Création de données

Cours M1 TAL - Inalco

Johanna Cordova

Programme

I. Formats de données

Structures de données Python, JSON, Dataframes, XML-TEI, formats spécifiques du TAL (conll, HF datasets)

Programme

I. Formats de données

Structures de données Python, JSON, Dataframes, XML-TEI, formats spécifiques du TAL

II. Obtenir des données et corpus

OCR, Corpora open source

Programme

I. Formats de données

Structures de données Python, JSON, Dataframes, XML-TEI, formats spécifiques du TAL

II. Obtenir des données et corpus

OCR, Corpora open source

III. Visualisations

Matplotlib, Seaborn, Bokeh

Partie 1 : Formats de données

Reconnaissance de la parole

(*Automatic Speech Recognition*, ASR)

Extraits audio (< 15 secondes) + leur transcription

Traduction automatique

Phrases alignées en 2 ou plusieurs langues

Vision par ordinateur

Corpus d’images annotées (objets présents, catégories, relations spatiales)

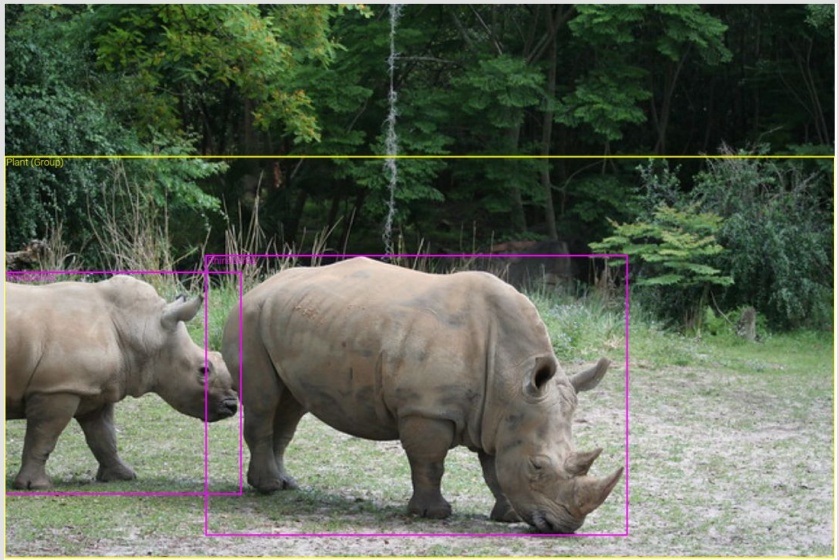

Image tirée du corpus Open Images Dataset V7

Formats de données courants

Fichiers texte brut (.txt)

Manipulables via un éditeur de texte (`gedit`, nano, vim, TextEdit, `cot`, etc.)

C’est dans des fichiers textes qu’on écrit le code informatique.

Formats de données courants

Fichiers tabulaires

Contiennent des données sous forme de lignes et colonnes.

Souvent manipulables sur des tableurs (Excel, Libre Office).

Ex : format CSV (*comma-separated values*),

TSV (*tab-separated values*)

Formats de données courants

Fichiers structurés

Contiennent à la fois les données et des métadonnées qui structurent le document.

Fichiers XML, structurés par des balises :

- pages HTML
- documents de la suite Office (docx, odt, xlsx, etc.)

Formats de données courants

Fichiers structurés, le langage des machines

Format JSON

```
{       "ville": "Paris",
        "temperature": 18,
        "conditions": "nuageux",
        "previsions":
                [ {"jour": "lundi", "temp_max": 20, "temp_min": 15},
                {"jour": "mardi", "temp_max": 22, "temp_min": 16} ]
}
```

Les données élémentaires : bits et octets

Comment sont encodées les données ?

L’unité de base de l’information est le **bit** (valeur 0 ou 1).

Dans l’ordinateur, les informations en mémoire sont stockées par groupes de 8 bits (= **octet,** *byte* en anglais).