Modulul String Matching

**Realizatori**

Baciu Ciprian

Matei Alexandru

Grierosu Claudiu

Bran Razvan

El-Nazli Cezar

Vitel Andreea

Luncasu Anca

**Grupa** B6

**Anul** 2

# Cuprins

1. Descriere
2. Domeniu
3. Actionari/Interese
4. Actori&Obiective
5. Scenarii de utilizare
6. **Descriere**

Se realizeaza modulul de aliniere si potrivire pentru proiectul de sincronizare si comparare de texte formatate in limbaj de marcare.

Modulul va primi doua texte in limbaj de marcare deja validate si cu etichetele irelevante eliminate, dupa care le va alinia si va calcula un procent de potrivire intre texte si intre paragrafele acestora.

Rezultatul compararii va fi furnizat modulului urmator.

1. **Domenii**

Se vor descrie scenariile de utilizare pentru calcularea procentajului in urma caruia vor fi sincronizate textele.

1. **Actionari/Interese**

**Modulul de validare**: pune la dispozitia modulului curent textele in format de marcare valid si cu etichetele irelevante eliminate.

**Modulul de comparare**:primeste doua texte valide in format mark-up, elimina spatiile redundante, alinieaza textele, dupa care le compara, furnizand rezultatul compararii modulului urmator.

**Modulul de sincronizare:** asteapta rezultatul modulului de comparare si se foloseste de rezultatul primit pentru a sincroniza cele doua texte.

1. **Actori & Obiective**

**Modulul de validare:** ofera un text valid modulului de comparare.

**Modulul de comparare:** alinieaza si compara textele si paragrafele continute de acestea, dupa care ofera o aliniere si un procentaj de potrivire.

**Modulul de sincronizare:** primeste rezultatele modulului anterior si le foloseste mai departe.

1. **Scenarii de utilizare**
   1. **Eliminarea spatiilor redundante**
      1. **Obiectiv/Context**

Eliminarea spatiilor redundante din jurul textului, dintre paragrafe si din interiorul textului, precum si aducerea documentului la un anumit stil (e.g. eliminarea spatiilor dinaintea semnelor de punctuatie, adaugarea spatiilor dupa semnele de punctuatie).

* + 1. **Scenariu/Pasi**

1. Sunt eliminate spatiile albe din jurul documentului (spatii, linii goale, tab-uri). De asemenea, se rezolva probleme tehnice precum carriage return-ul din documentele create pe Windows (prin eliminarea sa din text) si caracterele tab (prin inlocuirea lor cu un singur spatiu).
2. Sunt eliminate liniile redundante dintre paragrafe, acestea fiind delimitate printr-un singur newline. Una sau mai multe linii consecutive in interiorul frazelor sunt inlocuite printr-un spatiu, iar in cele din urma, grupurile de unul sau mai multe spatii consecutive sunt inlocuite cu unul singur.
3. Se corecteaza eventualele greseli de stil (e.g. spatiu inaintea unui semn de punctuatie, spatiu dupa o paranteza deschisa, lipsa unui spatiu intre o paranteza inchisa si un cuvant).
   * 1. **Extensii**
4. Corecturile se pot adapta in functie de structura documentului.
   1. **Impartirea textului pe paragrafe**
      1. **Obiectiv/Context**

Impartirea textului pe paragrafe.

* + 1. **Scenariu/Pasi**

1. Se imparte textul in paragrafe multiple, in functie de structura acestuia.

* + 1. **Extensii**

1a. Daca mai multe paragrafe sunt identice dar au o pozitie diferita in text, acestea vor fi aliniate (aliniere globala).

* 1. **Impartirea pe propozitii a paragrafelor**
     1. **Obiectiv/Context**

Impartirea pe propozitii a paragrafelor.

* + 1. **Scenariu/Pasi**

1. Se imparte fiecare paragraf in propozitii dupa semne de punctuatie.
   1. **Alinierea textelor**
      1. **Obiectiv/Context**

Textelor obtinute li se aplica o aliniere locala prin intermediul algoritmului Smith–Waterman, dupa care se determina un procent de potrivire. Daca acesta trece de un anumit prag determinat ulterior empiric, cele doua texte sunt trimise mai departe modulului de sincronizare. Pentru acest pas vom folosi o biblioteca externa ce ofera deja o parte din functionalitatea descrisa (JAligner). Biblioteca are nevoie de o matrice pentru calculul potrivirii, iar aceasta i se va furniza ulterior, in functie de natura textelor.

* + 1. **Scenariu/Pasi**

1. Cele doua texte initiale se convertesc intr-un format intern bibliotecii.
2. Se determina o aliniere locala, utilizand algoritmul descris si matricea suport.
3. Pentru aceasta aliniere se determina procentul de potrivire.
4. Textele sunt trimise modulului de sincronizare.
   * 1. **Extensii**

3a. Daca procentul de potrivire nu trece de un prag, textele nu vor fi trimise mai departe.