysis (NLTK)**

- Part of speech tagging and parsing methods using NLTK

Thursday, October 28: **Syntactic parsing continued: racist and abusive language**

- Clarke, Isobelle, and Jack Grieve. 2017. “Dimensions of abusive language detection on Twitter. *Proceedings of the First Workshop on Abusive Language Online.*

Week 10: Authorship attribution

Notes / assignments:

- Assignment 6: Identify author of mystery text (due Nov. 9)

Tuesday, November 2: **Techniques of authorship attribution**

- Harol, Corinne, Brynn Lewis, and Subhash Lele. 2020. "Who wrote it? *The Woman of Colour* and adventures in stylometry. *Eighteenth-Century Fiction*, 32(2): 341-353.

Thursday, November 4: **Identifying the authors of Obama's speeches**

- Herz, Jonathan, and Abdelghani Bellaachia. 2014. "The authorship of audacity: Data mining and stylometric analysis of Barack Obama speeches." *Unpublished manuscript.*

Week 11: Machine translation / word vectorization

Notes / assignments:

- *Project presentations begin Nov. 18!*

Tuesday, November 9: **Introduction to machine translation**

- Lewis-Kraus, Gideon. 2016. "The Great AI Awakening." *New York Times*, 14 Dec. https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html

Thursday, November 11: **Word vectorization as a solution to machine translation**

- Prates, Marcelo, Pedro Avelar, and Luis C. Lamb. 2018. "Assessing Gender Bias in Machine Translation: A Case Study with Google Translate." *Neural Computing and Applications* (doi: 10.1007/s00521-019-04144-6)

V. Presentations

Tuesday, November 16: Guest presentations by W&M faculty

Thursday, November 18: Student presentations
Tuesday, November 23: Student presentations

Thursday, November 25: *Thanksgiving Break (no class)*

Tuesday, November 30: Student presentations
Thursday, December 2: Student presentations
Tuesday, December 7: Student presentations
Thursday, December 9: Student presentations

Final project is due by midnight on Dec. 14 (in lieu of the final exam scheduled for that date)