

M2.851 - Tipología y ciclo de vida de los datos - Aula 2

Máster en Ciencia de Datos

Práctica 2 – Web scraping

Iván MASEDA ZURDO & Lucas REY PITALUGA

Código

El código fuente para la extracción de datos se puede encontrar en el repositorio de GitHub <https://github.com/lreyp/Water-Quality-Classification>. Además, en el mismo repositorio puede encontrarse también el dataset de origen y los datasets generados.

Tabla de contribuciones

Contribuciones	Firma
Investigación previa	IMZ, LRP
Redacción de respuestas	IMZ, LRP
Desarrollo código	IMZ, LRP

Licencia

Publicado bajo licencia CC BY-NC-SA 4.0

Este tipo de licencia ofrece libertad para **compartir, copiar y redistribuir** los datos a través de cualquier medio y bajo cualquier formato, y permite **adaptar los datos, mezclarlos, transformarlos y construir nuevos datos** en base a éstos, con cualquier fin, incluso comercialmente, siempre y cuando se cumplan las condiciones siguientes:

- **Atribución:** el usuario de los datos debe otorgar el crédito correspondiente, proporcionar un enlace a la licencia e indicar si se realizaron cambios. Puede hacerlo de cualquier manera razonable, pero no de ninguna manera que sugiera que el licenciante respalda al usuario o su uso.
- **Compartir por igual:** si el usuario de los datos remezcla, transforma o construye sobre el material, debe distribuir sus contribuciones bajo la misma licencia que el original.

De esta forma, permitimos que otras personas analicen este mismo conjunto de datos, reconociendo nuestro trabajo y pudiendo replicar la información. Además, en caso de realizar transformaciones o análisis con este conjunto de datos, deberán compartir bajo la misma licencia que lo hacemos nosotros, asegurando el acceso a nuevo conocimiento.

Bibliografía

- 1. Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- 2. Subirats, L., Pérez, D., Calvo, M. (2019) Introducción a la limpieza y análisis de los datos. Editorial UOC.
- 3. R Documentation <https://www.rdocumentation.org/>
- 4. Dealing with Missing Data using R
<https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17>
- 5. Flexible Imputation of Missing Data <https://stefvanbuuren.name/fimd/>
- 6. Root-Mean-Square Error in R Programming
<https://www.geeksforgeeks.org/root-mean-square-error-in-r-programming/>
- 7. MissForest - missing data imputation using iterated random forests
<https://rpubs.com/lmorgan95/MissForest>
- 8. Hmisc <https://www.rdocumentation.org/packages/Hmisc/versions/4.5-0>
- 9. Package 'Hmisc' <https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf>
- 10. Extra - Missing Data Tools
<https://unc-libraries-data.github.io/R-Open-Labs/Extras/Missing/missing.html>
- 11. Package 'mi' <https://cran.r-project.org/web/packages/mi/mi.pdf>
- 12. Package 'xgboost' <https://cran.r-project.org/web/packages/xgboost/xgboost.pdf>
- 13. trainControl: Control parameters for train
<https://www.rdocumentation.org/packages/caret/versions/6.0-88/topics/trainControl>
- 14. expand.grid: Create a Data Frame from All Combinations of Factor Variables
<https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/expand.grid>
- 15. Simple R - xgboost - caret kernel
<https://www.kaggle.com/nagsdata/simple-r-xgboost-caret-kernel>
- 16. XGBoost Multinomial Classification Iris Example in R
<https://rpubs.com/dalekub/XGBoost-Iris-Classification-Example-in-R>
- 17. Tune Machine Learning Algorithms in R
<https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/>
- 18. Tuning xgboost in R: Part I
<https://insightr.wordpress.com/2018/05/17/tuning-xgboost-in-r-part-i/>
- 19. Tuning xgboost in R: Part II
<https://www.r-bloggers.com/2018/07/tuning-xgboost-in-r-part-ii/>
- 20. Logistic Regression Essentials in R
<http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/>