June 7, 2024

[3]:
```
!pip install ja-ginza
```

```
Requirement already satisfied: ja-ginza in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (5.2.0)
Requirement already satisfied: spacy<4.0.0,>=3.4.4 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from ja-ginza)
(3.7.5)
Requirement already satisfied: sudachipy<0.7.0,>=0.6.2 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from ja-ginza)
(0.6.8)
Requirement already satisfied: sudachidict-core>=20210802 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from ja-ginza)
(20240409)
Requirement already satisfied: ginza<5.3.0,>=5.2.0 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from ja-ginza)
(5.2.0)
Requirement already satisfied: plac>=1.3.3 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
ginza<5.3.0,>=5.2.0->ja-ginza) (1.4.3)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (1.0.5)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (1.0.10)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (2.0.8)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (3.0.9)
Requirement already satisfied: thinc<8.3.0,>=8.2.2 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (8.2.4)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in
```

```
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (1.1.3)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (2.0.10)
Requirement already satisfied: weasel<0.5.0,>=0.1.0 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (0.4.1)
Requirement already satisfied: typer<1.0.0,>=0.3.0 in
/home/jun/.pyenv/versions/3.11.8/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/.pyenv/versions/3.11.8/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/.pyenv/versions/3.11.8/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/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (2.7.0)
Requirement already satisfied: jinja2 in
/home/jun/.pyenv/versions/3.11.8/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/.pyenv/versions/3.11.8/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/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (23.2)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (3.4.0)
Requirement already satisfied: numpy>=1.19.0 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
spacy<4.0.0,>=3.4.4->ja-ginza) (1.26.4)
Requirement already satisfied: language-data>=1.2 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
langcodes<4.0.0,>=3.2.0->spacy<4.0.0,>=3.4.4->ja-ginza) (1.2.0)
Requirement already satisfied: annotated-types>=0.4.0 in
/home/jun/.pyenv/versions/3.11.8/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==2.18.1 in
/home/jun/.pyenv/versions/3.11.8/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) (2.18.1)
Requirement already satisfied: typing-extensions>=4.6.1 in
```

```
/home/jun/.pyenv/versions/3.11.8/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/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.4.4->ja-ginza) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.4.4->ja-ginza) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
requests<3.0.0,>=2.13.0->spacy<4.0.0,>=3.4.4->ja-ginza) (1.26.18)
Requirement already satisfied: certifi>=2017.4.17 in
/home/jun/.pyenv/versions/3.11.8/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)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
thinc<8.3.0,>=8.2.2->spacy<4.0.0,>=3.4.4->ja-ginza) (0.7.11)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
thinc<8.3.0,>=8.2.2->spacy<4.0.0,>=3.4.4->ja-ginza) (0.1.5)
Requirement already satisfied: click>=8.0.0 in
/home/jun/.pyenv/versions/3.11.8/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/.pyenv/versions/3.11.8/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/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
typer<1.0.0,>=0.3.0->spacy<4.0.0,>=3.4.4->ja-ginza) (13.7.1)
Requirement already satisfied: cloudpathlib<1.0.0,>=0.7.0 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
weasel<0.5.0,>=0.1.0->spacy<4.0.0,>=3.4.4->ja-ginza) (0.18.1)
Requirement already satisfied: smart-open<8.0.0,>=5.2.1 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
weasel<0.5.0,>=0.1.0->spacy<4.0.0,>=3.4.4->ja-ginza) (7.0.4)
Requirement already satisfied: MarkupSafe>=2.0 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from
jinja2->spacy<4.0.0,>=3.4.4->ja-ginza) (2.1.5)
Requirement already satisfied: marisa-trie>=0.7.7 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from language-
data>=1.2->langcodes<4.0.0,>=3.2.0->spacy<4.0.0,>=3.4.4->ja-ginza) (1.2.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/home/jun/.pyenv/versions/3.11.8/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) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
/home/jun/.pyenv/versions/3.11.8/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: wrapt in
```

```
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from smart-
open<8.0.0,>=5.2.1->weasel<0.5.0,>=0.1.0->spacy<4.0.0,>=3.4.4->ja-ginza)
(1.16.0)
Requirement already satisfied: mdurl~=0.1 in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (from markdown-it-
py>=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)
```

[9]:

```python
import pandas as pd
import spacy
import itertools
from sklearn.feature_extraction.text import CountVectorizer
from typing import List, Tuple

# CSV
import pandas as pd
import spacy
import itertools
from sklearn.feature_extraction.text import CountVectorizer
from typing import List, Tuple

# CSV
data_path = './data/reviews_with_sentiment.csv'
df = pd.read_csv(data_path)

# spaCy
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:
    return list(itertools.chain.from_iterable(lists))

def remove_duplicates(l: List[Tuple[str, float]]) -> List[Tuple[str, float]]:
    return list({e[0]: e[1] for e in l}.items())

#
df["doc"] = [nlp(review) for review in df["review"]]

# Bag-of-Words
```

```
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 and token.has_vector]
        similarities = remove_duplicates(similarities)
        similarities = sorted(similarities, key=lambda sim: sim[0],␣
  ↪reverse=True)[:MAX_TERMS_IN_DOC]
        tokens.append([similarity[1] for similarity in similarities])
    flattened_tokens = [' '.join(token_list) for token_list in tokens]
    cv = CountVectorizer(ngram_range=(1, NGRAM), max_df=MAX_DF, min_df=MIN_DF,␣
  ↪max_features=NUM_VOCAB)
    bows[sentiment] = cv.fit_transform(flattened_tokens).toarray()
    cvs[sentiment] = cv
df = pd.read_csv(data_path)

# spaCy
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:
    return list(itertools.chain.from_iterable(lists))

def remove_duplicates(l: List[Tuple[str, float]]) -> List[Tuple[str, float]]:
    return list({e[0]: e[1] for e in l}.items())

#
df["doc"] = [nlp(review) for review in df["review"]]

# Bag-of-Words
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 and token.has_vector]
```

```
        similarities = remove_duplicates(similarities)
        similarities = sorted(similarities, key=lambda sim: sim[0],␣
 ↪reverse=True)[:MAX_TERMS_IN_DOC]
        tokens.append([similarity[1] for similarity in similarities])
    flattened_tokens = [' '.join(token_list) for token_list in tokens]
    cv = CountVectorizer(ngram_range=(1, NGRAM), max_df=MAX_DF, min_df=MIN_DF,␣
 ↪max_features=NUM_VOCAB)
    bows[sentiment] = cv.fit_transform(flattened_tokens).toarray()
    cvs[sentiment] = cv
```

[ ]:

```
import numpy as np

TOP_K = 20

#
vocabs = {}
term_frequencies = {}
for sentiment in df["sentiment"].unique():
    bow = bows[sentiment]
    cv = cvs[sentiment]
    vocab = cv.vocabulary_
    term_frequency = np.sum(bow, axis=0)
    vocabs[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(f"{sentiment} :")
    for w, c in zip(words, term_frequency[indices_topk]):
        print(w, ":", c)
```

[ ]:

```
!pip install python-dp
```

```
Requirement already satisfied: python-dp in
/home/jun/.pyenv/versions/3.11.8/lib/python3.11/site-packages (1.1.4)
```

[ ]:

```
import numpy as np
from pydp.algorithms.laplacian import Count
from typing import List, Tuple
```

[ ]:

```
def preprocess_for_private_counts(tf: np.ndarray) -> List[np.ndarray]:
    repeated_words = []
    for i, term in enumerate(tf):
        repeated_words.append(np.repeat(i, term))
    return repeated_words
```

[ ]:

```
def cal_private_count(
    epsilon: float,
    max_partition_contributed: float,
    max_contributions_per_partition: float,
    repeated_words: List[np.ndarray],
) -> List[int]:
    private_counts = []
    for repeated_word in repeated_words:
        counter = Count(epsilon, max_partition_contributed,␣
 ↪max_contributions_per_partition)
        count = counter.quick_result(repeated_word)
        private_counts.append(count)
    return private_counts

def top_k_words_and_counts(k: int, tf: np.ndarray, vocab: dict) ->␣
 ↪List[Tuple[str, int]]:
    indices_topk = np.argsort(tf)[::-1][:k]
    reverse_vocab = {vocab[key]: key for key in vocab.keys()}
    words = [reverse_vocab[i] for i in indices_topk]
    counts = [tf[i] for i in indices_topk]
    return list(zip(words, counts))
```

[ ]:

```
epsilons = [0.01, 0.05, 0.1, 0.3, 0.7, 1.0, 2.0, 3.0, 7.0, 10.0]
MAX_DUPLICATED_TERMS = 1

for eps in epsilons:
    print(f" : {eps}")
    for sentiment in df["sentiment"].unique():
        repeated_words =␣
 ↪preprocess_for_private_counts(term_frequencies[sentiment])
        private_counts = cal_private_count(eps, MAX_TERMS_IN_DOC,␣
 ↪MAX_DUPLICATED_TERMS, repeated_words)
        words_and_counts = top_k_words_and_counts(TOP_K, private_counts,␣
 ↪vocabs[sentiment])
        print(f"{sentiment} : {words_and_counts}")
```