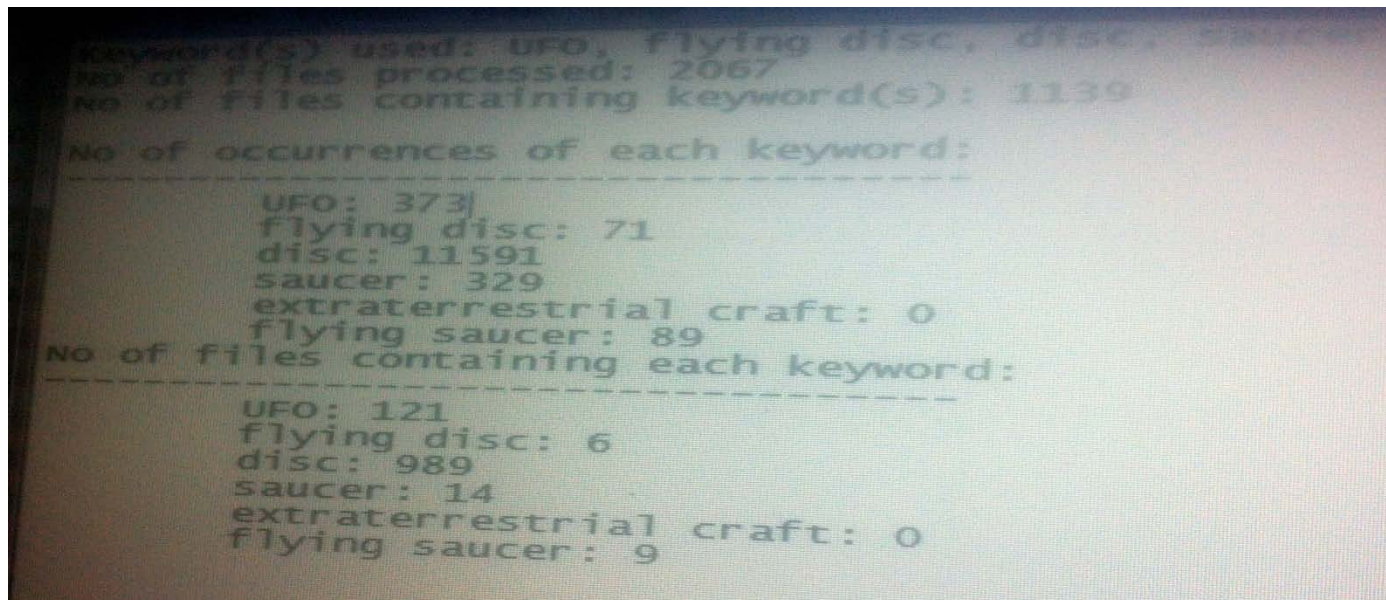


Observations:

The corpus contains a lot of files containing the given keywords and not just 23.

1) Initially searching the corpus for the files with the given keywords yields the following results:



```
Keyword(s) used: UFO, flying disc, disc, saucer
No of files processed: 2067
No of files containing keyword(s): 1139

No of occurrences of each keyword:
-----
UFO: 373
flying disc: 71
disc: 11591
saucer: 329
extraterrestrial craft: 0
flying saucer: 89
No of files containing each keyword:
-----
UFO: 121
flying disc: 6
disc: 989
saucer: 14
extraterrestrial craft: 0
flying saucer: 9
```

2) Refining the search to remove the words that could be substrings in some other word, a space is appended before and after the keyword. This reduces the number of files containing keywords by pretty big number.

```
Keyword(s) used: UFO, flying disc, disc, saucer, extraterrestrial craft, flying saucer, space ship, space craft, air ship, phantom, space probe
No of files processed: 2067
No of files containing keyword(s): 123

No of occurrences of each keyword:
-----
UFO: 156
flying disc: 8
disc: 70
saucer: 88
extraterrestrial craft: 0
flying saucer: 15
space ship: 4
space craft: 3
air ship: 0
phantom: 1
space probe: 0
No of files containing each keyword:
-----
UFO: 65
flying disc: 3
disc: 34
saucer: 8
extraterrestrial craft: 0
flying saucer: 6
space ship: 3
space craft: 3
air ship: 0
phantom: 1
space probe: 0
```

3) Changing the original String and the keywords to be searched to lower case and then searching the corpus again increases the counts of a few results and keeps others unchanged. The results are as shown below:

```
Keyword(s) used: UFO, flying disc, disc, saucer, extraterrestrial craft, flying saucer, space ship, space craft, air ship, phantom, space probe
No of files processed: 2067
No of files containing keyword(s): 120
```

```
No of occurrences of each keyword:
```

```
-----
UFO: 158
flying disc: 15
disc: 113
saucer: 120
extraterrestrial craft: 0
flying saucer: 30
space ship: 4
space craft: 3
air ship: 0
phantom: 1
space probe: 0
```

```
No of files containing each keyword:
```

```
-----
UFO: 65
flying disc: 4
disc: 55
saucer: 8
extraterrestrial craft: 0
flying saucer: 6
space ship: 3
space craft: 3
air ship: 0
phantom: 1
space probe: 0
```

Notes on Tika:

Apache Tika is fairly easy to use and good examples are available to understand how to implement it. It is a great tool that compiles