INF2209H: Human-Centred Topic Models

**Syllabus**

**Instructor:** Shion Guha

**E-mail:** [shion.guha@utoronto.ca](mailto:shion.guha@utoronto.ca)

**Github Repo:** [https://github.com/shionguha/inf2209-humancentredtopicmodels-fa22](https://github.com/shionguha/inf2209-humancentredtopicmodels-fa21)

**Office hours:** Wednesdays 10:00am – 11:00am (BL514)

**Location:** Bisell 538

**Class Times:** Tuesdays 2:00pm – 5:00pm

### Official Course Description:

Topic models have emerged as a powerful methodological approach in human-centered data science since they lie at the intersection of empirical, positivist and interpretive techniques. Data scientists like topic models because they can provide sophisticated insights about large amounts of text data by incorporating human context and knowledge. This course will provide students with the computational ability to uncover latent topics from text documents and use qualitative, thematic analysis to make sense of them.

All concepts will be accompanied by examples and the students will have practical exercises and projects in which they will demonstrate their knowledge.

### Prerequisites:

• INF 1340H: Programming for Data Science

• INF 2190H: Data Analytics: Introduction, Methods and Practical Approaches

### Course Structure

Three (3) in-class hours per week will be divided into lectures and tutorials, in which we discuss and further probe topics covered in the lectures and readings. Note that for every one (1) hour of class time, students can expect to do 3 hours of reading and preparation work on their own, outside class.

**All coursework will be available on Quercus with detailed instructions and submission deadlines (date/time). There will also be an announcement section, which students should be responsible for checking regularly.**

### Lecture and Tutorial Topics:

* **Week 1 (Sep 13):** Introduction to the course; understanding topic models
* **Week 2 (Sep 20):** Wrangling and cleaning text data
* **Week 3 (Sep 27):** word embeddings and word2vec
* **Week 4 (Oct 4):** introduction to LDA (for dense text corpuses)
* **Week 6 (Oct 18):** interpreting and evaluation LDA style algorithms
* **Week 7 (Oct 25):** introduction to BTM (for sparse text corpuses)
* **Week 8 (Nov 1):** interpreting and evaluating BTM style algorithms
* **Week 9 (Nov 8): No Class; Fall reading week (Nov 7-11)**
* **Week 10 (Nov 15):** applying and evaluatingtopic models in practice
* **Week 11 (Nov 22):** introduction to thematic analysis for topic interpretation
* **Week 12 (Nov 29):** semi-supervised topic modeling and topic classification
* **Week 13 (Dec 6):** introduction to large language models
* **Week 14 (Dec 13):** conclusion to the course; next steps in HCDS; revision

### Learning Objectives:

### Upon successful completion of this course, students will be able to:

1. First, students will be able to understand the importance of computational text analysis in modern data science projects, teams and environments. *(as demonstrated in all of the assignments)*
2. Second, students will become familiar with the concept of topic models and the typical algorithmic approaches with which topic models can be used to computationally analyze text. *(as demonstrated through projects and tutorial activities)*
3. Third, students will be able to use human-centred, interpretive approaches to evaluate the results of topic models in practice to provide actionable, data-driven outcomes. *(as demonstrated in all of the assignments)*
4. Fourth, students will learn how to make statistical inference from topic models to provide quantitative evidence about specific research questions from large amounts of text data. *(as demonstrated in project and tutorial activities)*
5. Fifth, students will be able to combine both qualitative and quantitative topic inferential approaches to holistically interpret results, understand and reduce potential biases within topic modeling approaches and produce corresponding reports about their analyses. *(as demonstrated in all of the assignments)*

**Relationship to Master of Information (MI) Program-Level Student Learning Outcomes:**

Master of Information Program-Level Student Learning Outcomes can be found [here](https://ischool.utoronto.ca/areas-of-study/master-of-information/).

The students in this class will be exposed to the main steps and theoretical foundations of defining the appropriate steps for performing computational text analysis using topic models (**Outcome 1**). The practical examples of the course will combine theoretical foundations with practical approaches, such that the students can respond to the changing parameters as well as the size and variety of the given data (**Outcome 4**). By employing statistical metrics and interpretive evaluation criteria they will be able to provide robust qualitative and quantitative interpretations of topic models. At the same time they will learn and apply the principles of providing reproducible solutions (**Outcome 5**). Finally, the course will allow students to develop their own goals and continue in life-long intellectual growth beyond graduation (**Outcome 6**).

**Class Format**

The course will consist of lectures, class discussions, and tutorials. Students are expected to attend the classes and to actively participate in the discussions and tutorials. For each class, a series of topics are provided to guide students through the readings and activities, and to frame the lectures, discussions, and studios.

Teaching and learning is a shared responsibility, influenced by individual knowledge and experience, and achieved through expanding our awareness of the different issues and approaches involved in information architecture. Commitment, preparation, and active participation are important ingredients to realize this goal. Your preparation and participation are important to your learning and the learning of your colleagues.

All the course materials will be available on the University of Toronto learning management system (Quercus) together with assignments and announcements.

**Weekly Readings:**

|  |  |
| --- | --- |
| **Week** | **Assigned Reading** |
| 2 | Technical:   * [Bits to Characters](https://nbviewer.jupyter.org/url/mimno.infosci.cornell.edu/info3350/notebooks/bits_to_chars.ipynb)   Discussion:   * [Stephen Ramsay, Toward an Algorithmic Criticism](https://mimno.infosci.cornell.edu/info3350/readings/Ramsay-algorithmic-criticism.pdf) * [Stephen Marche, Literature is not Data: Against Digital Humanities](https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/) |
| 3 | Technical:   * [David Zentgraf, What Every Programmer...To Work With Text](http://kunststube.net/encoding/) * [RegexOne: Learn Regular Expressions](http://regexone.com/) * [Christopher Potts, Sentiment Symposium Tutorial: Tokenizing](http://sentiment.christopherpotts.net/tokenizing.html)   Discussion:   * [Ted Underwood, Where to Start with Text Mining](http://tedunderwood.com/2012/08/14/where-to-start-with-text-mining/) * [Thomas Dimson, Emojineering Part 1: Machine Learning for Emoji Trends](http://instagram-engineering.tumblr.com/post/117889701472/emojineering-part-1-machine-learning-for-emoji) |
| 4 | Technical:   * Implement, and document [Monroe et al., *Fightin' Words*](http://languagelog.ldc.upenn.edu/myl/Monroe.pdf). [Stats vs. Data Science Stack Exchange data](https://mimno.infosci.cornell.edu/info6150/stats_vs_datascience.zip).   Discussion:   * Landauer et al., [An Introduction to Latent Semantic Analysis](https://mimno.infosci.cornell.edu/info6150/readings/dp1.LSAintro.pdf). * Brown et al., [Class-Based n-gram Models of Natural Language](http://aclweb.org/anthology/J/J92/J92-4003.pdf). |
| 6 | Technical + Discussion:   * Rosen-Zvi et al., [The Author-Topic Model for Authors and Documents](https://mimno.infosci.cornell.edu/info6150/readings/398.pdf). * Wallach, Mimno, McCallum, [Rethinking LDA: Why Priors Matter](https://mimno.infosci.cornell.edu/papers/NIPS2009_0929.pdf). * Wallach, [Topic Modeling: Beyond Bag-of-Words](https://mimno.infosci.cornell.edu/info6150/readings/wallach06topic.pdf). |
| 7 | Technical:   * [The five most common similarity measures implemented in python](https://dataaspirant.com/2015/04/11/five-most-popular-similarity-measures-implementation-in-python/) (with cute animals)   Discussion:  [Barron et al., Individuals, Institutions, and Innovation in the Debates of the French Revolution](https://arxiv.org/pdf/1710.06867.pdf) |
| 8 | Technical:   * [Testing Burrows's Delta](https://mimno.infosci.cornell.edu/info3350/readings/delta.pdf) by David L. Hoover. Focus on the introduction.   Discussion:  [How Patrick Juola identified J.K. Rowling's pseudonym](https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/) (also this [more technical version](http://languagelog.ldc.upenn.edu/nll/?p=5315)). |
| 8 | Technical:   * Matthew Jockers, [The LDA Buffet](http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/) * [Boyd-Graber, Hu, and Mimno chapter 1](https://mimno.infosci.cornell.edu/papers/2017_fntir_tm_applications.pdf) (read for intuition, don't worry too much about distributions).   Discussion:  [Boyd-Graber, Hu, and Mimno chapters 4 and 6](https://mimno.infosci.cornell.edu/papers/2017_fntir_tm_applications.pdf) |
| 10 | Technical + Discussion:   * [Muller, M., Guha, S., Baumer, E. P., Mimno, D., & Shami, N. S. (2016, November). Machine learning and grounded theory method: convergence, divergence, and combination. In *Proceedings of the 19th international conference on supporting group work* (pp. 3-8).](https://dl.acm.org/doi/abs/10.1145/2957276.2957280) * [Baumer, E. P., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?. *Journal of the Association for Information Science and Technology*, *68*(6), 1397-1410.](https://epublications.marquette.edu/cgi/viewcontent.cgi?article=1562&context=mscs_fac) |
| 11 | Technical + Discussion:   * [Guest, G., MacQueen, K. M., & Namey, E. E. (2012). Introduction to applied thematic analysis.](http://antle.iat.sfu.ca/wp-content/uploads/Guest_2012_AppliedThematicAnlaysis_Ch1.pdf) *[Applied thematic analysis](http://antle.iat.sfu.ca/wp-content/uploads/Guest_2012_AppliedThematicAnlaysis_Ch1.pdf)*[,](http://antle.iat.sfu.ca/wp-content/uploads/Guest_2012_AppliedThematicAnlaysis_Ch1.pdf) *[3](http://antle.iat.sfu.ca/wp-content/uploads/Guest_2012_AppliedThematicAnlaysis_Ch1.pdf)*[(20), 1-21.](http://antle.iat.sfu.ca/wp-content/uploads/Guest_2012_AppliedThematicAnlaysis_Ch1.pdf) * [Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.](https://biotap.utk.edu/wp-content/uploads/2019/10/Using-thematic-analysis-in-psychology-1.pdf.pdf) |
| 12 | Technical + Discussion:   * [Blei, D. M., & McAuliffe, J. D. (2010). Supervised topic models. *arXiv preprint arXiv:1003.0783*.](https://arxiv.org/abs/1003.0783) * [Ramage, D., Manning, C. D., & Dumais, S. (2011, August). Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 457-465).](https://dl.acm.org/doi/pdf/10.1145/2020408.2020481?casa_token=4mX70yypFXsAAAAA%3AIO1pMvS-MKrfWx9zbADw2CxkI7Ff16enQiFtulaM25OEQBl6vNvrglvroAjpJ3Uw1T-Y_gamb_OP-w) |
| 13 | Technical + Discussion:   * Peters et al., [Deep contextualized word representations](https://allennlp.org/elmo) (page links to arXiv PDF) * Devlin et al., [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](https://arxiv.org/abs/1810.04805), with [code](https://github.com/google-research/bert) * [Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).](https://dl.acm.org/doi/pdf/10.1145/3442188.3445922) |

**Deliverables and Evaluation:**

|  |  |  |
| --- | --- | --- |
| **Evaluations** | **Due Date** | **Weight** |
| Class Performance and Discussions | ongoing | 10% |
| Reading Responses: Most weeks students will be asked to submit a reading response by Mondays 9 am EST | Tuesdays 12 pm EST | 20% |
| Mid Term Project Submission (Modeling) | November 5 | 35% |
| Final Project Submission (Evaluation and Interpretation with Thematic Analysis) | December 15 | 35% |

The course requirements and weights are final and will not be modified throughout the term. Late submissions will not be considered.

**Communication Policy:**

If you have a question, there is a high chance that other students in the course have the same question or, at least, will benefit from the answer. Please post all the questions to the INF2209 Quercus Discussion Board so everyone in the course can benefit from your questions and our answers. Students are encouraged to post answers to the questions of other students where appropriate.

Emails to the instructor and TAs must have a subject that starts with "INF2209" and include some more details, e.g., "INF2209: book appointment March 4th", **and must be submitted from your *mail.utoronto.ca* student account**.

**Readings:**

It is important to complete the required readings before the lecture in order to fully benefit from the class activities. We will be discussing them in class.

**Grading:**

Please consult the Faculty of Information’s:

* Grade Interpretation Guidelines: <http://ischool.utoronto.ca/wp-content/uploads/2020/08/grade_interpretation_revised_August2020.pdf>
* The University Assessment and Grading Practices Policy: <http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/grading.pdf> (note if nothing will be handed back before the final date to drop without penalty, that should be stated in the syllabus.)
* The Guidelines on the Use of INC, SDF, & WDR: <https://www.sgs.utoronto.ca/policies-guidelines/inc-sdf-wdr/>

These documents will form the basis for grading in the course.

**Late Policy:**

There is no late policy. Late submissions are not accepted and will not be considered.

**Writing Support:**

As stated in the iSchool’s Grade Interpretation Guidelines, "work that is not well written and grammatically correct will not generally be considered eligible for a grade in the A range, regardless of its quality in other respects". With this in mind, please make use of the writing support provided to graduate students by the SGS Graduate Centre for Academic Communication (<http://www.sgs.utoronto.ca/currentstudents/Pages/English-Language-and-Writing-Support.aspx>). The services are designed to target the needs of both native and non-native speakers and all programs are free. Please consult the current workshop schedule (<http://www.sgs.utoronto.ca/currentstudents/Pages/Current-Terms-Courses.aspx>) for more information.

**Academic Integrity:**

Please consult the University’s site on Academic Integrity ([http://academicintegrity.utoronto.ca](http://academicintegrity.utoronto.ca/)). The iSchool has a zero-tolerance policy on plagiarism as defined in section B.I.1.(d) of the University’s Code of Behaviour on Academic Matters

(<http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/ppjun011995.pdf>). You should acquaint yourself with the Code. Please review the material in Cite it Right and if you require further clarification, consult the site How Not to Plagiarize (<http://advice.writing.utoronto.ca/using-sources/how-not-to-plagiarize/>).

Cite it Right covers relevant parts of the UofT [Code of Behaviour on Academic Matters (1995)](http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/ppjun011995.pdf). It is expected that all iSchool students take the Cite it Right workshop and the online quiz. Completion of the online Cite it Right quiz should be made prior to the second week of classes. To review and complete the workshop, visit the orientation portion of the iSkills site: <https://inforum.library.utoronto.ca/workshops/orientation>

The essence of academic life revolves around respect not only for the ideas of others, but also their rights to those ideas and their promulgation. It is therefore essential that all of us engaged in the life of the mind take the utmost care that the ideas and expressions of ideas of other people always be appropriately handled, and, where necessary, cited. For writing assignments, when ideas or materials of others are used, they must be cited. APA format is suggested, however you may use any formal citation format you are familiar with, as long as it is used consistently in your paper, the source material can be located and the citation verified. What is most important is that the material be cited. In any situation, if you have a question, please post it to QUERCUS. Such attention to ideas and acknowledgment of their sources is central not only to academic life, but life in general.

**Accommodations:**

Students with diverse learning styles and needs are, of course, welcome in this course. If you have a disability or a health consideration that may require accommodations, please feel free to approach Student Services and/or the Accessibility Services Office (<http://www.studentlife.utoronto.ca/as>) as soon as possible. The Accessibility Services staff are available by appointment to assess needs, provide referrals and arrange appropriate accommodations. The sooner you let them know your needs, the quicker they can assist you in achieving your learning goals in this course.

**Participation and Attendance:**

Students Discussion and interaction in the classes are an important ways to learn. Sharing your experiences and ideas with your classmates is central to your learning experience in this course. As such, you should attend and participate in every class. There will also be exercises and discussions that you will participate in within your groups in your class. Some of the activities will be very helpful in completing your assignments.

**Regrading Policy:**

This is primarily a project-based course and as such, usual re-grading policies regarding assignment submission do not apply. Students and/or groups may reach out to the instructor and TA on an ad hoc basis to inquiry about why their course

Instructors and TAs should ensure all communications with the student is in writing (e.g. follow-up e-mail) and keep a copy for later reference.

**Academic Dates**: <https://ischool.utoronto.ca/current-students/academic-resources/academic-calendar/>

**Statement of Acknowledgement of Traditional Land:**

The following is the University approved land acknowledgment statement for official ceremonies (Ceremonial Committee, Governing Council):

See: <https://www.provost.utoronto.ca/wp-content/uploads/sites/155/2018/05/Final-Report-TRC.pdf>

“I (we) would like to acknowledge this land on which the University of Toronto operates. For thousands of years it has been the traditional land of the Huron-Wendat, the Seneca, and most recently, the Mississaugas of the Credit River. Today this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to have the opportunity to work on this land.”

See also, the Faculty of Information’s Commitment to the Findings and Call for Action of the Truth and Reconciliation Commission (approved at the Feb. 4, 2016 Faculty Council): <https://ischool.utoronto.ca/wp-content/uploads/2017/11/iSchools-TRC-Commitment.pdf>

**Equity, Diversity and Inclusion:**

The University of Toronto is committed to equity, human rights and respect for diversity. All members of the learning environment in this course should strive to create an atmosphere of mutual respect where all members of our community can express themselves, engage with each other, and respect one another’s differences. U of T does not condone discrimination or harassment against any persons or communities.

**Information about Faculty of Information iSkills and co-curricular Workshops:**

The following workshop series are exclusively available to the Faculty of Information community. Faculty of Information professors, Inforum librarians, current students, alumni, and a collective of professionals and academics from each program and concentration, work together to create these unique rosters.

Together with the MMSt and MI curricula, these academic, professional, and technical iSkills workshops provide a robust information and heritage graduate educational experience.

**iSkills Workshops**: <https://inforum.library.utoronto.ca/workshops/iSkills>

In an effort to ensure your success at the Faculty of Information, key information and skills that all Faculty of Information students must possess, regardless of program or concentration, are covered in these online orientation workshops.

**Orientation Workshops**: <https://inforum.library.utoronto.ca/workshops/orientation>