Assumptions Made:

1. The keywords as well as the content of all pdfs were case folded, in order to get more hits (as UFO, ufo and UfO are same).
2. I have also considered the plural forms of the keywords. The proper way of doing it is through Stemming, but since the number of keywords are very less, i have changed the regular expression accordingly to count even the plurals. So for both UFO and UFOs, the count is increased and its value is written is against the keyword UFO in output.txt.

Observations:

1. When the documents in Vault folder was tested against the default Keywords list, the total number of files with the keywords came to be around 206. Out of which, there were 345 occurrences of word UFO.
2. When the documents in Vault folder was tested against the new Keywords list, the total number of files with the keywords came to be around 396. Out of which, there were 345 occurrences of word UFO and 2526 occurrences of word alien. Clearly, we can see the word 'alien' has more frequency than UFO. Generally speaking, we can imply the term alien is more used than the term UFO so that's why there were more hits.
3. When the plural situation was handled as per the Assumption 2, the number of files containing the keywords increased drastically.
4. When the case folding situation was handled against as per the Assumption 1, the number of files containing the keywords increased.

Reasons for discrepancies between the calculated figure and the reported figure:

1. When UFO is searched on the website, things like case folding and Stemming are not taken care of, that is why it shows so less results.
2. The Search on the website is done purely based on word, it is not done semantically. The results does not show files which contains words which is semantically similar to UFO(for example flying saucer or discs).
3. When I searched UFO without case folding and stemming, I found 111 files containing the keyword UFO which is way greater than 23, so I can conclude that the search engine on the FBI website does not check whether the content of the document contains the keyword but it only checks whether the metadata of the PDF contains the keyword.

Apache Tika :

It was easy to use and it makes the life of a programmer very easy through their collection of parser libraries. Many more parsing libraries can be included seeing the number of content type increasing everyday.