Find a Dork Tower Comic Strip I remember with OCR and agrep

I just wanted to link someone a dork tower comic strip I remember and ended up OCRing all the strips because search engines may be throwing out billions of dollars for AI but can't do text recognition on comic strips that have been online for two decades.

Foreword: This is kind of unkind, because it downloads all of Dork Tower. Don't do this at home. Except if you really need to find that one awesome strip and search engines only turn up old reddit threads.

Catch 'em all

This isn't nice. On the other hand, bandwidth these days has become really cheap, so it should not hurt anyone. We'll at least limit this to the actual structure (all the comic html pages are in subfolders based on the year, and the images are in `files` — thanks for good structuring!).

```
wget -e robots=off -N -mkE http://www.dorktower.com \
     http://www.dorktower.com/{1997..2024} \
     -I $(echo {1997..2024} | sed 's/ /,/g'),files
```

`-e robots=off` means "this is not a search engine, ignore whatever is in robots.txt". `-N` at least says "do not download anything twice" and `-I` limits the download to actual comic pages.

Now all of Dork Tower is in a folder named www.dorktower.com.

Next step: OCR with `tesseract-ocr`. English for Dork Tower is already included in the base package, but OSD needs the tessdata package. To get it in guix, use

```
guix shell tesseract-ocr tesseract-ocr-tessdata-fast
```

All other distros should have it, too.

Train to understand

Now OCR all images:

```
cd www.dorktower.com && \
  ls files/*/*/*.jpg | grep -v -- - | \
  xargs -P8 -I {} -d "\n" tesseract {} {} --psm 12
```

Adjust `-P8` to the number of processors you have. The second `{}` is intentional: that gives the output file (tesseract appends `.txt`). `--psm 12` searches for all text with orientation detection.

Search far and wide

Now we can search with grep or ripgrep or ag (the silver searcher) or similar for the strip we remember:

```
grep -i -C2 campaign files/*/*/*.txt
```

And if you want to quickly thumb through all the results because you actually remember how it **looked**:

```
grep -l -i campaign files/*/*/*.txt | sed 's/\.txt//' | \
  xargs gwenview
```

This is far from perfect, because all those scripts are lettered by hand and OCR for that is hard, but it's better than Google and `tre` / `agrep` can help with fuzzy searching (use `guix shell tre` to get `agrep`):

```
agrep -l -2 -k -i "fellowship" files/*/*/*txt  | sed "s/.txt//" | xargs gwenview
```

If search engines fail at the most basic tasks, we have to do them ourselves.

The friends along the way

But even though I can now search these strips efficiently, I failed, because I cannot find that one Dork Tower strip where they call the eagles to destroy the lord of the rings plot by dropping the ring into mount doom while the air defenses of Mordor are still weak.

I wonder whether I remember it from a printed version I have.

I'm sorry, John Kovalic, for annoying your server so much.

But thank you for great memories which are stronger than Google.

The titles may be inspired by the recent episode of moon channel.