1 Évaluation des approches (`semeval_executable.py`)

1.1 Méthodes EXPORT

`getpredfilename` permet de s'assurer que les noms des fichiers de prédiction sont tous construits de la même manière.

Un script shell est utilisé pour extraire le score MAP d'un fichier de prédiction :

```
prediction=$1

if echo $prediction | grep --quiet "2016"
then
    reference=scorer/SemEval2016-Task3-CQA-QL-test.xml.subtaskB.relevancy
else
    if echo $prediction | grep --quiet "2017"
    then
        reference=scorer/SemEval2017-Task3-CQA-QL-test.xml.subtaskB.relevancy
    else
        reference=scorer/SemEval2016-debug.relevancy
    fi
fi

python2 scorer/ev.py $reference $prediction | grep "^MAP" | sed 's/ \+/;/g' | cut -f 4 -d
 ↪  ';'
```

1.1.1 Méthodes bruteforce

Les méthodes bruteforce correspondent à tester toutes les combinaisons d'arbres de documents, de sacs de mots et de filtres. Les méthodes bruteforce sont crées en faisant le produit cartésien des dimensions envisagées.

Les méthodes précédemment générées sont exécutées et les scores produits sont écrits dans les fichiers correspondants.

année 2016

Sac de mots	Filtres	Score MAP
Lemmes	Mots outils, ≤ 2	0.7679
Textes des entités nommées	Mots outils	0.7493
Textes des entités nommées	Mots outils, ≤ 1	0.7493
Textes des entités nommées	Mots outils, ≤ 2	0.7493
Textes des entités nommées	Mots outils, ≤ 3	0.7493
Textes des entités nommées	Mots outils, ≤ 4	0.7493
Textes des entités nommées	≤ 1	0.7493
Textes des entités nommées	≤ 2	0.7493
Textes des entités nommées	≤ 3	0.7493
Textes des entités nommées	≤ 4	0.7493
Textes des entités nommées	Pas de filtre	0.7493
Lemmes	Mots outils, ≤ 1	0.7490
Tokens	Mots outils, ≤ 2	0.7488
Tokens	Mots outils, ≤ 1	0.7446
Lemmes	Mots outils	0.7424
Tokens	Mots outils, ≤ 3	0.7423
Tokens	Mots outils, ≤ 4	0.7422
Étiquettes des entités nommées	Mots outils, ≤ 4	0.7369
Étiquettes des entités nommées	≤ 4	0.7369
Tokens	Mots outils	0.7322
Lemmes	Mots outils, ≤ 3	0.7317
Lemmes	≤ 2	0.7293
Lemmes	≤ 4	0.7292
Tokens	≤ 4	0.7275
Tokens	≤ 2	0.7274
Lemmes	≤ 1	0.7269
Lemmes	≤ 3	0.7253
Tokens	≤ 3	0.7219
Lemmes	Mots outils, ≤ 4	0.7202
Étiquettes des entités nommées	Mots outils	0.7153
Étiquettes des entités nommées	Mots outils, ≤ 1	0.7153
Étiquettes des entités nommées	Mots outils, ≤ 2	0.7153
Étiquettes des entités nommées	≤ 1	0.7153
Étiquettes des entités nommées	≤ 2	0.7153
Étiquettes des entités nommées	Pas de filtre	0.7153
Lemmes	Pas de filtre	0.7148
Tokens	≤ 1	0.7135
Tokens	Pas de filtre	0.7098
Étiquettes des entités nommées	Mots outils, ≤ 3	0.7081
Étiquettes des entités nommées	≤ 3	0.7081

année 2017

Sac de mots	Filtres	Score MAP
Tokens	Mots outils, ≤ 1	0.4705
Lemmes	Mots outils, ≤ 1	0.4678
Tokens	Mots outils, ≤ 2	0.4653
Tokens	Mots outils, ≤ 3	0.4650
Lemmes	Mots outils, ≤ 2	0.4624
Tokens	≤ 2	0.4623
Tokens	Mots outils	0.4598
Lemmes	Mots outils	0.4581
Lemmes	Pas de filtre	0.4580
Tokens	Mots outils, ≤ 4	0.4521
Lemmes	≤ 1	0.4473
Lemmes	Mots outils, ≤ 3	0.4458
Tokens	≤ 1	0.4455
Lemmes	≤ 2	0.4425
Lemmes	Mots outils, ≤ 4	0.4419
Tokens	Pas de filtre	0.4418
Tokens	≤ 3	0.4392
Lemmes	≤ 3	0.4389
Lemmes	≤ 4	0.4249
Tokens	≤ 4	0.4153
Textes des entités nommées	Mots outils, ≤ 3	0.4139
Textes des entités nommées	Mots outils, ≤ 4	0.4139
Textes des entités nommées	≤ 3	0.4139
Textes des entités nommées	≤ 4	0.4139
Étiquettes des entités nommées	Mots outils	0.4123
Étiquettes des entités nommées	Mots outils, ≤ 1	0.4123
Étiquettes des entités nommées	Mots outils, ≤ 2	0.4123
Étiquettes des entités nommées	≤ 1	0.4123
Étiquettes des entités nommées	≤ 2	0.4123
Étiquettes des entités nommées	Pas de filtre	0.4123
Étiquettes des entités nommées	Mots outils, ≤ 3	0.4104
Étiquettes des entités nommées	≤ 3	0.4104
Textes des entités nommées	Mots outils	0.4083
Textes des entités nommées	Mots outils, ≤ 1	0.4083
Textes des entités nommées	Mots outils, ≤ 2	0.4083
Textes des entités nommées	≤ 1	0.4083
Textes des entités nommées	≤ 2	0.4083
Textes des entités nommées	Pas de filtre	0.4083
Étiquettes des entités nommées	Mots outils, ≤ 4	0.4063
Étiquettes des entités nommées	≤ 4	0.4063

Année	Score MAP baseline
2016	0.7475
2017	0.4185

1.1.2 Méthodes pondérées

Le but des méthodes pondérées est d'utiliser plusieurs indicateurs au sein d'une même méthode. Un exemple d'approche de pondération est d'utiliser les lemmes pour estimer la similarité de phrases, et de donner une plus grande importance aux lemmes communs qui sont également des entités nommées.

1. Recherche des pondérations optimales
2. Pondération par entités nommées

année 2016

Sac de mots	Filtres	Score MAP
Lemmes	Mots outils, ≤ 2	0.7743
Lemmes	Mots outils, ≤ 1	0.7716
Lemmes	Mots outils, ≤ 3	0.7551
Lemmes	Mots outils	0.7504
Tokens	Mots outils, ≤ 2	0.7489
Tokens	Mots outils, ≤ 1	0.7460
Tokens	Mots outils, ≤ 3	0.7460
Tokens	Mots outils	0.7450
Lemmes	≤ 2	0.7444
Tokens	Mots outils, ≤ 4	0.7431
Lemmes	≤ 1	0.7356
Lemmes	≤ 3	0.7354
Tokens	≤ 2	0.7350
Lemmes	Pas de filtre	0.7300
Tokens	≤ 4	0.7298
Tokens	≤ 1	0.7282
Tokens	≤ 3	0.7281
Lemmes	≤ 4	0.7274
Lemmes	Mots outils, ≤ 4	0.7250
Tokens	Pas de filtre	0.7203

année 2017

Sac de mots	Filtres	Score MAP
Lemmes	Pas de filtre	0.4741
Tokens	Mots outils, ≤ 3	0.4700
Tokens	Mots outils, ≤ 2	0.4686
Tokens	Mots outils, ≤ 1	0.4680
Tokens	Mots outils	0.4675
Lemmes	Mots outils, ≤ 1	0.4652
Tokens	≤ 2	0.4610
Lemmes	Mots outils, ≤ 2	0.4606
Lemmes	≤ 1	0.4594
Lemmes	Mots outils	0.4571
Tokens	Mots outils, ≤ 4	0.4516
Lemmes	Mots outils, ≤ 3	0.4460
Lemmes	≤ 2	0.4460
Lemmes	Mots outils, ≤ 4	0.4452
Tokens	≤ 1	0.4448
Tokens	Pas de filtre	0.4436
Lemmes	≤ 3	0.4386
Tokens	≤ 3	0.4376
Lemmes	≤ 4	0.4227
Tokens	≤ 4	0.4149