Cactolith

We've all been irritated by jargon — those collections of polysyllabic technical terms used in place of simple English, or worse still, everyday English words redefined into technical meanings. Below are two of my favorite, if ancient, definitions of geological terms, both of which may be just slightly tongue-in-cheek.

> "***Crocydite**, belonging to the group of vaguely bordered migmatites (dictyonite, nebulite, stictolite), may be genetically defined by the new terminology as an endomerismite with magmatic neosome in a palaeosome which is a stereogenic cyriosome.*" (de Waard D, 1950)

> "*A **cactolith** is a quasihorizontal chonolith composed of anastomosing ductoliths whose distal ends curl like a harpolith, thin like a sphenolith, or bulge discordantly like an akmolith or ethmolith.*" (Hunt CB, 1953)

The latter term and its associated definition were created by Charles B. Hunt, a researcher at the United States Geological Survey (USGS). Whilst he was in fact describing an actual geological feature — a laccolith which he saw as resembling a cactus — he was also, tongue-in-cheek, commenting on what he saw as an absurd number of "-lith" words in the field of Geology. Word Ways: The Journal of Recreational Linguistics chose cactolith as its word of the year for 2010.

In linguistics, the **Gunning-Fog index** is used to measure the readability of English writing. The index estimates the years of formal eduction needed to understand a given text on first reading. For example, a fog index of 12 requires the reading level of a U.S. high school senior (around 18 years old). The fog index is commonly used to confirm that a text can be read easily by the intended audience. Texts for a wide audience generally need a fog index less than 12. Texts requiring near-universal understanding generally need an index less than 8. Philip Chalmers of *Benefit from IT* provided the following typical fog index scores, to help ascertain the readability of documents.

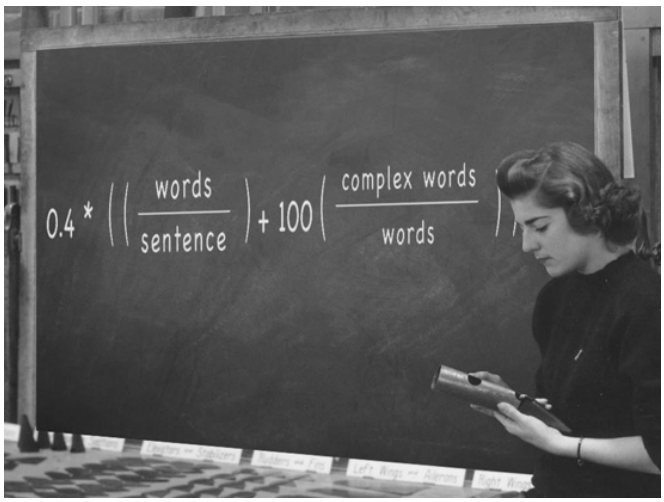
fog index	**examples**
6	TV guides, the Bible, Mark Twain
8	Reader's Digest
8-10	most popular novels
10	Time, Newsweek
11	Wall Street Journal
14	The Times, The Guardian
15-20	scientific papers
$\geq$ 20	only government sites can get away with this, because you can't ignore them
$\geq$ 30	the government is covering something up

The Gunning-Fog index is calculated using the following algorithm:

1. Select a passage (such as one or more full paragraphs) of around 100 words. Do not omit any sentences.

2. Determine the average sentence length by dividing the number of words by the number of sentences.
3. Count the complex words: these are the words with three or more syllables.
4. Add the average sentence length and the percentage of complex words.
5. Multiply the result from the previous step by 0.4.

Expressed as a formula, this becomes $$ 0.4 \left[\left(\frac{\textrm{words}}{\textrm{sentences}} \right) + 100 \left(\frac{\textrm{complex words}}{\textrm{words}} \right) \right] $$ While the fog index is a good sign of hard-to-read text, it has limits. Not all complex words are difficult. For example, "asparagus" is not generally thought to be a difficult word, though it has four syllables. A short word can be difficult if it is not used very often by most people.



The complete formula of the Gunning-Fog index.

Assignment

Your task is to compute the Gunning-Fog index for a series of text fragments, where each of these fragments has been stored in a text file. In order to do so, you proceed as follows.

- Write a function `syllables` that takes a word (string) containing letters only. The function must return an estimate of the number of syllables in the given word. Although not entirely correct, the function must determine the number of syllables as the number of vowel sequences (`a`, `e`, `i`, `o`, `u` or `y`). The function should make no distinction between uppercase and lowercase letters.
- Use the function `syllables` to write a function `statistics` that takes the location of a text file. This file must contain a text fragment, with each sentence on a separate line. In addition, the text file may contain empty lines (lines containing nothing or just whitespace characters (spaces and tabs)), for example to separate the sentences of successive paragraphs. These empty lines are not considered to be sentences. The function must return a tuple containing three integers, that respectively indicate how many sentences, words and complex words occur in the given text fragments. The words of a sentence are defined as the longest possible sequence of letters. In determining whether or not a word is complex, the number of syllables in the word must be determined using the function `syllables`.
- Use the function `statistics` to write a function `gunningfog` that takes the location of a text file. This files contains a text fragment that should be interpreted in the same way as with the function `statistics`. The function must return the computed Gunning-Fog index of the given text fragment as a floating point number.

Example

In the following interactive session, we assume that the text files crocydite.txt, cactolith.txt and wikipedia.txt are located in the current directory. The first two files contain the definitions of the geological terms as given in the introduction of this exercise. The third file contains the initial paragraphs of the Wikipedia article about Geology.

```
>>> syllables('cactolith')
3
>>> syllables('quasihorizontal')
6
>>> syllables('palaeosome')
4

>>> statistics('crocydite.txt')
(1, 34, 17)
>>> statistics('cactolith.txt')
(1, 29, 11)
>>> statistics('wikipedia.txt')
(5, 119, 37)

>>> gunningfog('crocydite.txt')
33.6
>>> gunningfog('cactolith.txt')
26.77241379310345
>>> gunningfog('wikipedia.txt')
21.956974789915968
```

Resources

de Waard D (1950). Palingenetic structures in augen gneiss of the Sierra de Guadarrama, Spain. *Bull. Comm. Géol. Finlande* **150(23)**, 51–66.

Hunt CB (1953). Geology and geography of the Henry Mountains region, Utah. *US Geological Survey Professional Paper* **228**, 234.

Iedereen ergert zich wel eens aan jargon — van die meerlettergrepige technische termen die gebruikt worden in plaats van eenvoudige woorden, of erger nog, alledaagse woorden die omgetoverd worden tot technische vaktaal die voor buitenstaanders vaak moeilijk te begrijpen is. Hieronder staan alvast twee van mijn favoriete (zij het ietwat oude) definities van geologische termen, die allebei lichtjes ironisch moeten opgevat worden.

> "***Crocydite**, belonging to the group of vaguely bordered migmatites (dictyonite, nebulite, stictolite), may be genetically defined by the new terminology as an endomerismite with magmatic neosome in a palaeosome which is a stereogenic cyriosome.*" (de Waard D, 1950)

> "*A **cactolith** is a quasihorizontal chonolith composed of anastomosing ductoliths whose distal ends curl like a harpolith, thin like a sphenolith, or bulge discordantly like an akmolith or ethmolith.*" (Hunt CB, 1953)

Deze laatste term en de bijhorende definitie zijn van de hand van Charles B. Hunt, onderzoeker aan de United States Geological Survey (USGS). Naast het feit dat hij daarmee een werkelijk

geologisch fenomeen beschreef — een laccoliet die hij had waargenomen en die de vorm had van een cactus — was het meteen ook een ironische commentaar op wat hij zag als een absurd aantal "-lith" woorden in het domein van de geologie. Het tijdschrift Word Ways: The Journal of Recreational Linguistics riep cactolith uit tot haar woord van het jaar voor 2010.

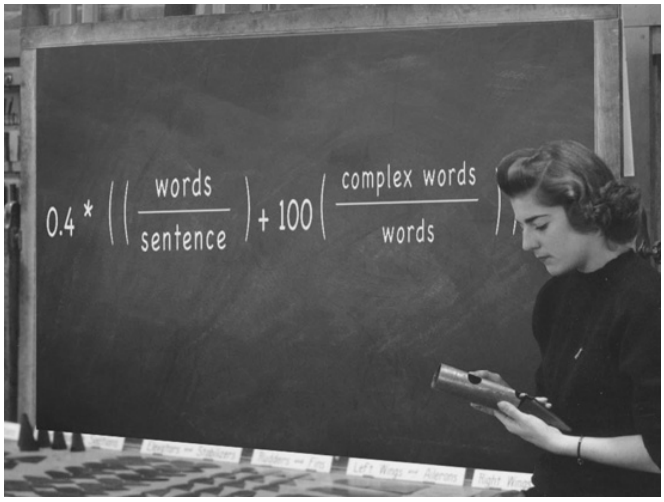
Om de leesbaarheid van teksten te bepalen, wordt in de taalkunde vaak gebruik gemaakt van de **Gunning-Fog index**. Deze index geeft een schatting van het aantal jaren onderwijs dat iemand moet genoten hebben om een bepaalde tekst te kunnen begrijpen bij eerste lezing. Een fog index van 12 vereist bijvoorbeeld het leesniveau die een beginnende universitair zou moeten hebben (leeftijd van ongeveer 18 jaar). De fog index wordt doorgaans gebruikt om te bevestigen dat een tekst makkelijk leesbaar is voor het beoogde doelpubliek. Onderstaande tabel geeft een overzicht van enkele typische fog index scores, samengesteld door Philip Chalmers van *Benefit from IT*.

fog index	**voorbeelden**
6	handleiding TV, Bijbel, Mark Twain
8	Reader's Digest
8-10	populaire boeken
10	Time, Newsweek
11	Wall Street Journal
14	The Times, The Guardian
15-20	wetenschappelijke tijdschriften
$\geq$ 20	enkel websites van de overheid komen hiermee weg, omdat je ze toch niet kunt negeren
$\geq$ 30	de overheid heeft iets te verbergen

De Gunning-Fog index wordt berekend aan de hand van het volgende algoritme:

1. Selecteer een tekstfragment (typisch één of meer paragrafen) van ongeveer 100 woorden. Laat geen zinnen weg.
2. Bepaal de gemiddelde lengte van een zin door het aantal woorden te delen door het aantal zinnen.
3. Tel het aantal complexe woorden: dit zijn woorden die bestaan uit minstens drie lettergrepen.
4. Tel de gemiddelde lengte van de zinnen en het percentage complexe woorden bij elkaar op.
5. Vermenigvuldig het resultaat uit de vorige stap met 0.4.

Uitgedrukt als een formule wordt dit $$ 0.4 \left[\left(\frac{\textrm{woorden}}{\textrm{zinnen}} \right) + 100 \left(\frac{\textrm{complexe woorden}}{\textrm{woorden}} \right) \right] $$ Ondanks het feit dat de Gunning-Fog index een goede maatstaf is voor de leesbaarheid van teksten, heeft hij toch zijn beperkingen. Niet alle complexe woorden zijn moeilijk. Zo wordt "lettergrepen" algemeen niet als een moeilijk woord beschouwd, terwijl het toch vier lettergrepen telt. Een kort woord kan toch moeilijk zijn als het niet vaak gebruikt wordt door de meeste mensen.



De formule van de Gunning-Fog index.

Opgave

Voor deze opgave vragen we je om de Gunning-Fog index te bepalen van enkele tekstfragmenten, waarbij we elk tekstfragment hebben opgeslaan in een tekstbestand. Hiervoor ga je als volgt te werk.

- Schrijf een functie `lettergrepen` waaraan een woord (string) dat enkel bestaat uit letters moet doorgegeven worden. De functie moet een schatting teruggeven van het aantal lettergrepen in het gegeven woord. Hoewel niet geheel correct, moet de functie het aantal lettergrepen bepalen als het aantal klinkergroepen. Een klinkergroep is de langst mogelijke opeenvolging van klinkers (`a`, `e`, `i`, `o`, `u` en `y`). Hierbij mag geen onderscheid gemaakt worden tussen hoofdletters en kleine letters.
- Gebruik de functie `lettergrepen` om een functie `statistieken` te schrijven waaraan de locatie van een tekstbestand moet doorgegeven worden. Dit bestand bevat een tekstfragment, waarbij elke zin op een afzonderlijke regel staat. Het tekstbestand kan voorts ook lege regels bevatten (regels waarop niets staat, of enkel witruimte (spaties en tabs)), bijvoorbeeld om de zinnen van opeenvolgende paragrafen van elkaar te scheiden. Deze lege regels worden echter niet als zin meegeteld. De functie moet een tuple van drie natuurlijke getallen teruggeven, die respectievelijk aangeven hoeveel zinnen, woorden, en complexe woorden er voorkomen in het tekstfragment. De woorden van een zin worden bepaald als de langst mogelijke opeenvolging van letters. Om te bepalen of een woord complex is, moet het aantal lettergrepen van het woord uiteraard geteld worden aan de hand van de functie `lettergrepen`.
- Gebruik de functie `statistieken` om een functie `gunningfog` te schrijven waaraan de locatie van een tekstbestand moet doorgegeven worden. Dit bestand moet een tekstfragmenten bevatten dat op dezelfde manier moet geïnterpreteerd worden als bij de functie `statistieken`. De functie moet de berekende Gunning-Fog index van het gegeven tekstfragment teruggeven als een *floating point* getal.

Voorbeeld

Bij onderstaande voorbeeldsessie gaan we ervan uit dat de tekstbestanden crocydite.txt, cactolith.txt en wikipedia.txt zich in de huidige directory bevinden. De eerste twee tekstbestanden bevatten de definities van de geologische termen uit de inleiding van deze opgave. Het derde tekstbestand bevat de inleidende paragrafen uit het Wikipedia artikel over geologie.

```
>>> lettergrepen('cactolith')
3
>>> lettergrepen('quasihorizontal')
6
>>> lettergrepen('palaeosome')
4

>>> statistieken('crocydite.txt')
(1, 34, 17)
>>> statistieken('cactolith.txt')
(1, 29, 11)
>>> statistieken('wikipedia.txt')
(5, 119, 37)

>>> gunningfog('crocydite.txt')
33.6
>>> gunningfog('cactolith.txt')
26.77241379310345
>>> gunningfog('wikipedia.txt')
21.956974789915968
```

Bronnen

de Waard D (1950). Palingenetic structures in augen gneiss of the Sierra de Guadarrama, Spain. *Bull. Comm. Géol. Finlande* **150(23)**, 51–66.

Hunt CB (1953). Geology and geography of the Henry Mountains region, Utah. *US Geological Survey Professional Paper* **228**, 234.