AI VIET NAM – COURSE 2024

Lập trình Python căn bản - Day 5

Ngày 15 tháng 6 năm 2024

Ngày thực hiện:	15/06/2024
Người thực hiện:	Đinh Thị Tâm
Nguồn:	AIO2024 - Weekly Reading
Nguồn dữ liệu (nếu có):	Link of Data Sources of Day 5
Từ khóa:	Bag of Words
Người tóm tắt:	Đinh Thị Tâm

1. **Mô tả thuật toán:**

   Bag of Words là một thuật toán hỗ trợ xử lý ngôn ngữ tự nhiên và mục đích của BoW là phân loại text hay văn bản. Ý tưởng của BoW là phân tích và phân nhóm dựa theo "Bag of Words" (corpus). Với test data mới, tiến hành tìm ra số lần từng từ của test data xuất hiện trong "Bag". Cách thức thực hiện như sau:

   - Bước 1: Chia nhỏ văn bản thành các từ riêng lẻ.
   - Bước 2: Tạo một tập hợp các từ xuất hiện trong văn bản. Tập hợp này không có phần tử trùng nhau.
   - Bước 3: Biểu diễn văn bản input ở dạng vector: Mỗi câu (mỗi input) được biểu diễn bằng một vector, với mỗi phần tử trong vector thể hiện số lần xuất hiện của từ đó trong input.

2. **Cài đặt thuật toán:**

```
   # Ham tao Bag of Words
def getVocabul(data):
    aText = ""
    lenght = len(data)
    for x in data:
        aText = aText + x+" "
    aText = aText.replace(",", "")
    aText = aText.replace(".", "")
    aText = aText.replace("_", " ")
    aText = aText.replace("-", " ")
    aText = aText.split()
    aSet = set(aText)
    aText = list(aSet)
    aText.sort()
    return aText
# Lay so lan xuat hien cua cac tu trong data


def getOccur(data, aVocalbu):
```

```
    result = []
    lenght = len(aVocalbu)
    for idx in range(lenght):
        result.append(data.count(aVocalbu[idx]))
    return result


# my main
corpus = ["T i  th ch  m  n  To  n.", "T i  th ch AI", "T i  th ch  m   n h c "]
myVocal = getVocabul(corpus)
myText = "T i  th ch AI  th ch  To  n"
mylist = myText.split()
print(f"- {myText}: {getOccur(mylist, myVocal)}")
print(f"- Bag-of-Words:{myVocal}")



```

3. **Output chương trình:**

- Tôi thích AI thích Toán: [1, 1, 1, 0, 0, 2, 0]
- Bag-of-Words:['AI', 'Toán', 'Tôi', 'môn', 'nhạc', 'thích', 'âm']