WIKIPEDIA

Mô hình túi từ

Bách khoa toàn thư mở Wikipedia

Mô hình túi từ (bag-of-words) là một biểu diễn đơn giản hóa được sử dụng trong xử lý ngôn ngữ tự nhiên và truy vấn thông tin (IR). Trong mô hình này, một văn bản (chẳng hạn như một câu hoặc một tài liệu) được thể hiện dưới dạng túi (multiset) chứa các từ của nó, không quan tâm đến ngữ pháp và thậm chí trật tự từ nhưng vẫn giữ tính đa dạng. Mô hình túi từ cũng đã được sử dụng cho thị giác máy tính.[1]

Mô hình túi từ thường được sử dụng trong các phương pháp phân loại tài liệu trong đó sự xuất hiện (tần suất) của mỗi từ được sử dụng như một đặc trưng để đào tạo máy phân loại [2].

Một tài liệu tham khảo đầu tiên về "túi từ" trong ngữ cảnh ngôn ngữ có thể được tìm thấy trong bài viết năm 1954 của Zellect Harris về *Cấu trúc phân phối*.[3]

Mục lục

Ví dụ áp dụng

Dưới đây mô hình một tài liệu văn bản bằng cách sử dụng túi từ. Đây là hai tài liệu văn bản đơn giản:

```
(1) Phúc thích xem phim. Đạt cũng thích xem phim.
```

```
(2) Bích cũng thích xem các trận bóng đá.
```

Dựa trên hai tài liệu văn bản này, một danh sách được xây dựng như sau cho mỗi tài liệu: (Ví dụ dưới đây coi mỗi tiếng là một từ, "bóng đá" là "bóng" và "đá". Cũng có thể xác định từ theo ngữ pháp, "bóng đá" là một từ.)

```
"Phúc", "thích", "xem", "phim", "Đạt" "cũng", "thích", "xem", "phim"

"Bích" "cũng" "thích" "xem" "các" "trận" "bóng" "đá"
```

Đại diện cho mỗi túi từ dưới dạng đối tượng JSON và quy cho biến Javascript tương ứng:

```
BoW1 = {"Phúc":1,"thích":2,"xem":2,"phim":2,"Đạt":1,"cũng":1};
BoW2 = {"Bích":1,"cũng":1,"thích":1,"xem":1,"các":1,"trận":1,"bóng":1,"đá":1};
```

Mỗi khóa là từ và mỗi giá trị là số lần xuất hiện của từ đó trong tài liệu văn bản đã cho.

Thứ tự của các từ bị bỏ qua, vì vậy, ví dụ `{"Phúc":1,"thích":2,"Đạt":1,"cũng":1,"xem":2,"phim":2}` cũng là *BoW1*.

Lưu ý: nếu một tài liệu khác giống như một kết hợp của hai văn bản trên,

```
(3) Phúc thích xem phim. Đạt cũng thích xem phim. Bích cũng thích xem các trận bóng đá.
```

thể diện Javascript của nó sẽ là:

```
BoW3 = BoW1 =
{"Phúc":1,"thích":3,"xem":3,"phim":2,"Đạt":1,"cũng":2,"Bích":1,"các":1,"trận":1,"bóng":1,"đá":1};
```

Vì vậy, như chúng ta thấy trong đại số túi, "liên kết" của hai văn bản trong cách biểu thị túi, chính thức là liên kết rời rạc, tổng hợp các bội số của từng từ.

$BoW3 = BoW1 \biguplus BoW2$.

Ứng dụng

Trong thực tế, mô hình túi từ chủ yếu được sử dụng như một công cụ tạo đặc trưng. Sau khi chuyển đổi văn bản thành một "túi từ", chúng ta có thể tìm các biện pháp khác nhau để mô tả văn bản. Loại đặc điểm hoặc tính năng phổ biến nhất được tính toán từ mô hình túi từ là tần số thuật ngữ, cụ thể là số lần một thuật ngữ xuất hiện trong văn bản. Đối với ví dụ trên, chúng ta có thể xây dựng hai danh sách sau để ghi lại tần số thuật ngữ của tất cả các từ riêng biệt (BoW1 và BoW2 được chứa như trong BoW3):

```
(1) [1, 2, 2, 2, 1, 1, 0, 0, 0, 0, 0]
(2) [0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1]
```

Mỗi mục trong danh sách đề cập đến số lượng mục tương ứng trong danh sách (đây cũng là biểu diễn biểu đồ). Ví dụ: trong danh sách đầu tiên (đại diện cho tài liệu 1), hai mục đầu tiên là "1,2":

- Mục đầu tiên tương ứng với từ "Phúc" là từ đầu tiên trong danh sách và giá trị của nó là "1" vì "Phúc" xuất hiện trong tài liệu đầu tiên 1 lần.
- Mục thứ hai tương ứng với từ "thích", đó là từ thứ hai trong danh sách và giá trị của nó là "2" vì "thích" xuất hiện trong tài liệu đầu tiên 2 lần

Danh sách (hoặc vectơ) đại diện này không bảo vệ thứ tự của các từ trong câu gốc. Đây chỉ là tính năng chính của mô hình Túi từ. Loại đại diện này có một số ứng dụng thành công, chẳng hạn như lọc email.[1]

Tuy nhiên, tần số thuật ngữ không nhất thiết là đại diện tốt nhất cho văn bản. Các từ phổ biến như "the", "a", "to" trong tiếng Anh hầu như luôn là các thuật ngữ có tần suất cao nhất trong văn bản. Vì vậy, có số lượng thô cao không nhất thiết có nghĩa là từ tương ứng là quan trọng hơn. Để giải quyết vấn đề này, một trong những cách phổ biến nhất để "bình thường hóa" tần số thuật ngữ là tính trọng số của một thuật ngữ bằng nghịch đảo của tần số tài liệu, hoặc **tf–idf**. Ngoài ra, với mục đích cụ thể

của phân loại, các lựa chọn thay thế được giám sát đã được phát triển để giải thích cho nhãn lớp của tài liệu.[1] Cuối cùng, trọng số nhị phân (hiện diện / vắng mặt hoặc 1/0) được sử dụng thay cho tần số cho một số vấn đề (ví dụ: tùy chọn này được triển khai trong hệ thống phần mềm học máy WEKA).

Mô hình N-gram

Mô hình túi từ là một đại diện tài liệu không có trật tự, chỉ có số lần xuất hiện của từ được coi trọng. Chẳng hạn, trong ví dụ trên "Phúc thích xem phim. Đạt cũng thích xem phim ", đại diện túi từ sẽ không tiết lộ rằng động từ " thích" luôn theo sau tên của một người trong văn bản này. Thay vào đó, mô hình n-gram có thể lưu trữ thông tin thứ tự này. Áp dụng cho ví dụ tương tự ở trên, một mô hình **bigram** sẽ phân tích văn bản thành các đơn vị sau và lưu trữ tần số thuật ngữ của từng đơn vị như trước đây.

```
[
  "Phúc thích",
  "thích xem",
  "xem phim",
  "Đạt cũng",
  "cũng thích",
  "thích xem",
  "xem phim",
]
```

Về mặt khái niệm, chúng ta có thể xem mô hình túi từ như một trường hợp đặc biệt của mô hình n-gram, với n = 1. Với n > 1, mô hình được đặt tên là w-shingling (trong đó *w* tương đương với *n* biểu thị số lượng từ được nhóm). Xem mô hình ngôn ngữ để thảo luận chi tiết hơn.

Triển khai Python

```
    Sentence_1 = ["Phúc thích xem phim. Đạt cũng thích xem phim."]
    tokenizer = Tokenizer()
    tokenizer.fit_on_texts(Sentence_1)
    sequences_1 = tokenizer.texts_to_sequences(Sentence_1)
    word_index_1 = tokenizer.word_index
    BoW_1={}
    for key in word_index_1:
        BoW_1[key]=sequences_1[0].count(word_index_1[key])
    print(BoW_1)
    print(f"Bag of word sentence 1:\n{BoW_1}")
    print(f'We found {len(word_index_1)} unique tokens.')
```

Thủ thuật băm

Một cách khác để sử dụng từ điển là thủ thuật băm, trong đó các từ được ánh xạ trực tiếp đến các chỉ mục có hàm băm [4]. Vì vậy, không có bộ nhớ được yêu cầu để lưu trữ một từ điển. Xung đột băm thường được xử lý thông qua giải phóng bộ nhớ để tăng số lượng băm. Trong thực tế, băm đơn giản hóa việc thực hiện các mô hình túi từ và cải thiện khả năng mở rộng.

Ví dụ sử dụng: lọc thư rác

Trong lọc thư rác Bayes, một thông điệp email được mô hình hóa như một tập hợp các từ được sắp xếp theo thứ tự được chọn từ một trong hai phân phối xác suất: một đại diện cho thư rác và một đại diện cho email hợp pháp ("ham"). Hãy tưởng tượng có hai túi chữ đầy chữ. Một túi chứa đầy các từ được tìm thấy trong tin nhắn rác và túi còn lại có các từ được tìm thấy trong e-mail hợp pháp. Mặc dù

bất kỳ từ nào có khả năng nằm ở đâu đó trong cả hai túi, túi "spam" sẽ chứa các từ liên quan đến spam như "chứng khoán", "Viagra" và "mua" thường xuyên hơn, trong khi túi "ham" sẽ chứa nhiều từ liên quan đến bạn bè hoặc nơi làm việc của người dùng.

Để phân loại thư e-mail, bộ lọc thư rác Bayes giả định rằng thư đó là một đống từ được đổ ngẫu nhiên từ một trong hai túi và sử dụng xác suất Bayesian để xác định túi nào có khả năng nằm trong túi đó.

Xem thêm

- Làm mịn cộng tính
- Mô hình túi từ trong thị giác máy tính
- Phân loại tài liệu
- Ma trận thuật ngữ tài liệu
- Trích xuất đặc trưng
- Thủ thuật băm
- Học máy
- MinHash
- n-gram
- Xử lý ngôn ngữ tự nhiên
- Mô hình không gian vector
- w-shingling

Ghi chú

1. ^ ***a*** ***ă*** ***â*** Chú thích trống (trợ giúp)
2. **^** McTear et al 2016, p. 167.
3. **^** Harris, Zellig (1954). “Distributional Structure”. *Word* **10** (2/3): 146–62. “And this stock of combinations of elements becomes a factor in the way later choices are made... for language is not merely a bag of words but a tool with particular properties which have been fashioned in the course of its use”
4. **^** Weinberger, K. Q.; Dasgupta A.; Langford J.; Smola A.; Attenberg, J. (2009). “Feature hashing for large scale multitask learning,”. *Proceedings of the 26th Annual International Conference on Machine Learning*: 1113–1120. Bibcode:2009arXiv0902.2206W (http://adsabs.harvard.edu/abs/2009arXiv0902.2206W). arXiv:0902.2206 (https://arxiv.org/abs/0902.2206).

Tham khảo

- McTear, Michael (et al) (2016). *Giao diện hội thoại*. Nhà xuất bản quốc tế Springer.