```
!pip install spacy matplotlib pandas
!python -m spacy download en_core_web_sm
```

```
Requirement already satisfied: spacy in /usr/local/lib/python3.12/dist-packages (3.8.11)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.12/dist-packages (3.10.0)
Requirement already satisfied: pandas in /usr/local/lib/python3.12/dist-packages (2.2.2)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.12/dist-packages (from spacy) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (1.0.15)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.12/dist-packages (from spacy) (2.0.13)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.12/dist-packages (from spacy) (3.0.12)
Requirement already satisfied: thinc<8.4.0,>=8.3.4 in /usr/local/lib/python3.12/dist-packages (from spacy) (8.3.10)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python3.12/dist-packages (from spacy) (1.1.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.12/dist-packages (from spacy) (2.5.2)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/python3.12/dist-packages (from spacy) (2.0.10)
Requirement already satisfied: weasel<0.5.0,>=0.4.2 in /usr/local/lib/python3.12/dist-packages (from spacy) (0.4.3)
Requirement already satisfied: typer-slim<1.0.0,>=0.3.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (0.21.1)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (4.67.1)
Requirement already satisfied: numpy>=1.19.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (2.0.2)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (2.32.4)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /usr/local/lib/python3.12/dist-packages (from spacy)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.12/dist-packages (from spacy) (3.1.6)
Requirement already satisfied: setuptools in /usr/local/lib/python3.12/dist-packages (from spacy) (75.2.0)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.12/dist-packages (from spacy) (25.0)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (1.3.3)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (4.61.1)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (1.4.9)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (11.3.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (3.3.1)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.12/dist-packages (from matplotlib) (2.9.0.post
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/dist-packages (from pandas) (2025.3)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!=1.8.1
Requirement already satisfied: pydantic-core==2.41.4 in /usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!=1.8.1,
Requirement already satisfied: typing-extensions>=4.14.1 in /usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!=1.
Requirement already satisfied: typing-inspection>=0.4.2 in /usr/local/lib/python3.12/dist-packages (from pydantic!=1.8,!=1.8
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-packages (from python-dateutil>=2.7->matplotlib) (
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.13.0->spacy)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.13.0->
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.12/dist-packages (from requests<3.0.0,>=2.13.0->
Requirement already satisfied: blis<1.4.0,>=1.3.0 in /usr/local/lib/python3.12/dist-packages (from thinc<8.4.0,>=8.3.4->spac
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.12/dist-packages (from thinc<8.4.0,>=8.3.4
Requirement already satisfied: click>=8.0.0 in /usr/local/lib/python3.12/dist-packages (from typer-slim<1.0.0,>=0.3.0->spacy
Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in /usr/local/lib/python3.12/dist-packages (from weasel<0.5.0,>=0.
Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in /usr/local/lib/python3.12/dist-packages (from weasel<0.5.0,>=0.4.
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.12/dist-packages (from jinja2->spacy) (3.0.3)
Requirement already satisfied: wrapt in /usr/local/lib/python3.12/dist-packages (from smart-open<8.0.0,>=5.2.1->weasel<0.5.0
Collecting en-core-web-sm==3.8.0
  Downloading https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.8.0/en_core_web_sm-3.8.0-py3-none
  ──────────────────────────────────── 12.8/12.8 MB 109.5 MB/s eta 0:00:00
✓ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
⚠ Restart to reload dependencies
If you are in a Jupyter or Colab notebook, you may need to restart Python in
order to load all the package's dependencies. You can do this by selecting the
'Restart kernel' or 'Restart runtime' option.
```

```
import pandas as pd
import spacy
from collections import Counter
import matplotlib.pyplot as plt
from spacy.matcher import Matcher
```

```
nlp = spacy.load("en_core_web_sm")
```

```
df = pd.read_csv("arxiv_data.csv", engine='python', on_bad_lines='skip')
```

```
df.columns
```

```
Index(['titles', 'summaries', 'terms'], dtype='object')
```

```
df = df.dropna(subset=["summaries"])
print(df.head())
print(df.info())
```

```
                                          titles  \
0  Survey on Semantic Stereo Matching / Semantic ...
1  FUTURE-AI: Guiding Principles and Consensus Re...
2  Enforcing Mutual Consistency of Hard Regions f...
3  Parameter Decoupling Strategy for Semi-supervi...
4  Background-Foreground Segmentation for Interio...

                                       summaries  \
0  Stereo matching is one of the widely used tech...
1  The recent advancements in artificial intellig...
2  In this paper, we proposed a novel mutual cons...
3  Consistency training has proven to be an advan...
4  To ensure safety in automated driving, the cor...

                          terms
0          ['cs.CV', 'cs.LG']
1  ['cs.CV', 'cs.AI', 'cs.LG']
2          ['cs.CV', 'cs.AI']
3                  ['cs.CV']
4          ['cs.CV', 'cs.LG']
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20815 entries, 0 to 20814
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   titles     20815 non-null  object
 1   summaries  20815 non-null  object
 2   terms      20815 non-null  object
dtypes: object(3)
memory usage: 488.0+ KB
None
```

```python
texts = df["summaries"].head(200).tolist()
```

```python
docs = [nlp(text) for text in texts]
```

```python
tokens = []
for doc in docs:
    for token in doc:
        if not token.is_stop and not token.is_punct:
            tokens.append(token.text.lower())

print(tokens[:30])
```

```
['stereo', 'matching', 'widely', 'techniques', 'inferring', 'depth', '\n', 'stereo', 'images', 'owing', 'robustness', 'speed
```

```python
noun_phrases = []

for doc in docs:
    for chunk in doc.noun_chunks:
        noun_phrases.append(chunk.text.lower())
```

```python
np_counts = Counter(noun_phrases)
top_nps = np_counts.most_common(10)

print("Top Noun Phrases:")
for np, count in top_nps:
    print(np, ":", count)
```

```
Top Noun Phrases:
we : 540
which : 172
that : 144
it : 120
this paper : 74
the-art : 72
our method : 50
image segmentation : 47
this work : 47
medical image segmentation : 37
```

```python
entities = []

for doc in docs:
    for ent in doc.ents:
        entities.append(ent.label_)
```

```
entity_counts = Counter(entities)
print(entity_counts)
```

```
Counter({'ORG': 525, 'CARDINAL': 284, 'ORDINAL': 67, 'PERSON': 60, 'PERCENT': 54, 'GPE': 44, 'NORP': 42, 'DATE': 28, 'PRODU(
```

```python
matcher = Matcher(nlp.vocab)

pattern1 = [
    {"POS": "ADJ"},
    {"POS": "NOUN"}
]

pattern2 = [
    {"POS": "NOUN"},
    {"POS": "NOUN"}
]

matcher.add("TECH_TERM", [pattern1, pattern2])
```
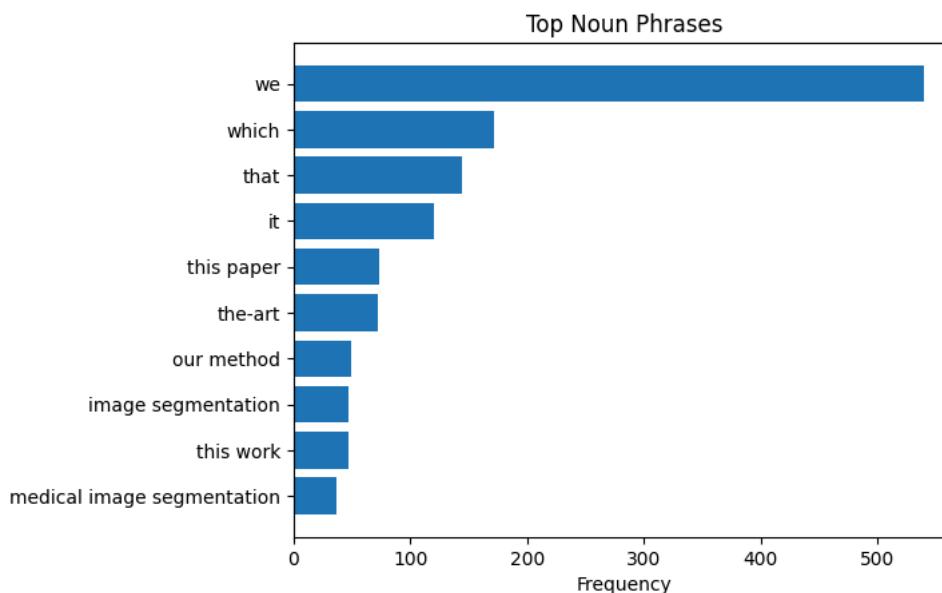
```python
matched_terms = []

for doc in docs:
    matches = matcher(doc)
    for match_id, start, end in matches:
        matched_terms.append(doc[start:end].text.lower())
```

```python
tech_counts = Counter(matched_terms)
tech_counts.most_common(10)
```

```
[('image segmentation', 206),
 ('medical image', 98),
 ('semantic segmentation', 49),
 ('deep learning', 39),
 ('segmentation tasks', 27),
 ('medical images', 25),
 ('training data', 21),
 ('neural networks', 21),
 ('experimental results', 21),
 ('computer vision', 20)]
```
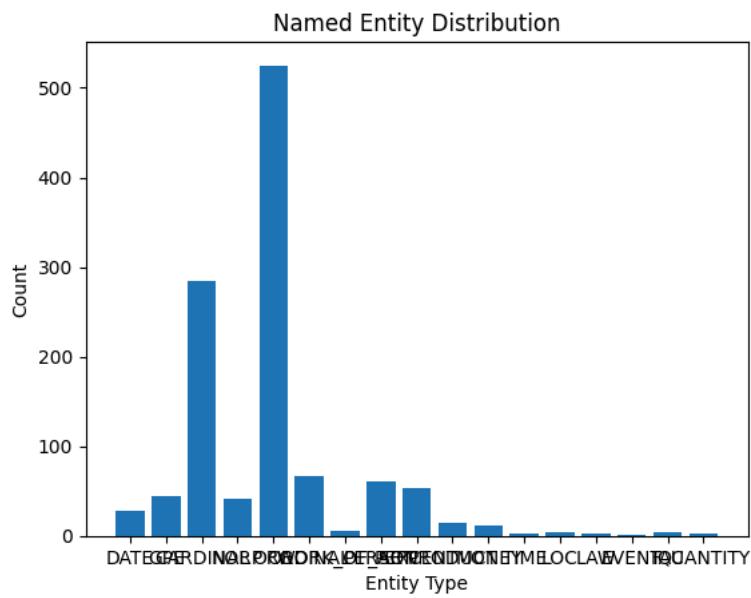
```python
labels, values = zip(*top_nps)

plt.figure()
plt.barh(labels, values)
plt.xlabel("Frequency")
plt.title("Top Noun Phrases")
plt.gca().invert_yaxis()
plt.show()
```

```
ent_labels = list(entity_counts.keys())
ent_values = list(entity_counts.values())

plt.figure()
plt.bar(ent_labels, ent_values)
plt.xlabel("Entity Type")
plt.ylabel("Count")
plt.title("Named Entity Distribution")
plt.show()
```


Start coding or generate with AI.