Reconnaissance automatique des dates
Exemple simplifié de rapport
Jacques Ladouceur

1 Objectif

Reconnaître automatiquement les dates exprimées sous forme de chiffres à l'intérieur d'un texte.

2 Description du texte

Extrait d'un article de Wikipédia qui porte sur la théorie de la relativité restreinte :

https://fr.wikipedia.org/wiki/Relativit%C3%A9_restreinte

Le texte est constitué de moins d'une page et compte 161 mots. Voici un extrait *(ici, comme il ne s'agit que d'un exemple avec un petit texte, je mets le texte au complet, mais normalement je mettrais de courts extraits avec des exemples de dates)* :

> Des formules de transformation pour passer d'un observateur à un autre furent établies par Hendrik Lorentz avant 1904 ; il s'agissait d'équations de compatibilité dont la signification n'était pas claire aux yeux de leur auteur. D'autres physiciens, tels que Woldemar Voigt (1887), avaient eu une démarche similaire plus tôt encore. Henri Poincaré a publié des articles pour en trouver une interprétation, peu de temps avant Albert Einstein. La répartition des rôles de tel ou tel savant dans l'émergence de la théorie de la relativité restreinte a fait l'objet d'une controverse, en particulier dans les années 2000. Comme la masse d'un muon est d'environ 100 MeV, l'énergie de la particule est 100/6 fois plus grande, soit d'environ 2000 MeV ou 2 GeV.
>
> Électromagnétisme et relativité restreinte.
>
> Dans l'espace newtonien à trois dimensions, une particule de charge "q" placée dans un champ électrique formula_205 et un champ magnétique formula_206 est soumise à la force de Lorentz et l'équation qui régit son mouvement est

3 Méthodologie

Nous allons examiner attentivement toutes les dates à l'intérieur du texte afin de définir une première méthode de reconnaissance automatique. Nous allons ensuite préparer un algorithme et un programme. Le résultat de ce programme sera évalué en termes de précision et taux de rappel.

Suite à l'évaluation, nous apporterons si nécessaire des modifications à notre méthode et nous procéderons à nouveau à une évaluation. Nous répéterons ces étapes jusqu'à ce que nous soyons satisfaits du résultat.

3.1 Observations

Il y a trois dates dans le texte *(normalement, il y aurait plus de cas alors je les mettrais dans un tableau pour améliorer la lecture des données)*:

1904 : établies par Hendrik Lorentz avant 1904 ;

1887 : que Woldemar Voigt (1887), avaient eu

2000 : en particulier dans les années 2000.

3.2 Analyse de départ

En retenant toutes les séquences de chiffres, on peut identifier la plupart des dates dans le texte *(ici je simplifie beaucoup parce que je ne veux pas faire l'analyse pour vous, ce n'est pas un exemple d'analyse, mais un exemple de rapport)*.

3.3 Premier algorithme

Mettre le texte dans la variable *texte*

Découper le texte en mots avec *split* du module *re* et mettre le résultat dans la variable *mots*

Pour chaque cellule du tableau *mots* :

Si la chaîne de caractères n'est constituée que de chiffres :

Afficher la chaîne à l'écran avec le mot qui précède

3.4 Premier programme

```
import re

texte = '''Des formules de transformation pour passer d'un observateur à un autre furent établies par Hendrik Lorentz avant 1904 ; il s'agissait d'équations de compatibilité dont la signification n'était pas claire aux yeux de leur auteur. D'autres physiciens, tels que Woldemar Voigt (1887), avaient eu une démarche similaire plus tôt encore. Henri
```

Poincaré a publié des articles pour en trouver une interprétation, peu de temps avant Albert Einstein. La répartition des rôles de tel ou tel savant dans l'émergence de la théorie de la relativité restreinte a fait l'objet d'une controverse, en particulier dans les années 2000. Comme la masse d'un muon est d'environ 100 MeV, l'énergie de la particule est 100/6 fois plus grande, soit d'environ 2000 MeV ou 2 GeV.

Électromagnétisme et relativité restreinte.

Dans l'espace newtonien à trois dimensions, une particule de charge "q" placée dans un champ électrique formule 205 et un champ magnétique formule 206 est soumise à la force de Lorentz et l'équation qui régit son mouvement est'''

```
mots = re.split('\W+',texte)

for no,m in enumerate(mots):

   if m.isnumeric():

      print(mots[no-1],mots[no])
```

3.5 Résultat obtenu

1. avant 1904
2. Voigt 1887
3. années 2000
4. environ 100
5. est 100
6. 100 6
7. environ 2000
8. ou 2
9. formule 205
10. formule 206

3.6 Évaluation

Taux de rappel (cas retenus parmi ceux à retenir): 3/3 (100%)

Précision (nombre de bons cas parmi tous ceux retenus): 3/10 (30%)

Le résultat n'est pas bon, la précision étant trop faible. Il faut analyser le résultat et le texte afin de trouver des éléments permettant d'améliorer la précision sans *trop* diminuer le taux de rappel.

4 Analyse des résultats et améliorations

On observe que les dates ont toujours quatre chiffres. Si on tient compte de la longueur, nous retenons les cas 1, 2, 3 et 7. Cependant, le cas 7 n'est pas bon.

Pour éliminer le cas 7, nous pouvons tenir compte de la présence de certains mots sémantiquement compatible avec des dates : avant et années. On pourrait alors dire qu'une séquence de quatre chiffres constitue une date si elle est précédée d'un de ces mots. Cependant, nous perdrons le cas 2.

Si on revient au texte afin d'examiner plus en détail le cas 2, on voit qu'il est placé à l'intérieur de parenthèses. On pourrait alors ajouter qu'une séquence de quatre chiffres à l'intérieur de parenthèses a de bonnes chances d'être une date. Mais il faudra découper le texte en conservant les parenthèses.

4.1 Algorithme (modification de l'itération)

Découper le texte en mots avec le module *re* en conservant les parenthèses et mettre le résultat dans la variable *mots*

Pour chaque cellule du tableau *mots* :

Si la chaîne de caractères a une longueur de 4, n'est constituée que de chiffres et est précédée de *avant* ou *années*:

Afficher la chaîne à l'écran avec le mot qui précède

Si la chaîne de caractères a une longueur de 6, commence par *'('*, se termine par *')'* et que la sous chaîne de 1 à -1 (ex. m[1 :-1]) n'est constituée que de chiffres :

Afficher la chaîne à l'écran avec le mot qui précède

4.2 Programme (uniquement la partie modifiée)

```
mots = re.split('[^\w+()]',texte)

for no,m in enumerate(mots):

  if (len(m) == 4) and m.isnumeric() and (mots[no-1] in ['avant','années']):

    print(mots[no-1],mots[no])

  if (len(m) == 6) and (m[0] == '(') and (m[-1] == ')') and (m[1:-1].isnumeric()):

    print(mots[no-1],mots[no])
```

4.3 Évaluation finale

Cas retenus :

avant 1904

Voigt (1887)

années 2000

Taux de rappel (cas retenus parmi ceux à retenir): 3/3 (100%)

Précision (nombre de bons cas parmi tous ceux retenus): 3/3 (100%)