

```
In [1]: import math
import pandas as pd
import numpy as np
from sklearn.datasets import fetch_rcv1
import time
from scipy import sparse
```

```
In [2]: def freq(x, prob: bool = True) -> list:
    if type(x) == sparse.csr_matrix or type(x) == sparse.csc_matrix:
        nonzero = x.nonzero()[0]
        uniques = set(nonzero)

        count_nonzero = len(nonzero)
        counts = {
            0: x.shape[0] - count_nonzero,
            1: count_nonzero
        }
        total = sum(counts.values())

    return [uniques, counts if prob is False else {key: val / total for key, val in counts.items()}]

    counts = {}
    uniques = []

    for val in x:
        if val not in uniques:
            uniques.append(val)

        if val in counts.keys():
            counts[val] += 1
            continue

        counts[val] = 1

    total = sum(counts.values())
    return [uniques, counts if prob is False else {key: val / total for key, val in counts.items()}]
```

Poniższa funkcja zwraca dla zadanych kolumn danych X i Y unikalne wartości atrybutów X oraz Y, a także w zależności od parametru **prob** zwraca łączny rozkład częstości lub prawdopodobieństwa łączne liczności. Funkcja obsługuje także macierze rzadkie, w których dane zostały binaryzowane - przedstawiają jedynie sam fakt wystąpienia. Możliwe przyspieszenie obliczeń dla tej funkcji po zredukowaniu rozwiązania do pojedynczej pętli zamiast podwójnej dla kolumn danych X i Y.

```
In [3]: def freq2(x, y, prob: bool = True) -> list:
    if (type(x) == sparse.csr_matrix or type(x) == sparse.csc_matrix) and (type(y) == sparse.csr_matrix or type(y) == sparse.csc_matrix):
        x_nonzero = x.nonzero()[0]
        y_nonzero = y.nonzero()[0]

        uniques_x = set(x_nonzero)
        uniques_y = set(y_nonzero)
        intersection_x_y = uniques_x.intersection(uniques_y)

        count_intersection = len(intersection_x_y)
        count_x_nonzero = len(x_nonzero)
        count_y_nonzero = len(y_nonzero)
        count_shared_zeros = x.shape[0] - count_x_nonzero + y.shape[0] - count_y_nonzero

        counts = {

    counts = {
```

```

        (0, 0): count_shared_zeros,
        (0, 1): count_y_nonzero - count_intersection,
        (1, 0): count_x_nonzero - count_intersection,
        (1, 1): count_intersection
    }
    total = sum(counts.values())

    return [uniques_x, uniques_y, counts if prob is False else {key: val / total: value}]

counts = {}
uniques = {'x': [], 'y': []}

for x_val in x:
    if x_val not in uniques['x']:
        uniques['x'].append(x_val)

    for y_val in y:
        key = (x_val, y_val)

        if key not in counts.keys():
            counts[key] = 1

            if y_val not in uniques['y']:
                uniques['y'].append(y_val)
            else:
                counts[key] += 1

        total = sum(counts.values())
    return [uniques['x'], uniques['y'], counts if prob is False else {key: val / total: value}]

```

Dla entropii łącznej został obsłużony przypadek, gdy prawdopodobieństwo łączne liczności byłoby równe 0 - np. brak wspólnych elementów w kolumnach danych X i Y będące zbinaryzowanymi macierzami rzadkimi (klucz (1, 1)).

```
In [4]: def entropy(x, y=None, conditional_reverse: bool = False):
    if y is None:
        uniques, probs = freq(x)
    else:
        uniques_x, uniques_y, probs = freq2(x, y)

        if conditional_reverse is True and y is not None:
            uniques_x, probs_x = freq(y)
            entropy_y = entropy(y)

            return sum(prob * entropy_y for prob in probs_x.values())

    return -sum(prob * math.log2(prob) if prob != 0 else 0 for prob in probs.values)
```

Poniżej wersja dla entropii infogain:

```
In [5]: def infogain(x, y, reverse: bool = False):
    if reverse is False:
        return entropy(x) + entropy(y) - entropy(x, y)
    return entropy(y) - entropy(x, y, conditional_reverse=True)
```

```
In [6]: def kappa(x, y):
    return infogain(x, y) / entropy(y)
```

W poniżej funkcji odpowiadającej wyliczaniu indeksu Giniego został obsłużony przypadek dla indeksu Giniego warunkowego ($Y|X$), bazując na funkcji freq2.

```
In [7]: def gini(x, y=None, conditional_reverse: bool = False):
    if y is None:
        uniques, probs = freq(x)
    else:
        uniques_x, uniques_y, probs = freq2(x, y)

    if conditional_reverse is True and y is not None:
        uniques, probs = freq(x)
        gini_y = gini(y)
        return sum(prob * gini_y for prob in probs.values())

    return 1 - sum(prob ** 2 for prob in probs.values())
```

```
In [8]: def ginigain(x, y):
    return gini(y) - gini(x, y, True)
```

Pomimo, że entropia i indeks Giniego mierzą podobne miary dotyczące informacji, to wyniki entropii są większe od wyników indeksu Giniego. Wskaźnik Giniego jest wykorzystywany przez algorytm CART (drzewo klasyfikacji i regresji), natomiast przyrost informacji poprzez redukcję entropii jest wykorzystywany przez algorytmy takie jak C4.5 (algorytmy do generowania drzewa decyzyjnego).

Poniższa funkcja jest odpowiedzialna za dokonanie eksperymentu na wczytanych danych ze zbioru danych **Reuters Corpus Volume I**. Dostarczane są tutaj kolumny danych jako sparse.csc_matrix (preferowane do operacji kolumnowych) lub sparse.csr_matrix (preferowane do operacji wierszowych). Czas jest mierzony dla obliczenia ilości informacji na temat wybranej zmiennej decyzyjnej `y` oraz dla sortowania wyników i wybierania 50 najlepszych rezultatów. Czasy są wypisywane na ekranie i zwracane są posortowane rezultaty.

```
In [9]: def experiment_best_50_words(x, y):
    info_gain_words = []

    time_info_gain1 = time.time()
    for i, word in enumerate(word_list['A']):
        gain = infogain(x[:, i], y)
        info_gain_words.append({'infogain': gain, 'word': word, 'id': i})
    time_info_gain2 = time.time()

    time_sort_and_select_the_best1 = time.time()
    info_gain_words = sorted(info_gain_words, key=lambda item: item['infogain'], reverse=True)
    best50 = info_gain_words[:50]
    time_sort_and_select_the_best2 = time.time()

    print(f'time of calculating info gain: {time_info_gain2 - time_info_gain1}s')
    print(f'time of sort results and select 50 of the best: {time_sort_and_select_the_best2 - time_sort_and_select_the_best1}s')

    return best50
```

Poniżej wczytanie zbioru autos oraz wypisanie posortowanych atrybutów od najwyższych wartości przyrostu informacji. Tutaj została zastosowana miara entropii:

```
In [10]: autos = pd.read_csv('D:\Programming\Python\computational-intelligence\machine-learning\autos.csv')
info_gains = {key: entropy(autos[key]) for key in autos.columns}
print(sorted(info_gains.items(), key=lambda x: x[1], reverse=True))
```

```
[('animal', 6.638409502553759), ('type', 2.390559682294039), ('legs', 2.0338113440641234), ('predator', 0.9914266810680207), ('catsize', 0.9880162151534646), ('hair', 0.9840304711717017), ('eggs', 0.9794662187017298), ('milk', 0.9743197211096903), ('toothed', 0.9685867165455516), ('aquatic', 0.9396846718728563), ('tail', 0.8228368841492257), ('airborne', 0.7910662980902585), ('breathes', 0.7374895672137456), ('feathers', 0.7179499765002912), ('backbone', 0.6761627418829198), ('fins', 0.653839880626333), ('domestic', 0.5538976334852962), ('venomous', 0.3993820824245975)]
```

Wczytanie danych i ich binaryzacja. Tutaj został wybrany atrybut decyzyjny pod indeksem 98, który odpowiada **M14 COMMODITY MARKETS** (rynki towarowe):

```
In [11]: rcv1 = fetch_rcv1()
X = rcv1['data'] > 0
Xr = X[:, 2]
Y = rcv1['target'][:, 98]
```

Wczytanie listy słów niezbędnych do wyznaczenia najlepszych słów w eksperymencie:

```
In [12]: word_list = pd.read_csv('D:\Programming\Python\computational-intelligence\machine-')
```

Wykonanie eksperimentu. Domyślnie dane z bazy danych **Reuters Corpus Volume I** wczytane za pomocą `fetch_rcv1()`, są wczytywane jako format wierszowy (sparse.csr_matrix).

```
In [13]: best50 = experiment_best_50_words(X, Y)
```

```
time of calculating info gain: 4253.585737466812s
time of sort results and select 50 of the best: 0.008523702621459961s
```

Czas wykonania eksperimentu dla danych ułożonych wierszowo jest znaczny. Wynika to z wykonywania operacji kolumnowych używanych w powyższych funkcjach na dostarczonych danych.

Zmiana formatu dla wczytanych danych z formatu wierszowego na kolumnowy i ponowne wykonanie eksperimentu:

```
In [14]: X = X.tocsc()
Y = Y.tocsc()
best50 = experiment_best_50_words(X, Y)
```

```
time of calculating info gain: 559.7162163257599s
time of sort results and select 50 of the best: 0.009001016616821289s
```

Czas wykonania eksperimentu dla danych ułożonych kolumnowo jest o wiele mniejszy. Jest on ponad 7 razy krótszy od czasu dla eksperimentu wykonanego na danych wierszowych. Wąskim gardłem wydajnościowym jak widać po powyższych czasach jest sposób ułożenia danych i ich przedstawienia pod daną strukturą danych.

Przygotowanie danych do wyświetlenia i wyświetlenie 50 zmiennych (słów) dostarczających najwięcej informacji na temat wybranej zmiennej decyzyjnej - w tym przypadku zmienna decyzyjna pod indeksem 98.

```
In [15]: data_to_display = [f'word: `{best["word"]}`', infogain: {best["infogain"]}], k={kappa}
print(np.array(data_to_display))
```

```
[ 'word: `trad`, infogain: 0.4872280008335079, k=0.9976219097905742'
'word: `pric`, infogain: 0.48174961945370587, k=0.9864046700479289'
'word: `market`, infogain: 0.47211915316869535, k=0.9666858440547796'
'word: `shar`, infogain: 0.460514696809748, k=0.9429251819107942'
'word: `high`, infogain: 0.45957283160081697, k=0.9409966692495932'
'word: `week`, infogain: 0.4593571723372246, k=0.9405550969137648'
'word: `compan`, infogain: 0.45861371716875565, k=0.9390328380047062'
'word: `newsroom`, infogain: 0.45734535087527284, k=0.9364357992428664'
'word: `month`, infogain: 0.45529170137032693, k=0.9322308567156311'
'word: `end`, infogain: 0.45500980413160386, k=0.9316536590562633'
'word: `expect`, infogain: 0.45332242093495645, k=0.9281986637680036'
'word: `stat`, infogain: 0.45295961314841093, k=0.9274557980125847'
'word: `million`, infogain: 0.4500940545086355, k=0.9215884339080194'
'word: `cent`, infogain: 0.44883310344006555, k=0.9190065781628094'
'word: `percent`, infogain: 0.44815444715431174, k=0.917616997968882'
'word: `report`, infogain: 0.44750059202576675, k=0.9162781992936221'
'word: `bank`, infogain: 0.44699268292353933, k=0.9152382318703678'
'word: `offic`, infogain: 0.4468949272347006, k=0.9150380725674843'
'word: `day`, infogain: 0.4435033617873948, k=0.9080936851494116'
'word: `tuesday`, infogain: 0.4433433995584578, k=0.9077661550730336'
'word: `low`, infogain: 0.4418515686066401, k=0.9047115620679291'
'word: `wednesday`, infogain: 0.4417049050375834, k=0.9044112616139143'
'word: `thursday`, infogain: 0.44134161395455274, k=0.9036674062865461'
'word: `clos`, infogain: 0.4403480309070217, k=0.9016329989542134'
'word: `govern`, infogain: 0.44022680615622933, k=0.9013847856594265'
'word: `early`, infogain: 0.4394743318959091, k=0.8998440597418588'
'word: `add`, infogain: 0.43772683504110554, k=0.8962659789528651'
'word: `billion`, infogain: 0.4373485429747803, k=0.8954914084170577'
'word: `monday`, infogain: 0.4371560853146327, k=0.8950973424394417'
'word: `friday`, infogain: 0.43571452606615213, k=0.8921456830765044'
'word: `produc`, infogain: 0.432930559928715, k=0.8864453834014631'
'word: `told`, infogain: 0.43111815566877143, k=0.8827344016926443'
'word: `time`, infogain: 0.4309080275533217, k=0.8823041546389506'
'word: `group`, infogain: 0.42962642635401527, k=0.8796800167941332'
'word: `lead`, infogain: 0.4289701783861206, k=0.878336318669421'
'word: `financ`, infogain: 0.42840790944035056, k=0.877185046947485'
'word: `sale`, infogain: 0.4267444627014415, k=0.8737790626656571'
'word: `rate`, infogain: 0.4254896473117499, k=0.871209770007277'
'word: `plan`, infogain: 0.4250036002690255, k=0.8702145661639402'
'word: `invest`, infogain: 0.42168054972928226, k=0.863410466194074'
'word: `stock`, infogain: 0.42166034819688003, k=0.863369102620344'
'word: `operat`, infogain: 0.4206884704708995, k=0.861379137939708'
'word: `deal`, infogain: 0.42054089454745225, k=0.8610769693978'
'word: `unit`, infogain: 0.42042397442666646, k=0.8608375700324274'
'word: `minist`, infogain: 0.419058010925605, k=0.8580406964654282'
'word: `make`, infogain: 0.41820862908552625, k=0.8563015477876199'
'word: `issu`, infogain: 0.41800029593696253, k=0.8558749760118038'
'word: `nation`, infogain: 0.41631750269512247, k=0.8524293788687103'
'word: `includ`, infogain: 0.41436274453616284, k=0.8484269209549944'
'word: `buy`, infogain: 0.4129776286812684, k=0.8455908320560045' ]
```

Wnioski:

- Patrząc na wyniki, słowa przynoszące najwięcej informacji dla tematu 'rynki towarowe' zostały wybrane dobrze, jednak mogły zostać wybrane lepiej. Słowa niosące najwięcej informacji zostały wybrane trafnie, jednak słowa takie jak 'plan', 'invest', 'stock', 'buy' powinny mieć wyższą entropię od słów takich jak 'tuesday', 'million', 'high'.
- Spoglądając na wyniki, współczynnik `k` liczony przez funkcję `kappa` jest ponad dwukrotnie większy od funkcji przyrostu informacji `infogain`.
- Zdecydowanie szyszne operacje w przypadku tego zadania laboratoryjnego są w przypadku reprezentacji danych kolumnowo. Dla wierszowej reprezentacji obliczenia

trwają bardzo długo.

- Długość obliczeń jest także zależna od ilości zagnieżdzonych pętli. Zredukowanie kodu w funkcji `freq2` z podwójnej zagnieżdzonej pętli do pojedynczej pętli bez zagnieżczeń mogłoby przyspieszyć obliczenia w przypadku danych nie będących reprezentacjami macierzy rzadkich - macierze rzadkie zostały obsłużone w możliwie najszybszy sposób.
- Pomimo, że entropia i indeks Giniego mierzą podobne miary dotyczące informacji, to wyniki entropii są większe od wyników indeksu Giniego. Wskaźnik Giniego jest wykorzystywany przez algorytm CART (drzewo klasyfikacji i regresji), natomiast przyrost informacji poprzez redukcję entropii jest wykorzystywany przez algorytmy takie jak C4.5 (algorytmy do generowania drzewa decyzyjnego).