Why text extracting doesn't work for all PDFs

You may have come across PDF files in the past, where you could not extract or copy'n'paste text from their pages.

One reason could be a deliberate "protection" (by setting a "user password" – even if empty) on the PDF file. Another reason could be that the "text" in reality is a raster image, or made up of vector shapes. These cases are not considered with the files contained in this sub directory (*"handcoded/textextract"*).

Here we take a look at PDF files which indeed use "fonts" to represent "text".

We'll have 5 different files here, all of them identically looking in all PDF viewers, and very similar in their PDF source code. Of course there are some differences in the source code which make them to behave differently whenever you try to access the textual content outside from *rendering* the pages:

1. **`textextract-good.pdf`**
   This file lets you extract or copy'n' paste all text correctly.
2. **`textextract-bad1.pdf`**
   This file lets you extract or copy'n' paste all text – but none of the strings appears correctly, all of them look like gobble-di-gook.
3. **`textextract-bad2.pdf`**
   This file lets you extract or copy'n' paste all text – but only the first half of the strings appears correctly, the other half is somehow garbled.
4. **`textextract-bad3.pdf`**
   This file lets you extract or copy'n' paste all text – but only the second half of the strings appears correctly, the first half is somehow garbled.
5. **`textextract-bad4.pdf`**
   This file lets you extract or copy'n' paste all text – but only the first half of the strings appears correctly, the second half may, superfically looking, appear correct, but in reality is *"rot13"* encoded.

If you compare the different files's source code, you'll easily find where they differ: it's the way how they do define and set up (or don't do it at all) their `/ToUnicode` tables for the respective font used by the text strings.

A missing or an incorrect or a corrupt `/ToUnicode` table is the number 1 reason why text is not completely or correctly extractable from (otherwise apparently correct and complete and spec-conforming) "unprotected" PDF files.

You are invited to try the following commands:

```
$ pdfinfo   textextract-good.pdf
$ pdffonts  textextract-good.pdf
$ pdftotext textextract-good.pdf -
```

Do the same for each of the `textextract-bad[1-4].pdf` files. Find out what the differences for each of them are when compared to the `textextract-good.pdf`. Also compare the `textextract-bad[1-4].pdf` files to each other.

You will see, that it is the `/ToUnicode` informations contained in each of the files which determines whether, and to what degree *correctly*, each text extraction (or copying) works. However, `/ToUnicode` does not have any influence on the *rendering* and readability of the PDF page in the PDF viewers.

Manipulated `/ToUnicode` tables can be (ab)used to confer hidden (by obscurity) messages to a receiver, which only reveal their intended meaning when extracted as text, but which use a "decoy" text when reviewed on a rendered PDF page.
