TEXT MINING FOR ECONOMICS AND FINANCE

Faculty: Michael Yeomans (m.yeomans@imperial.ac.uk)

Teaching Assistants:
Burint Bevis (**b.bevis20@imperial.ac.uk**) & Yaoxi Shi (**yaoxi.shi@imperial.ac.uk**)

Office Hours: By appointment (ideally right before/after a lecture)

INTRODUCTION

This course focuses on methods for quantitatively analysing text data, such as newspaper articles, social media posts, political speeches, and company product descriptions. The amount and availability of such data is growing rapidly, and extracting valuable information from it is an important challenge. In recent years, numerous machine learning methods have been developed for text. This course will introduce students to these methods, but of equal importance will be to discuss their application to problems in economics and finance. The course assignments and materials will primarily be conducted using the R software language, and extensions in Python will also be discussed. The course will begin by introducing different ways of representing text as data, as well as tools to evaluate the validity and generalisability of different models. The course will then turn to machine learning methods for studying text. We will discuss both unsupervised learning, where the goal is to uncover hidden structure in text, as well as supervised learning, where the goal is to make predictions based on text.

Some statistical and computational background material necessary for understanding these models. Strengths and weaknesses of different modeling approaches will be discussed in the context of specific applications. Students will work on programming problems that implement different methods for quantifying text as part of assignments. The goal is for students to learn a toolkit that they can apply in their future work, while also developing a mature understanding of the procedures they apply.

COURSE OBJECTIVES

- Represent text as data in a variety of ways
- Use text to describe and analyze the content of vast corpora
- Use text to predict variables such as the political ideology of media, the level and uncertainty in economic conditions, and the preferences of policy makers
- Connect algorithms for text data to other forms of unstructured data
- Program basic text algorithms in R and apply them to example datasets
- Appreciate how the use of text in economics and finance may differ from that in computer science

CLASS DATES

January 10 - March 7, 2024
Wednesdays 5-6:30pm
Thursdays 1-2:30pm
No classes week of Feb 5; or Feb 29 (Feb 28th from 5-8pm)
Final Week - March 6, 5-8pm; March 7, 1-4pm

All classes are mandatory. The module methodology is highly participative and utilizes class discussion and group work. This policy is necessary to ensure your own and your classmates' learning. If you have an excusable reason to miss class (own illness or illness of dependent, religious observance, military service), you must submit the excuse and the appropriate documentation in writing to the professor as soon as possible. We recognize that technical problems happen in an online learning environment. For those of you online, we strongly advise you to login a few minutes before class to ensure that everything is working properly and to email us *before the beginning of class* if it is not.

READINGS

Students are required to complete one reading each week. Without doing so, you will be ill-prepared to participate fully in the class discussion and exercises, which will hinder your own learning, and the learning of your classmates. I have tried to choose short articles. There is an optional reading in most weeks as well, that we will discuss in class, and you can read afterwards if you are interested. There is no textbook for the course, however if you are interested in pursuing the topic in more detail I highly recommend the following:

Jurafsky, D., & Martin, J. H. (2017). Speech and language processing. Vol. 3. It can be found online at https://web.stanford.edu/~jurafsky/slp3/

COURSE OUTLINE

Week 1: Text Analysis for Humans Jan 10-11

Required:

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261-266.

Week 2: Counting Words Jan 17-18

Required:

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*, 395-419.

Week 3: Supervised Learning Jan 24-25

Required:

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, *117*(19), 10165-10171.

Supplementary Reading:

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry*, *77*(5), 534-540.

Week 4: Interpretability Jan 31 - Feb 1

Required:

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16(3),* 31-57

Supplementary Reading:

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

Week 5: Categories Feb 14-15

Required:

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Supplementary Reading:

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., ... & Ungar, L. H. (2021). Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, *26*(4), 398.

Week 6: Embeddings Feb 21-22

Required:

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31-36.

Supplementary Reading:

Arora, S., Liang, Y., & Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

Week 7: Sentence & Dialogue Structure Feb 28 (double class on Wed, no class on Thu)

Required:

Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., ... & Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, *114*(25), 6521-6526.

Supplementary Reading:

Yeomans, M., Boland, K., Collins, H., Abi-Esber, N., and Brooks, A.W. (2022). A User's Guide to Conversation Research.

Week 8: Group Presentations and Review March 6-7

CODING ACTIVITIES

This class puts special emphasis on the practical application of the tools you will learn about. Most of the in-class lecture content will be paired with coding activities that will be completed during class time. You will spend roughly an hour a week completing an in-class activity in a group. For each activity you will submit something, and it will be graded essentially as a completion credit. This should not be stressful! The activities are all about learning together, and if you make a sincere attempt to follow along, you will get full credit.

For the first few weeks, we will assign groups randomly, but as you establish groups for your final projects, you will then be able to complete class activities with your project group. There are two reasons for random assignment. First, it will give you the chance to meet more of your classmates, and to brainstorm ideas for your group projects. More importantly, this will ensure that groups are mixed between people who are R experts, versus those who are newer. In all group work, it is essential that the people who are better coders take the time to teach their skills to those still catching up.

Please be prepared to code! Everyone should have RStudio installed on their laptop before the first week. If you cannot bring a laptop to class, please let me know immediately. You will still be able to complete the activities in a group.

CODING LANGUAGE

In truth, any decent analyst will be able to at least modify code in several languages, even if they specialize in one language. However, any class like this can quickly descend into a debate over which programming language is most suitable for the task. These debates are usually unproductive, and often stoked by the least informed. Accordingly, this will not be a point of discussion at any time in this class.

This class will be primarily taught in R. I am an R expert, and frequently write production-ready code (e.g. R packages on CRAN). Thus, I can most fluidly teach you the basics of data manipulation, plotting, etc. in R. R is also a better language for this kind of work - the packages are well-vetted, well-documented, and written by academic experts. If you are newer to programming (which was true for many students last year) then following my instruction in R is almost surely your best approach.

While Python is also popular, it is more chaotic, with difficult installations, little documentation or version control, and other traditional computer science problems that have little relevance to the content of the class itself. There are also many different kinds of Python code editor software - though ironically, I've found that RStudio is also the best available software for Python, as well as R.

However, I acknowledge most date science teams in industry rely on a mix of both languages, and all of the methods I will teach in class work well enough in both languages. More importantly, some of you are already proficient in Python. I will allow groups to submit activities and group presentation code in Python, if they prefer. But be warned, as this is a risky strategy. I have had groups attempt this before and quickly get in over their heads! This is not a python class, it is an NLP class. So if you are not very familiar with the basics of python (including packages like pandas, scikit-learn, nltk, matplotlib) then you will learn much more about NLP if you are following along with R.

For each coding activity, I will briefly walk through R code in class before you work in your groups. We have also prepared a companion guide to the class in Python, mirroring every line of my assignment R code, though I will not teach this in the lecture proper. I will still give feedback on Python assignments, and my TA is a Python expert, and can also provide assistance. If there is persistent interest in using Python for this class, we will develop a more robust strategy. For reference, last year two final group presentations were conducted in Python, and no one used python for any of the weekly activities. I expect a similar mix this year.

ASSESSMENT

- Participation in class activities (20%)
- Group Project Presentation (30%)
- Final Exam (50%)

Class Participation (20%)

This includes completing all out-of-class polls and preparing readings for discussions in class. This also includes in-class attendance, punctuality, timely and conscientious completion of group exercises, and contributing thoughtfully to class discussions. We do not have time for every single person to participate in every single discussion, but you should prepare to contribute to all discussions, and anticipate that you should contribute at least once during the module. By doing this, and by completing all of the in-class activities on time, you will be sure to get a good grade.

Every week of the class will conclude with a group coding exercise. You must submit this to the course email address, textminingimperial@gmail.com to get credit. If you complete it as a group, you can submit a single response and your whole group will get credit for the work. This submission must have every person's name at the top of the page, and everyone in the group must be CC'ed in the email so that we can provide feedback to the whole group. Each assignment is due at the beginning of the next week, on Monday at 12pm. This will allow us to give feedback before the next class.

Completion of the assignment will give you a 7/10 on that assignment. Each week a small number of groups who go above and beyond will get a full 10/10 grade, but these are not common. The important thing is not the grade itself, but the feedback you get, and the opportunity to practice your skills.

Group Projects (30%)

During the final week of the class, you will get a chance to present a project to the rest of your classmates. You will complete this project in groups of 5-6. However, unlike the coding activities, you will choose who you work with, based on mutual interest in topics. You must decide your groups by February 16th! Please start these discussions early, during your other group activities and outside of class. We will discuss the group projects in more detail during the first few classes. You can choose to analyse a dataset from one of the examples we work on in our class activities, or another dataset from a list I will provide. You can choose your own dataset as well, but it must be suitable for the project - if you choose your own, you must get approval from me (by email) beforehand.

We will have a special exercise in class on February 15th to encourage group formation, but you should have an idea of what you want to work on before then. If you do not have a group by then, I will assign students to join other groups that have formed. Please try to include people with a mix of abilities - I may adjust groups by hand if I notice all of the most (or least) experienced students are joining together.

The projects will be evaluated in parts. First, there will be an 8-minute presentation in the last week of class, followed by up to 5 min of Q&A. The presentation should include a statement of the research question, a description of the empirical strategy, some analyses (including at least three figures and one table), and a discussion of next steps. Every person from the group must be present, and speak once during the presentation. Each group will also submit commented R or Python scripts that reproduce all the analyses. The presentation will make up half the grade (split equally into presentation and content), while the Q&A will be another quarter and the submitted scripts will be the final quarter.

Exam (50%)

The finance program requires that all class have at least 50% of the grade be determined by a closed-book exam. So I will prepare an exam that will test some basic concepts in the class. Again, if you come to class and do all the activities, you will be sure to pass the exam. During the last week we will discuss previous exams and conduct a review session, to cover any questions you have about the material.