Tag cloud

Nowadays, you can find a tag cloud at every turn on the Internet. This is a visual representation of the frequency with which words appear in a text. Words that are more common are displayed in a larger font. Thus, the key words of the text immediately catch the eye. The visual effect can be enhanced by working with different colors and placement of the words. Usually a list of stop words is used as well. These are words that occur frequently in most texts. Words that appear in the stop word list, do not appear in the word cloud.

Example of a tag cloud.

Assignment

Write a function tag cloud that returns the frequency of words that occur in a given text file in the form of a dictionary. Each word from the text that is not in a given stop word list is used as a key in the dictionary, and the corresponding value is the number of occurrences of that word in the text. The words of the text are formed by the longest possible sequences of letters and the apostrophe ('). The text fragment "Don't say that word!" thus consists of four words.

To the function takes the locations of two text files as its arguments. The first text file contains a few lines of text. The second text file contains a list of stop words, with each stop word on a separate line.

```
def tagcloud(textFile, stopWordsFile)
```

Note: If you correctly implement the function tagcloud, it will be used to generate a tag cloud based on the script of a famous film. Can you figure out which film?

Example

```
>>> tagcloud('tagcloud.txt', 'stopWords.txt')
{'say': 1, 'word': 2, "don't": 1}
```

Vandaag de dag zie je om de haverklap wel ergens een woordenwolk (*tag cloud*) opduiken op het Internet. Dit is een visuele weergave van de frequentie waarmee woorden voorkomen in een tekst. Woorden die vaker voorkomen worden in een groter lettertype weergegeven. Op die manier springen de belangrijkste woorden van de tekst direct in het oog. Het visuele effect kan nog versterkt worden door te werken met verschillende kleuren en plaatsing van de woorden.

Meestal wordt ook nog een lijst van stopwoorden gebruikt. Dit zijn woorden die veelvuldig in de meeste teksten voorkomen. Woorden die in deze stopwoordenlijst voorkomen, worden niet weergegeven in de woordenwolk.

Voorbeeld van een woordenwolk.

Opgave

Schrijf een functie woordenwolk die de frequentie van de woorden die voorkomen in een gegeven tekstbestand teruggeeft onder de vorm van een dictionary. Elk woord uit de tekst dat niet in een gegeven stopwoordenlijst voorkomt wordt als sleutel gebruikt in de dictionary, en de corresponderende waarde is het aantal voorkomens van dat woord in de tekst. De woorden uit de tekst worden gevormd door de langst mogelijke opeenvolgingen van letters en de apostrof (').
Het tekstfragment "Don't say that word!" bestaat dus uit vier woorden.

Aan de functie moeten de locaties van twee tekstbestanden doorgegeven worden. Het eerste tekstbestand bevat enkele regels tekst. Het tweede tekstbestand bevat een stopwoordenlijst, waarbij elk stopwoord op een afzonderlijke regel staat.

```
def woordenwolk(tekstBestand, stopwoordenBestand)
```

Opmerking: Als je de functie woordenwolk correct hebt geïmplementeerd, dan wordt die gebruikt om een woordenwolk te genereren op basis van het script van een bekende langspeelfilm. Kan je achterhalen over welke film het gaat?

Voorbeeld

```
>>> woordenwolk('woordenwolk.txt', 'stopwoorden.txt')
{'say': 1, 'word': 2, "don't": 1}
```