The Hong Kong Polytechnic University

Subject Description Form

Subject Code	ENGL4026
Subject Title	Language and social data analytics
Credit Value	3
Level	4
Pre-requisite/ Co-requisite/ Exclusion	N.A.
Objectives	Digital data such as social media, news archives, and various kinds of sentiments, trends, and demographics pervade our world. Such data play an increasingly important role in the media, marketing, and education industries in today's digital economy. Analysed and presented appropriately, they can provide insights into human behaviour and optimise professional decision-making. This subject provides hands-on training on how to derive and communicate insights from such data. Students will learn how to write simple programming codes to annotate, analyse, and model textual data. Ethical issues arising from the rapid growth of the digital economy will also be critically discussed. Some background in programming, data analytics, or statistics will be an advantage. The subject meets the following objectives: 1. Introduce basic data analytic techniques and their implementation in Python, R or equivalent; 2. Train students to collect, annotate, analyse, interpret, and communicate data-driven insights for problem-solving and decision-making; 3. Raise awareness of ethical challenges faced by today's digital economy. A flipped classroom approach will be adopted with balanced assessment tasks. Classroom and independent learning will be supported by open-source software and other online resources.
Intended Learning Outcomes *(Note 1)*	Upon completion of the subject, students will be able to: **Category A**: Professional/academic knowledge and skills a. Understand basic data analytic techniques; b. Collect and analyse data in linguistic and social contexts with self-written programming code;

	c. Communicate data analytic insights to the general public and professional audiences; **Category B:** Attributes for all-roundedness d. Appreciate the growing importance and relevance of machine-assisted data analysis in linguistic and social contexts; e. Weigh the benefits and potential pitfalls of data analytics along practical and ethical dimensions.
Subject Synopsis/ Indicative Syllabus *(Note 2)*	*Introduction* • Contemporary data in linguistic and social contexts • The nature and scope of data analytics • Programming languages *Data management, visualisation, and communication* • Data collection • Textual annotation • Visualising patterns and relationships in data *Data analytics #1: Working with numbers* • Linear regression • Generalised linear model and logistic regression • Prediction, classification, and clustering *Data analytics #2: Working with words* • Natural Language Processing and text analytics • Sentiment analysis • Topic modelling *Data ethics: A critical perspective* • Privacy, discrimination, and social inequalities
Teaching/Learning Methodology *(Note 3)*	Each weekly session will combine lecture and hands-on practice. During lecture activities, the instructor will impart concepts and facilitate class activities. This will be the main channel for transmitting professional and academic knowledge (ILO a, b, e). During hands-on activities, students will work in groups to analyse data and present solutions to practical problems. These prepare students for the workplace and allow them to develop higher order thinking skills and life-long learning skills (ILO c, d, e). Teaching and learning are supported by online resources. Students will be expected to read the prescribed materials and revise previous lessons before each session.

Assessment Methods in Alignment with Intended Learning Outcomes

(Note 4)

Specific assessment methods/tasks	% weighting	Intended subject learning outcomes to be assessed				
		a	b	c	d	e
1. Assignment (Individual)	40%	✓	✓	✓		✓
2. Quiz (Individual)	40%	✓	✓		✓	✓
3. Participation	20%	✓	✓	✓	✓	✓
Total	100 %					

The assignment and quiz assess students' understanding of data analytic techniques and their ability to write basic programming code to process data and interpret the results in context. Both assessments are individual work.

Participation may consist of both in-class and online discussions and activities.

Student Study Effort Expected

Class contact:	
▪ Lecture (3 hours x 13 weeks)	39 Hrs.
Other student study effort:	
▪ Independent reading (2 hours x 13 weeks)	26 Hrs.
▪ Independent learning (5 hours x 13 weeks)	65 Hrs.
▪ Assignments (1 hour x 13 weeks)	13 Hrs.
Total student study effort	143 Hrs.

Reading List and References

Arnold, T., & Tilton, L. (2015). *Humanities data in R: Exploring networks, geospatial data, images, and text*. Springer.

Beysolow, T. (2018). *Applied natural language processing with Python: Implementing machine learning and deep learning algorithms for natural language processing*. Apress.

Bruce, P. C., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientist: 50+ essential concepts using R and Python* (2 ed.). O'Reilly.

Buckland, W. (2023). *Who wrote Citizen Kane?: Statistical analysis of disputed co-authorship*. Springer Nature.

Clark, A., Fox, C., & Lappin, S. (2013). *The handbook of computational linguistics and natural language processing*. Wiley-Blackwell.

Desagulier, G. (2017). *Corpus linguistics and statistics with R: Introduction to quantitative methods in linguistics*. Springer Nature.

Jockers, M. L., & Thalken, R. (2020). *Text analysis with R: For students of literature* (2 ed.). Springer.

Kroonenberg, P. M. (2021). *Multivariate humanities*. Springer.

	Sarkar, D. (2019). *Text analytics with Python: A practitioner's guide to natural language processing* (2nd ed.). Apress. Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly. Speelman, D., Heylen, K., & Geeraerts, D. (2018). *Mixed-effects regression models in linguistics*. Springer. Other resources • datacamp.com • kaggle.com • towardsdatascience.com

Revised by Phoebe Lin, March 2024