TEXT ANALYSIS

Period: a.y. 2021/22 – 2 sem.

Class times: Tue and Wed, 16:50–18:20 TBD. Room: TBD

Instructor:

Prof. **DIRK HOVY**
Dept. of MARKETING. - Room 4-C1-16
dirk.hovy@unibocconi.it

Course description

Text is an abundant data source, capturing information that is of interest to management and marketing scholars. Every day, an estimated 656 million tweets, 23 billion emails, and 224 billion text messages are sent. This data comes in addition to billions of sentences of quarterly earnings reports, press releases, patents, newspaper articles, blog posts, reviews, and dozens of other genres. However, due to its unstructured nature, size, and sparsity, it can be difficult to incorporate this information into social science studies. Luckily, there exist a variety of tools that can address these issues, both to help us make sense of the abundance of information, and to learn more about latent attributes in the data. This course will give students an overview of the applied text analysis methods for social science scholars.
The goal of the course is to introduce participants into the research field of data science and text analysis. This involves the study of applied methods, models, and algorithms in natural language processing and machine learning research. Participants will familiarize themselves with programming in Python, and learn to be able to code as well as read and understand simple scripts.
The course is structured in two parts, where the first part focuses on the fundamentals of working in Python. The second part looks at working with text, including basic concepts and data handling, to uncover hidden structures to understand large amounts of text, and predictive modeling, i.e., the use of patterns in the data in order to predict characteristics of the text (e.g., sentiment) or the author (e.g., age and gender).

This advanced course is organized as an interactive seminar, rather than a traditional lecture course, to give students the tools and practical understanding of the matter. Therefore, substantial in-class participation is expected and required. This class will use a flipped-classroom format. *What does that mean?* Before the lessons, students will get access to a notebook that implements and explains the core concepts. Students are expected to *read*

and work through this notebook *before class* at their own pace, and write down any questions they have.
In class, I will answer those questions, and we will work together on an applied practice set to deepen and apply the new concepts.

Course Material

To prepare for each week, I recommend you make sure they understood the previous lesson's approach, and are comfortable with the coding. Please carefully re-read the materials and take note of important information, concepts, and ideas. All students are expected to bring their own laptop, have a working installation of Python 3.6+ (preferably through Anaconda), and knowledge of how to use Jupyter notebooks.
All relevant reading will be contained in the notebooks, with optional references for individual in-depth exploration of research articles.

Assessment Methods

Effective class participation includes preparation of the lessons and active attendance, and having a positive attitude toward learning. The class finishes with an individual project (due after 2 weeks) where students apply what they have learned practically.

Faculty Bio

Dirk Hovy is associate professor of computer science at Bocconi University in Milan, specializing on Natural Language Processing. Before that, he was a postdoc and then faculty in the Computer Science department at the University of Copenhagen. He received his PhD in Computer Science from the University of Southern California (USC), and a linguistics masters in Germany.
Dirk's research lies at the interaction of language, society, and machine learning, or what language can tell us about society, and what computers can tell us about language. He has authored over 70 articles on these topics, including 3 best paper awards. He recently received an ERC Starting Grant for a project on demographic factors and bias in NLP, and has headed grant projects on work with Twitter, the Swedish Riksbank, and the Eurostars project. His first textbook appeared 2020 at Cambridge University Press, with the second part due at the end of 2021.

Tentative schedule

DATE	HOUR	TOPIC
Tue, Feb 01	16:50–18:20	1: Python - Variables and Strings
Wed, Feb 02	16:50–18:20	2: Python - Lists and Sets
Wed, Feb 09	16:50–18:20	3: Python - Dictionaries
Tue, Feb 15	16:50–18:20	4: Python - If-statements
Wed, Feb 16	16:50–18:20	5: Python - For-loops
Tue, Mar 01	16:50–18:20	6: Python - Functions and Objects
Tue, Mar 08	16:50–18:20	7: pandas and data formats
Wed, Mar 09	16:50–18:20	8: Data Scraping
Tue, Mar 15	16:50–18:20	9: Text representations
Wed, Mar 16	16:50–18:20	10: RE and basic analysis
Tue, Mar 22	16:50–18:20	11: Topic Models
Wed, Mar 23	16:50–18:20	12: Text Classification

Suggested complementary readings

All necessary reading will be provided in the notebooks and as lecture notes. Additional reading:

- Hovy, Dirk. *Text Analysis in Python for Social Scientists: Discovery and Exploration*. Cambridge University Press, 2020.
- Hovy, Dirk. *Text Analysis in Python for Social Scientists: Prediction and Classification*. Cambridge University Press, 2021.
- Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing.* Prentice Hall, 2009.
- Manning, Christopher, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.