Objectif du TP implémenter quelques-unes des mesures de précision et de rappel que l'on peut définir quand il s'agit de comparer un ensemble de liens de coréférence calculés avec un ensemble de liens qui sert de référence.

Principe les *mentions* (ou *entités*) sont représentées par des entiers, et l'ensemble des liens de co-référence est représenté par une liste de couples (i, j) [1].
On appellera G (gold) la liste de référence, et R (réponse) la liste produite par le système qu'il s'agit d'évaluer.

Manipulations Le fichier fourni `canevas.py` contient un tuple de couples correspondant au gold G, et une séries de tuples de couples correspondant à différentes réponses possibles de systèmes à évaluer R_1, R_2, etc.

1. Implémenter la mesure « couples » qui définit une précision et un rappel en ne considérant que les couples (et pas les chaînes).
2. Ecrire une fonction qui crée, à partir d'une liste de couples, un ensemble de *chaînes* (classes d'équivalence). On ajoutera systématiquement sous forme de singletons les mentions qui ne sont pas dans une relation anaphorique [2].
3. Proposer une variante de la mesure « couples », dans laquelle les chaînes sont prises en compte, c'est-à-dire qu'un bon antécédent est un antécédent dans la bonne chaîne.

▷ *Pour la suite, on supposera que G et R sont des suites (ou des ensembles) de classes d'équivalence.*

4. Implémenter la mesure de MUC qui est basée sur le comptage du nombre de liens dans les chaînes.

   Les réponses aux questions précédentes peuvent être envoyées par email à la fin de la séance (`pascal.amsili@gmx.fr`) pour un éventuel **bonus**.

5. Implémenter la mesure B^3 qui est basée sur le comptage des entités dans les chaînes de référence.
6. Implémenter la mesure CEAF présentée dans [Luo, 2005], basée sur un appariement des classes de G et des classes de R.
7. Implémenter la mesure BLANC présentée dans [Recasens and Hovy, 2011].
8. Utiliser vos différentes métriques pour comparer les performances des réponses proposées, étant donnée la référence G.

   Fournir les scores de rappel, précision et F-score retournées par les trois métriques. Quelle métrique vous semble la plus adéquate, et pour quelles raisons ?

1. On supposera, sans perte de généralité, que $i < j$ (par convention, l'ordre croissant des mentions représente l'ordre linéaire du texte).
2. On suppose qu'on a un inventaire fixe de mentions, par exemple les entiers entre 1 et 12.