Représentation de textes, reconnaissance d'entités nommées et fouille de motifs

Exercice 1 : représentation de textes

Soit la collection composée des 5 documents suivants :

Doc. 1 *Entrée en lice des deux champions d'Europe en titre*
Doc. 2 *Championnats d'Europe : le défi de Brice*
Doc. 3 *Championnat d'Europe de badminton en Vendée : les champions 2016 sont...*
Doc. 4 *Viktor prive Jan d'un doublé aux championnats d'Europe*
Doc. 5 *Entrée réussie pour l'équipe de France aux championnats d'Europe*

(a) Pour chaque document, donnez le document obtenu après les pré-traitements suivants :
   1. Suppression des majuscules et de la ponctuation ;
   2. Lemmatisation ;
   3. Suppression des mots outils de la liste suivante : {*en, de, le, être, à, un, pour*}.

(b) Dessinez l'index inversé de cette collection pré-traitée, en indiquant le `df` de chaque terme ainsi que son `tf` dans chaque document.

(c) Donnez la représentation vectorielle de chaque document pré-traité D_i, selon le modèle `tf.idf` [1].

Exercice 2 : reconnaissance d'entités nommées

(a) Soit le texte suivant :

```
Alors que le nombre d'abonnés augmente en France à un " très très bon rythme
" selon l'opérateur, Netflix s'apprête à augmenter ses tarifs en 2018. Le
ticket d'entrée au site va rester à 7,99 euros pour un seul écran mais aug-
mentera de 1 euro, à 10,99 euros, pour les abonnements à deux écrans per-
sonnalisés, et de 2 euros, à 13,99 euros, pour 4 écrans. Les tarifs ont déjà
augmenté en Amérique Latine et en Scandinavie cet été. Sont concernés cette
fois les prix aux Etats-Unis et dans plusieurs pays européens, dont également
l'Allemagne et le Royaume-Uni. Une hausse de prix prévue selon Yann Lafargue,
porte-parole du groupe pour l'Europe qui précise qu'il " ne s'agit pas d'une
réaction à la taxe Netflix de 2% annoncée récemment ".
```

   i. Encadrez les entités nommées de type `ENAMEX` du texte et indiquez leur catégorie.

   ii. Existe-t-il des ambiguïtés ? Quelle est leur nature ?

1. Nous supposerons les approximations suivantes : $\log(2) = 0,3$; $\log(3) = 0,5$; $\log(4) = 0,6$; $\log(5) = 0,7$.

(b) Soit l'annotation suivante réalisée par un système automatique :

```
Alors que le nombre d'abonnés augmente en <LOCATION>France</LOCATION> à un "
très très bon rythme " selon l'opérateur, Netflix s'apprête à augmenter ses
tarifs en <DATE>2018</DATE>. Le ticket d'entrée au site va rester à 7,99
euros pour un seul écran mais augmentera de 1 euro, à 10,99 euros, pour
les abonnements à deux écrans personnalisés, et de 2 euros, à 13,99 euros,
pour 4 écrans. Les tarifs ont déjà augmenté en <LOCATION>Amérique</LOCATION>
Latine et en Scandinavie cet été. Sont concernés cette fois les prix aux
<LOCATION>Etats-Unis</LOCATION> et dans plusieurs pays européens, dont égale-
ment l'<LOCATION>Allemagne</LOCATION> et le <LOCATION>Royaume-Uni</LOCATION>.
Une hausse de prix prévue selon <PERSON>Yann Lafargue<PERSON>, porte-parole
du groupe pour l'<LOCATION>Europe</LOCATION> qui précise qu'il " ne s'agit
pas d'une réaction à la taxe Netflix de 2% annoncée récemment ".
```

i. Quelles sont les erreurs de reconnaissance des entités nommées par rapport à votre référence de la question précédente ? (parmi les 5 types d'erreurs utilisées dans le calcul du SER)

ii. Quel est alors le SER obtenu, pour ce système automatique ? On considère que le poids est de 0,5 pour les erreurs de type ainsi que pour les erreurs de frontière et qu'il est de 1 pour les erreurs de suppression, d'insertion et de type+frontière) ?

Exercice 3 : représentation de textes

Soit le texte suivant, dans une collection de 500 documents :

```
Nick, qui est membre d'un obscur groupe de rock indépendant, vient de vivre
une rupture difficile. De son côté, Norah a du mal à donner un sens à sa vie
et à sa relation épisodique avec un musicien trop égoïste. Les deux jeunes
gens n'ont rien en commun, sauf leurs goûts en matiÃ¨re de musique. Leur
rencontre fortuite va les entraîner toute une nuit à New York vers le lieu
mystérieux où doit se produire leur groupe préféré. Au cours de cette nuit
de surprises et d'aventures, ils vont découvrir qu'ils ont peut-être plus en
commun que leur seul amour de la musique...
```

(a) Si nous souhaitons détecter les entités nommées de ce texte, quels sont les pré-traitements à éviter ?

(b) Donnez le texte obtenu après les pré-traitements choisis à la question précédente.

(c) Calculez le $tf.idf$ des mots *musique* et *rencontre*, sachant que le premier mot apparaît 10 fois dans la collection et que le second mot y apparaît 50 fois. Selon vous, lequel de ces mots décrirait le mieux le texte ?

Exercice 4 : fouille de motifs séquentiels

Les 3 séquences suivantes, correspondent à des phrases extraites d'articles biomédicaux :

1. *we conclude that GENE is essential for maturation of ubiquitin containing autophagosomes and that defect in this function may contribute to DISEASE pathogenesis.*
2. *somatic mutation in isocitrate dehydrogenase 1 GENE and GENE occur in glioma and acute myeloid leukaemia DISEASE.*
3. *DISEASE is normally caused by an autosomal dominant mutation in the type i collagen gene GENE and GENE.*

À partir des séquences précédentes, quel(s) motif(s) d'itemsets vous paraîtraient intéressant(s) pour identifier les relations entre une maladie et un gène ? Quelles contraintes utiliseriez-vous ? Nous considérons que chaque mot est décrit par sa forme de base, son lemme et sa catégorie morpho-syntaxique et que les noms de maladies et de gènes ont été remplacés par les étiquettes *DISEASE* et *GENE*, respectivement.

Exercice 5 : représentation de textes

Soit la collection composée des 4 documents suivants :

Doc. 1 *Fête de la musique dans le Sézannais : on prend note !*
Doc. 2 *Les auditeurs prennent l'antenne pour la fête de la musique.*
Doc. 3 *Fête de la musique. Premières notes samedi.*
Doc. 4 *Les Allumettes vont mettre le feu à la fête de la musique.*

(a) Pour chaque document, donnez le document obtenu après les pré-traitements suivants :
   1. Suppression des majuscules et de la ponctuation ;
   2. Lemmatisation ;
   3. Suppression des mots outils de la liste suivante, aprs lemmatisation : {*de, le, dans, on, pour*}.

(b) Donnez la représentation vectorielle du premier document pré-traité D_i, selon le modèle `tf.idf` [2], l'idf étant calculé par

$$idf(t) = \log_{10} \frac{N}{df(t)},$$

avec t le terme considéré, N le nombre de documents de la collection et $df(t)$ la fréquence de document du terme t, dans la collection considérée.

2. Nous supposerons les approximations suivantes : $\log(2) = 0,3$; $\log(3) = 0,5$; $\log(4) = 0,6$; $\log(5) = 0,7$.

Exercice 6 : réflexion

Nous souhaitons travailler sur des résumés d'articles scientifiques provenant du domaine bioinformatique (voir figures ci-aprés). Le but est de regrouper les articles mentionnant les mêmes organismes biologiques. Les noms des organismes peuvent apparaître sous différentes dénominations. Par exemple, le terme *Dinoflagellates* peut également apparaître sous le terme *dinoflagellates* mais aussi *Dinophyceae*. De plus, un organisme peut également apparaître sous le nom plus général de son groupe.

(a) Nous souhaitons tout d'abord extraire automatiquement les noms des organismes présents dans les résumés.

i. Quels pré-traitements sont à réaliser sur les résumés ? Justifiez votre réponse.

ii. Quelle approche peut-on utiliser pour détecter les noms des organismes, à partir des résumés pré-traités ?

(b) Nous souhaitons ensuite regrouper les articles faisant référence aux mêmes organismes. Quelle approche vous paraît la plus adaptée pour résoudre ce problème ? Dessinez les différentes étape de la chaîne de traitement qui prend un corpus de résumés d'articles bioinformatique, en entrée, et qui produit le regroupement des résumés par organisme, en sortie. Vous préciserez également les paramètres à prendre en compte ou le principe de la phase d'apprentissage du système, s'il y en a une.

(c) Comment peut-on évaluer les performances du système obtenu, pour pouvoir le comparer à d'autres systèmes (existants ou non) ?

Toward understanding the genetic diversity and distribution of copepod-associated symbiotic ciliates and the evolutionary relationships with their hosts in the marine environment, we developed a small subunit ribosomal RNA gene (18S rDNA)-based molecular method and investigated the genetic diversity and genotype distribution of the symbiotic ciliates on copepods. Of the 10 copepod species representing six families collected from six locations of Pacific and Atlantic Oceans, 9 were found to harbor ciliate symbionts. Phylogenetic analysis of the 391 ciliate 18S rDNA sequences obtained revealed seven groups (ribogroups), six (containing 99% of all the sequences) belonging *Vorticella gracilis*. Among the Apostomatida groups, Group III were essentially identical to *Vampyrophrya pelagica*, and the other five groups represented the undocumented ciliates that were close to *Vampyrophrya/Gymnodinioides/Hyalophysa*. Group VI ciliates were found in all copepod species but one (*Calanus sinicus*), and were most abundant among all ciliate sequences obtained, indicating that they are the dominant symbiotic ciliates universally associated with copepods. In contrast, some ciliate sequences were found only in some of the copepods examined, suggesting the host selectivity and geographic differentiation of ciliates, which requires further verification by more extensive sampling. Our results reveal the wide occurrence and high genetic diversity of symbiotic ciliates on marine copepods and highlight the need to systematically investigate the host- and geography-based genetic differentiation and ecological roles of these ciliates globally.

FIGURE 1 – Résumé de l'article *« Prevalent Ciliate Symbiosis on Copepods : High Genetic Diversity and Wide Distribution Detected Using Small Subunit Ribosomal RNA Gene »*

Blastodinium is a genus of dinoflagellates that live as parasites in the gut of marine, planktonic copepods in the World's oceans and coastal waters. The taxonomy, phylogeny, and physiology of the genus have only been explored to a limited degree and, based on recent investigations, we hypothesize that the morphological and genetic diversity within this genus may be considerably larger than presently recognized. To address these issues, we obtained 18S rDNA and ITS gene sequences for *Blastodinium* specimens of different geographical origins, including representatives of the type species. This genetic information was in some cases complemented with new morphological, ultrastructural, physiological, and ecological data. Because most current knowledge about *Blastodinium* and its effects on copepod hosts stem from publications more than half a century old, we here summarize and discuss the existing knowledge in relation to the new data generated. Most *Blastodinium* species possess functional chloroplasts, but the parasitic stage, the trophocyte, has etioplasts and probably a limited photosynthetic activity. Sporocytes and swarmer cells have well-developed plastids and plausibly acquire part of their organic carbon needs through photosynthesis. A few species are nearly colorless with no functional chloroplasts. The photosynthetic species are almost exclusively found in warm, oligotrophic waters, indicating a life strategy that may benefit from copepods as microhabitats for acquiring nutrients in a nutrient-limited environment. As reported in the literature, monophyly of the genus is moderately supported, but the three main groups proposed by Chatton in 1920 are consistent with molecular data. However, we demonstrate an important genetic diversity within the genus and provide evidences for new groups and the presence of cryptic species. Finally, we discuss the current knowledge on the occurrence of *Blastodinium* spp. and their potential impact on natural copepod populations.

FIGURE 2 – Résumé de l'article *« The Parasitic Dinoflagellates Blastodinium spp. Inhabiting the Gut of Marine, Planktonic Copepods : Morphology, Ecology, and Unrecognized Species Diversity »*