

Kata 5: Word Cloud - Infographic

Breakdown:

- Use Hash Tables $\langle \text{String}, \text{int} \rangle$, where string is the word (Key), and value, the proposed font size for the word cloud.
- The higher the value, the higher the frequency.

Input: "After beating the eggs, Dana read the next step"
"Add milk and eggs, then add flour and sugar"

What can be done?

A hash table **must not** be designed such that "Add" and "add" are different words. In a higher level, it could be to learn words. For example:

hola = ola = holis = qubo =

But we need to clean the string first:

- Remove unnecessary whitespaces
- Remove cases HOLA → hola

Furthermore, we can add some word recognition and get rid of articles, prepositions or repetitive words or tokens.

But how?

Steps to follow:

① Clean string: Remove spaces n

② Change all to lowercase n

// This is done so we can avoid errors on the hash

③ For every token in string: n

- Insert into hashtable 1

- Set initial size to 12. 1

- If there are previous occurrences, then increase the size of the font. 1

④ Iterate over the HT to print all keys and values n

Average complexity: $O(n)$

For this kata, the hash table implementation will be the 'class' implementation. So, the code added will correspond to a string tokenizer and the functions needed to "clean" the string.