Text normalization – true casing

Determining the correct case of a given word in a text. We distinguish between three types of word case: lowercase, UPPERCASE, and Titlecase.

Much of the content that can be found today on the web today – especially on social media outlets, e.g., Reddit and Twitter – are written in an unformal, and often ungrammatical and noisy manner.

Using incorrect word case is one of the phenomena characterizing this language.
The task of true-casing is important since the correct case carries over clues for various NLP tasks, as POS-tagging, Named Entity Recognition (NER), and word-sense disambiguation.

Our goal in this assignment is to determine the correct case of each word in a given text, with highest possible accuracy.

Input: (1) a large well-formed English corpus, (2) a text file (a sample test file with text collected from Reddit and slightly case-distorted is attached to the assignment).
Output: a text file with exactly the same format as (2), where each word is true-cased. Note that you are not likely to achieve perfect results, you can only do as well as your dataset (1) is.

For example, these input paragraphs:

```
There is a prior LIFE experience of support that matters here, too.
A Person who was raised in a stable environment, with adults that were
nurturing, protective and engaged, is more likely to come out of a
traumatic situation less traumatized.
The US president would speak shortly, let us wait a few more minutes.
```

Will result in the following output paragraphs:

```
There is a prior life experience of support that matters here, too.
A person who was raised in a stable environment, with adults that were
nurturing, protective and engaged, is more likely to come out of a
traumatic situation less traumatized.
The US president would speak shortly, let us wait a few more minutes.
```