References

[Chri2012] - *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Peter Christen, Springer-Verlag 2012.

[ConfMatr] - https://en.wikipedia.org/wiki/Confusion_matrix, *Confusion Matrix*, Wikipedia, the free encyclopedia.

[FeatWiki] - http://www.swissbib.org/wiki/index.php?title=Features_Deduplication, Wikipedia site created for documentation of features for deduplication.

[PropRepo] - https://github.com/epfl-extension-school/capstone-proposal-ads-ml-c5-s1-1365-585, github repository to the capstone project's proposal, Andreas Jud.

[ProjRepo] - https://github.com/epfl-extension-school/capstone-project-ads-ml-c5-s2-1365-585of, github repository to the capstone project's implementation, Andreas Jud.

[HanK2012] - *Data Mining - Concepts and Techniques*, Jiawei Han, Micheline Kamber, Jian Pei, 3rd Edition, Morgan Kaufmann Publishers 2012.

[JudACaps] - *Deduplication of Swissbib raw data - Capstone proposal by Andreas Jud*, 15.03.2020.

[KeraRand] - https://keras.io/getting-started/faq/#how-can-i-obtain-reproducible-results-using-keras-during-development, *How can I obtain reproducible results using Keras during development?*, Keras Documentation.

[MARC] - https://www.loc.gov/marc/bibliographic/, MARC 21 format description for bibliographic data, Library of Congress.

[matp] - https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.hist.html, *matplotlib.pyplot.hist*, matplotlib.

[Padm2012] - *An approach based on artificial neural network for data deduplication*, M. Padmanaban, T. Bhuvaneswari, International Journal of Computer Science and Information Technologies, Vol. 3(4), 2012, 4637-4644.

[PCAWiki] - https://en.wikipedia.org/wiki/Principal_component_analysis, *Principal component analysis*, Wikipedia, the free encyclopedia.

[rocauc] - https://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics, *Receiver operating characteristic (ROC)*, scikit learn.

[ScalRepo] - https://github.com/guenterh/andreas_cluster_features, Repository to scala code for data extraction, Günter Hipler.

[SmWa] - https://gist.github.com/nornagon/6326a643fc30339ece3021013ed9b48c, *Smith-Waterman Python implementation* in github.

[StSi] - https://github.com/luozhouyang/python-string-similarity, *python-string-similarity* – A library implementing different string similarity and distance measures.

[svc] - https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html, *sklearn.svm.SVC*, scikit learn.

[Swis] - https://www.swissbib.ch/, Access web site of Swissbib's online catalogue.

[SwRe] - https://github.com/guenterh/andreas_cluster_features, Repository to scala code for data extraction.

[tSNE] - https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding, *t-distributed stochastic neighbor embedding*, Wikipedia, the free encyclopedia.

[TeDi] - https://pypi.org/project/textdistance/, *TextDistance* – Python library for comparing distance between two or more sequences by many algorithms.

[WiCo2001] - Communication by Silvia Witzig, Universität Basel, Universitätsbibliothek, Projekt swissbib, 24.01.2020.

```
In [ ]:
```