Ćwiczenie Bag of Words

Przepisz i przeanalizuj poniższy program.

```
from sklearn.feature_extraction.text import CountVectorizer

# Przykładowe dokumenty tekstowe
documents = [
    "To jest pierwszy dokument.",
    "Dokument numer dwa.",
    "Ostatni dokument w tym zbiorze."
]

# Inicjalizacja obiektu CountVectorizer
vectorizer = CountVectorizer()

# Przetwarzanie i transformacja dokumentów za pomocą Bag of Words
X = vectorizer.fit_transform(documents)

# Wyświetlenie słownika (unikalnych słów)
print("Słownik (unikalne słowa):", vectorizer.get_feature_names_out())

# Wyświetlenie macierzy Bag of Words
print("Macierz Bag of Words:")
print(X.toarray())
```

Na podstawie powyższego programu uzupełnij kolejny program, aby otrzymać taki sam rezultat (z dokładnością do słów, a nie ich kolejności)

```python
import re


# Przykładowe dokumenty tekstowe
documents = [
    "To jest pierwszy dokument.",
    "Dokument numer dwa.",
    "Ostatni dokument w tym zbiorze."
]

# Tokenizacja i oczyszczanie tekstu (użyj metody split)
def preprocess_text(text):
    # Usunięcie znaków interpunkcyjnych i zamiana na małe litery
    text = re.sub(r'[^\w\s]', '', text.lower())
    # Tokenizacja po spacjach
    […]
    return tokens

# Stworzenie słownika i policzenie częstości słów
word_counts = {}
for document in documents:
    tokens = preprocess_text(document)
    #użyj słownika do zliczenia tokentów w dokumencie (key=token, val=ilość
wystąpień)
    […]

# Utworzenie słownika (unikalne słowa)
# utwórz listę z kluczy słownika (użyj list() oraz slownik.keys())
vocabulary = […]

# Utworzenie macierzy Bag of Words
bow_matrix = []
for document in documents:
    tokens = preprocess_text(document)
    # dla danego dokumentu utwórz wektor BoW na podstawie vocabulary
    bow_vector = […]
    # dodaj wektor do macierzy BoW
    bow_matrix.append(bow_vector)

# Wyświetlenie słownika (unikalne słowa)
print("Słownik (unikalne słowa):", vocabulary)

# Wyświetlenie macierzy Bag of Words
print("Macierz Bag of Words:")
for bow_vector in bow_matrix:
    print(bow_vector)
```