Manipulations

1. Avec `nltk`, charger les jeux de données RTE (1, 2 et 3) avec l'interface `nltk.download()`. [1]

   Les manipulations suivantes seront faites avec le jeu de développement de RTE3 (`rte3_dev.xml`) et les tests avec le jeu de test (`rte3_test.xml`).

```
>>> from nltk.corpus import rte
>>> l = rte.pairs('rte3_dev.xml')
>>> l[8].text
"Mrs. Bush's approval ratings have remained very high, above 80%, (...)
>>> l[7].hyp
'80% approve of Mr. Bush.'
>>> l[7].value
0
```

2. Ecrire un premier script pour réaliser une baseline : on va mesurer pour chaque paire le recouvrement lexical en se basant uniquement sur les formes présentes (pas de lemmatisation). Sur le jeu de développement, calculer le seuil α tel que en répondant `YES` chaque fois que le recouvrement est $> \alpha$, et `NO` sinon, on maximise le score. Faire ensuite tourner cette baseline sur le *test set*, et noter la performance.
3. Version 2 de la baseline : réaliser la même mesure après une lemmatisation des paires.
4. Version 3 de la baseline : remplacer le recouvrement lexical strict par une mesure de similarité lexicale (par exemple similarité par chemin dans WordNet, ou celle de Resnik).
5. Faire une analyse d'erreur : proposer, sur la base d'une étude manuelle d'un nombre significatif d'erreurs dans le *test set* (≈ 40), des heuristiques qui, sans passer par une analyse syntaxique et une représentation sémantique, devraient permettre d'améliorer sensiblement la baseline.

On demande le code des trois baselines, et un fichier texte (ou pdf) pour la discussion et l'analyse d'erreurs.

1. Ces corpus sont aussi accessibles sur le site suivant de l'ACL, mais les versions nltk sont légèrement différentes, elles ont fait l'objet d'une normalisation.
`http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool`