1. General Information

Course Subject	FINA
Course Number	4350
Course Title	Text Analytics and Natural Language Processing in Finance and Fintech
Academic Years	2024-2025
Grading Method	Letter

2. Instructors

Professor BUEHLMAIER, Matthias
Office: Room 1106 11/F K.K. Leung Building
Email: buehl-teaching@hku.hk
Office: 2219 4177
Subclasses: 2A

4. Course Description

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.

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 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.) |
| Free Elective | Yes |

5. 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.

6. Faculty Learning Goals

Goal 1: Acquisition and internalization of knowledge of the programme discipline

Goal 2: Application and integration of knowledge

Goal 3: Inculcating professionalism

Goal 4: Developing global outlook

Goal 5: Mastering communication skills

Goal 6: Cultivating leadership

7. Course Learning Outcomes

Course Teaching and Learning Activities	Aligned Faculty Learning Goals					
	1	2	3	4	5	6
CLO1. Acquire solid understanding of quantitative textual analysis with financial applications.	✔	✔				
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.		✔	✔		✔	✔
CLO3. Demonstrate, display, and visualize the results and insights obtained from NLP and text analytics.				✔	✔	

8. 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	28.5
T&L2. Blogging incl. Illustrative code examples	30	21.9
T&L3. Group project	30	21.9
T&L4. Presentation(s)	8	5.8
T&L5. Self-study	30	21.9
	Total: 137	Total: 100

9. Assessment Methods

Assessment	Description	Weight %	Aligned Course

9. Assessment Methods

Methods			Learning Outcomes
A1. Midterm		20%	1,2,3
A2. Group project presentation	Each student's presentation skills are evaluated individually, i.e. independently of his/her group.	15%	1,3
A3. Group project: Project report, presentation slides, and code		25%	3
A4. Blog post(s) including illustrative code		10%	1,2,3
A5. Final Exam		30%	1,2,3

Assessment Rubrics

A1. Midterm	
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.
A2. Group project 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.
A3. Group project: Project report, presentation	

Assessment Rubrics	
slides, and code	
A+,A,A-	The project is extremely well done, demonstrating exceptional research, analysis, organization, and presentation. The code is highly efficient, functional, and well-documented, reflecting a deep understanding of the subject matter.
B+,B,B-	The project is well done, with a good balance of research, organization, and presentation. The code is mostly functional and efficient, with minor improvements needed. The overall project meets expectations but lacks the depth and polish of an A-grade project.
C+,C,C-	The project is satisfactory, with adequate research and organization, but lacking in-depth analysis and presentation quality. The code may have some issues with efficiency or functionality, and documentation could be improved. The project partially meets expectations.
D+,D	The project is below average, with limited research, poor organization, and weak presentation. The code is poorly structured and may have significant functionality issues. The project does not meet expectations and requires significant improvement.
F	The project is poorly executed and fails to meet expectations. Research, organization, and presentation are inadequate or irrelevant, and the code is non-functional or poorly documented. The project demonstrates a lack of understanding and effort.
A4. Blog post(s) including illustrative code	
A+,A,A-	The blog post is exceptionally well written, demonstrating original and insightful content, clear organization, and engaging tone. Thorough research and accurate information are evident, and the post encourages reader interaction and effectively communicates with the audience.
B+,B,B-	The blog post is well written, with relevant and original content. The organization and tone are generally good, with minor improvements needed. Research and accuracy are mostly solid, and the post encourages some reader interaction and communicates well with the audience.
C+,C,C-	The blog post is satisfactory, with adequate content and organization, but lacking in originality or depth. The tone may not be as engaging, and there could be room for improvement in research and accuracy. The post may have limited reader interaction and communication.
D+,D	The blog post is below average, with limited content, poor organization, and weak tone. Research and accuracy may be lacking, and the post does not encourage reader interaction or communicate effectively with the audience.
F	The blog post is poorly executed and fails to meet expectations. Content is irrelevant or unoriginal, organization is incoherent, and the tone is unacceptable. Research and accuracy are insufficient, and the post does not engage or interact with the audience.
A5. Final Exam	
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.

10. 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.

11. Course Content and Tentative Teaching Schedule

Topic/ Session	Content
1	Introduction
2	Data sources and applications of NLP in finance
3	Data workflow
4	Python programming
5	Sentiment analysis
6	Machine learning for NLP
7	Text processing
8	Basic concepts of NLP and pre-processing
9	Representing text in numerical format
10	Web scraping
11	Document similarity
12	Dimensionality reduction in NLP

12. Required/Recommended Readings & Online Materials

Textbook	Albrecht, Ramachandran, and Winkler, *Blueprints for Text Analytics Using Python: Machine Learning-Based Solutions for Common Real World (NLP) Applications*, O'Reilly Media, 2020

13. 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

14. 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.

15. 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.