Computer aided simulations and performance evaluation

Lab 9 - Fingerprinting for movie titles

Gabriele Cuni 277957

9.1 Input Parameters

- **seed**: It is used to set the random library seed, which is used to shuffle two times the list of all the titles.
- **words**: It is a python list of 170265 unique italian movie titles coming from the IMDb [1]

9.2 Output Parameters

- **bExpMin:** It is the minimum number of bytes of the fingerprint for which there are no collisions in the **fingerprintSet** given the input list of titles.
- **bTeo:** It is the theoretical number of bytes of the fingerprint for which there are no collisions in the **fingerprintSet** given the input list of titles.
- **pFalsePositive:** It is the probability of false positives given the number of titles and the fingerprintSet range.
- **fingerprintSetBytesSize:** It is the data structure size of **fingerprintSet** computed by means of pympler [2].
- **wordsSetBytesSize:** It is the data structure size of **wordsSet** computed by means of pympler [2].

9.3 Main Data Structures

- **experiments_result:** it is the python dictionary which contains all the simulation results.
- **fingerprintSet:** It is the python set used to store all the fingerprints and it is used to assess if a collision occurs.
- **wordsSet:** It is the python set of all titles used to be compared with **fingerprintSet**. **wordsSet** is used, instead of **words** (which is a list), because given the same content, python sets take up more memory space than python lists.
- **bExp:** it is used to find the **bExpMin**. It is an incremental variable.
- **collision:** It is a flag variable used to assess if a collision is occured.

9.4 Formulas

- $bTeo = \log_2(\frac{m}{\epsilon})$ where m is the number of titles and ϵ is 0.5
- $P("False\ positive") = 1 - (1 - \frac{1}{n})^m$ where m is the number of titles and n is $2^{bExpMin}$
- $m \cdot b$ where m is the number of titles and b is **bTeo**. It is computed inside plot9.py by plotting

9.5 Python Functions

- **pympler.asizeof.asizeof():** It is used to assess the memory occupancy in bytes of **fingerprintSet** and **wordsSet**. [2]
- **random.shuffle():** It is used two times to shuffle the titles inside the list **words,** in order to make them not in the alphabetic order.
- **hashlib.md5():** It is used to obtain the hexadecimal md5 string.

9.6 Charts and Reasoning

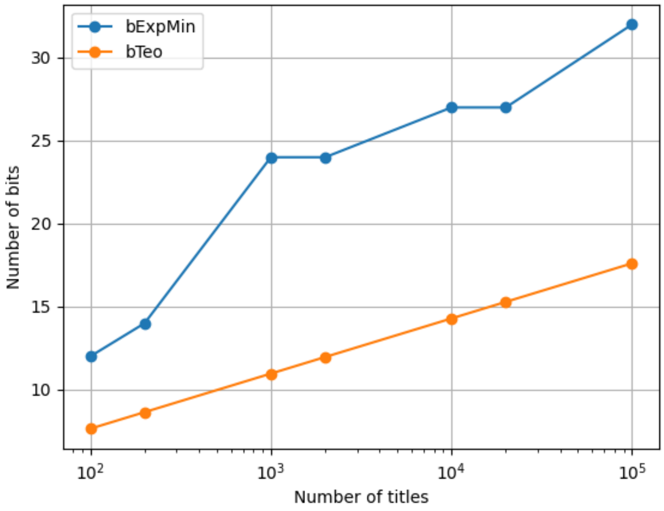

Figure 1 shows that **bTeo** is about half of **bExpMin** which it seems right given that **bTeo** is computed with $\epsilon = 0.5$, instead **bExpMin** is simulated as if $\varepsilon \rightarrow 0$.
The experiments are done with [100, 200, 1000, 2000, 10000, 20000, 100000] titles.

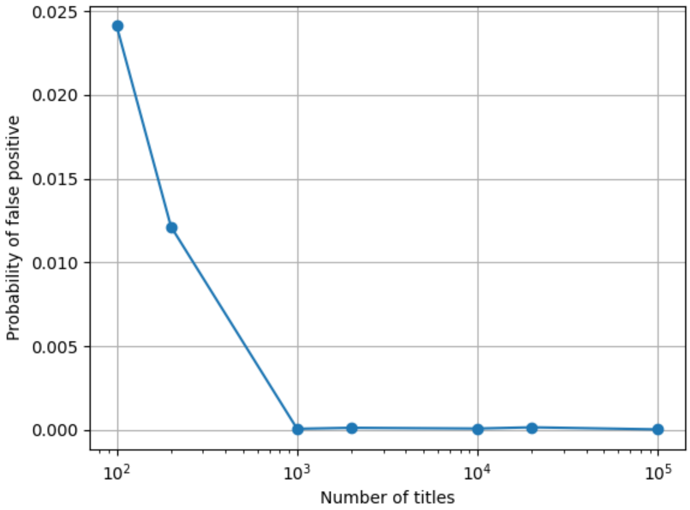

Figure 2 shows the probability of collision computed as shown in section 8.4 v.s. the number of words as before. As shown the probability of collision tends to zero by increasing the number of titles, because by doing so, the fingerprint range increases a lot faster than the number of titles and the wasted space is higher.

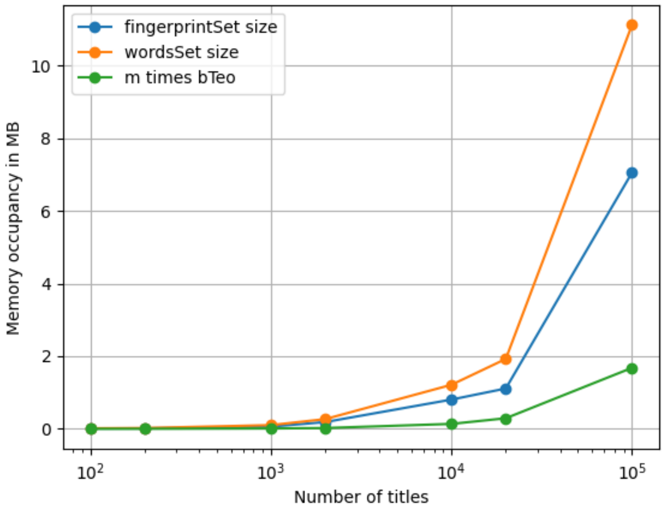

Figure 3 shows a comparison between the **fingerprintSet**, the **wordsSet** memory occupancy and the theoretical occupancy given by $m \cdot b$.
As can be shown it is worth using the fingerprint also to save memory.
The memory space of the python sets is much larger than the theoretical one because the python set object stores more data other than its content.
The important thing to notice is that all follow the same exponential trend given the number of titles.

9.7 References

[1] https://datasets.imdbws.com/%C2%A0
[2] https://pympler.readthedocs.io/en/latest/