- *L'objectif de ce TP est d'effectuer une étude sur un corpus étiqueté en entités nommées afin de caractériser le format et la fréquence de chaque type d'entités. Pour cela vous disposerez d'un grand corpus annoté, déjà tokenisé, disponible sous un format tabulaire en colonne où chaque colonne correspond à une information ou trait (feature en anglais) attachée à chaque mot. Vous pouvez télécharger ce corpus sur le lien suivant: corpus entité nommés*
- *Entités nommées : définition*
    1. *La définition la plus générale des entités nommées est la suivante : Toute entité faisant référence à un identificateur unique (par rapport à un contexte) est une entité nommée.*
    2. *Exemple:*
        - *Quantités numériques : 1 litre, 12Hz, 1000 euros, 6 heures, etc.*
        - *Dates : 12 décembre 2006, 6 janvier, demain, etc.*
        - *Entités noms propres*
            - *Noms de personnes : Mireille Mathieu, Lucky Luke*
            - *Noms de lieu : Les Cotes d'Armor, Montfavet*
            - *Noms d'organisation : SNCF, Les Déménageurs Bretons*
            - *Noms d'oeuvres : Les mystères de Paris, la tour Eiffel*
    3. *Il n'y a pas de jeu d'étiquette absolu pour les entités nommées, chaque corpus, chaque application peut avoir un jeu différent avec des granularités plus ou moins fine. Par exemple la classe personne peut être générique ou bien détailler les personnes fictives (personnage de roman par exemple) des personnes réelles.*
    4. *Dans ce TP nous nous focaliserons sur les entités de type noms propres et utiliserons un corpus annoté selon 4 types d'entités :*
        - *Entités Géographique (geoloc) : La boutique ouvre à Paris en France*
        - *Entités Organisation (org) : La France s'est qualifiée pour la finale.*
        - *Entités Personne (person) : Il a été opposé à Rafael Nadal en finale.*
        - *Entités Produit (product) : L'album The dark side of the moon a été réédité*
- *Corpus d'entités nommées*
- *Le corpus que vous allez manipuler contient des transcriptions d'émissions radiophonique, en français, annoté selon les 4 types d'entités nommées présentées précédemment. Voici un exemple de phrases annotées au format XML:*

```
Investiture aujourd'hui à ‹en type="geoloc"›Bamako‹/en› , ‹en
type="geoloc"›Mali‹/en›, du président ‹en type="personne"›Amadou Toumani Touré‹/en›
```

    1.
    2. *Format en colonnes : Plutôt que de travailler sur des textes bruts avec des annotations XML, nous allons utiliser un format en colonne où chaque mot est représenté sur 1 ligne, et où chaque colonne représente une information sur le mot. De plus un traitement de découpage en phrase, de tokenisation, de suppression des majuscules de début de phrase (quand cela est pertinent) a été effectué.*
- *Voici un exemple de corpus en colonne sur la phrase précédente :*

```
0       investiture     nc      O
1       aujourd'hui     adv     O
2       à               prep    O
3       Bamako          np      B-geoloc
4       ,               ponctw  O
5       Mali            np      B-geoloc
6       ,               ponctw  O
7       du              prep    O
8       président       nc      O
9       Amadou          np      B-person
10      Toumani         np      I-person
11      Touré           np      I-person
```

    1.
    2. *Comme on peut le voir, le corpus contient 4 colonnes. Chaque ligne décrit un mot, ou un signe de ponctuation. Les annotations en entités nommées, contenues dans la 4e colonne,*

se retrouve avec l'encodage B,I,O, où la lettre B correspond au début d'une entité nommées, la lettre I au milieu ou à la fin d'une entité nommée, et enfin la lettre O correspond aux mots hors entités nommées. De plus le corpus est découpé en phrase, chaque phrase étant séparée par une ligne vide.

3. *Chaque colonne contient une information sur le mot. Voici la signification de chacune d'entre elle :*
    - *colonne 1 : indice du mot dans la phrase, en commençant à la valeur 0*
    - *colonne 2 : forme du mot*
    - *colonne 4 : étiquette morpho syntaxique (Part-Of-Speech - POS) du mot donné par un étiqueteur automatique (peut contenir des erreurs)*
    - *colonne 4 : étiquette d'entité nommée à prédire.*
4. *Les étiquettes à prédire sont :*
    - *B-geoloc I-geoloc : début et milieu (ou fin) d'une entité géographique*
    - *B-org I-org : début et milieu (ou fin) d'une entité organisation*
    - *B-product I-product : début et milieu (ou fin) d'une entité organisation*
    - *I-person : début et milieu (ou fin) d'une entité personne*
    - *O : tout mot n'appartenant pas à une entité nommée*

- *Travail à faire : étude sur corpus*
    1. *Ambiguité lexicale*
        - *Faites un programme, dans le langage de votre choix, avec la structure de votre choix, qui permet de lire phrase à phrase ce corpus en colonne et de stocker des informations provenant de chaque phrase. Grâce à ce programme vous devrez répondre aux questions suivantes :*
        - *Exercice 1 : Quel est le nombre d'entités de chaque type dans le corpus donné en exemple ?*
        - *Exercice 2 : Calculer l'ambigüité de chaque mot du corpus. L'ambigüité d'un mot se calcule en mesurant le nombre d'étiquettes (entités ou O) qu'il peut recevoir. Plus ce nombre est grand, plus le mot est ambigu. Si un mot ne reçoit qu'une seule étiquette, son ambigüité est de 1. Par exemple, dans notre corpus le mot, France peut recevoir 5 étiquettes : geoloc, org, person, product, O, son ambigüité est donc de 5. Vous afficherez tous les mots dont l'ambigüité est supérieure à 1, classés par ambigüité décroissante.*
    2. *Patrons d'entités nommées*
        - *En vous basant sur le programme précédent, vous allez maintenant produire des patrons d'entités nommées, à partir des étiquettes morphosyntaxiques attachées aux mots du corpus.*
        - *Exercice 3 : Faite un programme qui affiche toutes les entités nommées d'un corpus avec leur fréquence. Par exemple : Organisation des Nations Unies 126*
        - *Exercice 4 : Vous allez maintenant rajouter le nombre d'occurrence de chaque entité dans les 5 étiquettes geoloc, org, person, product, O . Exemple : France geoloc 629 org 377 person 0 product 0 O 24*
        - *Exercice 5 : Mêmes questions que les deux précédentes, mais en considérant maintenant les patrons d'entités nommées en remplaçant les mots par les étiquettes morphosyntaxiques. Par exemple, le patron de l'entité France est np, celui de l'entité association des producteurs de pétrole africains est nc prep nc prep nc adj.*
-