HW 9: Draconic Parsimony

Your job is to deduce a consensus gene from a large collection of modern gene samples. This notebook will be called `parsimony-calculator.ipynb`.

In your dragon-breeding business, you try to breed dragons that breathe interesting jets of flame (in fancy colors and shapes), to compete for the coveted Best of Show trophy in the prestigious annual *Minas Morgul Wyrm Club Dragon Show*^{TM}. You have isolated one important fire-breath-related gene from several of your dragons, and you are curious what gene their common ancestor might have had.

The input file contains many lines of DNA data, each of which contains the gene in question. These genes are largely the same, but will have small subtle differences. They are already aligned with each other, and each one has the exact same length. Each character is one of 5 characters: `'A'`, `'C'`, `'G'`, `'T'`, or `'-'`. The first four indicate one of four the nucleotides, while the dash indicates that this dragon's gene is shorter than the others, and doesn't contain a nucleotide analogous to that location. (For example, if two dragons have the sequences `AAATTT` and `AAACCTTT`, they would be aligned in the file as `AAA--TTT` and `AAACCTTT`.)

You will first ask the user for the file to open. Then you will calculate and output a consensus sequence, which is the hypothesized ancestor sequence. Each character of the sequence is considered separately, and some letters of the sequence will be unknown. If every dragon contained a `'-'` at a given location, then that nucleotide is completely unknown. If every dragon contained a `'-'` or the same letter, then the consensus sequence has that letter. If every dragon contains a `'-'` or one of multiple letters, than it is assumed that the ancestor contained one of those, but you can't know which. The consensus sequence contains all of the observed letters.

What you must output is four lines, each one of equal length, representing the consensus sequence. The first line will contain only `'A'`s and periods. If the consensus sequence holds an `'A'` (that is, at least one dragon had an `'A'` in that location), it will be `'A'`. Otherwise, when no dragon had an `'A'` at that location, it will be a period. The same is true for the next three lines, but with the other nucleotides.

For example, let us say that the analyzed dragons have these sequences:

```
AGT----TAG
AGTCAC-TAG
AGTCCC-TAG
AGTCCT-TAG
```

Then, your output should be:

```
A...A...A.
...CCC....
.G.......G
..T..T.T..
```

Note that to do this assignment properly, you should be using sets. You will be graded down if you do not. A good way to divide up this assignment is as follows:

- a function to open a file, and return a list of strings
- a function to analyze the list of strings, turning it into a list of sets
- a function to output the list of sets as needed

A sample file (`dragons.gene`) has been provided for you. Good luck!