| | |
|---|---|
| **Subject Code** | FH6058 |
| **Subject Title** | Data Analytics for Applied Language Scientists |
| **Credit Value** | 3 |
| **Level** | 6 |
| **Pre-requisite / Co-requisite/ Exclusion** | N/A |
| **Objectives** | Data analytic skills are increasingly valued in applied language sciences. These include the analysis of natural language (i.e., natural language processing) as well as language-related data (e.g., experimental, survey, demographic data), and the implementation of relevant techniques with programming languages.<br><br>This subject complements *FH6051 Computational Linguistics*, which focuses on natural language processing, to further develop competencies in programming, data visualization, and statistical modeling using the *Python* language. It adopts a thematic and problem-based approach to meet the following objectives.<br><br>• Train students to import, manage, and visualize data relevant to language professionals in *Python*<br>• Impart working knowledge of basic statistical and machine-learning techniques for language and social research<br>• Develop critical interpretation skills for insights and decision-making<br>• Raise awareness of ethical challenges in today's digital economy, and their implications for language professionals<br><br>An interactive pedagogical approach will be adopted with balanced assessment tasks and a variety of learning materials. No background in programming, data analytics, or statistics is assumed. |
| **Intended Learning Outcomes**<br><br>*(Note 1)* | Upon completion of the subject, students will be able to:<br><br>a. Independently write *Python* code for research and other work<br><br>b. Use key *Python* libraries like pandas, matplotlib, and statsmodels to manage, visualize, and analyze datasets<br><br>c. Understand and use basic statistical and machine learning techniques<br><br>d. Apply critical thinking skills to interpret quantitative results for humanistic inquiry<br><br>e. Weigh the pros and cons of data analytics along practical and ethical dimensions |

<table>
<tr>
<td>Subject Synopsis/ Indicative Syllabus<br><br>(Note 2)</td>
<td>
Introduction<br>
• Contemporary data in linguistic and social contexts<br>
• The nature and scope of data analytics<br><br>
Data management, visualization, and communication<br>
• Spreadsheets and dataframes<br>
• Visualizing patterns and relationships<br><br>
Data analytics and machine learning<br>
• Techniques and methods<br>
  ○ General(ized) linear model<br>
  ○ Machine learning models<br>
  ○ Monte Carlo simulations<br>
• Data analytic objectives<br>
  ○ Prediction, classification, clustering<br><br>
Data ethics: a critical perspective<br>
• Privacy, discrimination, and social inequalities
</td>
</tr>
<tr>
<td>Teaching/Learning Methodology<br><br>(Note 3)</td>
<td>
Each weekly session will comprise a 2-hour lecture and a 1-hour seminar. During the lecture, the instructor will impart concepts and facilitate/supervise class activities. The seminar, on the other hand, is student-led with groupwork to practice programming skills, analyze data, and present solutions. Students are expected to bring their personal laptops.<br><br>
Teaching and learning is supported by the open-source programming language Python and other online resources. Students should read the prescribed materials and revise previous lessons before each session.<br><br>
There are two individual take-home assignments and an individual project. The take-home assignments require conceptual understanding of data analytic techniques, writing basic programming code, and interpreting the analyses in context. The project requires students to apply these techniques and communicate insights and solutions in simulated professional settings, such as an academic or business conference.
</td>
</tr>
<tr>
<td>Assessment Methods in Alignment with Intended Learning Outcomes<br><br>(Note 4)</td>
<td>
<table>
<tr><td rowspan="2">Specific assessment methods/tasks</td><td rowspan="2">% weighting</td><td colspan="5">Intended subject learning outcomes to be assessed (Please tick as appropriate)</td></tr>
<tr><td>a</td><td>b</td><td>c</td><td>d</td><td>e</td></tr>
<tr><td>1. Take-home assignment 1</td><td>35</td><td>✓</td><td>✓</td><td>✓</td><td></td><td></td></tr>
<tr><td>2. Take-home assignment 2</td><td>35</td><td>✓</td><td>✓</td><td></td><td>✓</td><td></td></tr>
<tr><td>3. Individual project</td><td>30</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td></tr>
<tr><td>Total</td><td>100 %</td><td colspan="5"></td></tr>
</table>
</td>
</tr>
</table>

| | | |
|---|---|---|
| | The take-home assignments (35% each) require students to demonstrate the intended learning outcomes by combining data analytic and programming with critical thinking skills. The project (30%) requires the same skills with more emphasis on communication and presentation. | |
| **Student Study Effort Expected** | Class contact: | |
| | ▪ Lecture (2 hours x 13 weeks) | 26 Hrs. |
| | ▪ Seminar (1 hour x 13 weeks) | 13 Hrs. |
| | Other student study effort: | |
| | ▪ Independent reading (3 hours x 13 weeks) | 39 Hrs. |
| | ▪ Independent research (2 hours x 13 weeks) | 26 Hrs. |
| | ▪ Assignments (2 hours x 13 weeks) | 26 Hrs. |
| | Total student study effort | 130 Hrs. |

| | |
|---|---|
| **Reading List and References** | Recommended text and reference books<br><br>Norton, P. et al. (2005). *Beginning Python.* Indianapolis: Wiley.<br><br>Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python.* Sebastopol: O'Reilly.<br><br>Hai-Jew, S. (Ed.). (2017). *Data Analytics in Digital Humanities*. Cham: Springer.<br><br>O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Books.<br><br>Tay, D., & Pan, M. X. (Eds.). (2022). *Data Analytics in Cognitive Linguistics: Methods and Insights*. Berlin: De Gruyter Mouton.<br><br>Reference papers<br><br>Coupé, C. (2018). Modeling linguistic variables with regression models: Addressing non-gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers in Psychology*, *9*(APR), 1–21.<br><br>Kroese, D. P., Brereton, T., Taimre, T., & Botev, Z. I. (2014). Why the Monte Carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, *6*(6), 386–392.<br><br>Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLoS ONE*, *1*, 1–34.<br><br>Tay, D. (2020). A computerized text and cluster analysis approach to psychotherapy talk. *Language & Psychoanalysis*, 9(1), 1–22.<br><br>*Python* user guides<br><br>• Pandas: https://pandas.pydata.org/<br>• Scikit-learn: https://scikit-learn.org/stable/<br>• Seaborn: https://seaborn.pydata.org/index.html<br>• Statsmodels: www.statsmodels.org/stable/user-guide.html<br>• Numpy: numpy.org/doc/stable/<br>• Scipy: https://www.scipy.org/docs.html<br><br>Other online resources<br><br>• datacamp.com<br>• kaggle.com<br>• towardsdatascience.com |