HKU BUSINESS SCHOOL
港大經管學院

Inspire · Empower · Lead

FINA4350 Text Analytics and Natural Language Processing in Finance and Fintech

Second Semester of 2022-23

GENERAL INFORMATION

Instructor: Dr. BUEHLMAIER, Matthias

Email: buehl-teaching [at] hku [dot] hk
Office: KK1106, 11/F, K.K. Leung Building
Phone: +852 2219 4177
Consultation times (tentative): Wednesdays from 4:00 p.m. to 7:00 p.m. To allow for better preparation, students should email the instructor a brief description of the consultation topics they wish to discuss at the latest on the day before the consultation. Students who cannot attend the consultation hours due to a time clash with other courses/tutorials may request a separate consultation time.

Tutor: TBD

Course website: Moodle via HKU Portal
Other important details:

1. This course uses open-source software that is freely available on the internet. Students taking this course are responsible for installing and setting up all software used in this course. Instructions on software installation will be provided during the course.
2. This course involves computer programming using the Python programming language.
3. No previous Python programming knowledge is required before this course. Nonetheless, if some students would like to "brush up" their Python programming skills before the course begins, they may contact the course instructor by email (buehl [at] hku [dot] hk) to receive a free e-booklet containing an introduction to Python programming. (This is entirely optional and students may take this course without having gone through the e-booklet before the course.)

COURSE DESCRIPTION

This course covers the main elements of natural language processing (NLP), text analytics, and text mining, providing students with a foundation in collecting, managing, and analyzing textual data with financial and economic applications in mind, such as FinTech. Examples of potential applications include understanding and responding to sentiment in financial newspapers and social media, using social media to improve performance in asset/investment management, due diligence, Fed watching, monitoring of company events, and detecting insider trading. Although students write their own computer programs in this course, they are not required to implement most algorithms from scratch. Instead, the focus of this course is on how to use existing state-of-the-art open-source software libraries and how to apply them in a financial context. This course consists of three parts. In the first part, we work with real-world textual data sets to obtain proficiency in collecting, importing, organizing, and cleaning textual data from sources related to finance and economics. Among others, we cover web scraping, textual corpora, text processing, tokenization, stemming, and stop word removal. In the second part we delve into a more detailed analysis of NLP, text analytics, and machine learning with a particular focus on FinTech. For instance, we examine bag-of-words, word weighting schemes, document classification, document clustering, sentiment analysis, and topic models. The third part consists of summarizing, displaying, and visualizing results obtained from NLP and text analytics for applications in finance and economics.

Pre-requisite(s): None
Co-requisite(s): None
Mutually exclusive: None

COURSE OBJECTIVES

1. Cultivate a deep and rich understanding of typical NLP and text analytics workflows in finance and economics.
2. Sharpen analytic competence and programming skills using real-world textual data sets from finance and economics.
3. Establish a firm grasp of common pitfalls in NLP and text analytics and how to avoid them.
4. Foster awareness of the capabilities and limitations of NLP and text analytics in FinTech

applications.

FACULTY LEARNING GOALS (FLGs)

FLG1: Acquisition and internalization of knowledge of the programme discipline
FLG2: Application and integration of knowledge
FLG3: Inculcating professionalism
FLG4: Developing global outlook
FLG5: Mastering communication skills
FLG6: Cultivating leadership

COURSE LEARNING OUTCOMES (CLOs)

Course Learning Outcomes	**Aligned Faculty Learning Goals (FLGs)**
CLO1: Acquire solid understanding of quantitative textual analysis with financial applications.	Goals 1-2
CLO2: Develop/create new financial applications of NLP and text analytics by fostering thought leadership and by using a collaborative approach to social learning, e.g. group work.	Goals 2, 3, 5, 6
CLO3: Demonstrate, display, and visualize the results and insights obtained from NLP and text analytics.	Goals 4-5

COURSE TEACHING AND LEARNING ACTIVITIES

Course Teaching and Learning Activities	**Expected Study Hours**	**Study Load (% of study)**
T&L1: Lectures or supervised individual or group work	39	28%
T&L2: Blogging incl. Illustrative code examples	30	22%
T&L3: Group project	30	22%
T&L4: Presentation(s)	8	6%
T&L5: Self-study	30	22%
Total	137	100%

Assessment Methods		**Weight**	**Aligned Course Learning Outcomes**
A1: Midterm		20%	CLOs 1-3
A2: Group project presentation (each student's presentation skills are evaluated individually, i.e. independently of his/her group)		15%	CLOs 1 & 3
A3: Group project: Project report, presentation slides, and code		25%	CLOs 3
A4: Blog post(s) including illustrative code		10%	CLOs 1-3
A5: Final		30%	CLOs 1-3
	Total	100%	

Coursework / Examination Ratio: 70 % / 30 %

STANDARDS FOR ASSESSMENT

Course Grade Descriptors

A+, A, A-	Exhibited high level of understanding of the course materials through excellent

	performance in class discussion, assignments, presentations and exams.
B+, B, B-	Exhibited reasonably high level of understanding of the course materials through good performance in class discussion, assignments, presentations and exams.
C+, C, C-	Exhibited fair level of understanding of the course materials.
D+, D	Evidence of basic familiarity with the subject.
F	Candidate has demonstrated a poor grasp of the subject with evidence of largely inaccurate understanding of principles, concepts and arguments presented within this course.

Assessment Rubrics for Each Assessment

In-Class Performance
A+, A, A-: Extremely well prepared for class discussion, very active in sharing views and attended almost all lectures and
tutorials.
B+, B, B-: Partially prepared for class discussion, quite active in sharing views and attended most of the lectures and tutorials.
C+, C, C-: Not well prepared for class discussion, limited active in sharing views and attended many of the lectures and tutorials.
D+, D: Not well prepared for class discussion, no sharing of views and attended some of the lectures and tutorials.
F: Poorly prepared for class discussion and no sharing of views and experience and rarely attended lectures and tutorials.

Group Project: Individual Presentation
A+, A, A-: Professional presentation style, comprehensive content coverage, well-articulated on critical issues, effective use of
management concepts, and quality interaction with audience.
B+, B, B-: Decent presentation style, appropriate content coverage, clear discussion of critical issues, moderately effective use of management concepts, and acceptable interaction with audience.
C+, C, C-: Mediocre presentation style, limited content coverage, marginally acceptable discussion of critical issues, infrequent use of management concepts, and limited interaction with audience.
D+, D: Weak presentation style, key content omitted, unclear focus on critical issues, very limited use of management concepts, and poor interaction with audience.
F: Unacceptable presentation style, questionable content coverage, omitting critical issues, zero use of management concepts, and no interaction with audience.

Group Project: Group Report
A+, A, A-: Student presented extremely well.
B+, B, B-: Student presented very well.
C+, C, C-: Student presented in a mediocre way.
D+, D: Student did not presented well.
F: Nonsensical or flawed presentation.

Midterm, Final, and Problem Sets
A+, A, A-: Idea development is insightful and sophisticated; Supporting evidence is convincing, accurate and detailed. Well
written with clear focus.
B+, B, B-: Idea development is clear and thoughtful; Supporting evidence is sufficient and accurate. Well written.
C+, C, C-: Idea development is simplistic and lacking in relevance; Supporting evidence insufficient but accurate. Somewhat well written.
D+, D: Idea development is superficial and ineffective; Supporting evidence is insufficient and inaccurate. Writing is unclear.
F: Idea development is absent; Supporting evidence is vague or missing. Poorly written.

COURSE CONTENT AND TENTATIVE TEACHING SCHEDULE

This timeline is tentative and subject to change. Depending on the progress of the group projects, we might move through these topics nonlinearly. Student presentations will take place throughout the course. More detailed instructions are given in the lecture notes.

- Primer on Python programming (interweaved throughout beginning of course, with the pace depending on student's prior programming knowledge)
- Overview of NLP and text analytics (finish in week 1)
- Primer on textual programming (finish in week 2)
- Textual data collection and organization (finish in week 3)
- Text processing and regular expressions (finish in week 4)
- Cleaning and preprocessing of textual data (finish in week 5)
- Text analytics (finish in week 7)
- Natural language processing (finish in week 9)
- Machine learning for NLP and text analytics (finish in week 10)
- Data visualization and presentations (finish in week 11)
- Presentations (week 12)

RECOMMENDED READINGS

Manning, Christopher D. and Schütze, Hinrich, Foundations of Statistical Natural Language Processing, MIT Press, 1999

MEANS/PROCESSES FOR STUDENT FEEDBACK ON COURSE

☒ Conducting mid-term survey in additional to SETL around the end of the semester

☐ Online response via Moodle site

☐ Others: _____ (please specify)

COURSE POLICY

Class Conduct
Students are required to attend all classes on time. If you miss a class, it is entirely your responsibility for what you have missed. In case you have to leave the class early, please inform the instructor beforehand and leave quietly. No use of mobile phone or chatting is allowed when the class is in session. Remember to turn off or mute the phone before each session. The instructor has the discretion to give penalty in case of class misconduct. Respect your instructors and your fellow students. Be considerate to others.

Academic Dishonesty
Plagiarism and misconduct cases will be permanently recorded in the Faculty of Business and Economics for future reference.

The University Regulations on academic dishonesty will be strictly enforced! Please check the University Statement on plagiarism on the web: http://www.hku.hk/plagiarism/

Academic dishonesty is behavior in which a deliberately fraudulent misrepresentation is employed in an attempt to gain undeserved intellectual credit, either for oneself or for another. It includes, but is not necessarily limited to, the following types of cases:

- Plagiarism: The representation of someone else's ideas as if they are one's own. Where the arguments, data, designs, etc., of someone else are being used in a paper, report, oral presentation, or similar academic project, this fact must be made explicitly clear by citing the appropriate references.
- Cheating on In-class Exams: The covert gathering of information from other students, the use of unauthorized notes, unauthorized aids, etc.
- Academic dishonesty is any act that misrepresents a person's own academic work or that compromises the academic work of another. It includes (but not limited to) cheating on assignments or examinations; plagiarizing, i.e., representing someone else's ideas as if they are one's own; sabotaging another's work.

If you are caught in an act of academic dishonesty or misconduct, you will receive an “F” grade for the subject. The relevant Board of Examiners may impose other penalties in relation to the seriousness of the offense.

Plagiarism and copying of copyright materials are serious offences and may lead to disciplinary actions. You should read the chapters on “Plagiarism” and “Copyright” in the Undergraduate/Postgraduate Handbook for details. You are strongly advised to read the booklet entitled “What is Plagiarism?” which was distributed to you upon your admission into the University, a copy of which can be found at www.hku.hk/plagiarism. A booklet entitled “Plagiarism and How to Avoid it” is also available from the Main Library.

To avoid intellectual property and copyright infringement, and/or violation of the Personal Data (Privacy) Ordinance, **DO NOT upload** HKU teaching-related materials including but not limited to course materials, marking schemes, examination papers, etc. to websites. If you have done so in the past, you are asked to take steps to take down relevant materials immediately.

ADDITIONAL COURSE INFORMATION

- Announcement, assignments, and lecture slides will be posted on the course website. Hard copy of lecture notes will not be provided.
- No late assignments will be accepted.
- Special examinations are not granted to students taking up summer internships. Please avoid starting your internships before the end of the examination period.