GenBank

The GenBank sequence database is an open access, annotated collection of all publicly available nucleotide sequences and their protein translations. This database receives sequences produced in laboratories throughout the world from more than 100.000 distinct organisms. In the more than 30 years since its establishment, GenBank has become the most important and most influential database for research in almost all biological fields, whose data are accessed and cited by millions of researchers around the world. GenBank continues to grow at an exponential rate, doubling every 18 months. In February 2013, the database contained over 150 billion nucleotide bases in more than 162 million sequences.

GenBank records sometimes contain long genome sequences or sequence fragments. Look at this example GenBank record containing a gene of *Saccharomyces cerevisiae* (baker's yeast). At the bottom of the record, you can see a clear overview of the sequence, using a format that prints the sequence across multiple lines. This is done by breaking the sequence into blocks of size m, and putting n successive blocks separated by spaces on a single line. The letters of the sequence are always converted into lower case in the GenBank format. Depending on the length of the sequence, the last line does not necessarily contain n blocks, and the last block on that line does not necessarily has size m. Each line starts with an integer that indicates the position of the first letter on the line in the original sequence. Positions in the sequence are indexed from 1. The integer that indicates the position is right-aligned over k positions and followed by a space. The integer k is the number of digits of the integer that is at the start of the last line in the GenBank formatted sequence.

Assignment

Write a function genbank that takes three arguments: *i*) the size of a block m, *ii*) the maximal number of blocks per line n and *iii*) a string s that only contains letters of the alphabet (both upper case and lower case letters are allowed). The function must return a string that contains a version of the string s in GenBank format. To do so, the string s must be broken into blocks of size m, and each line must contain up to n blocks, preceded by the position of the first letter of the first block within the string s. Make sure that the last line of the string returned by the function does not end with a newline character.

Example

```
>>> print(genbank(4, 3, 'AGGCTGTCAATGCTAGGCATAgaagtcgTGCTGTAGagatagTCTGATAGTCGC'))
 1 aggc tgtc aatg
13 ctag gcat agaa
25 gtcg tgct gtag
37 agat agtc tgat
49 agtc gc
>>> print(genbank(5, 2, 'GGATGCTGGTAGATCGATAT'))
 1 ggatg ctggt
11 agatc gatat
>>> print(genbank(10, 6, 'AGCT' * 252))
  1 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
 61 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
121 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
181 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
```

241 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
301 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
361 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
421 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
481 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
541 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
601 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
661 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
721 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
781 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
841 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
901 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
961 agctagctag ctagctagct agctagctag ctagctagct agctagct

De GenBank sequentiedatabank is een open en geannoteerde verzameling van alle publiek beschikbare nucleotidesequenties en hun eiwitvertalingen. GenBank ontvangt sequenties die geproduceerd worden in laboratoria over de hele wereld, en dit voor meer dan 100.000 verschillende soorten organismen. Doorheen zijn dertigjarige bestaansgeschiedenis is GenBank uitgegroeid tot de belangrijkste en meest invloedrijke databank voor onderzoek in bijna alle biologische disciplines, met data die wereldwijd opgevraagd en geciteerd worden door miljoenen onderzoekers. GenBank blijft exponentieel groeien, en verdubbelt ongeveer elke 18 maanden van omvang. In december 2012 bevatte de databank meer dan 150 miljard nucleotidebasen in meer dan 162 miljoen sequenties.

In GenBank worden vaak lange genoomsequenties of sequentiefragmenten opgeslaan. Bekijk hier bijvoorbeeld een GenBank record met een gen van *Saccharomyces cerevisiae* (bakkersgist). Onderaan deze record zie je een overzichtelijke weergave van de sequentie, die gebruik maakt van een formaat waarbij de sequentie over verschillende regels uitgeschreven wordt. Hierbij wordt de sequentie opgedeeld in stukken van lengte m, en bevat elke regel n opeenvolgende stukken die van elkaar gescheiden worden door een spatie. De letters van de sequentie worden bij het uitschrijven steeds omgezet naar kleine letters. Afhankelijk van de lengte van de sequentie hoeft de laatste regel niet noodzakelijk n stukken te bevatten, en heeft het laatste stuk op die regel niet noodzakelijk lengte m. Elke regel begint met een natuurlijk getal dat de positie binnen de sequentie aangeeft van de eerste letter op die regel. De posities worden hierbij genummerd vanaf 1. Het getal dat de positie aangeeft wordt rechts uitgelijnd over k posities en gevolgd door een spatie. Hierbij staat k voor het aantal cijfers in het getal dat de grootste positie aangeeft die in de eerste kolom moet uitgeschreven worden.

Opgave

Schrijf een functie genbank waaraan drie argumenten moeten doorgegeven worden: *i*) de lengte van een blok m, *ii*) het maximaal aantal blokken per regel n en *iii*) een string s die enkel bestaat uit letters van het alfabet (zowel hoofdletters als kleine letters zijn toegelaten). De functie moet als resultaat een string teruggeven waarin de string s werd opgemaakt in GenBank formaat. De string s moet daarbij dus opgesplitst worden in stukken van lengte m, en elke regel moet tot maximaal n stukken bevatten voorafgegaan door de positie van de eerste letter van het eerste stuk binnen de string s. De laatste regel van de string die door de functie teruggegeven wordt, mag niet eindigen op een newline.

Voorbeeld

```
>>> print(genbank(4, 3, 'AGGCTGTCAATGCTAGGCATAgaagtcgTGCTGTAGagatagTCTGATAGTCGC'))
```

```
 1 aggc tgtc aatg
13 ctag gcat agaa
25 gtcg tgct gtag
37 agat agtc tgat
49 agtc gc
>>> print(genbank(5, 2, 'GGATGCTGGTAGATCGATAT'))
 1 ggatg ctggt
11 agatc gatat
>>> print(genbank(10, 6, 'AGCT' * 252))
  1 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
 61 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
121 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
181 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
241 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
301 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
361 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
421 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
481 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
541 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
601 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
661 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
721 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
781 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
841 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
901 agctagctag ctagctagct agctagctag ctagctagct agctagctag ctagctagct
961 agctagctag ctagctagct agctagctag ctagctagct agctagct
```