Web Mining: Übung 2

Lösungsvorschlag

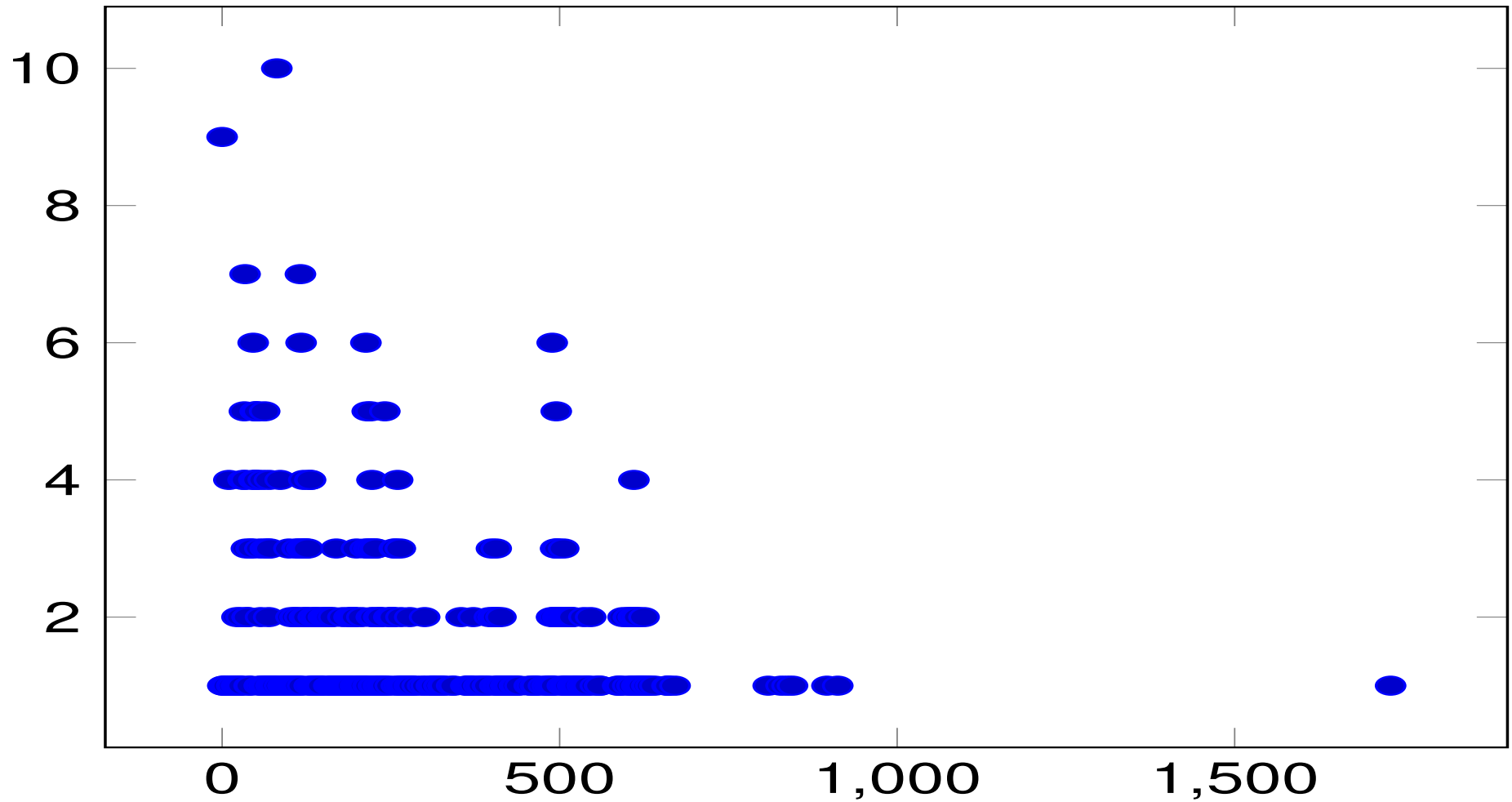
TECHNISCHE UNIVERSITÄT DARMSTADT

Inhaltsverzeichnis

TECHNISCHE UNIVERSITÄT DARMSTADT

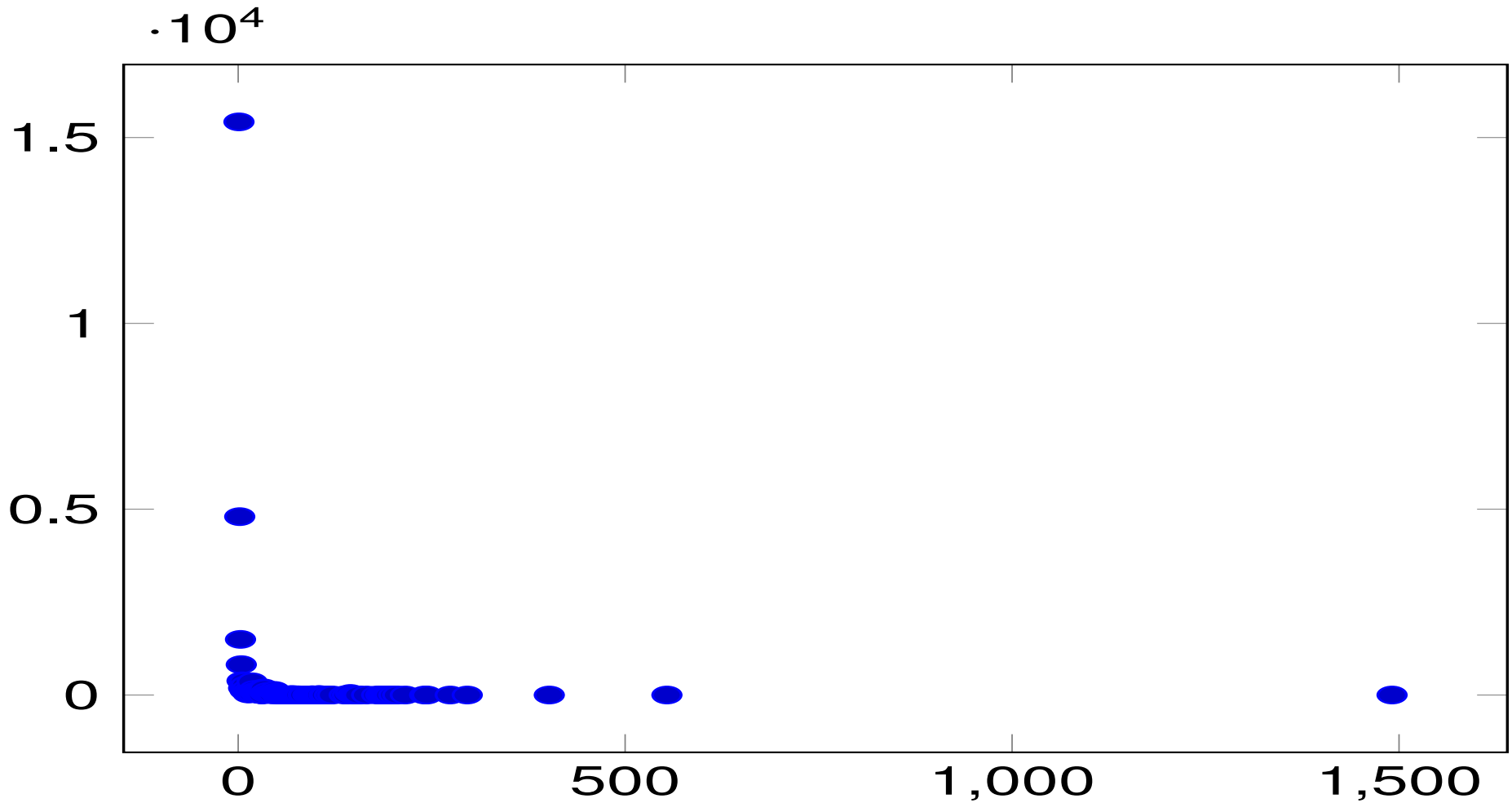
Aufgabe 2: Entwicklung eines Crawlers
Histogramm: Es gibt *y* Seiten mit *x* Links

TECHNISCHE UNIVERSITÄT DARMSTADT



Aufgabe 2: Entwicklung eines Crawlers
Histogramm: Es gibt *y* Links, die *x*-mal auftraten

TECHNISCHE UNIVERSITÄT DARMSTADT



Aufgabe 2: Entwicklung eines Crawlers
Erkennung wiederkehrender Links

TECHNISCHE UNIVERSITÄT DARMSTADT

- Entfernung aller Listen:

```
1 <ul>
2         <li><a href="start.html">Start</a></li>
3         <li><a href="news.html">News</a></li>
4         <li><a href="impressum.html">Impressum</a></li>
5 </ul>
```

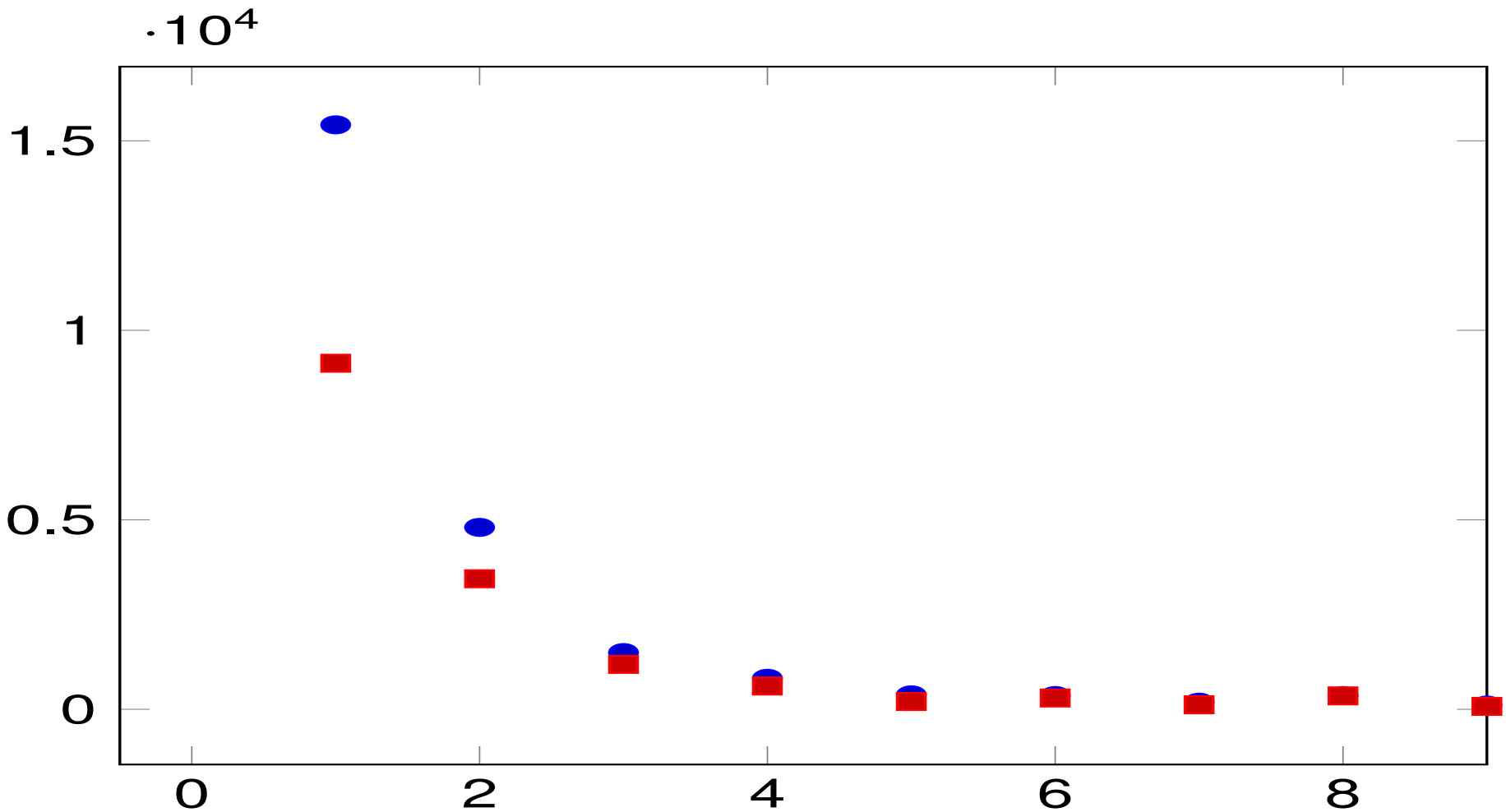
- Entfernung mit CSS-Selektor: `[class*='nav']`:

```
1 <div class="main-navigation">
2         ...
3 </div>
```

Aufgabe 2: Entwicklung eines Crawlers
Wirksamkeit der Duplikaterkennung
Histogramm: Es gibt *y* Links, die *x*-mal auftraten

TECHNISCHE UNIVERSITÄT DARMSTADT

TECHNISCHE
UNIVERSITÄT
DARMSTADT


·10^4
1.5
1
0.5
0
0
2
4
6
8

Aufgabe 2: Entwicklung eines Crawlers
Wie oft wurden Hosts besucht?

TECHNISCHE UNIVERSITÄT DARMSTADT