Heidelberg University
Database Systems Research Group
Prof. Dr. Michael Gertz
Satya Almasian, Dennis Aumiller

November 16, 2020
Text Analytics
Winter Semester 2020/21

Assignment 1: "Text Conversion and Regular Expressions"

Due: Monday 2pm, November 30, 2020, via Moodle

Submission guidelines

- For working on the assignments handed out in this class, please form **teams of three or four students**. Student teams are fixed and are identical to the teams that collaborate for the group project.
- Written solutions need to be **uploaded as a single PDF**.
- Source code of programming exercises needs to be provided in a separate folder. Preferably your code should be in the form of a **Jupyter notebook** that is able to run in **Google Colab**. If you do not want to use Google Colab, we require that your code can be run using **Python 3.7 on a Linux machine**. In this case you also need to provide a **requirements.txt** that lists all dependencies.
- Zip your written solutions and source code before you upload them.
- It is sufficient if one person per group uploads the solution **to Moodle**, but make sure that **the names of all team members are given on the PDF and in the source code**.
- Justify all of your answers.

Problem 1-1 Text Analytics Pipeline 1 + 2 + 2 = 5 points

1. Identify the drivers, as given on slide `2-3` of the lecture, for the following problems.
   (i) Your company has collected written user feedback for a new product. To efficiently incorporate it into the further development of the product, you want to automatically summarize the most important arguments of the users.
   (ii) You attended a text analytics conference and want to incorporate a new method into your existing pipeline to improve the classification of scanned documents.
2. Outline the advantages and disadvantages of stratified sampling. Give a scenario in which stratified sampling is unsuitable.
3. Name at least two advantages and disadvantages for each of the following concepts. Illustrate your points by giving examples.
   (i) Stop word removal
   (ii) Stemming

Problem 1-2 PDF Conversion and Regular Expressions 1 + 3 + 2 + 8 + 1 = 15 points

In this assignment you have to convert a set of PDFs to raw text and extract information from the files using regular expressions.

1. You have a set of HTML files and need to parse their content. Would you use regular expressions to parse the files? Justify your answer.
2. Download the PDFs provided via Moodle. Since PDFs cannot be directly processed using regular expressions, you first have to convert their contents to plain text. Find at least three methods to convert PDFs to text, and compare their performance. You have to provide a quantitative as well as a qualitative analysis. For comparison of two generated files, you should use Python's `SequenceMatcher.ratio()`. With this analysis as a basis, decide for one of the methods, and provide the processed raw text files in your submission. Justify your decision!
3. Why is a high quality conversion from PDF to plain text hard? Your answer does not need to be exhaustive but should outline some of the most important reasons.
4. The provided files are split into three groups. For each of the individual parts, make sure to print out all extracted results **into a dedicated text file** which should be included in your submission. Each line should contain only a single instance in a normalized format. For phone numbers, e.g., you could remove all spaces and special characters, etc.:
   (i) From the PDF files in the folder `phone_numbers/` extract as many valid phone numbers as possible.
   (ii) Extract valid URLs and Email addresses from the files in the folder `phone_numbers/`.
   (iii) From the PDF file in the folder `isbn/` extract all valid ISBN numbers.
   (iv) In the folder `unit_conversion/` you can find a file containing a list of quantities. Convert all values given in milliliters to liters **using only functions of the Python `regex` module**. Converted values should be truncated to three decimals in the liter representation.
5. Apply your solution from task 4-(i) to the files in the folder `scans/` which consists of pages scanned from a phone book. Analyse how well your solution performs by giving examples.