# Apache Spark
# Übungsaufgabe 16

**Analyse von Patenttexten**

- Zu Übung 34-Apache-Spark-ML-Claims
- In der Beispieldatei „patent_claims_excerpt.csv“ wird jeder Claim (Anspruch) eines Patents (pat_no) in einer Zeile dargestellt.

```
pat_no,claim_no,claim_txt,dependencies,ind_flg,appl_id
3930271,1,"1. A golf glove comprising at least an index
3930271,4,4. A golf glove adapted for use on one hand c
3930271,3,"3. A glove comprising an index finger recept
3930271,2,"2. A golf glove in accordance with claim 1 w
3930272,1,"1. In combination with a height adjustable c
3930272,3,3. The lock defined in claim 1 wherein the pi
3930272,2,2. The lock defined in claim 1 wherein the br
```

- In Zeile 20 der Datei „claims.py“ wird die csv-Datei eingelesen.
- Filtern Sie die Daten so, dass für jedes Patent nur der erste Claim (also claim_no=1) berücksichtigt wird.
- Nach der Filterung sehen die Daten folgendermaßen aus:

```
pat_no,claim_no,claim_txt,dependencies,ind_flg,appl_id
3930271,1,"1. A golf glove comprising at least an index
3930272,1,"1. In combination with a height adjustable c
3930273,1,"1. A bed arrangement comprisinga bed frame,a
3930274,1,"1. An assembly for use in recreational activ
3930275,1,"1. A method of manufacturing slippers each h
```

- Führen Sie die Analyse der Patenttexte mit der erweiterten Filterung durch.
- Erstellen Sie eine kurze Präsentation, in der Sie die Aufgabenstellung, die Vorgehensweise und das Ergebnis darstellen.

Quelle: https://www.informatik-aktuell.de/entwicklung/methoden/einfuehrung-in-spark-ein-text-mining-projekt.html; https://bulkdata.uspto.gov/data/patent/claims/economics/2014/patent_claims_fulltext.csv.zip