PDF papers containing empirical studies contain statistical information that can be used in research to analyze statistical values. The aim of the Bachelor's thesis is to extract all relevant statistical information found in SOUPS Technical Sessions | USENIX by using Natural Language Processing Method (Tokenization and Part-of-Speach (tagging)).

The first problem in the Bachelor's thesis was converting PDF documents into machine-readable text. 20 PDF documents from Soups 2022 and 2021 were examined using three different libraries (pdfplumber, PyPDF2 and pdfminer) in order to find the best library that can convert PDF-Dokument to machine-readable text. eventually, it was found that pdfminer was the most effective library for solving common issues that we will encounter while converting PDF-Dokument in machine-readable text.

The second problem was the extraction of statistical information from machine-readable texts. Statistical information like the famous statistical Test names, results, and terms was collected. These statistical information were analyzed to create Regex-tagged patterns and they are responsible for extracting the statistical information. Using 314 Regex-tagged patterns, 866 statistical sentences were extracted from Soups 2022 and 2021. The method was to break the extracted text from PDF-Dokument into sentences by using the tokenization Method and to focus on sentences that they include only statistical terms and ignore the other sentences. After applying the tokenization Method Regex and POS were used in order to extract the statistical sentences and tag all founded statistical Terms in the sentence.

The effectiveness of the 314 Regex-tagged patterns was tested on 228 PDF papers from Soups from 2010 to 2020. A Python application was developed, consisting of five essential functions. Using this application, all 228 PDF papers were quickly converted into machine-readable texts, and the statistical information was extracted. The number of extracted statistical sentences from Soups from 2010 to 2020 was 2715. The efficiency of the 314 tagged patterns was then tested on 11 selected PDF papers from each year from Soups 2010 to 2020, revealing that only 10.2% of statistical sentences were missed during extraction using this Python application.