NYU | Center for Data Science

This syllabus is subject to updates!
Last updated February 18, 2025

DS-GA 1015: Text as Data

Spring 2025

Lecture: Mondays 10:00–11:40 AM, CDS (60 5th Ave) Room 150
Lab: Thursdays 12:30–1:20 PM, Silver (31 Washington Pl) Room 401

1) The Instructional Team:

Instructor:	Byung-Doh Oh	**Office Hours:**	Mondays 12:30–1:30 PM
Email:	oh.b@nyu.edu	**Room:**	CDS 607
TA:	Dias Akhmetbekov	**Office Hours:**	Thursdays 11:00 AM–12:00 PM
Email:	da2669@nyu.edu	**Room:**	CDS 242
TA:	Sarthak Khandelwal	**Office Hours:**	Tuesdays 1:00–2:00 PM
Email:	sarthak.k@nyu.edu	**Room:**	CDS 242

2) Course-at-a-Glance:

1. **Reading:** While not burdensome, there will be some required instructional reading.
2. **Homework (48%):** There will be four homework assignments, for which you will apply the lecture material to conduct modeling and interpretation of text data.
3. **Final Project (44%):** There will be a final project, for which you will formulate original research questions and (try to) answer them using text data.
4. **Participation (8%):** There will be occasional "minute papers" asking you to reflect on what you have learned and provide a general vibe check of the course in 3-4 sentences.

3) Course Description: The availability of text data has exploded in recent times, and so has the demand for analysis of that data. This course introduces students to the quantitative analysis of text from a social science perspective, with a special focus on political science. The course is applied in nature, and while we will give some theoretical treatment of the topics at hand, the primary aim is to help students understand the types of questions we can ask with text, and how to go about answering them. With that in mind, we first explain how texts may be modeled as quantitative entities and discuss how they might be compared. We then move to both supervised and unsupervised techniques in some detail, before dealing with some 'special topics' that arise due

to recent advances in generative AI. Ultimately, the goal is to help student conduct their own text as data research projects, and this class provides the foundations on which more focused, technical research can be built.

While many of the techniques we discuss have their origins in computer science or statistics, this is not a CS class: we will spend relatively little time on traditional natural language processing issues (such as machine translation, optical character recognition, part-of-speech tagging). Other offerings in the university cover those matters more than adequately. Similarly, this class will not deal much with obtaining text data: again, there are excellent classes elsewhere dealing with e.g. web-scraping.

4) Course Objectives: The objectives of this course are to:

1. Become well-versed with various methods for representing and analyzing text
2. Gain hands-on skills for conducting text analysis and interpreting the results
3. Learn to critically evaluate the strengths and weaknesses of each method, and make informed decisions for particular use cases
4. Practice formulating research questions that can be answered using text data, designing analyses/experiments to answer them, and clearly communicating the results
5. Conduct honest academic work, learn to manage time wisely, treat colleagues professionally, maintain healthy work-life balance, ...

5) Mode of Instruction: This course consists of two meetings on a weekly basis:

1. A 100-minute lecture led by the instructor
2. A 50-minute lab session led by a TA

Enrolled students are expected to attend every lecture and lab session. At the moment, we have no plans to livestream or record anything. The instructional team will provide the information and skills that you need to complete your homework assignments and final project. If you will be missing class on a given day, you are still responsible for submitting homework on time via Brightspace, keeping up in the readings, studying the posted lessons, and beginning any new homework assignment. If you need any help with the course material, we highly encourage you to drop by our office hours or reach out to us by email (see above).

6) Assessment: The grades of all assignments will be converted from points to percentage, and then a weighted sum of the percentages will be calculated as described above. The final percentage will then be converted to letter grades according to the following rubric (The bracket '[' includes the endpoint and the parenthesis ')' excludes the endpoint).

7) Homework (48%): There will be four homework assignments, for which you will apply the lecture material to conduct modeling and interpretation of text data. There will also be a few questions designed to help you prepare for the final project in advance. All homework must be

A:	[93, 100]	**B:**	[83, 87)	**C:**	[73, 77)
A-:	[90, 93)	**B-:**	[80, 83)	**C-:**	[70, 73)
B+:	[87, 90)	**C+:**	[77, 80)		

submitted electronically as a single IPython notebook with both text answers and code. Intellectual honesty is important at NYU: you may confer with colleagues, but all work on the homework must be your own. If you copy work or allow another student to copy your work, the homework will be graded zero and your case will be passed to appropriate authorities in the university.

8) Final Project (44%): There will be a final project, for which you will formulate original research questions and (try to) answer them using text data. The project need not be confined to research questions about political science (i.e. they can be in any field, as long as they're communicated clearly) and can be technical in nature. You are expected to motivate the research question(s), describe the methods for collecting text data and analyzing it, and present the results and conclusions in a final paper of no more than 10 double-spaced pages. You are encouraged to work in teams of up to two people on this paper. The deadline for the paper will be May 9, 2025, with no extensions or exceptions.

9) Participation (8%): There will be occasional (approximately every two weeks) "minute papers" asking you to reflect on what you have learned and provide a general vibe check of the course in 3-4 sentences. These are meant to help you keep up with the lessons and help the instructor identify any parts of the lessons that are unclear. Each minute paper should include at least the following:

1. One question, either about something you didn't understand, or something you want to learn more about.
2. One important thing you learned that week.
3. Anything the instructor can do better to make the class more enjoyable.

In addition to these minute papers, any surveys administered by the instructor (such as the initial one about student backgrounds) will count toward the participation grade.

10) Late Homework Policy: All assignments must be turned in on time through Brightspace. They will not be accepted by email, which is unreliable. Each 24-hour interval after the deadline will incur a 10% deduction from the maximum obtainable score: a 'perfect' assignment will earn 90% of the total points if it is submitted 3 hours after the deadline, an 80% of the total points if it is submitted 27 hours after the deadline, and so on. If you are sick or have some other unavoidable problems that interfere with your ability to submit your homework on time, please let the instructor (and not the TAs) know before the deadline.

11) Academic Integrity Policy: All students are expected to do their own work. Students may discuss assignments with each other, as well as with the instructional team. Any discussion with others must be noted on the student's submitted assignment. Excessive collaboration (i.e. beyond discussing the assignment) will be considered a violation of academic integrity. Questions regarding

acceptable collaboration should be directed to the instructor prior to the collaboration. Needless to say, it is a violation of the academic integrity policy to copy or derive solutions from other students (or anyone at all), textbooks, previous instances of this course, or other courses covering the same topics. Finally, a good point to keep in mind is that you must be able to explain and/or re-derive anything that you submit. This is particularly important if you should adapt solutions from online sources. For the purposes of the final paper, these rules are intended to apply to each project group, rather than to each student individually.

12) Generative AI Policy: We live in the age of viable generative AI, and banning the use of these tools is neither realistic nor desirable. However, it is unwise to rely on them too much, as there are high-stakes situations where you won't have access to them, like technical interviews. In addition, working without AI assistance is a good way to develop/refine skills expected from a Data Science major. That said, you can use generative AI tools to help complete the assignments for this class, but if you use an AI tool to guide you in completing an assignment, you have to disclose which parts were generated by the AI tool (not doing so is a violation of academic integrity).

13) Academic Accommodations: Academic accommodations are available to assist student accessibility. Please contact the Moses Center for Student Accessibility (212-998-4980; mosescsa@nyu.edu) for further information. We recommend that students requesting academic accommodations reach out to the Moses Center as early as possible in the semester for assistance.

14) Course Website: Our primary course site is hosted on NYU Brightspace. On this site you will find the most recent version of the syllabus, as well as links to lecture slides and other course material. All your assignments should be submitted through the Brightspace site. We will also make important announcements via Brightspace.

15) Programming Language: The main programming language of choice will be Python, although we may have to rely on R if there are much better resources (this is why the initial course description mentioned both Python and R, but this is looking more and more unlikely). For any homework assignment, we will allow students to submit work done in another programming language, with prior permission from the instructional team. But we also think this course will be a great opportunity to gain familiarity with Python and advise students to take advantage of it.

The above guidelines are here to help students achieve the learning objectives, adhere to university policies, and to set a mutual agreement with the instructional team. I ask that you respect these guidelines and treat the instructional team with professionalism and honesty.

16) Course Materials: Due to the applied nature of the course, the lecture materials will come from a wide variety of sources, such as textbooks, academic papers, and other online resources. Therefore, somewhat unfortunately, most planned lectures don't 'break up' nicely into a handful of readings. I am still thinking about how to assign the readings for each lecture most effectively, hence the current TBAs. However, as the readings will be assigned below well in advance to each lecture, please check back for updates to the schedule below!

Readings:

[IIR] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*, Cambridge University Press. [LINK]

[SLP] Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. [LINK]

[DS18] Matthew J. Denny and Arthur Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2):168–189. [LINK]

[M+20] Reagan Mozer, Luke Miratrix, Aaron Russell Kaufman, and L. Jason Anastasopoulos. 2020. Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*, *28*(4):445–468. [LINK]

[B+19] Kenneth Benoit, Kevin Munger, and Arthur Spirling. 2019. Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science*, *63*(2):491–508. [LINK]

[S09] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3):538–556. [LINK]

[PL08] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2):1–135. [LINK]

[TP10] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1):24–54. [LINK]

[SP08] Jonathan B. Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3):705–722. [LINK]

[GK11] Justin Grimmer and Gary King. 2011. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, *108*(7):2643–2650. [LINK]

[B+03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*:993–1022. [LINK]

[BJ06] David M. Blei and Michael I. Jordan. 2006. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, *1*(1):121–143. [LINK]

[G10] Justin Grimmer. 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*:1–35. [LINK]

17) Course Schedule:

Week	Date	Lecture/Reading	Deadlines/Notes
1	1/27 (M)	Intro to course, background survey	
2	2/3 (M)	Representing text for computers – Preprocessing – Term frequency and weighting – Vector space model Reading: IIR 1.1, 2.2, 6.2-4 Optional: DS18	
3	2/10 (M)	Describing text properties 1 – Word distributions/collocations – Distance metrics – Identifying key words Reading: IIR 5.1, SLP 6.1-7 Optional: M+20	*HW1 (Week 1-3) out 2/14 (F)*
4	**2/18 (T)**	Describing text properties 2 – Lexical diversity – Complexity/readability – Style/author attribution Reading: B+19 (pp. 491-499), S09 (pp. 538-545)	*Legislative Monday* **Don't go to your Tuesday classes!**
5	2/24 (M)	Supervised methods 1 – Supervised learning basics – Evaluation workflow – Dictionary-based methods Optional: PL08 (pp. 1-27), TP10	**HW1 due 2/28 (F)** Byung-Doh away ('asynchronous learning')
6	3/3 (M)	Supervised methods 2 – 'Simple' classifiers – Support vector machines – Other kernel methods Reading: IIR 14.1-3, 15.1,3	*HW2 (Week 4-6) out 3/7 (F)*
7	3/10 (M)	Unsupervised methods 1 – Unsupervised learning basics – Dimensionality reduction – Clustering methods Reading: SP08, GK11	

8	3/17 (M)	Unsupervised methods 2 – Bayesian modeling basics – Latent Dirichlet Allocation – Topic modeling Reading: B+03, BJ06 Optional: G10	**HW2 due 3/21 (F)** *HW3 (Week 6-8) out 3/21 (F)* *...or some time during Spring Break*
9	3/24 (M)	Spring Break	
10	3/31 (M)	Very quick intro to deep learning – Deep learning basics – Sequence models Reading: TBA	
11	4/7 (M)	Quick intro to language models – Typical training setup – Text generation – RLHF Reading: TBA	**HW3 due 4/11 (F)** *HW4 (Week 10-11) out 4/11 (F)*
12	4/14 (M)	LLM-related themes 1 – Fact verification Reading: TBA	
13	4/21 (M)	LLM-related themes 2 – Bias/toxicity mitigation Reading: TBA	**HW4 due 4/25 (F)**
14	4/28 (M)	LLM-related themes 3 – Detecting machine-generated text Reading: TBA	
15	5/5 (M)	Course review	Or more LLM stuff if it rolls over **Final paper due 5/9 (F)**