TP Génie logiciel - Scrum
2023
Parseur d'articles scientifiques en format texte
*** Version 1

Le Département de contrôle de qualité demande une **évaluation de la qualité de votre système**. Cette évaluation peut être réalisée en fonction de la **précision** obtenue sur un **corpus de référence** (découpage de sections fait par des humains). Ainsi, les fragments de texte extraits présents dans les références seront comptabilisés positivement; leur absence sera pénalisée (voir l'exemple en Annexe). Le corps de l'article est exclue de la comparaison, compte tenu que c'est la section la plus difficile.
La précision sera calculée comme suit :

```
PrécisionTotal = Σ (Precison_Par_Sections)
```

Le titre de chaque article ainsi que les auteurs seront chacun considérés comme des sections qui doivent être retrouvées. Chaque section sera évaluée séparément par le système et la précision totale est indiquée à la fin. Le site pour tester la précision de vos résultats :

http://inf1603.alwaysdata.net/ParserResultComparator.php

La précision sera mesurée sur un **corpus de test** de **10 nouveaux articles PDF**. Dans un deuxième temps, sur le prochain sprint, je vous fournirais les 10 articles du test.

Le rapport, le système final et les versions intermédiaires doivent être placés sur ENT et sur Github. Chaque version dans une branche différente. Il faut documenter tous les différents événements de la méthode Scrum en ajoutant sur l'ENT des photos de votre travail (réunions de planification de sprint, mêlée quotidienne, revue de sprint, rétrospective du sprint).

Consignes

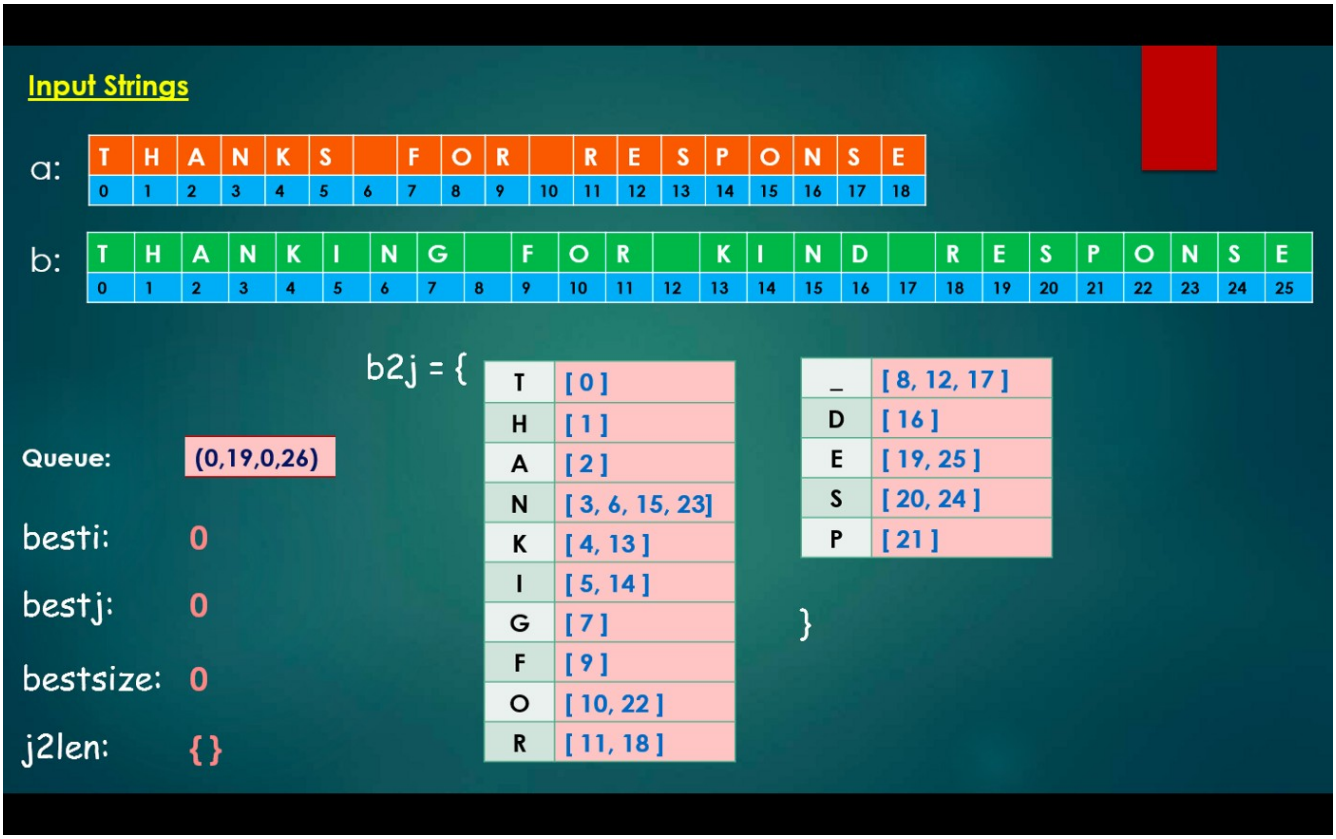
- Déposer vos rapports et codes sur ENT
- Équipes : Fixée en séance 1
- Méthodologie agile : SCRUM
- Plateforme : le projet doit fonctionner sous GNU Linux en ligne de commande.
- Soutenance et rendu prévisionnel : 4 mai 2021
- Soyez agiles !

Vous êtes en compétition ! Attention aux fuites des idées ou de logiciel !

Annexe : calcul de la précision

Le package utilisé pour calculer la précision est sequence matcher. SequenceMatcher est une classe disponible dans le module python nommé "difflib". Il peut être utilisé pour comparer des paires de séquences d'entrée. Dans le cadre de ce projet, une comparaison est effectuée entre chaque section de la référence et votre document XML. Dans l'image exemple ci dessous, une comparaison est effectué entre deux string a et b. Plus d'informations sur :

https://towardsdatascience.com/sequencematcher-in-python-6b1e6f3915fc


Input Strings
a:
T H A N K S _ F O R _ R E S P O N S E
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
b:
T H A N K I N G _ F O R _ K I N D _ R E S P O N S E
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
b2j = {
T [0]
H [1]
A [2]
N [3, 6, 15, 23]
K [4, 13]
I [5, 14]
G [7]
F [9]
O [10, 22]
R [11, 18]
_ [8, 12, 17]
D [16]
E [19, 25]
S [20, 24]
P [21]
}
Queue: (0,19,0,26)
besti: 0
bestj: 0
bestsize: 0
j2len: { }