جامعة هواري بومدين
للعلوم والتكنولوجيا
U S T H B

Extraction d'information

Cours 1

Nassim ZELLAL

Objectifs du cours

- Apprendre à maîtriser la plateforme de Traitement Automatique du Langage Naturel (TALN, en anglais NLP) « UNITEX » pour l'extraction d'information à partir de corpus. Les techniques d'extraction d'information proposées par « UNITEX » sont : les dictionnaires DELAF/DELAS, les graphes (grammaires) d'extraction et les expressions régulières.
- Apprendre à construire automatiquement des ressources linguistiques pour l'analyse et l'extraction d'information à partir du Web.
- Manipuler des chaînes de caractères pour traiter des problématiques liées à la langue et à l'analyse de fichiers textuels.
- Apprendre à maîtriser les encodages de fichiers et savoir mettre en œuvre des solutions, dans les situations les plus fréquemment rencontrées en pratique, pour la gestion et le traitement informatique des encodages.
- Exploiter les connaissances acquises dans le cours « Python », notamment la gestion de fichiers, les expressions régulières et l'analyse statistique de données textuelles pour l'« Extraction d'information ».

Le TALN

- Le traitement automatique du langage naturel (TALN) ou traitement automatique des langues (TAL) est la capacité pour un programme informatique de comprendre le langage humain. Il fait partie des technologies d'intelligence artificielle.
- En Intelligence Artificielle (IA), le TALN est une discipline qui a pour objectif donc de modéliser, grâce à l'informatique, le langage qu'il soit écrit ou parlé. Les technologies TALN sont présentes dans divers systèmes, comme Google, IBM Watson, Facebook, Apple Siri, Amazon Alexa, etc.).
- Le traitement automatique des langues, TAL, est une discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle.

Extraction d'information (EI)

- L'EI est utilisée dans le TALN, qui appartient au domaine de l'Intelligence Artificielle (IA).
- Intelligence artificielle -> Data mining (documents structurés) -> Text mining -> Extraction d'information (documents non structurés).
- L'EI ne cherche pas à comprendre les textes dans leur ensemble, mais vise à extraire d'un texte des informations (éléments) pertinents, afin de comprendre la sémantique du texte.

Informations extraites par un système d'EI

- Parmi les informations extraites par un système d'Extraction d'Information :
  - Reconnaissance d'entités nommées (NER : Named Entities Recognition). Une entité nommée peut être un événement ou un nom propre, e.g. nom de personne, nom de lieu, nom d'organisation, etc.
  - Reconnaissance de relations sémantiques entre entités nommées, e.g. relation d'acquisition, relation de contact, relation de déplacement, etc.

Extraction d'information (EI)

- Entrée :
  - L'**Algérie** a battu le **Niger** **4 à 0** lors du match de la 4e journée du deuxième tour des qualifications à la Coupe du monde **2022** au **Qatar**.

  Sortie :
  - Algérie, Niger et Qatar ➔ nom de pays ➔ nom de lieu ➔ nom propre
  - 4 à 0 ➔ score ➔ expression numérique
  - 2022 ➔ date ➔ expression temporelle
  - Coupe du monde 2022 au Qatar ➔ événement

Extraction d'information (EI)

- L'extraction d'information consiste à analyser des corpus (données textuelles), afin d'en extraire des informations (ou connaissances) pertinentes, en vue d'une application précise.
- L'EI permet de produire automatiquement une représentation structurée (e.g. format **XML**) du contenu non structuré (texte brut) d'un corpus.
- L'**<PAYS>**Algérie**</PAYS>** a battu le **<PAYS>**Niger**</PAYS>** **<SCORE>**4 à 0**</SCORE>** lors du match de la 4e journée du deuxième tour des qualifications à la **<EVEN>**Coupe du monde **<DATE>**2022**</DATE>** au **<PAYS>** Qatar **</PAYS>** **</EVEN>**

À quoi servent ces informations ?

Applications nombreuses :

- indexation sémantique de documents pour les moteurs de recherche;
- enrichissement automatiquement d'une base de données;
- analyse sémantique et enrichissement d'ontologies;
- inférence automatique (moteur sémantique);
- anonymisation de documents;
- informatique décisionnelle, aide à la décision (business intelligence);
- analyse en temps réel de l'information (veille intelligente et renseignement);
- systèmes de questions-réponses (moteurs de recherche);
- extraction automatique d'opinions (analyse d'opinions/sentiments);
- résolution d'anaphore coréférentielle et non coréférentielle en TAL.

Comment extraire l'information?

- Dictionnaires.
- Grammaires (graphes) d'extraction.
- Expressions régulières (voir les chapitres 5 et 6 du cours Python - L2 Acad - S4).

Extraction d'information - Plateformes de TAL

- **UNITEX.**
- **NOOJ.**
- **HST (THALES).**

Remarque : ces plateformes sont multilingues.

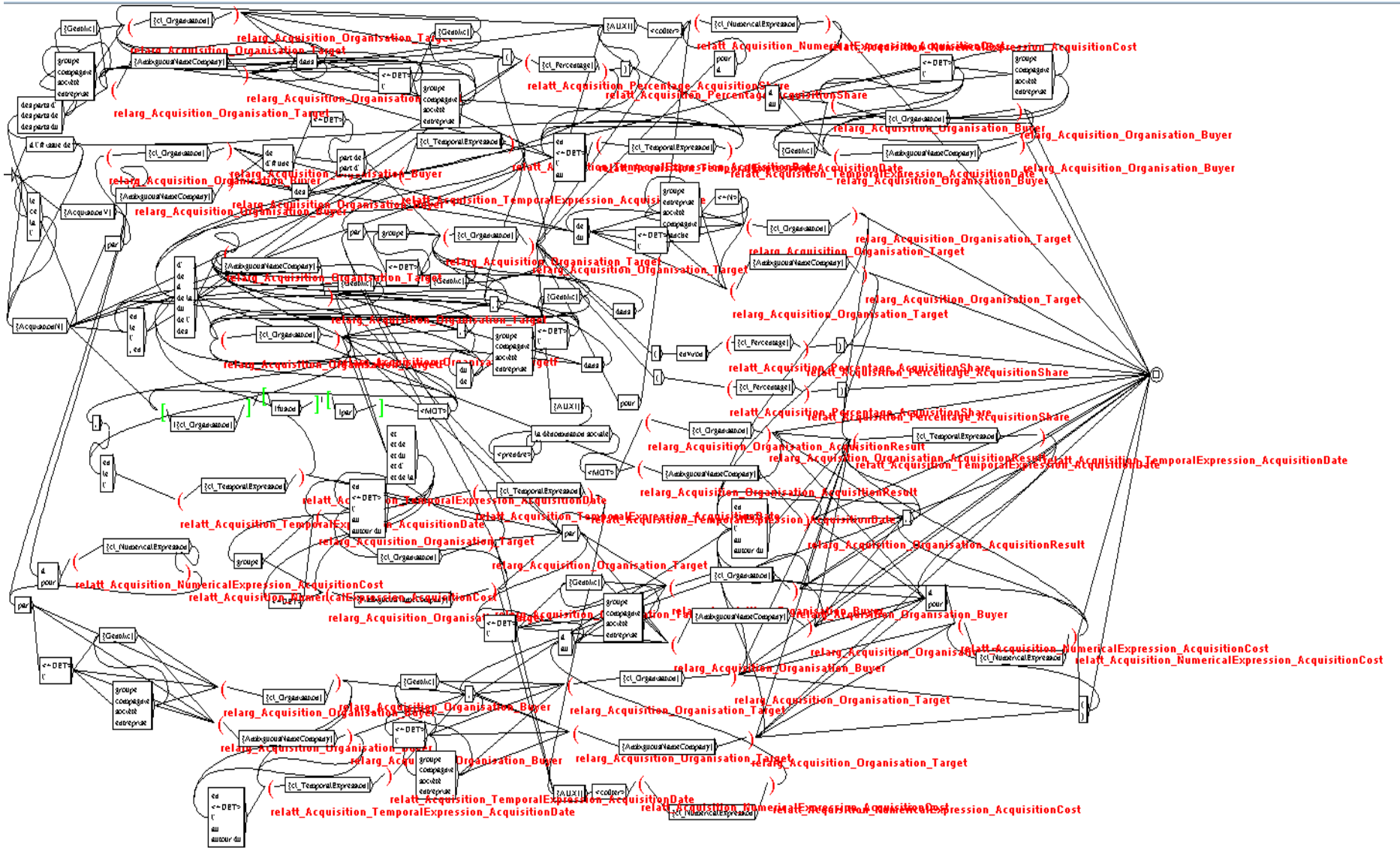
Qu'est-ce qu'UNITEX

- UNITEX est une plateforme/suite logicielle d'analyse de corpus et de TALN basée sur des dictionnaires et des grammaires (graphes).
- Les grammaires et les dictionnaires sont des ressources linguistiques permettant l'extraction d'information, e.g. l'extraction d'entités nommées (noms de personnes, d'organisations, etc.), ou de relations entre entités nommées, e.g. relation d'acquisition.

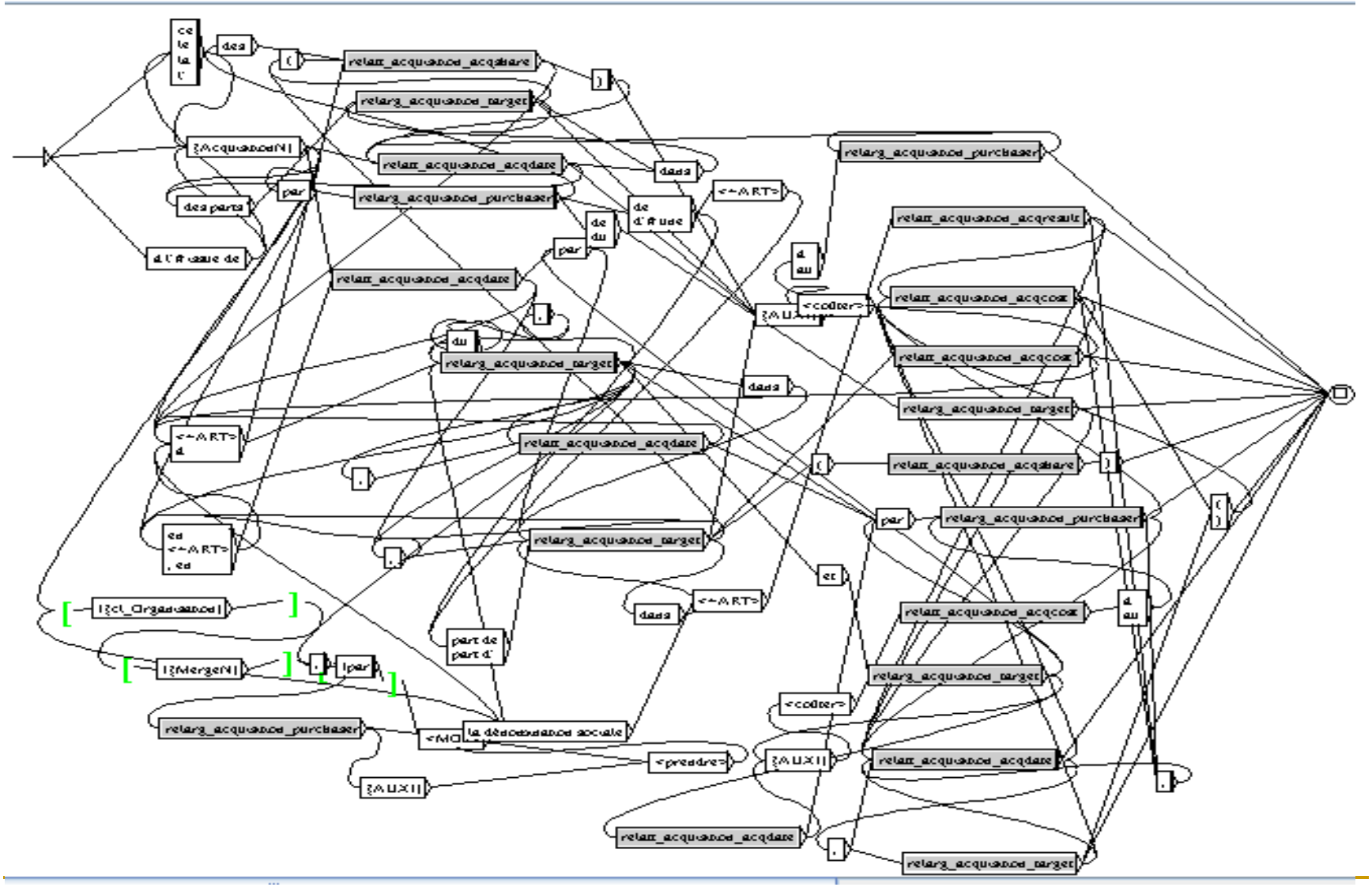
Exemple d'un dictionnaire DELAF (UNITEX)

```
abacavir,.N+subst
abatacept,.N+subst
abciximab,.N+subst
abiratérone,.N+subst
acamprosate,.N+subst
acarbose,.N+subst
acébutolol,.N+subst
acéclofénac,.N+subst
acémétacine,.N+subst
acénocoumarol,.N+subst
acépromazine,.N+subst
acéprométazine,.N+subst
acétazolamide,.N+subst
acétohexamide,.N+subst
acétylcholine chlorure,.N+subst
```

Exemple d'un graphe d'extraction (UNITEX)



Graphe d'extraction après factorisation



Résultat de l'extraction

de paris sur les événements organisés par ASO (Amaury Sport Organisation), à savoir le Tour de France et le Dakar ;– **<relcl_Acquisition>**Winamax qui a été racheté par le chanteur Patrick Bruel et par Marc Simoncini**</relcl_Acquisition>**, le fondateur et dirigeant du site de rencontres Meetic.[.] II.– UNE LIBÉRALISATION RÉGULÉE...[.] Ce projet de [.]

Sortie structurée XML

```
- <SemTag value="cl_Company" text="Winamax" offset="893" length="7">
    <SemTag value="cl_Organisation"/>
  </SemTag>
  <SemTag value="relcl_Acquisition" text="Winamax qui a été racheté par le chanteur Patrick Bruel et par Marc Simoncini" offset="893" length="77"/>
  <SemTag value="relarg_Acquisition_Organisation_Target" text="Winamax" offset="893" length="7"/>
  <SemTag value="cl_Person" text="Patrick Bruel" offset="935" length="13"/>
  <SemTag value="relarg_Acquisition_Person_Buyer" text="Patrick Bruel" offset="935" length="13"/>
  <SemTag value="cl_Person" text="Marc Simoncini" offset="956" length="14"/>
  <SemTag value="relarg_Acquisition_Person_Buyer" text="Marc Simoncini" offset="956" length="14"/>
  <SemTag value="relcl_Leadership" text="Marc Simoncini, le fondateur et dirigeant du site de rencontres Meetic" offset="956" length="70"/>
  <SemTag value="relarg_Leadership_Person" text="Marc Simoncini" offset="956" length="14"/>
- <SemTag value="cl_Company" text="Meetic" offset="1020" length="6">
    <SemTag value="cl_Organisation"/>
  </SemTag>
  <SemTag value="relarg_Leadership_Organisation" text="Meetic" offset="1020" length="6"/>
- <SemTag value="cl_Newspaper" text="La Tribune" offset="1114" length="10">
  - <SemTag value="cl_MediaCompany">
    - <SemTag value="cl_Company">
        <SemTag value="cl_Organisation"/>
      </SemTag>
```

Installation d'UNITEX

- Lien de téléchargement :
- **http://unitexgramlab.org/fr**
- La plateforme Unitex est multilingue.
- Elle est également multiplateforme : Windows, Linux et OS X.
- Choisir la version 64 bits ou 32 bits.
- Vérifiez que JAVA est bien installé.
- Utiliser « **Unitex Visual IDE** » (Unitex/GramLab IDE 3.2) dans le cadre du cours.
- Exploiter le manuel d'UNITEX.

Installation d'UNITEX

- Sous Linux :
- 1- téléchargez le fichier :
- Unitex-GramLab-3.2-linux-i686.run (32 bits)
- ou bien
- Unitex-GramLab-3.2-linux-x86_64.run (64 bits)

- 2-donnez lui les droits d'exécution par exemple :
- chmod +x Unitex-GramLab-3.2-linux-x86_64.run

- 3-lancez le fichier :
- ./Unitex-GramLab-3.2-linux-x86_64.run

- 4-lancez le jar "Unitex.jar" se trouvant dans le dossier "Unitex-GramLab>App".

- Sous Windows :
- Téléchargez l'exécutable Unitex-GramLab-3.2_win64-setup.exe (version 64 bits) ou Unitex-GramLab-3.2_win32-setup.exe (version 32 bits).
- Ensuite, lancez l'exécutable à partir du raccourci sur votre bureau ou bien à partir du dossier "Unitex-GramLab>App", en y ouvrant une invite de commandes puis tapez: java -jar Unitex.jar

Mon courriel

zellal.nassim@gmail.com