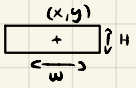
Word cloud

Input :
- A list of words with :
  - a ID
  - a text
  - a confidence level
- A container width and height (if not default value)

Output :
- A word cloud

Implementation :

- Set coordinate of the container
- Sort the words by the confidence level. The biggest in first.
- Define each word as a rectangle 

with (x,y) center and w width, h height

The size of the rectangle corresponds to the confidence level. The larger the confidence level, the larger the rectangle.

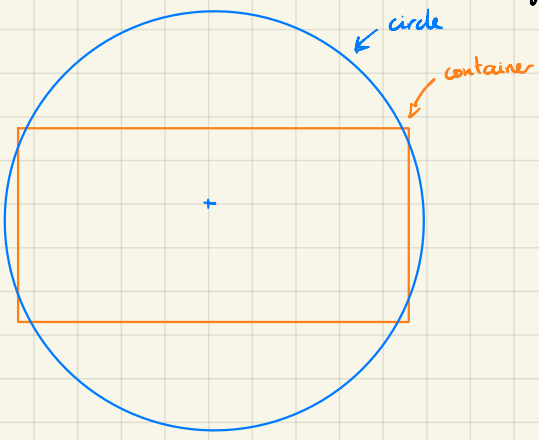
The front size is define as :

$$(\text{confidence level} - \text{cut\_off}) \cdot \left(\frac{1}{1-\text{cut\_off}}\right)^2 \cdot (\text{max\_font\_size} - \text{min\_font\_size}) + \text{min\_font\_size}$$

coefficient — max and min of rectangle size

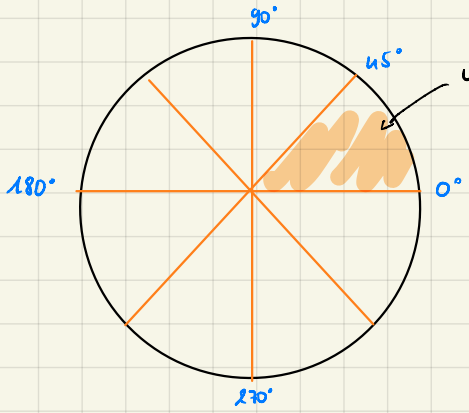
There are many words with a high coefficient and few with low coefficient. The aim is to be able to differentiate words with high coefficients and less differentiate words with lower coefficients

- Place the first word of the list in center of the container (so the larger word)
- Place the other words, a loop for each word :
  - Place the word in a random position on the container
    - Define a circle with centre the centre of the container



- if there are already some words placed in the word cloud the centre of the circle is the sum of mass of already placed words.

We cut the circle in multiple intervals, each of this intervals correspond to degrees



By creating intervals in this circle, we draw an interval at random and place the word on the circle in the corresponding interval.
Because the word will then be moved to the words already placed.
This circle then allows to make a cloud by placing uniformly the words, and to obtain a shape of word cloud.

We then have a weight vector that counts the number of words placed in a certain interval

weight = [1, 4, 0, 3]

↑ interval [0°, 90°[

So that the intervals with the fewest words have the best chance to being drawn, we subtract the maximum height from all the weights

invert weight = [3, 0, 4, 1]

max 4 - 1

The cumulative sum is then calculated

3+0 → 3 ; 3+0+4+1 → 8

cumulative weight = [3, 3, 7, 8]

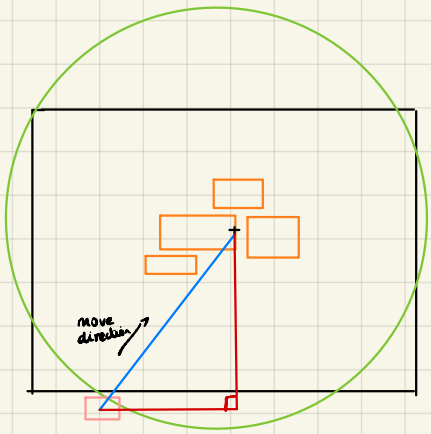
A number is drawn at random between 0 and the maximum of the cumulative sum

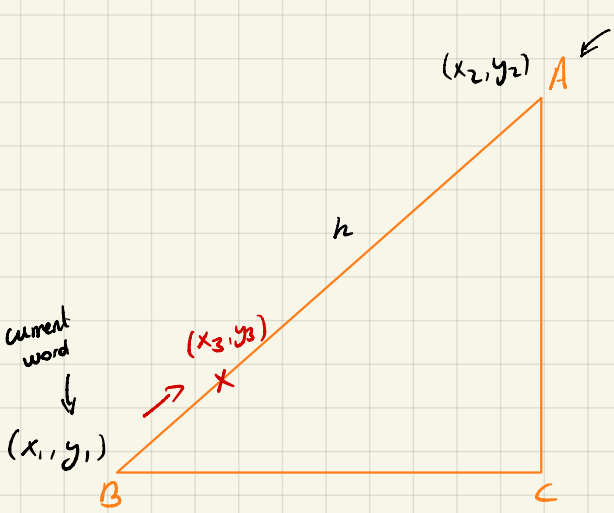
From 0 to 3 it correspond to [0°, 90°[

So intervals with fewer words will have a better chance of being chosen

- Now that we have placed our word on the circle, we will move it closer to the words already placed.

We compute the move direction, by sum the differences between the already placed words and the current word.





$$dx = \frac{step}{h} \cdot x_2$$

$$dy = \frac{step}{h} \cdot y_2$$

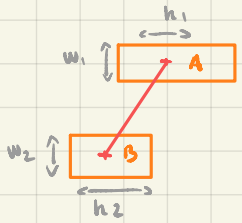
$$x_3 = x_1 + dx$$

$$y_3 = y_1 + dy$$

We move with the previous calculation our word, but care must be taken to avoid collisions with other words.
We calculate whether the move creates a collision on x and y:

- If not, we move the current word and start again
- If collision on x, we move the y position of the current word and we start agin
- If collision on y, we move the x position of the current word and we start again
- If collision on both side, the word is placed and we are done

To calculate collisions, we check if the distance between the center of the current word and the center of already placed words is not smaller than the height and width of each words:



$$A.x - B.x > \frac{A.w_1}{2} + \frac{B.w_2}{2}$$

$$A.y - B.y > \frac{A.h_1}{2} + \frac{B.h_2}{2}$$

When we have placed all the words, we calculate the boundary box of our word cloud. Thanks to this box, we can move all the words in our word cloud at once

The is also a padding on each word (added when calculating the size of the rectangle), this makes the word cloud more uniform.