# Introducere

Intr-un proces de digitalizare a informatiei in cadrul departamentului financiar-contabil al unei companii, facturile primite prin posta de la furnizori sunt scanate si tranformate in documente PDF ce se prelucreaza ulterior in vederea extragerii informatiei utile din aceste documente, informatie care este apoi trimisa catre aplicatiile financiar/contabile ale companiei.

Procesul de extragere a informatiei presupune doi pasi importanti: recunoasterea optica a caracterelor (OCR), sau faza de detectie, in urma careia rezulta un set de cuvinte recunoscute asezate la o anumita pozitie in pagina (coordonatele x si y precum si lungimea si latimea blocului text), urmata de identificarea si validarea informatiei utile din lista de texte recunoscute (numarul facturii, numele furnizorului, data factura, totalul, etc).

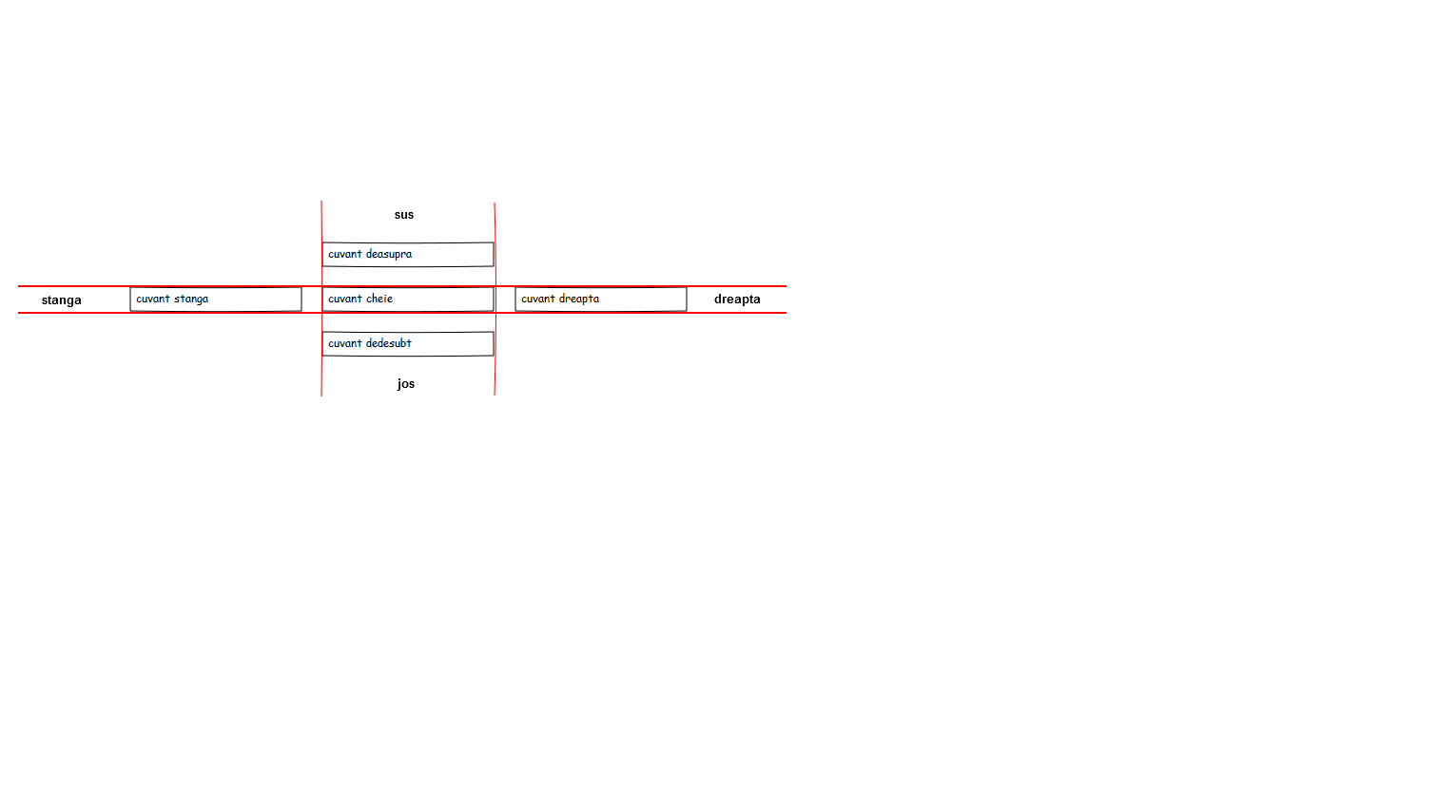
# Enunt

Pentru faza de identificare/validare se foloseste un mecanism pe baza de template-uri in care se identifica semnficatia unui anumit cuvant, in functie de cuvintele din jurul acestuia. Mai exact prezenta unor anumite cuvinte cheie, in imediata vecinatate a unui text, garanteaza intr-o anumita masura (pondere) semnificatia acestuia (de exemplu prezenta cuvantului „Numar Factura:” in stanga unui numar semnifica cu un grad mare de acuratete ca acest numar reprezinta numarul facturii)

Se cere implementarea unui algoritm care identifica semnificatia un cuvant din setul rezultat in urma fazei de detectie pe baza unui set de template-uri.

Template-urile sunt date sub forma unei liste de 5 valori din care prima reprezinta un identificator de template (cel care, in practica, determina semnificatia cuvantului recunoscut cu acest template) dat ca numar intreg, celelalte fiind, in aceasta ordine, cuvantul de sus, dreapta, jos, respectiv stanga lui. Atunci cand cuvantul vecin nu este specificat, in input se va afla valoarea “-”, iar acesta nu va mai fi urmat de o pondere. Fiecare din cele 4 cuvinte, are asociata ca pondere, un numar zecimal. Ponderea validitatii template-ului aplicat pe un anumit cuvant se calculeaza insumand ponderile cuvintelor aflate la pozitiile de sus, dreapta, jos si stanga relativ fata de cuvant, considerandu-se valoarea ponderii unui cuvant doar atunci cand acesta se identifica cu cuvantul aflat la pozitia corespunzatoare din template, relativ fata de cuvantul dat (in caz contrar, se va considera o pondere de 0).

Definim un cuvant ca fiind aflat deasupra unui alt cuvant daca se afla pe aceeasi coordonata X cu el si este cel mai apropiat pe axa Y de dreptunghiul care incadreaza cuvantul, fara a-l intersecta. Similar, cuvintele din stanga si dreapta sunt cele care au aceeasi coordonata Y si sunt cele mai apropiate pe axa X de cuvantul dat, fara a-l intersecta.



Definitia aliniamentului unui cuvant sus, jos, stanga sau dreapta

# Descriere intrare

Inputul este citit dintr-un fisier de intrare “input.txt” sub forma urmatoare:

* Pe prima linie: n (numarul de cuvinte din textul documentului scanat)
* Pe liniile 2 pana la (n+1), cuvintele din text cu structura *c x y w h* unde:
  + c – textul cuvantului aflat la pozitia (x, y)
  + x – pozitia pe axa Ox a coltului din stanga sus al textului
  + y – pozitia pe axa Oy a coltului din stanga sus al textului
  + w - latimea cuvantului
  + h - inaltimea cuvantului
* Pe linia (n+2): t (numarul de template-uri)
* Pe liniile (n+3) pana la (n+t+3), template-urile cu structura *id ct pt cr pr cb pb cl pl* unde:
  + id – identificator template
  + ct – cuvantul de sus
  + pt - ponderea cuvantului de sus
  + cr – cuvantul din dreapta
  + pr - ponderea cuvantului din dreapta
  + cb – cuvantul de jos
  + pb - ponderea cuvantului de jos
  + cl – cuvantul din stanga
  + pl - ponderea cuvantului din stanga
* Pe linia (n+t+4): indicele cuvantului de identificat din lista de cuvinte detectate (primul cuvant din lista are indicele 0)

# Descriere iesire

Programul trebuie sa afiseze in consola id-ul template-ului care are cea mai mare pondere de recunoastere pentru cuvantul dat. In cazul in care sunt mai multe template-uri cu aceasi pondere trebuie afisat id-ul primului template din lista.

# Exemplu:

## Input

9

text1 6 18 5 2

text2 0 15 5 2

text3 6 15 5 2

text4 0 12 5 2

text5 6 12 4 2

text6 12 12 5 2

text7 6 9 5 2

text8 0 6 4 2

text9 6 3 5 2

4

1 - - text7 0.5 -

2 text3 0.9 - - -

3 - text4 0.3 - -

4 text 0.6 text6 0.1 text3 0.2 -

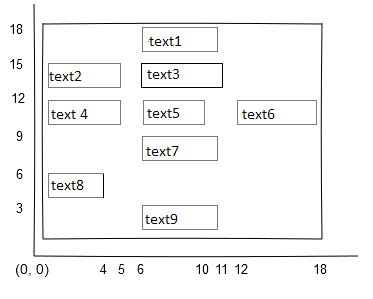
4

## Output:

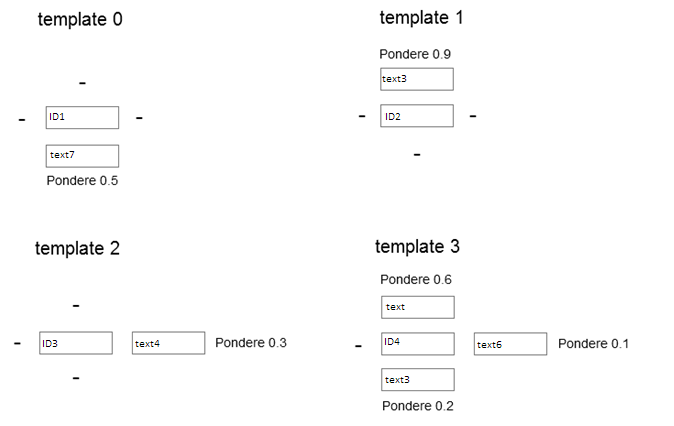
2

## Explicatie:

In poza de mai jos se poate vedea un sketch al cuvintelor in pagina si a template-urilor:



Cuvintele din pagina



Template-urile date

Cuvantul dat ca input este cel cu indicele 4, adica text5 aflat la coordonatele (6,12) de lungime 4 si latime 2.

Prin calcularea pozitiei celorlalte cuvinte din pagina, se determina sus cuvantul *text3*, in dreapta cuvantul *text6*, jos cuvantul *text7* iar in stanga cuvantul *text4.* Prin suprapunerea celor 4 template-uri date ca input se determina urmatoarele ponderi:

* + Template nr. 0 cu ponderea 0.5 (cuvantul text7 gasit la pozitia (6, 9))
  + Template nr. 1 cu ponderea **0.9** (cuvantul text3 gasit la pozitia (6, 15))
  + Template nr. 3 cu ponderea 0.1 (cuvantul text6 gasit la pozitia (12, 12))

Rezultatul final este template-ul nr. 1 cu ID-ul 2 care are o pondere de 0.9