

Unternehmensklassifikation

anhand von gecrawlten Snippets

Author

STEPHAN LENERT
PETER BEHRENS
TOM MÜLLER

Date

06.07.2021

Agenda

- Ziel des Projektes
- Datenquellen & Transformation
- Datenverteilung
- Theorie des Lösungsansatz
- Klassifikationsalgorithmen
- Herausforderungen
- Ergebnisse
- Verbesserungsansatz
- Demo

Was war das Ziel?

AUSGANGSSITUATION

Klassifikation der Unternehmen in verschiedene Segmente mithilfe von Snippets der Suchmaschine Bing.

Python Crawler zum Download der Snippets.

The screenshot shows a Microsoft Bing search results page for the query "mercedes-benz". The search bar at the top contains the text "mercedes-benz". Below the search bar, there are tabs for "ALL", "IMAGES", "VIDEOS", "MAPS", "NEWS", and "SHOPPING". The "ALL" tab is selected. Below the tabs, it says "33.900.000 Results" and "Date". Below this, there is a horizontal bar with the Mercedes-Benz logo and several buttons: "Mercedes-Benz", "Models", "Dealerships near me", "Forums", "Concept cars", and "Apparel". The main search results are displayed below. The first result is a sponsored link (Ad) for "Mercedes-Benz - mercedes-benz.de" with the URL "https://www.mercedes-benz.de/pkw". The ad text says: "Das Auge fährt mit: Entdecken Sie jetzt das innovative Design von Mercedes-Benz! Luxus, Sportlichkeit & Leistung vereint. Ob Limousine, T-Modell, Coupé, Cabrio, ... Broschüre bestellen · Händler kontaktieren · Probefahrt vereinbaren Leugnen bringt Dir Nichts, ein Junger Stern das Beste." Below the ad, there are several other search results, including "Das neue CLS Coupé", "CLA Coupé", "Händlersuche", "A-Klasse Probefahrt", "Mercedes-Benz Newsletter", and "Mercedes 8.100 € sparen - Top-Rabatte auf Ihren Neuwagen". The "Mercedes 8.100 € sparen" result is also a sponsored link with the URL "https://www.carwow.de/mercedes/neuwagen" and a button "Angebot anzeigen".

Microsoft Bing

mercedes-benz

ALL IMAGES VIDEOS MAPS NEWS SHOPPING

33.900.000 Results Date

Mercedes-Benz Models Dealerships near me Forums Concept cars Apparel

Mercedes-Benz - mercedes-benz.de
<https://www.mercedes-benz.de/pkw>
Ad Das Auge fährt mit: Entdecken Sie jetzt das innovative Design von **Mercedes-Benz**! Luxus, Sportlichkeit & Leistung vereint. Ob Limousine, T-Modell, Coupé, Cabrio, ... Broschüre bestellen · Händler kontaktieren · Probefahrt vereinbaren Leugnen bringt Dir Nichts, ein Junger Stern das Beste.

Das neue CLS Coupé
Lassen sich vom faszinierenden CLS Coupé begeistern. Alle Infos hier!

CLA Coupé
Rock. Star. Entdecken Sie hier das neue **Mercedes-Benz** CLA Coupé!

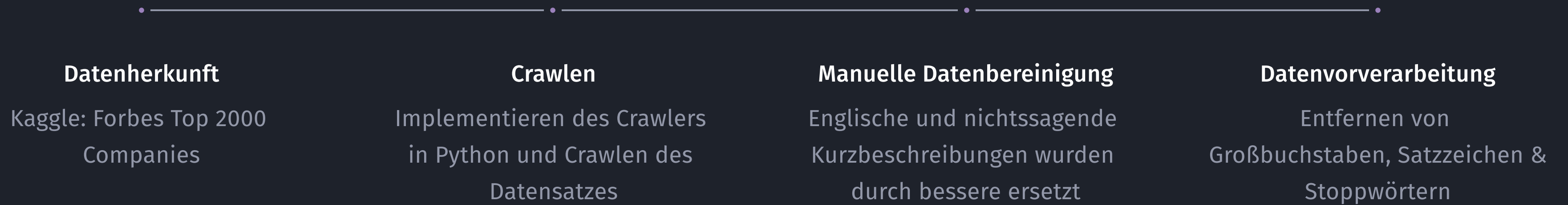
Händlersuche
Finden Sie hier einen **Mercedes-Benz** Partner in Ihrer Nähe. Zur Suche!

A-Klasse Probefahrt
Bereit für eine neue Generation: Einfach einsteigen & Probe fahren!

Mercedes-Benz Newsletter
Jetzt immer die aktuellsten Infos erhalten. Gleich hier registrieren!
See results only from mercedes-benz.de

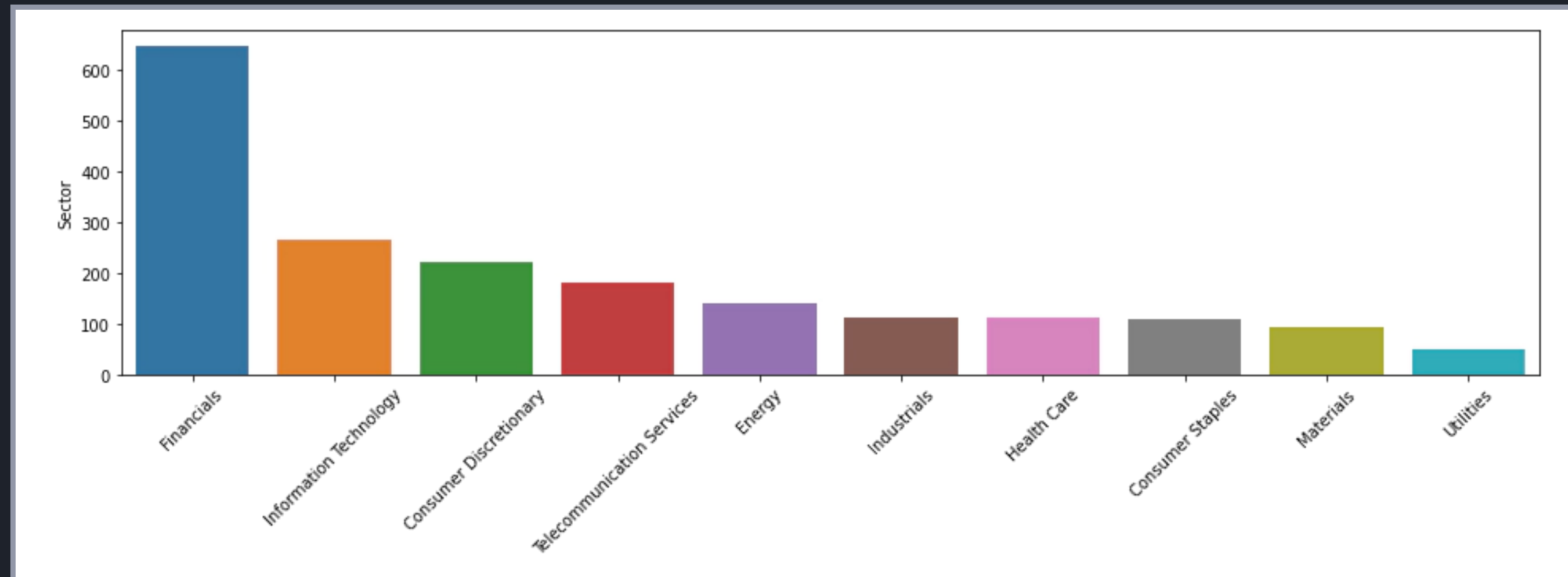
Mercedes 8.100 € sparen - Top-Rabatte auf Ihren Neuwagen
<https://www.carwow.de/mercedes/neuwagen>
Ad Angebote direkt von lokalen Händlern. Durchschnittlich sparen Sie 8.100 €. Konfigurieren Sie ihr Wunschauto & sichern Sie sich jetzt den besten Preis mit carwow. carwow.de has been visited by 100K+ users in the past month
Angebot anzeigen

DATENQUELLEN & -TRANSFORMATION



Datenverteilung

"Financials" ist überrepräsentiert im Datensatz. Alle anderen Sektoren sind ausgeglichen.



Theorie des Lösungsansatz

1. DATEN-VORVERARBEITUNG

```
def prepare_snippets(snippet, raw_string_return = False, remove_int = True, lowercase = True, stopwords =  
True, punctuations = True, only_nouns_n_adjs = True, lemmatize = True, reduce = True, word_embeddings = True):  
...
```

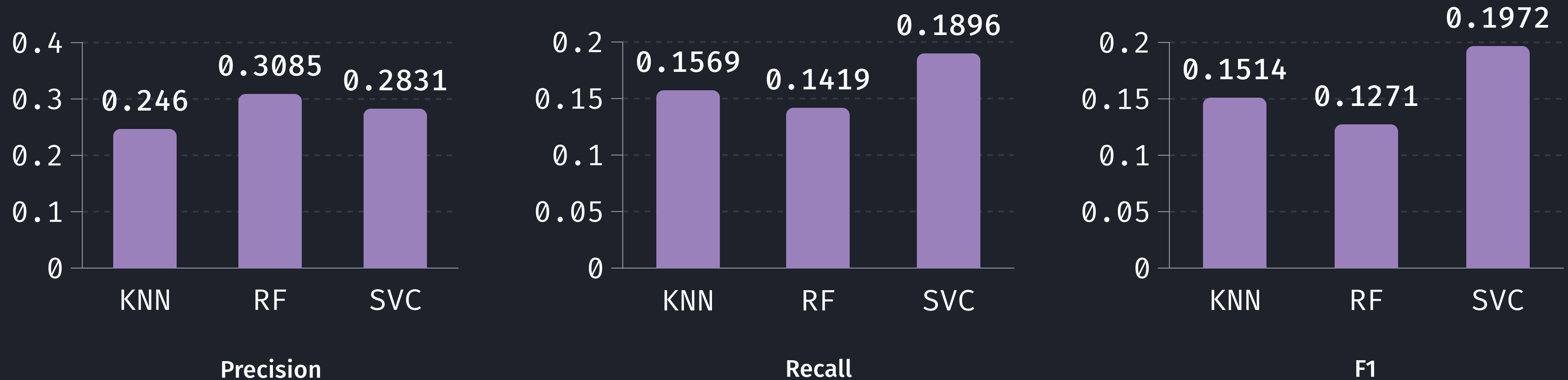
2. MODEL HYPERPARAMETER TUNING

```
from sklearn.model_selection import RandomizedSearchCV  
...  
search = RandomizedSearchCV(rf, space, n_iter=500, scoring='accuracy', n_jobs=-1, cv=cv, verbose = 1)
```

3. STACKING + CROSSVALIDATION

```
from sklearn.ensemble import StackingClassifier  
stacked_clf = StackingClassifier(estimators=classifiers, final_estimator=LogisticRegression(), cv = 10 )
```

VERWENDETE KLASSIFIKATIONSALGORITHMEN



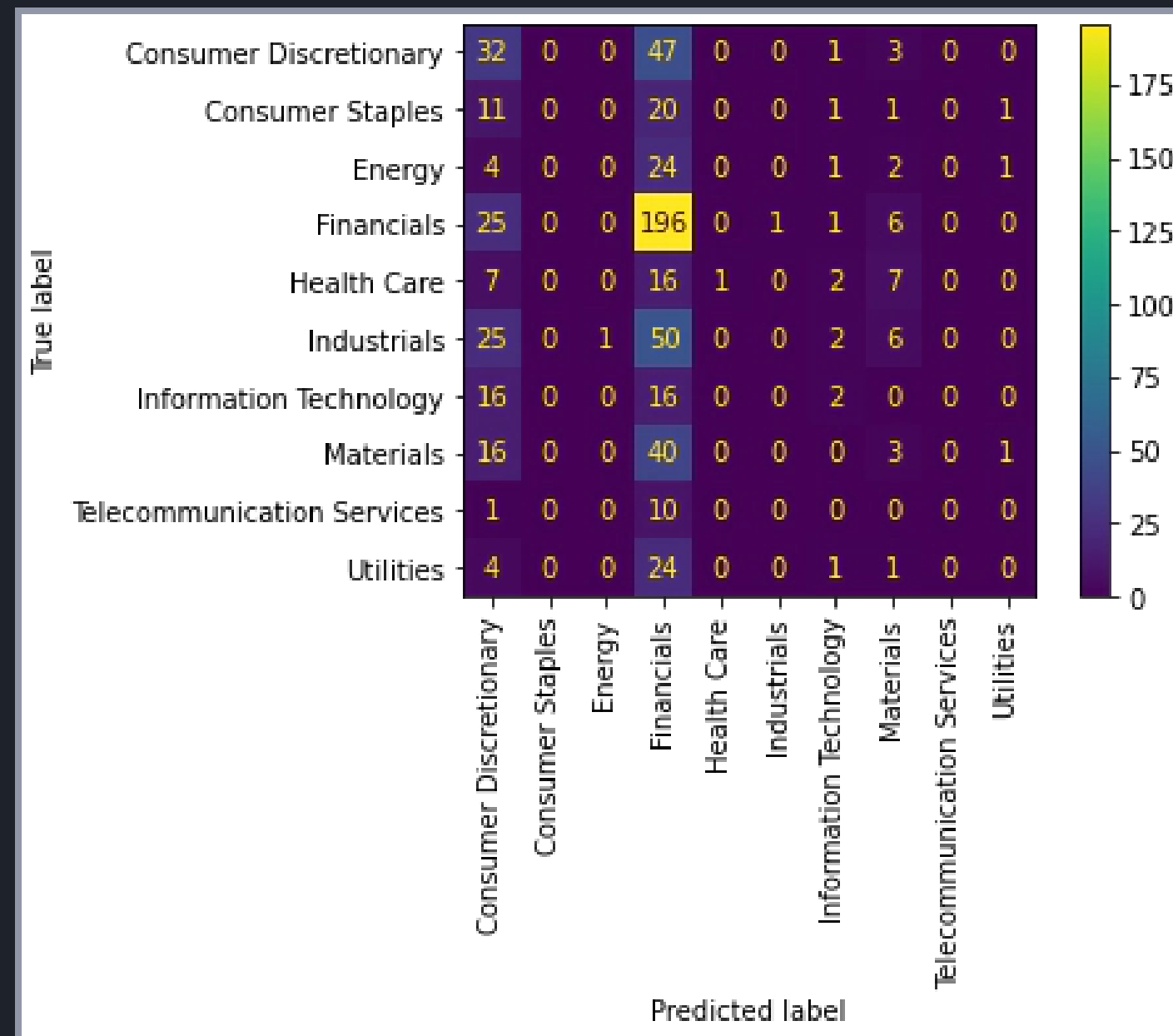
```
knn = Pipeline([('scaler', StandardScaler()), ('clf', KNeighborsClassifier(weights="distance", n_neighbors=22))])
rf = Pipeline([('scaler', StandardScaler()), ('clf', RandomForestClassifier(n_estimators = 200, criterion = "gini", max_depth = 10,
    max_features = "auto", min_samples_leaf = 0.005,
    min_samples_split = 0.005, n_jobs = -1, random_state = 1000, bootstrap = False))])
svc = Pipeline([('scaler', StandardScaler()), ('clf', SVC(C = 12.450980939125662, degree = 4, gamma = 0.008983921439908003,
    kernel = "rbf", probability = True, class_weight=None))])
```

Unsere Herausforderungen:

- Datensatz
- Crawler
- Modell Performance Verbessern
- Deployment

Ergebnisse

- Vorhersage beruht auf finalem stacked Modell auf gesamten Testdatensatz
- Einsehen und Klassifizierung der Unternehmenssnippets in der App



Test Set Accuracy Score

0,3714

Train Set Accuracy Score

0,705

Macro Average Precision

0,20

Macro Average Recall

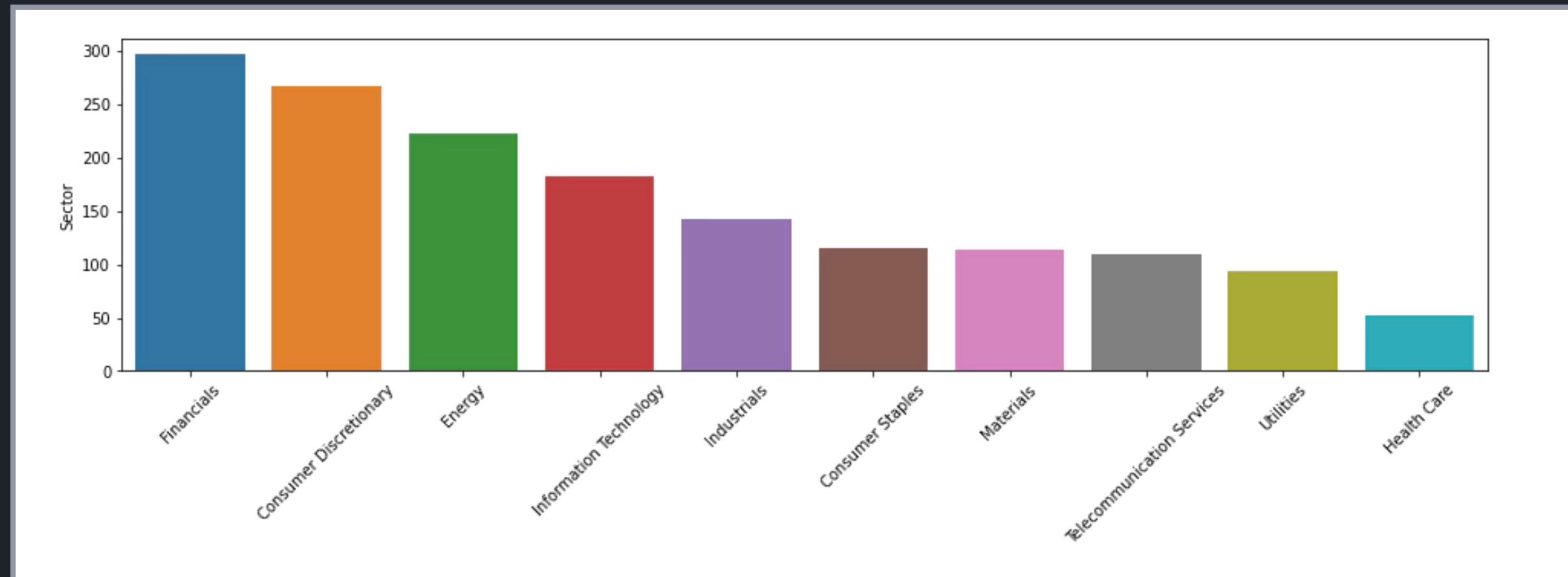
0,14

Macro Average F1-Score

0,11

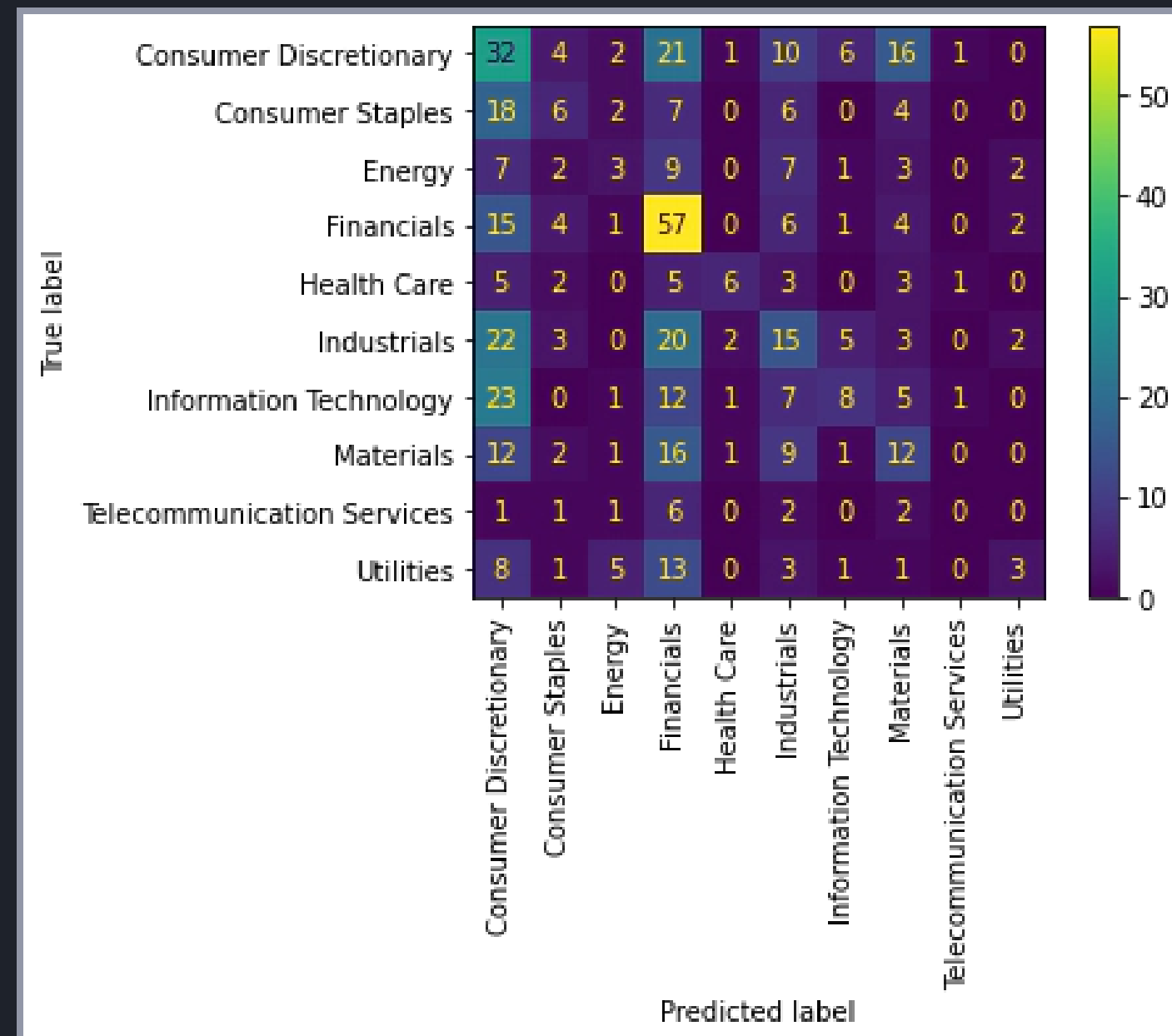
Datenverteilung - Verbesserung

"Financials" wurde angepasst, sodass der Datensatz ausgeglichener ist.



Ergebnisse

- Nach Ausgleich des Datensatzes
- Finales Model trainiert auf ganzem Datensatz



Test Set Accuracy Score

0,2747

Train Set Accuracy Score

0,8953

Macro Average Precision

0,27

Macro Average Recall

0,21

Macro Average F1-Score

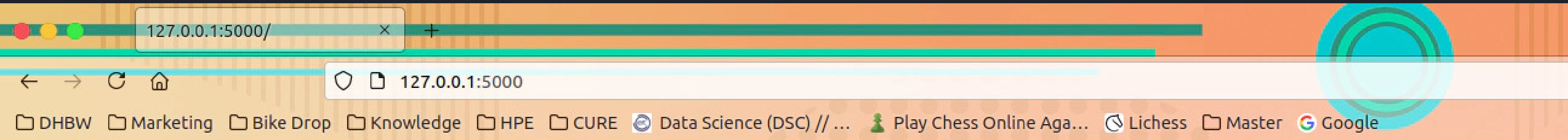
0,21

< / >

Demo

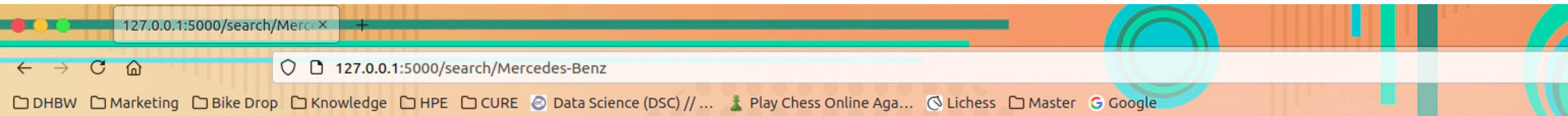
LOCAL FLASK APP

Mercedes-Benz



Enter Company

Mercedes-Benz



Company:

Mercedes-Benz

Search-Snippet:

<https://www.mercedes-benz.de>Luxus, Sportlichkeit & Leistung vereint. Ob Limousine, T-Modell, Coupé, Cabrio, Roadster, SUV & mehr. Erleben Sie die Produkte von Mercedes-Benz.

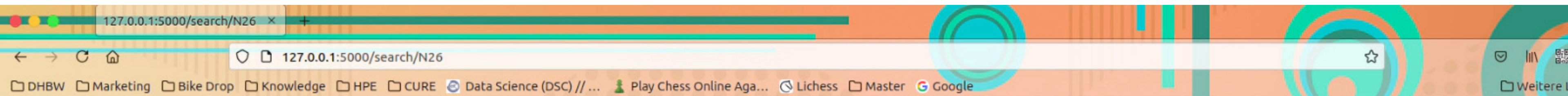
Predicted Sector:

Consumer Discretionary

N26



Enter Company

Company:

N26

Search-Snippet:

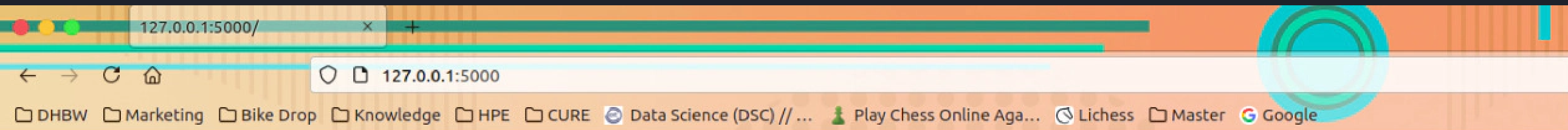
<https://n26.com/de-de>N26 ist die mobile Bank. Verwalte dein Girokonto 100 % mobil und verfolge deine Einnahmen und Ausgaben übersichtlich in Echtzeit. Eröffne jetzt dein N26 Girokonto in wenigen Minuten auf deinem Smartphone und nutze es, schon bevor deine physische Karte eintrifft. Girokonto eröffnen.

Predicted Sector:

Financials



Gazprom



Enter Company

Gazprom



Company:

Gazprom

Search-Snippet:

GAZPROM Germania GmbH. Markgrafenstraße 23 . 10117 Berlin. Telefon +49 (0)30 20195-0. Telefax +49 (0)30 20195-313. E-Mail: info@gazprom-germania.de -

Predicted Sector:

Financials

Pitch

Zurück

ARBEITSTEILUNG

Name	Rolle	Aufgabe
Tom	Der Mann fürs Schöne	Datenverarbeitung, Präsentation und Dokumentation, Flask Frontend
Peter	Backend	Recherche (Algorithmen, Datenquellen), Datenvorverarbeitung, Crawler, Dokumentation
Stephan	Backend	Implementierung, Evaluierung und Verbesserung der Algorithmen, Aufsetzen der Flask-App

"teamwork makes the dream work"

< / >

Vielen Dank für Ihre Aufmerksamkeit!

PETER, STEPHAN, TOM