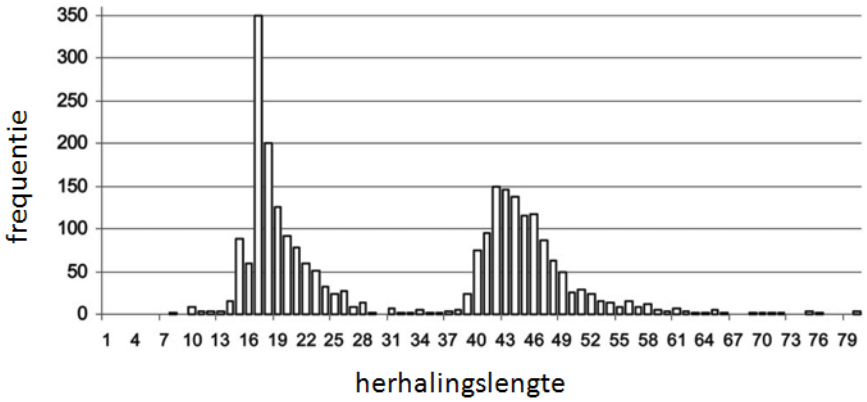
Huntingtons disease

Huntington's disease, which was reported thoroughly by the American doctor George Huntington in 1872, is a dominant hereditary condition which affects certain parts of the brain. The first symptoms of HD are mostly displayed between the age of 35 and 45, and, among others, consist of uncontrolled (choreatic) movements that slowly get worse, mental deterioration and a variety of psychiatric disorders. On average, the disease is fatal after eighteen years, mainly because of additional causes such as pneumonia.

HD is classified under the *trinucleotide repeat disorders*, which are caused by a repeated section of a gene, causing a deviation from the normal length. With HD it concerns a deviation in the *Huntingtin gene*. The Huntingtin gene exhibits a sequence of three base pairs at the 5' end of the DNA — cytosine-adenine-guanine (CAG) coding for the amino acid glutamine — that is repeated multiple times (...CAGCAGCAG...). This region is called a trinucleotide repeat. The figure below shows a distribution of the normal and expanded length of the HD trinucleotide repeat.



Normal and expanded repeat length HD

The distribution of repeats for the Huntington's disease can be divided into four categories. Repeats of 26 or less are perfectly normal. repeats between 27 and 35 are rare and are not associated with expression of the disease, but occasionally fathers with repeats will transfer a repeat to their heirs which is expanded within the interval for expression of the disease. Repeats between 36 and 39 are associated with reduced penetrance, in which case some people will develop HD and others will not. Repeats of 40 or more are associated with the expression of HD. Individuals who carry repeats in this category will develop HD, assuming they do not die earlier on in their life by other causes.

repeatlength	**diagnosis**
<27	normal
27-35	low risk
36-39	increased risk
>39	absolute risk

Assignment

1. Write a function repeatlength, that determines the maximum sequence of repeats of a given

string *B* for a given string *A*. The length of the two given strings is variable, and the comparison of both strings has to be executed without differentiating between uppercase and lowercase letters. The table below shows the number of examples of parameter values, and the matching result that should be generated by the function.

string A	string B	result
AATCGTCGTCGTAGCTTCGTGGTGAAGATAG	CTGTA	0
AATCGTCGTCGTAGCTTCGTGGTGAAGATAG	gtg	2
aatcgtcgtcgtagcttcgtggtgaagatag	TCG	3

If string *B* does not occur in string *A*, then the function should return the value zero. If the string *A* contains several subsequences consisting of repeats of the string *B*, then the number of repeats of the longest subsequence should be returned. Repeats never overlap, which means, for example, the string `TTTT` contains two repeats of the string `TT` and not three.

2. Write a function `HuntingtonDiagnosis`, that gives a diagnosis for the given DNA sequence of a Huntington gene concerning the possible risk for the development of Huntington's disease. This diagnosis is of course dependent on the number of repeats of the trinucleotide `CAG` (use the function `repeatlength`), and should be returned as a string similar to the table of diagnosis above. A DNA sequence is represented by a string containing only the lettes `A`, `G`, `C` and `T` (both uppercase and lowercase letters are allowed).

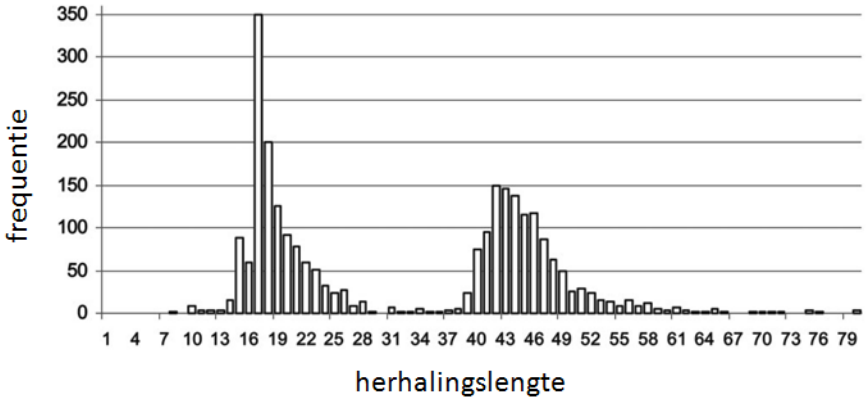
Example

```
>>> repeatlength("AATCGTCGTCGTAGCTTCGTGGTGAAGATAG","CTGTA")
0
>>> repeatlength("AATCGTCGTCGTAGCTTCGTGGTGAAGATAG","gtg")
2
>>> repeatlength("aatcgtcgtcgtagcttcgtggtgaagatag","TCG")
3
>>> HuntingtonDiagnosis('CAG' * 20)
'normal'
>>> HuntingtonDiagnosis('CAG' * 35)
'low risk'
>>> HuntingtonDiagnosis('CAG' * 38)
'increased risk'
>>> HuntingtonDiagnosis('CAG' * 52)
'absolute risk'
```

De ziekte van Huntington (*Huntington's disease*; HD), die in 1872 uitvoerig werd beschreven door de Amerikaanse arts George Huntington, is een dominante erfelijke aandoening die bepaalde delen van de hersenen aantast. De eerste symptomen van HD openbaren zich meestal tussen het 35e en 45e levensjaar, en bestaan onder andere uit ongecontroleerde (choreatische) bewegingen die langzaam verergeren, verstandelijke achteruitgang en een verscheidenheid aan psychische afwijkingen. De ziekte leidt gemiddeld na achttien jaar tot de dood van de patiënt(e), veelal door bijkomende oorzaken zoals longontsteking.

HD behoort tot de zogenaamde klasse van trinucleotideherhalingsziekten (*trinucleotide repeat disorders*), die veroorzaakt worden doordat een repeterende sectie van een gen de normale lengte overschrijdt. Bij HD gaat het om een afwijking in het *Huntingtingen*. Het Huntingtingen vertoont aan het 5' uiteinde van het DNA een reeks van drie baseparen — cytosine-adenine-guanine (`CAG`) coderend voor het aminozuur glutamine — die verschillende keren na elkaar

herhaald worden (...CAGCAGCAG...). Dit fenomeen staat bekend als een trinucleotideherhaling. Onderstaande figuur toont een verdeling van de normale en geëxpandeerde lengte van de HD trinucleotideherhaling.



Normale en geëxpandeerde herhalingslengte van HD.

De verdeling van de herhalingslengte voor het Huntingtingen kan worden onderverdeeld in vier categoriën, zoals aangegeven in onderstaande tabel. Herhalingen van 26 of minder zijn normaal. Herhalingen tussen 27 en 35 komen slechts sporadisch voor en worden niet geassocieerd met het tot expressie komen van de ziekte, maar in zeldzame gevallen zullen vaders met dergelijke herhalingen een herhaling overdragen naar hun nakomelingen die geëxpandeerd wordt tot binnen het bereik waarin de ziekte tot expressie komt. Herhalingen tussen 36 en 39 worden geassocieerd met een verhoogd risico, waarbij sommige individuen HD zullen ontwikkelen en andere niet. Herhalingen van 40 of langer worden geassocieerd met het tot expressie komen van HD. Personen die herhalingen vertonen in dit bereik zullen de ziekte van Huntington ontwikkelen, in de veronderstelling dat ze niet aan andere oorzaken sterven voordat de ziekte zich manifesteert.

herhalingslengte	**diagnose**
<27	normaal
27-35	laag risico
36-39	verhoogd risico
>39	absoluut risico

Opgave

1. Schrijf een functie herhalingslengte, die voor een gegeven string *A* de maximale reeks van herhalingen van een andere gegeven string *B* bepaalt. De lengte van de twee gegeven strings is variabel, en de vergelijking tussen beide strings moet uitgevoerd worden zonder verschil te maken tussen hoofdletters en kleine letters. Onderstaande tabel geeft een aantal voorbeelden van parameterwaarden, en het bijhorende resultaat dat door de functie moet gegenereerd worden.

tekenreeks A	**tekenreeks B**	**resultaat**
AATCGTCGTCGTAGCTTCGTGGTGAAGATAG	CTGTA	0
AATCGTCGTCGTAGCTTCGTGGTGAAGATAG	gtg	2

aatcgtcgtcgtagcttcgtggtgaagatag TCG 3

Indien de tekenreeks *B* niet voorkomt in de tekenreeks *A*, dan moet de functie de waarde nul als resultaat teruggeven. Indien de tekenreeks *A* verschillende deelreeksen bevat die bestaan uit herhalingen van de tekenreeks *B*, dan moet het aantal herhalingen van de langste deelreeks als resultaat worden teruggegeven. Herhalingen overlappen elkaar nooit, zodat bijvoorbeeld de tekenreeks TTTT twee herhalingen bevat van de tekenreeks TT en geen drie.

2. Schrijf een functie HuntingtonDiagnose, die voor de gegeven DNA sequentie van een Huntingtingen de diagnose stelt omtrent het mogelijke risico op het ontwikkelen van de ziekte van Huntington. Deze diagnose hangt uiteraard af van het aantal herhalingen van de trinucleotide CAG (gebruik hiervoor dus de functie herhalingslengte), en moet als string worden teruggegeven overeenkomstig de eerder vermelde diagnosetabel. Een DNA sequentie wordt hierbij voorgesteld door een string die enkel de lettes A, G, C en T bevat (zowel hoofdletters als kleine letters zijn toegelaten).

Voorbeeld

```
>>> herhalingslengte("AATCGTCGTCGTAGCTTCGTGGTGAAGATAG","CTGTA")
0
>>> herhalingslengte("AATCGTCGTCGTAGCTTCGTGGTGAAGATAG","gtg")
2
>>> herhalingslengte("aatcgtcgtcgtagcttcgtggtgaagatag","TCG")
3
>>> HuntingtonDiagnose('CAG' * 20)
'normaal'
>>> HuntingtonDiagnose('CAG' * 35)
'laag risico'
>>> HuntingtonDiagnose('CAG' * 38)
'verhoogd risico'
>>> HuntingtonDiagnose('CAG' * 52)
'absoluut risico'
```