★ ✎ 819 Most Common Word

2018年5月5日   20:00

Given a paragraph and a list of banned words, return the most frequent word that is not in the list of banned words. It is guaranteed there is at least one word that isn't banned, and that the answer is unique.

Words in the list of banned words are given in lowercase, and free of punctuation. Words in the paragraph are not case sensitive. The answer is in lowercase.

Example:

Input:

paragraph = "Bob hit a ball, the hit BALL flew far after it was hit."

banned = ["hit"]

Output: "ball"

Explanation:

"hit" occurs 3 times, but it is a banned word.

"ball" occurs twice (and no other word does), so it is the most frequent non-banned word in the paragraph.

Note that words in the paragraph are not case sensitive,

that punctuation is ignored (even if adjacent to words, such as "ball,"),

and that "hit" isn't the answer even though it occurs more because it is banned.

Note:

- 1 <= paragraph.length <= 1000.
- 1 <= banned.length <= 100.
- 1 <= banned[i].length <= 10.
- The answer is unique, and written in lowercase (even if its occurrences in paragraph may have uppercase symbols, and even if it is a proper noun.)
- paragraph only consists of letters, spaces, or the punctuation symbols !?',;.
- Different words in paragraph are always separated by a space.
- There are no hyphens or hyphenated words.
- Words only consist of letters, never apostrophes or other punctuation symbols.

来自 <https://leetcode.com/problems/most-common-word/description/>

给定一个段落 (paragraph) 和一个禁用单词列表 (banned)。返回出现次数最多，同时不在禁用列表中的单词。题目保证至少有一个词不在禁用列表中，而且答案唯一。

禁用列表中的单词用小写字母表示，不含标点符号。段落中的单词不区分大小写。答案都是小写字母。

示例:

输入:

paragraph = "Bob hit a ball, the hit BALL flew far after it was hit."

banned = ["hit"]

输出: "ball"

解释:

"hit" 出现了3次，但它是一个禁用的单词。

"ball" 出现了2次 (同时没有其他单词出现2次)，所以它是段落里出现次数最多的，且不在禁用列表中的单词。

注意，所有这些单词在段落里不区分大小写，标点符号需要忽略（即使是紧挨着单词也忽略，比如 "ball,"），

"hit"不是最终的答案，虽然它出现次数更多，但它在禁用单词列表中。

说明:

- 1 <= 段落长度 <= 1000.
- 1 <= 禁用单词个数 <= 100.
- 1 <= 禁用单词长度 <= 10.
- 答案是唯一的, 且都是小写字母 (即使在 paragraph 里是大写的，即使是一些特定的名词，答案都是小写的。)
- paragraph 只包含字母、空格和下列标点符号!?',;.
- paragraph 里单词之间都由空格隔开。
- 不存在没有连字符或者带有连字符的单词。
- 单词里只包含字母，不会出现省略号或者其他标点符号。

Solution for Python3：

```python
class Solution1:
    def mostCommonWord(self, paragraph, banned):
        """
        :type paragraph: str
        :type banned: List[str]
        :rtype: str
        """
        cnt = collections.Counter()
        paragraph = paragraph.lower().split(" ")
        for word in paragraph:
            if word and not word[-1].isalpha():
                cnt[word[:-1]] += 1
            elif word:
                cnt[word] += 1
        for word in cnt.most_common():
            if word[0] not in banned:
                return word[0]

class Solution2:
    def mostCommonWord(self, paragraph, banned):
        """
        :type paragraph: str
        :type banned: List[str]
        :rtype: str
        """
        count = collections.Counter(word.strip("!?',;.") for word in
paragraph.lower().split())
        for word in count.most_common():
            if word[0] not in banned:
                return word[0]
```

Solution for C++:

```cpp
class Solution {
public:
    string mostCommonWord(string paragraph, vector<string>& banned) {
        unordered_set<string> ban(banned.begin(), banned.end());
        unordered_map<string, int> cnt;
        for (auto& c : paragraph)
            c = isalpha(c) ? tolower(c) : ' ';
        istringstream iss(paragraph);
        string w;
        pair<string, int> res("", 0);
        while (iss >> w)
            if (ban.find(w) == ban.end() && ++cnt[w] > res.second)
                res = make_pair(w, cnt[w]);
        return res.first;
    }
};
```

Counter.most_common([n])

返回一个TopN列表。如果n没有被指定，则返回所有元素。当多个元素计数值相同时，排列是无确定顺序的。