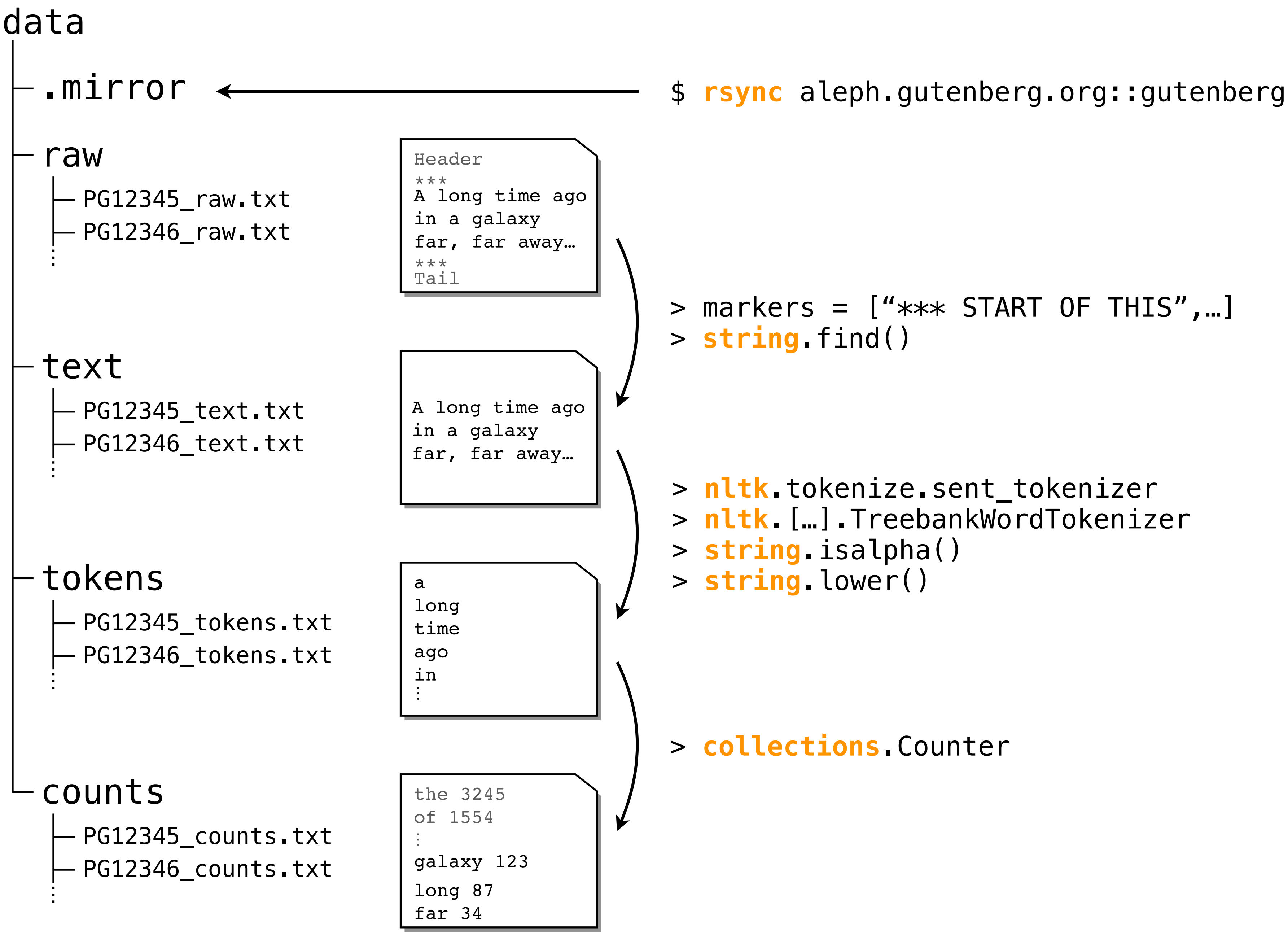

```
data
├ .mirror
├ raw
│  ├ PG12345_raw.txt
│  ├ PG12346_raw.txt
│  ⋮
├ text
│  ├ PG12345_text.txt
│  ├ PG12346_text.txt
│  ⋮
├ tokens
│  ├ PG12345_tokens.txt
│  ├ PG12346_tokens.txt
│  ⋮
└ counts
   ├ PG12345_counts.txt
   ├ PG12346_counts.txt
   ⋮
```

```
$ rsync aleph.gutenberg.org::gutenberg
```

```
Header
***
A long time ago
in a galaxy
far, far away…
***
Tail
```

```
> markers = ["*** START OF THIS",…]
> string.find()
```

```
A long time ago
in a galaxy
far, far away…
```

```
> nltk.tokenize.sent_tokenizer
> nltk.[…].TreebankWordTokenizer
> string.isalpha()
> string.lower()
```

```
a
long
time
ago
in
⋮
```

```
> collections.Counter
```

```
the 3245
of 1554
⋮
galaxy 123
long 87
far 34
```