TAL - Laboratoire 4

Auteurs: Johanna Melly & Yohann Meyer

Problème 4.1

Le code de cet exercice est disponible dans le fichier `NE_tags.ipynb`.

L'output de la première cellule montre la fréquence d'apparition des entités nommées sans tenir compte de leur label, avec le NE tagger d'**NLTK**. Nous avons utilisé le texte "Moby Dick" que nous avons tronqué afin d'en retirer l'en-tête. Nous avons choisi d'utiliser uniquement les 500000 premiers caractères de ce texte afin d'avoir un résultat relativement rapide mais suffisamment représentatif.

Les 10 entités nommées les plus fréquentes sont les suivantes:

```
 (('Queequeg', 'NNP'), 143),
 (('Ahab', 'NNP'), 119),
 (('Captain', 'NNP'), 110),
 (('Stubb', 'NNP'), 59),
 (('Bildad', 'NNP'), 59),
 (('Pequod', 'NNP'), 56),
 (('Jonah', 'NNP'), 53),
 (('Starbuck', 'NNP'), 46),
 (('Flask', 'NNP'), 44),
 (('Peleg', 'NNP'), 39),
 (('Nantucket', 'NNP'), 39)
```

L'output de la deuxième cellule montre en premier temps les mêmes données que dans la première cellule, mais avec les labels en plus:

```
((('Ahab', 'NNP'), 'PERSON'), 92)
((('Queequeg', 'NNP'), 'PERSON'), 87)
((('Captain', 'NNP'), 'PERSON'), 75)
((('Stubb', 'NNP'), 'PERSON'), 51)
((('Queequeg', 'NNP'), 'GPE'), 47)
((('Jonah', 'NNP'), 'PERSON'), 41)
((('Bildad', 'NNP'), 'GPE'), 36)
((('Starbuck', 'NNP'), 'PERSON'), 36)
((('Peleg', 'NNP'), 'PERSON'), 35)
((('White', 'NNP'), 'FACILITY'), 33)
```

Nous avons séparé ces données car elles sont différentes. En effet, le NE tagger d'NLTK pose parfois un label différent sur une même entité. C'est la raison pour laquelle on trouve seulement 87 fois "Queequeg" au lieu des 143 fois précédentes. On le retrouve en effet 47 autres fois labélisé comme GPE et, hors des 10 plus fréquents, 9 fois comme ORGANIZATION.

Dans la deuxième cellule, on trouve aussi ces données mais comptées et labélisées à l'aide du NE tagger de **Stanford**:

```
(('Queequeg', 'LOCATION'), 165),
(('Ahab', 'PERSON'), 163),
(('Peleg', 'PERSON'), 67),
(('Bildad', 'PERSON'), 58),
(('Jonah', 'PERSON'), 58),
(('Nantucket', 'LOCATION'), 50),
(('Starbuck', 'PERSON'), 36),
(('Stubb', 'ORGANIZATION'), 35),
(('Dick', 'PERSON'), 33),
(('Stubb', 'PERSON'), 31)
```

Le tagger de Stanford semble plus précis; du moins il ne tag pas une même entité avec plusieurs labels différents.

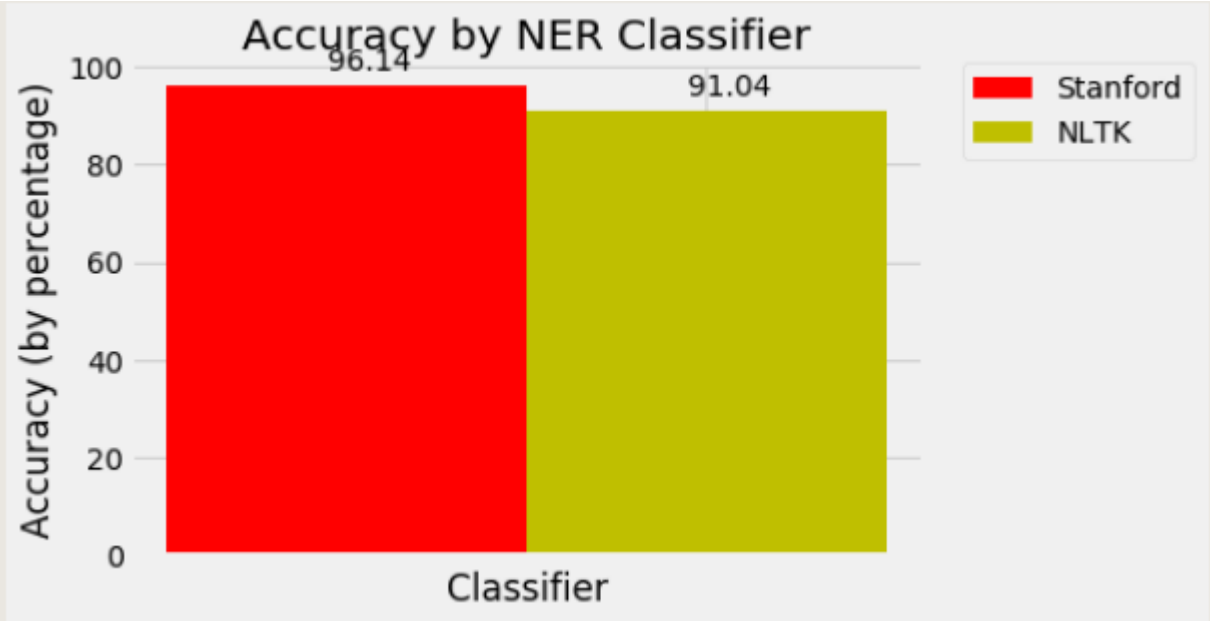
En observant les 50 entités nommées les plus fréquentes des deux taggers, on se rends compte de ceci:

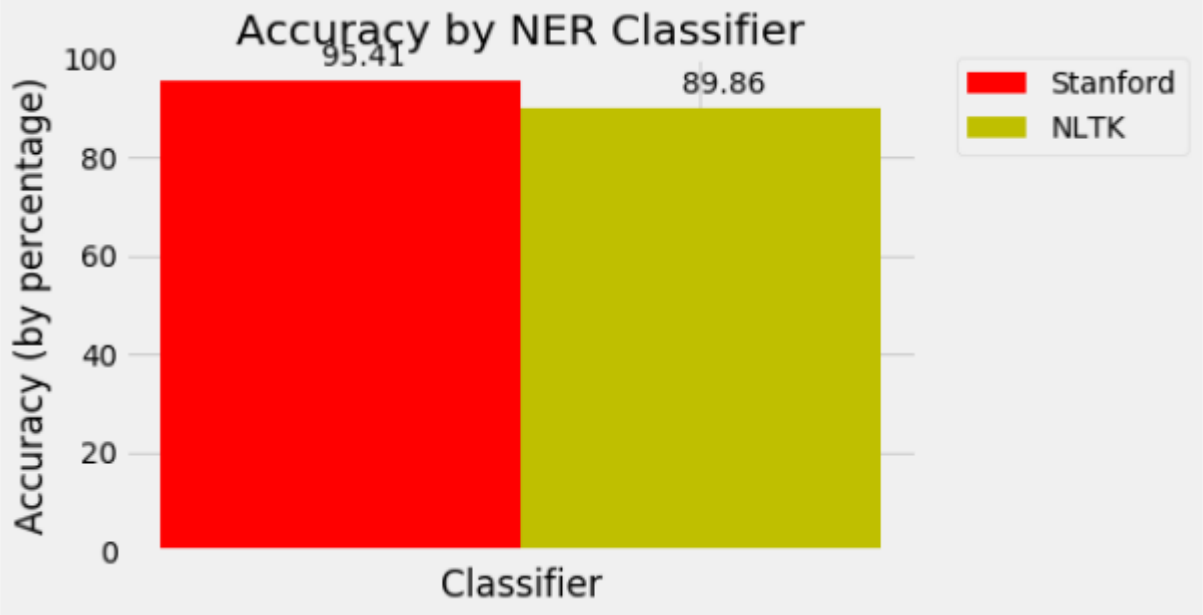
1. En comparant les NE du tagger de Stanford avec les NE trouvées avec NLTK et *non labélisées*: les mêmes entités ont un nombre d'apparition plus grand avec le NE tagger de Stanford, ce qui signifie qu'il les détecte mieux (plus souvent).
2. En les comparant avec les NE trouvées avec NLTK labélisées *et* non labélisées: le NE tagger de Stanford va évincer certaines entités avec une majuscule et qui ne sont pas des entités nommées, lorsque le NE tagger de NLTK va les prendre pour d'authentiques entités nommées (par exemple: "Mr." ou "BOOK").
3. En les comparant avec les NE trouvées avec NLTK et labélisées: le NE tagger de Stanford est plus correct sur les labels que celui d'NLTK (par exemple: NLTK donne le label "PERSON" à "Cape" et Stanford lui donne le label "LOCATION"; il s'agit effectivement d'une location puisqu'on parle bien de Cape Horn dans le roman).

Problème 4.2

Pour cette deuxième partie, nous avons suivi le tutoriel proposé et le code qui s'y trouvait. Le code pour cet exercice se trouve dans les fichiers `second_exercise.ipynb` et `second_exerciseB.ipynp`. Le code est identique dans les deux fichiers, mais les textes utilisés ne sont pas les mêmes, l'output et les graphiques sont donc différents.

Les graphiques montrant la précision des NE tagger de NLKT et de Standford sur les textes de test A et B sont les suivants (dans l'ordre):





Dans les deux cas, on voit une différence assez nette entre les deux NE tagger, qui confirme les observations faites à l'étape 4.1: le NE tagger de Stanford est plus performant.