Eksploracja danych internetowych

Laboratorium 3

Prowadzący: pracownik UR
Wykonał: Piotr Rojek, pr125159

Zadanie 1
Na wybranym przez siebie zbiorze danych wykonaj i udokumentuj operacje wektoryzacji zbioru tekstowego przy użyciu:

- **Count Vectorizer**
- **N-Gram**
- **TF-IDF**

```
import pandas as pd
import nltk
import re
import string
from nltk.stem import PorterStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer

data = pd.read_csv( filepath_or_buffer: "SMSDataset.csv", sep=',')
data = data.drop(columns=data.columns[-3:])
data.columns = ['label', 'body_text']
print("Początkowe dane:")
print(data.head(), "\n")

# nltk.download('stopwords')
stopwords = nltk.corpus.stopwords.words("english")

def clean_text(text):  1 usage
    if not isinstance(text, int) and not isinstance(text, float):
        no_punctuation = "".join([char for char in text if char not in string.punctuation])
        lower = no_punctuation.lower()
        tokens = re.split( pattern: r'\W+', lower)
        text = tokens
    text = [word for word in text if word not in stopwords]
    return text
```

```python
ps = PorterStemmer()

def steaming(tokenized_text):  1 usage
    text = [ps.stem(word) for word in tokenized_text]
    return text

data_clear_text = data.apply(lambda text: text.apply(clean_text))
data_clear_text.columns = ['label', 'body_clear_text']
print("Usuwanie znaków interpunkcyjnych, tokenizacja, usuwanie stopwords:")
print(data_clear_text.head(), "\n")

data_text_stemmed = data_clear_text.apply(lambda text: text.apply(steaming))
data_text_stemmed.columns = ['label', 'body_text_stemmed']
print("Stemming:")
print(data_text_stemmed.head(), "\n")

data_text_stemmed['body_text_stemmed'] = (
    data_text_stemmed['body_text_stemmed']
    .apply(lambda text: ' '.join(text) if isinstance(text, list) else ''))
```

```python
# Count Vectorizer

count_vect = CountVectorizer()
X_counts_1 = count_vect.fit_transform(data_text_stemmed['body_text_stemmed'])
print("Kształt 'Count Vectorizer':", X_counts_1.shape)
X_counts_1_array = X_counts_1.toarray()
count_vect_result = pd.DataFrame(X_counts_1_array, columns=count_vect.get_feature_names_out())
print("Macierz 'Count Vectorizer':\n", count_vect_result, "\n")

# N-Gram

ngram_vect = CountVectorizer(ngram_range=(2,2))
X_counts_2 = ngram_vect.fit_transform(data_text_stemmed['body_text_stemmed'])
print("Kształt 'N-Gram':", X_counts_2.shape)
X_counts_2_array = X_counts_2.toarray()
ngram_vect_result = pd.DataFrame(X_counts_2_array, columns=ngram_vect.get_feature_names_out())
print("Macierz 'N-Gram':\n", ngram_vect_result, "\n")

# TF-IDF

tfidf_vect = TfidfVectorizer()
X_counts_3 = tfidf_vect.fit_transform(data_text_stemmed['body_text_stemmed'])
print("Kształt 'TF-IDF':", X_counts_3.shape)
X_counts_3_array = X_counts_3.toarray()
tfidf_vect_result = pd.DataFrame(X_counts_3_array, columns=tfidf_vect.get_feature_names_out())
print("Macierz 'TF-IDF':\n", tfidf_vect_result, "\n")
```

Testy:

```
Początkowe dane:
      label                                          body_text
0       ham  Your opinion about me? 1. Over 2. Jada 3. Kusr...
1       ham  What's up? Do you want me to come online? If y...
2       ham                        So u workin overtime nigpun?
3       ham  Also sir, i sent you an email about how to log...
4  Smishing  Please Stay At Home. To encourage the notion o...

Usuwanie znaków interpunkcyjnych, tokenizacja, usuwanie stopwords:
        label                                  body_clear_text
0       [ham]  [opinion, 1, 2, jada, 3, kusruthi, 4, lovable,...
1       [ham]  [whats, want, come, online, free, talk, someti...
2       [ham]                      [u, workin, overtime, nigpun]
3       [ham]  [also, sir, sent, email, log, usc, payment, po...
4  [smishing]  [please, stay, home, encourage, notion, stayin...

Stemming:
     label                                body_text_stemmed
0    [ham]  [opinion, 1, 2, jada, 3, kusruthi, 4, lovabl, ...
1    [ham]   [what, want, come, onlin, free, talk, sometim, ]
2    [ham]                      [u, workin, overtim, nigpun]
3    [ham]  [also, sir, sent, email, log, usc, payment, po...
4  [smish]  [pleas, stay, home, encourag, notion, stay, ho...
```

```
Kształt 'Count Vectorizer': (5971, 8888)
Macierz 'Count Vectorizer':
       000  008704050406  0089mi  01143065228  ...  zouk  zyada  üll  ≦ud
0        0             0       0            0  ...     0      0    0    0
1        0             0       0            0  ...     0      0    0    0
2        0             0       0            0  ...     0      0    0    0
3        0             0       0            0  ...     0      0    0    0
4        0             0       0            0  ...     0      0    0    0
...    ...           ...     ...          ...  ...   ...    ...  ...  ...
5966     0             0       0            0  ...     0      0    0    0
5967     0             0       0            0  ...     0      0    0    0
5968     0             0       0            0  ...     0      0    0    0
5969     0             0       0            0  ...     0      0    0    0
5970     0             0       0            0  ...     0      0    0    0

[5971 rows x 8888 columns]
```

```
Kształt 'N-Gram': (5971, 33643)
Macierz 'N-Gram':
      000 sqft  008704050406 sp  0089mi last  ...  üll submit  üll take  ≟ud even
0            0                0            0  ...           0         0         0
1            0                0            0  ...           0         0         0
2            0                0            0  ...           0         0         0
3            0                0            0  ...           0         0         0
4            0                0            0  ...           0         0         0
...        ...              ...          ...  ...         ...       ...       ...
5966         0                0            0  ...           0         0         0
5967         0                0            0  ...           0         0         0
5968         0                0            0  ...           0         0         0
5969         0                0            0  ...           0         0         0
5970         0                0            0  ...           0         0         0

[5971 rows x 33643 columns]
```

```
Kształt 'TF-IDF': (5971, 8888)
Macierz 'TF-IDF':
      000  008704050406  0089mi  01143065228  ...  zouk  zyada  üll  ≟ud
0     0.0           0.0     0.0          0.0  ...   0.0    0.0  0.0  0.0
1     0.0           0.0     0.0          0.0  ...   0.0    0.0  0.0  0.0
2     0.0           0.0     0.0          0.0  ...   0.0    0.0  0.0  0.0
3     0.0           0.0     0.0          0.0  ...   0.0    0.0  0.0  0.0
4     0.0           0.0     0.0          0.0  ...   0.0    0.0  0.0  0.0
...   ...           ...     ...          ...  ...   ...    ...  ...  ...
5966  0.0           0.0     0.0          0.0  ...   0.0    0.0  0.0  0.0
5967  0.0           0.0     0.0          0.0  ...   0.0    0.0  0.0  0.0
5968  0.0           0.0     0.0          0.0  ...   0.0    0.0  0.0  0.0
5969  0.0           0.0     0.0          0.0  ...   0.0    0.0  0.0  0.0
5970  0.0           0.0     0.0          0.0  ...   0.0    0.0  0.0  0.0

[5971 rows x 8888 columns]
```

Zadanie 2
Odpowiedz na poniższe pytania:

- **Która z metod mogłaby dać najlepsze wyniki w przypadku identyfikacji słów kluczowych w różnych dokumentach? Odpowiedz uzasadnij.**
- **Scharakteryzuj pojęcie Bag-of-Words.**
- **Jakie niedogodności niesie użycie macierzowej reprezentacji dokumentów tekstowych?**

Która z metod mogłaby dać najlepsze wyniki w przypadku identyfikacji słów kluczowych w różnych dokumentach?
W przypadku identyfikacji słów kluczowych w różnych dokumentach najlepsze wyniki może dać metoda TF-IDF. Dzieje się tak, ponieważ ta metoda uwzględnia częstotliwość słów w dokumencie, ale jednocześnie kara za zbyt częste występowanie danego słowa w wielu dokumentach. Metoda Count Vectorizer nie odnosi się do innych dokumentów, a metoda N-Gram bardziej identyfikuje grupy słów niż pojedyncze słowa.

Scharakteryzuj pojęcie Bag-of-Words.
Jest to model tekstu, który wykorzystuje nieuporządkowaną kolekcję słów („bag”). Ignoruje kolejność słów, ale uwzględnia wielość występowania słowa. Model ten jest stosowany w metodach klasyfikacji dokumentów, gdzie częstotliwość występowania każdego słowa jest wykorzystywana jako cecha do trenowania klasyfikatora.

Jakie niedogodności niesie użycie macierzowej reprezentacji dokumentów tekstowych?
Użycie macierzowej reprezentacji dokumentów tekstowych posiada kilka niedogodności. Jedną z nich jest wysoka wymiarowość macierzy. Liczba kolumn rośnie wraz z liczbą unikalnych słów, co może prowadzić do dużej liczby obliczeń. Kolejną niedogodnością jest duża liczba zer w macierzy, to powoduje marnowanie pamięci do obliczeń i spowalnia te obliczenia. Inną niedogodnością jest brak uwzględniania kolejności słów, przez co w macierzy dwa przeciwne znaczenia jednego zdania mogą być uwzględnione w tej samej kolumnie.