Task 1: String-based data matching

Data

Amazon and Google product descriptions, based on the study:

Köpcke, H., Thor, A., & Rahm, E. (2010). Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, *3*(1-2), 484-493.

Example:

Amazon	Google	Match
upg sgms 1000 incremental node	sonicwall gms 1000 upgrade	1
clickart 950 000 - premier image pack (dvd-rom)	microsoft excel 2007 (pc)	0

Subtasks and points

1. Describe the data. (10)
2. Use algorithms from the String2string library:
    a. Levenshtein distance (at the character level) (10) tutorial link
    b. Jaccard distance (at the word level) (10)
    c. Jaro distance (at the character level) (10) tutorial link
    d. Jaro distance (at the word level) (10)
    Note: Jaro and Jaccard distances are inversed similarities (distance = 1 – similarity)
3. Plot ROC curves for the methods above. (10) definition sklearn
4. Analyze results. (10)
5. Propose string pre-processing that can improve results, report results. (15)
6. Propose a combination of character- and word-level distances that can improve results, experiment with further methods, report results. (15)
   Note that subtasks 5 and 6 imply indirect learning from the test data.