UNIB20005: Language and Computation

Project 1: Truecasing Text

Some text and speech input applications are able to guess the correct casing of text. Your task is to write a program to simulate this behaviour. For example, it will take the input:

```
we collect and preserve the abc radio and television recordings that have
documented the cultural life of australians since the first radio broadcast in 1932.
```

And produce the following output:

```
We collect and preserve the ABC Radio and Television recordings that have
documented the cultural life of Australians since the first Radio broadcast in 1932.
```

You are to implement this behaviour using a function `truecase(s)` which takes a string `s` as its argument, and which returns a string.

Define another function `evaluate(s)` to evaluate your truecase function. The evaluate function should take a string `s` as its argument, lowercase it (to discard case distinctions), then guess the correct casing, then compare the input and output versions. It should return the percentage of words that were correctly cased, as a floating point number.

For example, if our `truecase()` function produced the following output, you would report that 25/26 words were correctly cased, i.e. 96.15%.

```
We collect and preserve the ABC Radio and Television recordings that have
documented the cultural life of Australians since the first radio broadcast in 1932.
```

In this case, your evaluate function would return `96.15`. Note that a word is only scored correct if it is capitalised perfectly, otherwise it is scored incorrect. Note that it is difficult to get a perfect score on unrestricted text.

We should be able to run your program on the command line, specifying a filename, as follows:

```
$ python 4291381.py mytext.txt
73.56
```

Please post any questions about the project on the LMS Discussion Forum.

Submission: Your project is due at the end of week four (10pm, Friday 23 August). Please submit it by email to `sbird@unimelb.edu.au, nj@unimelb.edu.au`. Please use the subject line: `L&C Project 1`. All submissions will be acknowledged, so if you do not receive acknowledgement, please contact staff directly. Please name your file using your student id, for example: `2093183.py`. Any additional files should have the same prefix (e.g. `2093183.data`). Use the following code as a template: `http://lp20.org/lac/projects/truecasing.py`

The project must be original work. It is worth 10% of the total marks for this subject. Your work will be assessed for correctness, clarity, and insight:

correctness (3 marks): how well does your program perform on our test data?

clarity (3 marks): how clearly is your code set out?

insight (3 marks): how insightful is your discussion?

extra effort (1 mark): a bonus mark for anything particularly original or impressive.