BE103/512

THE UNIVERSITY OF HONG KONG
FACULTY OF BUSINESS AND ECONOMICS

MFIN7036 Text Analytics and Natural Language Processing in Finance and FinTech

Module 3, Academic Year 2025/2026

GENERAL INFORMATION

Instructor: Prof. Matthias BUEHLMAIER, PhD
Email: buehl-teaching[at]hku[dot]hk
Office: KK1106
Phone: +852 2219 4177

Consultation times: Wednesdays from 4:00 p.m. to 7:00 p.m. (tentative). 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

Other important details: This course uses Python, which is open-source software, freely available online. Students are responsible for installing Python and all Python libraries used in this course by themselves.

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 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 fintech. 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 finance and 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 fintech.

COURSE OBJECTIVES

1. Cultivate a deep and rich understanding of typical NLP and text analytics workflows in finance and fintech.
2. Sharpen analytic competence and programming skills using real-world textual data sets from finance and fintech.
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 finance and fintech applications.

Programme Learning Outcomes

PLO1. Acquisition of the techniques and advanced knowledge in the interdisciplinary fields of financial technology including but not limited to financial services, technology and law

PLO2. Application and integration of interdisciplinary knowledge and skills to identify and tackle practical problems, and design innovative products and systems with international standards and global vision

PLO3. Inculcating leadership, professional ethics and competence in the interdisciplinary fields of financial technology

PLO4. Mastering communication skills

COURSE LEARNING OUTCOMES

Course Learning Outcomes	Aligned Programme Learning Outcomes
CLO1: Acquire a solid understanding of quantitative textual analysis with financial applications.	PLO1 & PLO2
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.	PLO2 & PLO3
CLO3: Demonstrate, display, and visualize the results and insights obtained from NLP and text analytics.	PLO3 & PLO4

COURSE TEACHING AND LEARNING ACTIVITIES

Course Teaching and Learning Activities	Expected contact hour	Study Load (% of study)
T&L1. Lectures or supervised individual or group work	30	22%
T&L2. Blogging incl. illustrative code examples	30	22%
T&L3. Group project	34	25%
T&L4. Presentations	12	9%
T&L5. Self-study	30	22%
Total	136	100%

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

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.

Presentations
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.

Midterm, Blog Posts, Problem Sets, and Individual/Group Reports
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. 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 2)
- Text processing and regular expressions (finish in week 2)
- Cleaning and preprocessing of textual data (finish in week 3)
- Text analytics (finish in week 3)
- Natural language processing (finish in week 4)
- Machine learning for NLP and text analytics (finish in week 5)
- Data visualization and presentation (finish in week 6)

RECOMMENDED READING

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

MEANS/PROCESSES FOR STUDENT FEEDBACK ON COURSE

- Conduct mid-term survey
- Conduct SFTL

COURSE POLICY

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.

Any student caught in an act of academic dishonesty or misconduct 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.

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.

According to the MFFinTech regulations, students are required to attend at least 70% of classes for this course; otherwise they may be treated as having failed the whole course. Coming late to class is discouraged. If students are 30 minutes late, they will be regarded as not having attended the class.

ADDITIONAL COURSE INFORMATION

1. 1. Plagiarism and copying of copyright materials are serious offences and may lead to disciplinary actions. Students should read the chapters on “Plagiarism” and “Copyright” in the Undergraduate/Postgraduate Handbook for details. Students are strongly advised to read the booklet entitled “What is Plagiarism?” which was distributed upon students’ 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.
2. 2. 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.
3. Announcement, assignments, and lecture slides will be posted on the course website. Hard copy of lecture notes will not be provided.
4. No late assignments will be accepted.
5. Special examinations are not granted to students taking up internships. Please avoid starting your internships before the end of the examination period.