

JavaScript: trovare gli hapax in un testo

In questo articolo vedremo come trovare gli hapax in un testo con JavaScript.

Un hapax (dal greco ἁπαξ, "una volta") è una parola o una forma linguistica che appare solo una volta all'interno di un corpus di testo o di una raccolta di testi. In altre parole, un hapax è una parola che non è mai ripetuta all'interno del corpus di testo in questione. In linguistica, gli hapax sono spesso utilizzati per analizzare la diversità lessicale di un corpus di testo o per identificare parole o forme linguistiche insolite o rare. In generale, gli hapax sono utilizzati per analizzare un testo e capire la sua distribuzione lessicale, per esempio per capire se un autore utilizza un vocabolario molto ampio o se si concentra su un numero limitato di parole, oppure per capire se ci sono parole che si ripetono più di quanto ci si aspetterebbe in un testo scritto in una lingua naturale.

La soluzione è la seguente:

```
'use strict';

function findHapaxes(text) {
  const words = text.split(' ');
  let frequency = {};
  let hapaxes = [];

  for (let word of words) {
    frequency[word] = frequency[word] ?
frequency[word] + 1 : 1;
  }
}
```

```
    for (let word in frequency) {  
        if (frequency[word] === 1) {  
            hapaxes.push(word);  
        }  
    }  
  
    return hapaxes;  
}
```

Esempio d'uso:

```
const text = 'This is a test text. It is just a test  
text.';  
console.log(findHapaxes(text)); // Output: ['This',  
'is', 'a', 'test', 'text.', 'It', 'just', 'text.']
```

Nota: l'esempio considera come parola anche i segni di punteggiatura, quindi se si vuole evitare questo comportamento si può utilizzare una libreria per tokenizzare il testo in parole oppure scrivere una funzione che elimini i segni di punteggiatura dalle parole prima di essere conteggiate.