

# recruit

June 7, 2024

DP

```
[1]: !pip install ja-ginza
```

DEPRECATION: Loading egg at /home/jun/anaconda3/lib/python3.11/site-packages/python\_dlshogi2-0.0.0-py3.11.egg is deprecated. pip 24.3 will enforce this behaviour change. A possible replacement is to use pip for package installation.. Discussion can be found at <https://github.com/pypa/pip/issues/12330>

Collecting ja-ginza

Downloading ja\_ginza-5.2.0-py3-none-any.whl.metadata (5.8 kB)

Collecting spacy<4.0.0,>=3.4.4 (from ja-ginza)

Downloading

spacy-3.7.5-cp311-cp311-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl.metadata (27 kB)

Collecting sudachipy<0.7.0,>=0.6.2 (from ja-ginza)

Downloading SudachiPy-0.6.8-cp311-cp311-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl.metadata (12 kB)

Collecting sudachidict-core>=20210802 (from ja-ginza)

Downloading SudachiDict\_core-20240409-py3-none-any.whl.metadata (2.5 kB)

Collecting ginza<5.3.0,>=5.2.0 (from ja-ginza)

Downloading ginza-5.2.0-py3-none-any.whl.metadata (448 bytes)

Collecting plac>=1.3.3 (from ginza<5.3.0,>=5.2.0->ja-ginza)

Downloading plac-1.4.3-py2.py3-none-any.whl.metadata (5.9 kB)

Collecting spacy-legacy<3.1.0,>=3.0.11 (from spacy<4.0.0,>=3.4.4->ja-ginza)

Downloading spacy\_legacy-3.0.12-py2.py3-none-any.whl.metadata (2.8 kB)

Collecting spacy-loggers<2.0.0,>=1.0.0 (from spacy<4.0.0,>=3.4.4->ja-ginza)

Downloading spacy\_loggers-1.0.5-py3-none-any.whl.metadata (23 kB)

Collecting murmurhash<1.1.0,>=0.28.0 (from spacy<4.0.0,>=3.4.4->ja-ginza)

Downloading murmurhash-1.0.10-cp311-cp311-manylinux\_2\_5\_x86\_64.manylinux1\_x86\_64.manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl.metadata (2.0 kB)

Collecting cymem<2.1.0,>=2.0.2 (from spacy<4.0.0,>=3.4.4->ja-ginza)

Downloading

cymem-2.0.8-cp311-cp311-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl.metadata (8.4 kB)

Collecting preshed<3.1.0,>=3.0.2 (from spacy<4.0.0,>=3.4.4->ja-ginza)  
 Downloading preshed-3.0.9-cp311-cp311-manylinux\_2\_5\_x86\_64.manylinux1\_x86\_64.manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl.metadata (2.2 kB)  
 Collecting thinc<8.3.0,>=8.2.2 (from spacy<4.0.0,>=3.4.4->ja-ginza)  
 Downloading thinc-8.2.4-cp311-cp311-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl.metadata (15 kB)  
 Collecting wasabi<1.2.0,>=0.9.1 (from spacy<4.0.0,>=3.4.4->ja-ginza)  
 Downloading wasabi-1.1.3-py3-none-any.whl.metadata (28 kB)  
 Collecting srsly<3.0.0,>=2.4.3 (from spacy<4.0.0,>=3.4.4->ja-ginza)  
 Downloading srsly-2.4.8-cp311-cp311-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl.metadata (20 kB)  
 Collecting catalogue<2.1.0,>=2.0.6 (from spacy<4.0.0,>=3.4.4->ja-ginza)  
 Downloading catalogue-2.0.10-py3-none-any.whl.metadata (14 kB)  
 Collecting weasel<0.5.0,>=0.1.0 (from spacy<4.0.0,>=3.4.4->ja-ginza)  
 Downloading weasel-0.4.1-py3-none-any.whl.metadata (4.6 kB)  
 Requirement already satisfied: typer<1.0.0,>=0.3.0 in /home/jun/anaconda3/lib/python3.11/site-packages (from spacy<4.0.0,>=3.4.4->ja-ginza) (0.12.3)  
 Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /home/jun/anaconda3/lib/python3.11/site-packages (from spacy<4.0.0,>=3.4.4->ja-ginza) (4.66.2)  
 Requirement already satisfied: requests<3.0.0,>=2.13.0 in /home/jun/anaconda3/lib/python3.11/site-packages (from spacy<4.0.0,>=3.4.4->ja-ginza) (2.31.0)  
 Requirement already satisfied: pydantic!=1.8,!1.8.1,<3.0.0,>=1.7.4 in /home/jun/anaconda3/lib/python3.11/site-packages (from spacy<4.0.0,>=3.4.4->ja-ginza) (2.7.1)  
 Requirement already satisfied: jinja2 in /home/jun/anaconda3/lib/python3.11/site-packages (from spacy<4.0.0,>=3.4.4->ja-ginza) (3.1.3)  
 Requirement already satisfied: setuptools in /home/jun/anaconda3/lib/python3.11/site-packages (from spacy<4.0.0,>=3.4.4->ja-ginza) (69.5.1)  
 Requirement already satisfied: packaging>=20.0 in /home/jun/anaconda3/lib/python3.11/site-packages (from spacy<4.0.0,>=3.4.4->ja-ginza) (23.1)  
 Collecting langcodes<4.0.0,>=3.2.0 (from spacy<4.0.0,>=3.4.4->ja-ginza)  
 Downloading langcodes-3.4.0-py3-none-any.whl.metadata (29 kB)  
 Requirement already satisfied: numpy>=1.19.0 in /home/jun/anaconda3/lib/python3.11/site-packages (from spacy<4.0.0,>=3.4.4->ja-ginza) (1.26.4)  
 Collecting language-data>=1.2 (from langcodes<4.0.0,>=3.2.0->spacy<4.0.0,>=3.4.4->ja-ginza)  
 Downloading language\_data-1.2.0-py3-none-any.whl.metadata (4.3 kB)  
 Requirement already satisfied: annotated-types>=0.4.0 in

```

/home/jun/anaconda3/lib/python3.11/site-packages (from
pydantic!=1.8,!1.8.1,<3.0.0,>=1.7.4->spacy<4.0.0,>=3.4.4->ja-ginza) (0.6.0)
Requirement already satisfied: pydantic-core==2.18.2 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
pydantic!=1.8,!1.8.1,<3.0.0,>=1.7.4->spacy<4.0.0,>=3.4.4->ja-ginza) (2.18.2)
Requirement already satisfied: typing-extensions>=4.6.1 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
pydantic!=1.8,!1.8.1,<3.0.0,>=1.7.4->spacy<4.0.0,>=3.4.4->ja-ginza) (4.11.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.4.4->ja-ginza) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.4.4->ja-ginza) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.4.4->ja-ginza) (2.2.1)
Requirement already satisfied: certifi>=2017.4.17 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.4.4->ja-ginza) (2024.2.2)
Collecting blis<0.8.0,>=0.7.8 (from
thinc<8.3.0,>=8.2.2->spacy<4.0.0,>=3.4.4->ja-ginza)
  Downloading
blis-0.7.11-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata
(7.4 kB)
Collecting confection<1.0.0,>=0.0.1 (from
thinc<8.3.0,>=8.2.2->spacy<4.0.0,>=3.4.4->ja-ginza)
  Downloading confection-0.1.5-py3-none-any.whl.metadata (19 kB)
Requirement already satisfied: click>=8.0.0 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
typer<1.0.0,>=0.3.0->spacy<4.0.0,>=3.4.4->ja-ginza) (8.1.7)
Requirement already satisfied: shellingham>=1.3.0 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
typer<1.0.0,>=0.3.0->spacy<4.0.0,>=3.4.4->ja-ginza) (1.5.4)
Requirement already satisfied: rich>=10.11.0 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
typer<1.0.0,>=0.3.0->spacy<4.0.0,>=3.4.4->ja-ginza) (13.3.5)
Collecting cloudpathlib<1.0.0,>=0.7.0 (from
weasel<0.5.0,>=0.1.0->spacy<4.0.0,>=3.4.4->ja-ginza)
  Downloading cloudpathlib-0.18.1-py3-none-any.whl.metadata (14 kB)
Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
weasel<0.5.0,>=0.1.0->spacy<4.0.0,>=3.4.4->ja-ginza) (5.2.1)
Requirement already satisfied: MarkupSafe>=2.0 in
/home/jun/anaconda3/lib/python3.11/site-packages (from
jinja2->spacy<4.0.0,>=3.4.4->ja-ginza) (2.1.5)
Collecting marisa-trie>=0.7.7 (from language-
data>=1.2->langcodes<4.0.0,>=3.2.0->spacy<4.0.0,>=3.4.4->ja-ginza)

```

Downloading marisa\_trie-1.1.1-cp311-cp311-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl.metadata (8.6 kB)  
 Requirement already satisfied: markdown-it-py<3.0.0,>=2.2.0 in /home/jun/anaconda3/lib/python3.11/site-packages (from rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy<4.0.0,>=3.4.4->ja-ginza) (2.2.0)  
 Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /home/jun/anaconda3/lib/python3.11/site-packages (from rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy<4.0.0,>=3.4.4->ja-ginza) (2.17.2)  
 Requirement already satisfied: mdurl~=0.1 in /home/jun/anaconda3/lib/python3.11/site-packages (from markdown-it-py<3.0.0,>=2.2.0->rich>=10.11.0->typer<1.0.0,>=0.3.0->spacy<4.0.0,>=3.4.4->ja-ginza) (0.1.2)  
 Downloading ja\_ginza-5.2.0-py3-none-any.whl (59.1 MB)  
     59.1/59.1 MB  
 56.8 MB/s eta 0:00:00:00:0100:01  
 Downloading ginza-5.2.0-py3-none-any.whl (21 kB)  
 Downloading  
 spacy-3.7.5-cp311-cp311-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl (6.6 MB)  
     6.6/6.6 MB  
 92.5 MB/s eta 0:00:00:00:0100:01  
 Downloading SudachiDict\_core-20240409-py3-none-any.whl (72.0 MB)  
     72.0/72.0 MB  
 29.7 MB/s eta 0:00:0000:0100:01m  
 Downloading  
 SudachiPy-0.6.8-cp311-cp311-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl (2.6 MB)  
     2.6/2.6 MB  
 81.6 MB/s eta 0:00:00  
 Downloading catalogue-2.0.10-py3-none-any.whl (17 kB)  
 Downloading  
 cymem-2.0.8-cp311-cp311-manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl (46 kB)  
     46.3/46.3 kB  
 9.9 MB/s eta 0:00:00  
 Downloading langcodes-3.4.0-py3-none-any.whl (182 kB)  
     182.0/182.0 kB  
 22.5 MB/s eta 0:00:00  
 Downloading murmurhash-1.0.10-cp311-cp311-manylinux\_2\_5\_x86\_64.manylinux1\_x86\_64.manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl (29 kB)  
 Downloading plac-1.4.3-py2.py3-none-any.whl (22 kB)  
 Downloading preshed-3.0.9-cp311-cp311-manylinux\_2\_5\_x86\_64.manylinux1\_x86\_64.manylinux\_2\_17\_x86\_64.manylinux2014\_x86\_64.whl (157 kB)  
     157.2/157.2 kB  
 19.8 MB/s eta 0:00:00  
 Downloading spacy\_legacy-3.0.12-py2.py3-none-any.whl (29 kB)  
 Downloading spacy\_loggers-1.0.5-py3-none-any.whl (22 kB)  
 Downloading

```

srsly-2.4.8-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (490 kB)
490.9/490.9 kB
68.4 MB/s eta 0:00:00
Downloading
thinc-8.2.4-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (920 kB)
920.1/920.1 kB
75.6 MB/s eta 0:00:00
Downloading wasabi-1.1.3-py3-none-any.whl (27 kB)
Downloading weasel-0.4.1-py3-none-any.whl (50 kB)
50.3/50.3 kB
11.4 MB/s eta 0:00:00
Downloading
blis-0.7.11-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (10.2 MB)
10.2/10.2 MB
66.9 MB/s eta 0:00:000:010:01
Downloading cloudpathlib-0.18.1-py3-none-any.whl (47 kB)
47.3/47.3 kB
10.7 MB/s eta 0:00:00
Downloading confection-0.1.5-py3-none-any.whl (35 kB)
Downloading language_data-1.2.0-py3-none-any.whl (5.4 MB)
5.4/5.4 MB
90.1 MB/s eta 0:00:00ta 0:00:01
Downloading
marisa_trie-1.1.1-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl
(1.4 MB)
1.4/1.4 MB
81.1 MB/s eta 0:00:00
Installing collected packages: sudachipy, plac, cymem, wasabi,
sudachidict-core, spacy-loggers, spacy-legacy, murmurhash, marisa-trie,
cloudpathlib, catalogue, blis, srsly, preshed, language-data, langcodes,
confection, weasel, thinc, spacy, ginza, ja-ginza
Successfully installed blis-0.7.11 catalogue-2.0.10 cloudpathlib-0.18.1
confection-0.1.5 cymem-2.0.8 ginza-5.2.0 ja-ginza-5.2.0 langcodes-3.4.0
language-data-1.2.0 marisa-trie-1.1.1 murmurhash-1.0.10 plac-1.4.3 preshed-3.0.9
spacy-3.7.5 spacy-legacy-3.0.12 spacy-loggers-1.0.5 srsly-2.4.8 sudachidict-
core-20240409 sudachipy-0.6.8 thinc-8.2.4 wasabi-1.1.3 weasel-0.4.1

```

```

[2]: import pandas as pd
data_path = &quot;&lt;&lt;YOUR&gt;&gt;/&lt;&lt;PATH&gt;&gt;/&lt;&lt;TO&gt;&gt;/reviews.csv&quot;
df = pd.read_csv(data_path)
df

```

Cell In[2], line 2

```

data_path = &quot;&lt;&lt;YOUR&gt;&gt;/&lt;&lt;PATH&gt;&gt;/&lt;&lt;TO&gt;&gt;/reviews.csv&quot;
~

```

SyntaxError: invalid syntax

```
[ ]: from sklearn.feature_extraction.text import CountVectorizer
import itertools
import spacy
from typing import List, Tuple
nlp = spacy.load('ja_ginza')
POS = ['ADJ', 'ADV', 'INTJ', 'PROPN', 'NOUN', 'VERB']
MAX_TERMS_IN_DOC = 5
NGRAM=1
MAX_DF=1.0
MIN_DF=0.0
NUM_VOCAB=10000
def flatten(*lists) -> list:
    res = []
    for l in list(itertools.chain.from_iterable(lists)):
        for e in l:
            res.append(e)
    return res
def remove_duplicates(l: List[Tuple[str, float]]) -> List[Tuple[str, float]]:
    d = {}
    for e in l:
        d[e[0]] = e[1]
    return list(d.items())
df['doc'] = [nlp(review) for review in df['review']]
bows = {}
cvs = {}
for sentiment in df['sentiment'].unique():
    tokens = []
    for doc in df[df['sentiment'] == sentiment]['doc']:
        similarities = [(token.similarity(doc), token.lemma_) for token in doc if
        token.pos_ in POS]
        similarities = remove_duplicates(similarities)
        similarities = sorted(similarities, key=lambda sim: sim[1], reverse=True)[:
        MAX_TERMS_IN_DOC]
        tokens.append([similarity[1] for similarity in similarities])
    cv = CountVectorizer(ngram_range=(1,NGRAM), max_df=MAX_DF, min_df=MIN_DF,
    max_features=NUM_VOCAB)
    bows[sentiment] = cv.fit_transform(flatten(tokens)).toarray()
    cvs[sentiment] = cv
```

```
[4]: TOP_K = 20
vocab = {}
term_frequencies = {}
for sentiment in df['sentiment'].unique():
    bow = bows[sentiment]
```

```

cv = cvs[sentiment]
vocab = cv.vocabulary_
term_frequency = np.sum(bow, axis=0)
vocab[sentiment] = vocab
term_frequencies[sentiment] = term_frequency
indices_topk = np.argsort(term_frequency)[:,-1][:TOP_K]
bow_topk = np.take(bow, indices_topk, axis=1)
reverse_vocab = { vocab[k]:k for k in vocab.keys() }
words = [reverse_vocab[i] for i in indices_topk]
print(sentiment, &quot;;&quot;)
for w, c in zip(words, term_frequency[indices_topk]):
    print(w, &quot;;&quot;, c)

```

Cell In[4], line 4

```

for sentiment in df[&quot;sentiment&quot;].unique():
    ^

```

SyntaxError: invalid syntax