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[7].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). Bien sûr, il faut normaliser la mesure de recouvrement par rapport à la longueur des textes. Sur le jeu de développement, calculer le seuil α tel que en répondant `YES` chaque fois que le taux de recouvrement est $> \alpha$, et `NO` sinon, on maximise le score. Faire ensuite tourner cette baseline sur le *test set*, et noter la performance.
3. Variantes de la baseline : suppression de *stop words*, lemmatisation des paires, voire prise en compte des entités nommées.
4. 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.
5. Changement de jeu de données : "Stanford NLI", accessible @ `https://nlp.stanford.edu/projects/snli/snli_1.0.zip` Attention, cette fois-ci on manipule non plus environ 800 paires, mais 570 000 paires.
   — Récupération des données au format `json` (voir ci-dessous),
   — Sélection des données dans le jeu "dev" qui ont comme gold_label `"neutral"` ou `"entailment"`,
   — Tokenisation (grossière) des phrases,
   — Mesure de la performance avec α,
   — Analyse d'erreur sur quelques exemples


```
{ "annotator_labels": [ "neutral", "entailment", "neutral", "neutral", "neutral"],
  "captionID": "4705552913.jpg#2",
  "gold_label": "neutral",
  "pairID": "4705552913.jpg#2r1n",
  "sentence1": "Two women are embracing while holding to go packages.",
  "sentence1_binary_parse": "( ( Two women ) ( ( are ( embracing ( while ( holding ( to ( go package
  "sentence1_parse": "(ROOT (S (NP (CD Two) (NNS women)) (VP (VBP are) (VP (VBG embracing) (SBAR (IN
  "sentence2": "The sisters are hugging goodbye while holding to go packages after just eating lunch
  "sentence2_binary_parse": "( ( The sisters ) ( ( are ( ( hugging goodbye ) ( while ( holding ( to
  "sentence2_parse": "(ROOT (S (NP (DT The) (NNS sisters)) (VP (VBP are) (VP (VBG hugging) (NP (UH g
}
```


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`