Rapport TP4 TAL

SMAIL Aghilas

Q1 : Quels sont les 20 mots les plus ambigus (par rapport à l'étiquetage en entités nommées) du corpus `corpus_en_200k.train.txt`:

```
s21219997@V-PP-47-098:~/Bureau/S2/TAL/TP4$ cat
corpus_en_200k.train.txt | cut -f2,4 | sort -k1,1 | python3
ambig.py | sort -n -r | head -20
5	l'	org product geoloc person O
5	Le	geoloc org person product O
5	la	geoloc org product person O
5	*	geoloc org person product O
5	du	geoloc org person product O
5	d'	org geoloc person product O
5	de	geoloc org person product O
5	Al	geoloc org person product O
4	vingt	geoloc org product O
4	un	geoloc org product O
4	Saint	org person product O
4	Robert	person geoloc org O
4	Paris	geoloc org product O
4	Mohamed	person geoloc product O
4	Madrid	geoloc org product O
4	les	org product geoloc O
4	le	geoloc org product O
4	La	geoloc org product O
4	Hassan	person geoloc product O
4	,	geoloc person product O
```

Q2 : Quels sont maintenant les 20 entités nommées les plus ambigus du corpus `corpus_en_200k.train.txt`:

```
s21219997@V-PP-47-098:~/Bureau/S2/TAL/TP4$ cat corpus_en_200k.train.txt |
python3 extrait_entite_nomme.py | cut -f1,3 | sort -k1,1 | python3 ambig.py |
sort -n -r | head -20
2	Yougoslavie	geoloc org
2	Washington	geoloc org
2	Wall Street	geoloc org
```

```
2       URSS geoloc org
2       Unesco      geoloc org
2       Turquie     geoloc org
2       Tunisie     geoloc org
2       Tripoli     geoloc org
2       Togo geoloc org
2       Thorgal     person product
2       Téhéran     geoloc org
2       Taïwan      geoloc org
2       Syriegeoloc org
2       Suisse      geoloc org
2       Sénégal     geoloc org
2       Rwanda      geoloc org
2       Russie      geoloc org
2       Royaume_Unigeoloc org
2       Rome geoloc org
2       Roland_Garros     geoloc org
```

Q3 : Calculez l'ambiguité moyenne des entités nommées sur le même corpus.

```
s21219997@V-PP-47-098:~/Bureau/S2/TAL/TP4$ cat
corpus_en_200k.train.txt | python3 extrait_entite_nomme.py | cut
-f1,3 | sort -k1,1 | python3 ambig.py -m
Ambiguite moyenne :  1.0392982456140352
```

Q4 : Vous allez maintenant calculer l'ambiguité moyenne des entités nommées par rapport au nombre de mot les composant (ambiguité moyenne des entités composées de 1 mot, puis de 2 mots, etc.). Afficher la courbe reliant l'ambiguité moyenne et le nombre de mots :

```
s21219997@V-PP-47-098:~/Bureau/S2/TAL/TP4$ cat
corpus_en_200k.train.txt | python3 extrait_entite_nomme.py | cut
-f1,3 | sort -k1,1 | python3 ambiguite.py -mw
1       1.0846343467543138
6       1.0
7       1.0
3       1.010204081632653
5       1.0
2       1.0046253469010176
4       1.0080645161290323
8       1.0
10      1.0
9       1.0
16      1.0
11      1.0
23      1.0
```

```
34    1.0
13    1.0
```

Q5 :
En remarque l'ambiguïté est presque la même pour tout les entité sauf quelque une on dépassé le 1.

Q6 : En utilisant les résultats de l'exercice 5 du TP vous allez maintenant afficher les 50 patrons d'étiquettes morphosyntaxiques les plus fréquents de votre corpus qui peuvent représenter des entités nommées :

```
m21219997@V-PP-47-098:~/Bureau/S2/TAL/TP4$ cat
corpus_en_200k.train.txt | python3 extrait_entite_nomme.py | sort
| cut -f2,3 | sort -s -t$'\t' -k1,1 | python3 entite_nomme_freq.py
| sort -t$'\t' -k2,2 -r -n | head -50
np    5255
np np      1986
np np np    159
nc    140
np prep np      114
np adj      104
nc prep np      77
nc np      50
nc adj      48
det np      47
nc prep nc      40
ponctw np   24
np nc      20
ponctw      16
np prep nc      16
nc adj adj      16
det np prep det np    16
adj nc      13
np adj prep nc   12
np ponctw np     11
np adj prep det nc    11
np np prep np    10
np det np   10
nc prep nc prep np    10
nc prep det np   10
nc prep     10
det nc prep nc   10
adj   10
np prep det np   9
np np np np      9
nc adj prep np   9
```

```
np prep nc prep det np      8
np ponctw nc ponctw    8
np adj prep nc prep det nc       8
nc ponctw   7
nc np np    7
nc adj prep nc   7
adj adj prep nc       7
ponctw np np     6
np adj prep np   6
nc np np prep np       6
np v  5
np adj adj       5
nc prep nc prep det nc      5
nc prep nc adj   5
nc prep det nc   5
adj adj     5
prep np     4
np prep det nc   4
np np np np np   4
```

Q7 :