

Retour d'expérience : Apprentissage automatique et web sémantique

Méthodologie d'alignement automatique et de désambiguïsation d'entités géographiques

Pascal Cuxac

INIST-CNRS

Vandoeuvre lès Nancy

pascal.cuxac@inist.fr



Contexte

ISTEX : Initiative d'Excellence en Information Scientifique et Technique

Offrir, à l'ensemble de la communauté de l'ESR, un accès en ligne aux collections rétrospectives de la littérature scientifique dans toutes les disciplines (<http://www.istex.fr>).

2 principaux objectifs:

- Un vaste programme d'acquisition de contenus électroniques pour les scientifiques
- Mettre en place un système permettant d'agréger toutes les données achetées et d'offrir des données normalisées et enrichies via plusieurs canaux

Contexte

ISTEX-RD : Intégrer dans les données ISTEX des enrichissements complémentaires à partir du plein texte et à l'aide de plusieurs outils ou méthodes issus de la recherche pour les mettre à disposition d'autres projets ou initiatives.

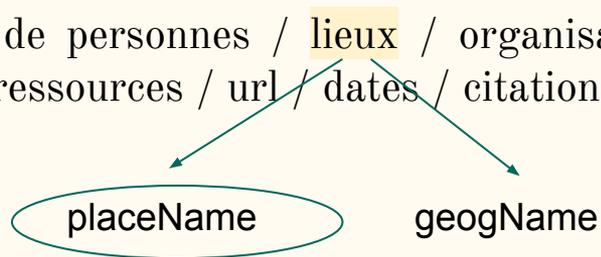
4 axes de travail :

- identification des références citées et structuration,
- indexation automatique,
- reconnaissance d'entités nommées,
- catégorisation des documents.

Contexte

ISTEX-RD :

- reconnaissance d'entités nommées :
 - en collaboration avec le LI de Tours
 - avec l'outil Unitex/CasSys
 - extraction de 8 types d'EN : nom de personnes / lieux / organisations / projets financés / organismes hébergeur de ressources / url / dates / citations



Contexte

DATA-ISTEX : exposition des données ISTEEX produites et/ou transformées à l'Inist selon les normes du web sémantique, modélisées via une ontologie dédiée.

data.istex.fr expose le Triple Store des données

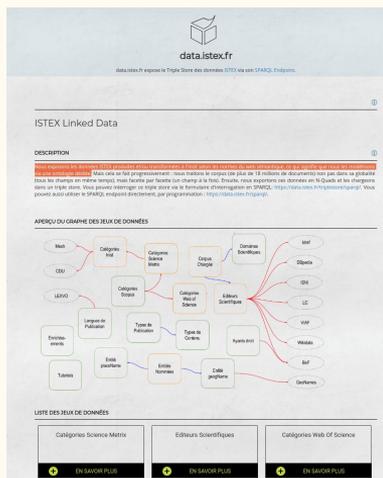
[ISTEX](#) via son [SPARQL Endpoint](#)

<https://data.istex.fr/>

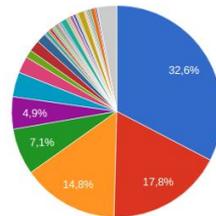
données exportées en N-Quads et chargées dans un triple store

<https://data.istex.fr/triplestore/sparql/>

<https://data.istex.fr/sparql/>



```
SELECT ?libellecatinist (count(?libellecatinist) as ?totallibellecatinist) (count (?libellecatinist)*100/?totalcontentTypePDR as ?percent)
where
{
  select ?libellecatinist
  where
  {
    ?docistex <https://data.istex.fr/ontology/istex#contentType> ?conttype.
    ?conttype <http://www.w3.org/2004/02/skos/core#altLabel> "papier de recherche"@fr.
    ?docistex <https://data.istex.fr/ontology/istex#subjectInist> ?catinist.
    ?catinist <http://www.w3.org/2004/02/skos/core#prefLabel> ?libellecatinist.
    filter(lang(?libellecatinist)="fr")
  }
  select (count(?docistex) as ?totalcontentTypePDR)
  where
  {
    ?docistex <https://data.istex.fr/ontology/istex#contentType> ?conttype.
    ?conttype <http://www.w3.org/2004/02/skos/core#altLabel> "papier de recherche"@fr.
  }
}
order by desc (?percent)
```



Contexte

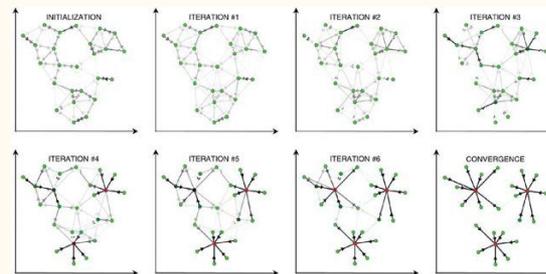
DATA-ISTEX

- quelques chiffres : 10 263 903 documents
55 348 674 placeName
- alignement des placeName avec Geonames →
- problèmes :
 - une même entité peut avoir plusieurs orthographes
 - une même entité peut correspondre à plusieurs lieux différents (Paris, France / Paris, Texas)
→ désambiguïsation nécessaire

The screenshot displays the 'data.istex.fr' interface for the entity 'Toulouse'. It features several sections: 'ENTITÉ NOMMÉE' with the value 'Toulouse', 'LABEL GEONAMES' with the value 'toulouse', and 'LOCALISATION' which includes a map of Europe with a red dot indicating the location of Toulouse in France. Below the map, the 'PAYS' section shows 'France'. At the bottom, the 'ENRICHISSEMENT WIKIDATA' section provides a Wikidata ID and a thumbnail image of a bridge over a river in Toulouse.

Projet ISTEEX-LOD

- Une même entité peut avoir plusieurs orthographes et/ou peut comporter des erreurs (OCR/ fautes d'orthographe)
 - Une méthode de clustering (classification non supervisée) peut-elle résoudre ce problème ?
 - Expérimentation :
 - Programme python :
 - calcul de similarités entre entités (Levenshtein / Jaccard / Hamming...)
 - clustering par propagation d'affinité



Cabanes, Guénaél. (2010). Two-level Unsupervised Clustering driven by neighborhood and density.

Projet ISTE_X-LOD

➤ Clustering par propagation d'affinité

Algorithme itératif reposant sur le partage des « affinités » :

- Chaque élément c repère dans son voisinage un élément qui lui ressemble suffisamment, et augmente son affinité pour cet élément ;

Les étapes suivantes consistent à « propager » cette affinité :

- Chaque élément c repère celui pour qui il a la plus grande affinité, noté m ;
- Il ajoute à ses propres affinités celles de m ;
- Cette étape est répétée un certain nombre de fois, jusqu'à ce que le nombre d'éléments passe en dessous d'un certain seuil (ou quand il n'y a plus aucun changement).

Il y a alors trois cas :

- L'élément considéré possède une affinité maximale pour un autre élément : il lui ressemble ;
- L'élément considéré possède une affinité maximale pour lui-même : il est « exemplaire » (exemplar) ;
- L'élément considéré possède une affinité nulle : il est « isolé ».

On obtient à l'issue de l'algorithme un arbre complet, reliant les éléments semblables qui ont pu être identifiés comme tels.

Projet ISTEEX-LOD

➤ Un regroupement “d’écritures” à l’aide de clustering

- 24- **"Oeste City"** - Oeste City
- 25- **"Oestenstad"** - Oestenstad
- 26- **"Oestrogenicity"** - Oestrogenicity
- 27- **"Oestrone Chetwynd Bridge"** - Oestrone Chetwynd Bridge
- 28- **"Oestrusgrad"** - Oestrusgrad
- 29- **"Oetmantown"** - Oetmantown
- 30- **"Oettersdorf"** - Oetsdorf, Oettcrsdorf, Oettersdorf
- 31- **"Saint Michels"** - Saint Micheh, Saint Michei, Saint Michel, Saint Micheland, Saint Micheld, Saint Michele, Saint Michell, Saint Michelle, Saint Michels, Saint Michiel
- 32- **"Uylenburgh"** - Uyicnburgh, Uylcnburgh, Uylenburgh, Uyllenburgh
- 33- **"Valium City"** - Valium City
- 34- **"Valkinburgh"** - Valkburgh, Valkcnburgh, Valkenbourg, Valkenburgh, Valkinburgh
- 35- **"Vandoeuvres-Lès-Nancy"** - Vandoeuvre-Les-Nancy, Vandoeuvre-Lès-Nancy, Vandoeuvres-Nancy, Vandoeuvres-Lès-Nancy, Vandoeuvre-Lès-Nancy, VandœUvre-Lès-Nancy
- 36- **"Vandorf"** - Vandorf
- 37- **"Verbocentricity"** - Verbocentricity
- 38- **"Verbofstad"** - Verbofstad
- 39- **"Verbotenheitsgrad"** - Verbotenheitsgrad, Verbotenhextsgrad, Verbundenheitsgrad
- 40- **"Verbrennungsgrad"** - Verbraunungsgrad, Verbreitungsgrad, Verbrennungsgrad, Verbriiunungsgrad

n° cluster

nom du cluster

composition
du cluster

Projet ISTE_X-LOD

- Un regroupement “d’écritures” à l’aide de clustering
 - problèmes :
 - tendance à la surestimation du nombre de classes
 - complexité de calcul quadratique

Projet ISTEEX-LOD

➤ une même entité peut correspondre à plusieurs lieux différents
(Paris, France / Paris, Texas)

➤ programme python utilisant la librairie “geopy”, permet d’interroger :

ArcGIS (USA) / Baidu Maps (Chine) / Bing Maps Locations (Microsoft) /
DataBC (Canada) / GeocodeFarm (Allemagne) / GeocoderDotUS (USA) /
GeoNames / **Google Maps v3** / IGN France GeoCoder OpenLS / LiveAddress
de SmartyStreets (USA) / NaviData (Canada) / **OpenStreetMap** / Open Cage
Data (UK) / OpenMapQuest / YahooPlaceFinder / Yandex (Russie) / ...

Projet ISTEEX-LOD

➤ une même entité peut correspondre à plusieurs lieux différents

- Algeciras <http://sws.geonames.org/3690160>
- Algeciras <http://sws.geonames.org/1731531>
- Bordeaux <http://sws.geonames.org/3031582>
- Bordeaux <http://sws.geonames.org/6420629>
- Bordeaux <http://sws.geonames.org/4795320>
- Bordeaux City Of London None
- Cassio Bridge None
- Cassis <http://sws.geonames.org/3028431>
- Cassis <http://sws.geonames.org/3728181>
- Casson <http://sws.geonames.org/3028428>
- Chigaco <http://sws.geonames.org/1049480>

Projet ISTEEX-LOD

➤ La Représentation Continue des Données : les méthodes “*Word2Vec*”

- Words embeddings (Word2Vec, Doc2Vec, FastText, Glove...):
 - prendre en compte le contexte d'apparition des mots
 - utiliser de gros volumes de données
 - construire une représentation vectorielle “continue” (dense) des mots

C'est quoi un vecteur de mots ?

Ah....
ok...!

```
0.81272 -1.0838 1.6314 2.183 0.88717 3.4032 -0.63171 -0.18799 1.1778 1.3269 -0.14901 1.9045 -0.82711 1.3089 -0.92383
0.85292 -1.3412 0.4911 0.5666 0.22239 1.7081 -1.9718 -0.22039 0.6157 0.68077 0.34799 -0.68727 -2.2208 -1.7079
27763 pursuing 0.1814 1.1396 -0.63954 0.53234 -1.5122 0.32583 -0.92983 -0.11104 -0.42462 0.6394 1.1556 -0.1322 1.0312 -
0.6212 -2.3034 -0.02958 -0.75447 1.1307 0.49887 -1.8479 1.2272 -0.6398 0.57017 0.85979 0.41118 -0.01866 1.5142
1.4477 -0.25592 -0.75293 -0.95643 0.69413 1.1009 0.602935 -0.37265 0.689 -0.68494 1.527 0.24741 0.27364 0.73745
0.11361 -0.62895 0.636582 -1.4827 0.89253 0.11218 -1.545 -0.38575 0.33598 -0.26433 0.31123 0.88516 0.26979 1.0297
0.67281 -0.29572 0.71344 -0.63047 0.82829 0.29336 2.1812 1.2466 2.864 0.32493 0.30274 0.88884 0.61155 0.88249 0.8495
0.68713 1.1251 -0.86267 0.51923 0.49477 1.0857 0.6618 0.6938 -0.98493 1.631 1.4415 -0.57286 -1.2931 0.48723 0.88866
2.4948 0.40932 0.76954 1.6533 -0.67864 0.79310 -0.10976 0.17449 -0.64651 -1.3529 -0.28863 0.56986 -0.8215 -0.
072794 -1.3649
27764 downward 0.665615 -2.3961 -0.803237 0.52187 0.49642 1.9361 -0.78846 -0.847227 1.206 1.0888 0.38067 -1.1656 1.6117 -0.
05176 0.48368 0.1878 -0.048939 -0.14583 0.63105 -0.57502 0.52289 0.57923 1.3293 -2.54455 -0.15868 -1.8628 0.60816
0.38320 -1.6135 0.9185 2.1224 1.1757 -0.844673 0.98873 0.48587 0.618942 0.43658 0.56358 -1.7883 0.78794 -0.38481 0.
03362 1.2486 -0.37043 -1.2286 0.2805 -0.84383 -0.84151 0.64212 -0.21208 -0.47805 -1.4384 0.28502 -0.71285 -0.38660
0.46832 -2.3843 1.1568 1.7884 -0.78994 -0.653413 0.66446 0.893119 0.023447 0.4886 -0.21264 -0.83842 -0.5382 -0.55733
0.7851 0.3843 0.3158 0.020589 1.6259 1.4288 -0.29089 -0.18405 0.17272 0.68828 0.25996 0.3613 0.28579 1.7473
0.28558 0.88845 0.12665 -0.0046637 -1.1268 0.74703 1.2159 -0.998028 0.63396 -0.26367 -0.25701 -0.27963 0.10115 1.1599
1.1073 -0.48877 0.63286
27765 low-activation 0.33553 0.32679 0.14132 0.6821 -0.73857 0.36626 0.45739 0.7956 -0.3197 1.2522 1.2808 0.88495 0.32463
0.37441 -2.229 0.37471 0.882989 0.1872 0.85429 -2.3875 0.20169 0.74811 1.1869 2.2773 1.6011 0.6386 0.85465 0.5938 0.
08182 1.2433 0.66692 0.38727 0.28242 0.74280 -1.1874 -1.1331 -0.55866 -0.831516 0.84831 0.14821 -1.1808 0.20843
0.28832 0.23431 -2.4965 0.36279 -0.0881374 -1.2379 -0.935735 -1.1075 -0.78529 0.89251 -0.69137 -0.50933 3.1696 0.36815
1.4824 0.3845 -0.082187 0.78317 0.85888 1.107 0.87895 0.86285 1.8285 0.68483 1.28941 -1.817 0.85994 0.3388 1.3833
0.27296 -0.3894 1.7252 0.4859 0.885933 0.95227 -0.63567 0.66814 0.91886 0.49945 1.9138 1.1416 0.57434 0.89904
-1.6443 -0.14428 -1.3177 -0.36643 1.2273 1.7445 -0.49936 0.0851118 -1.2809 -0.85999 -1.2977 0.78892 0.35613 0.88175
-1.1534
27766 danger 0.081921 -0.82785 -0.61853 -0.75085 0.11406 2.6327 -1.0893 -0.10661 1.5483 0.6112 0.67144 -2.1416 -0.955882
0.21958 1.0648 -1.8058 0.021109 0.4023 -0.68809 -0.87229 0.41285 0.81223 -0.26016 1.1937 0.36789 0.63197 -0.68087
0.185 0.33478 0.48542 0.73892 -1.4848 0.640553 0.39135 1.4814 0.878589 1.8337 1.705 -0.68826 0.55532 -0.04028 -0.1974 -
0.27888 -0.64762 -0.86282 -0.21108 -2.4508 0.08881 0.71881 -0.52168 0.66587 -0.88828 0.88889 -0.23489 1.509 0.20932
-2.2529 -0.49832 -0.63682 -0.97206 -1.2338 0.0347 1.3723 1.0471 1.7798 0.91481 1.6315 0.30381 -0.47404 -0.28024
0.80217 0.14975 0.41984 -0.16374 2.7256 -1.7366 0.89582 -1.1629 -0.98818 -0.39288 -0.212 1.8238 1.835 -1.171 0.37059
-1.1828 0.74843 0.4712 1.8638 -0.78959 -1.5242 -0.6553 0.02844 -0.9468 0.92895 -1.0455 0.66893 0.0128 -0.46784
0.27897
27767 hydroline 0.28803 -0.52882 -0.69147 0.46952 0.61777 1.7023 0.19598 -0.63481 -0.26455 0.71079 -0.06806 0.06806
1.5759 0.819236 -0.680251 0.85668 0.14172 0.43888 -0.19333 -0.3618 0.30287 0.69936 -0.18346 0.26809 0.60381
0.6688 1.2807 -0.8714 -0.66578 -0.5886 0.803888 -0.67296 2.0495 0.7737 -0.89831 -1.791 1.1888 0.345
0.46666 -0.3014 -0.26964 -0.88397 -0.48848 -0.60897 -0.44702 -0.421 -0.438 0.7068 0.26612 0.71344 -0.21609 -1.2654
0.821686 1.5382 -0.36997 -0.6971 -0.28073 -0.2938 0.29111 -0.43154 -1.2993 0.50726 1.2882 -1.2384 0.74432 -1.2914 0.
052787 -0.066938 0.67264 0.81862 -0.26868 -0.6182 1.2899 -0.59177 -0.21891 -0.03342 -0.77258 -1.1515 0.6111
```

Ben c'est ça !

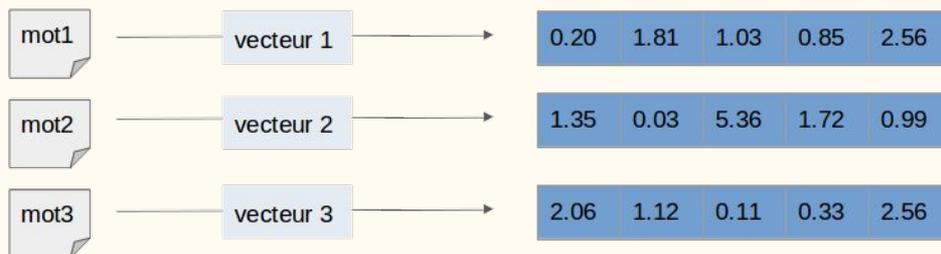
Projet ISTE_X-LOD

➤ Word2Vec : == auto-encodeur

Efficient Implementation of Word Representations in Vector Space,
T. Mikolov, K. Chen, G. Corrado, and J. Dean, 2013.

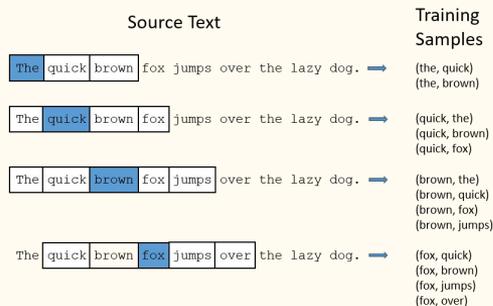
<http://arxiv.org/pdf/1301.3781.pdf>

Va permettre de passer de l'espace des mots à une représentation vectorielle continue. Nous n'avons plus une matrice creuse mais bien une matrice pleine !!

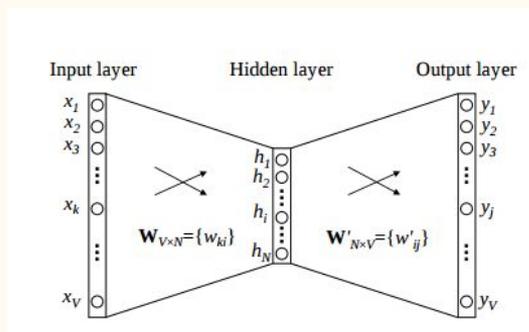


Projet ISTEEX-LOD

- Fixer une fenêtre de prise en compte du contexte



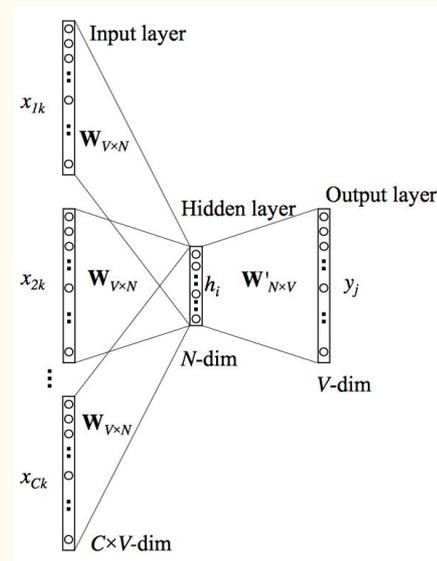
- Fixer la dimension de la couche cachée (dimension des vecteurs de sortie)



Projet ISTEEX-LOD

➤ Words embeddings :

- Réseau Neuronal avec 1 couche cachée,
- En entrée : une description du contexte d'un mot,
- En sortie : le mot qui apparaît dans ce contexte (ou l'inverse),
- Le système va "apprendre" (apprentissage non supervisé) que certains mots sont prédits par les mêmes contextes,
- C'est un réseau de neurones : on récupère à la fin les pondérations de chaque mot avec la couche cachée,
- On ne s'intéresse qu'à l'état du modèle prédictif, pas à sa capacité à prédire.



Projet ISTEEX-LOD

- Words embeddings :
 - Avec cette représentation, les mots se regroupent par **similarité de contexte** qui reflète à la fois une similarité syntaxique et une similarité sémantique,
 - On constate une forme d'**additivité**, par exemple la représentation la plus proche du résultat du calcul $[v\text{Madrid} - v\text{Spain} + v\text{France}]$ est $v\text{Paris}$,
 - Contrairement aux représentations BoW pondérées (TF-IDF), les représentations Word2Vec sont d'assez **faible dimension** (par ex. 200 à 300) et **denses**,
 - Demande un **apprentissage (non supervisé)** sur la base de ressources textuelles très **volumineuses**.

Projet ISTE_X-LOD

Tout ceci permet donc de **représenter des mots** (ou des documents) dans un **espace** et donc de calculer des **distances** entre eux (similarités) → les mots/documents les plus similaires.

Mais quid de la désambiguïsation des mots ?

Méthodologie

AdaGram :

[Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, Dmitry Vetrov. Breaking Sticks and Ambiguities with Adaptive Skip-gram. *International Conference on Artificial Intelligence and Statistics \(AISTATS\) 2016*](#)

Adaptation du modèle Skip-Gram : utilise une approche bayésienne non paramétrique pour “apprendre” plusieurs prototypes associés à un terme.

Apprentissage non supervisé comme Word2Vec, avec en plus un paramètre fixant le nombre de prototypes maximum.

Méthodologie

AdaGram : programme en Julia - Pourquoi Julia ?

- Excellentes performances (compilation à la volée),
- Code facile à lire,
- “*Julia has the performance of a statically compiled language while providing interactive dynamic behavior and productivity like Python, LISP or Ruby.*” in Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). Julia: A fast dynamic language for technical computing.

Méthodologie

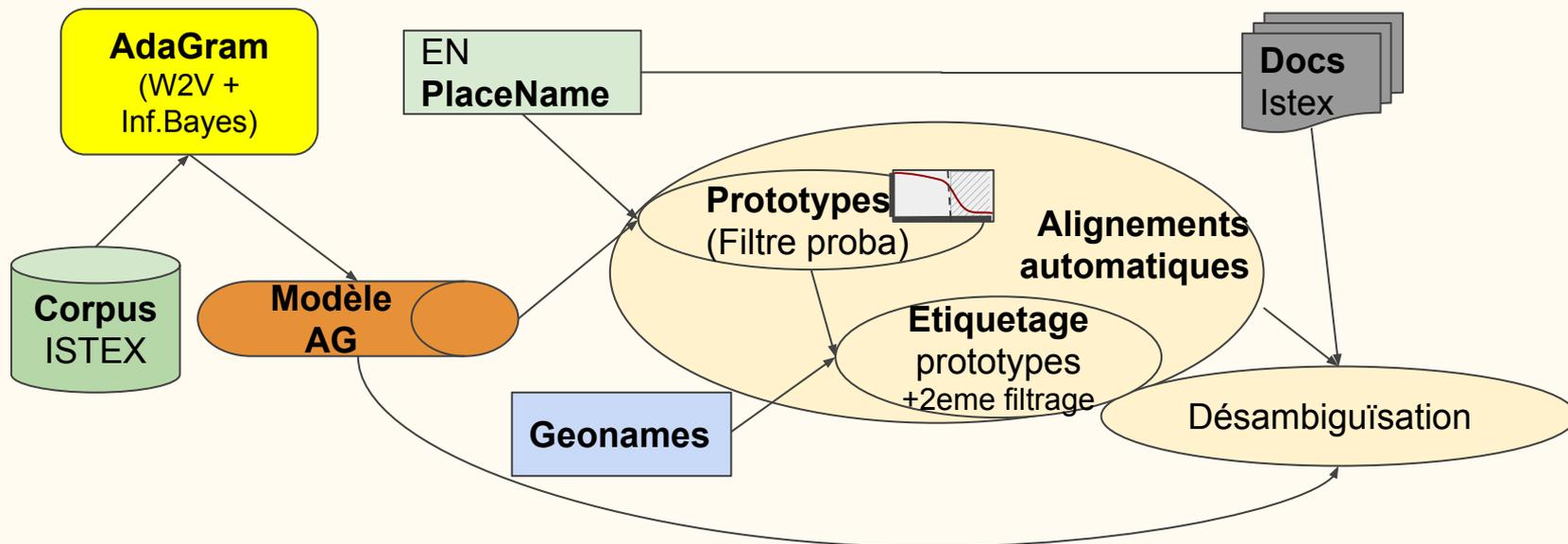
AdaGram - Programme en Julia : <https://github.com/sbos/AdaGram.jl>



Méthodologie

Alignement et Désambiguïsation :

- Alignement sur geonames et sélection de prototypes
- Désambiguïsation dans les documents



Méthodologie

Exemple : Paris

- - Corpus d'apprentissage (full text ISTEEX) → placeName **Paris**
- - AdaGram → 5 prototypes + seuil probabilités (dynamique) :
 - p1 île de france, tour eiffel, montmartre p=0.86
 - p2 dallas, désert, pétrole p=0.84
 - p3 blonde, mannequin, héritière p=0.78
 - ~~○ p4 PMU, cheval, tiercé p=0.23~~
 - ~~○ p5 psg, foot, qatar p=0.01~~

Méthodologie

Exemple : Paris

- - AdaGram → 5 prototypes + seuil probabilités :
 - p1 île de france, tour eiffel, montmartre p=0.86
 - p2 dallas, désert, pétrole p=0.84
 - p3 blonde, mannequin, héritière p=0.78
- - **Etiquetage** “géo” avec geonames (à partir des 10 premiers termes contenus dans le prototype) :
 - p1 France
 - p2 USA, Texas
 - p3 []
- - Au final : 2 **prototypes validés** pour le corpus d’apprentissage utilisé et alignés avec Geonames:
 - - Paris, France [geonames.org/2988507](https://www.geonames.org/2988507)
 - - Paris, USA [geonames.org/4717560](https://www.geonames.org/4717560)

Méthodologie

Exemple : Paris

- - Désambiguïsation dans les textes :
 - 2 prototypes pour Paris dans le corpus d'apprentissage :
 - -p1 France
 - -p2 USA

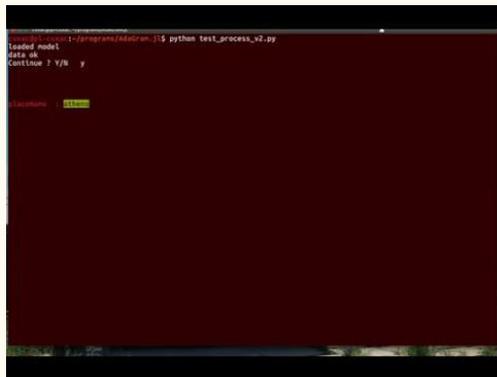
- Pour un document donné Doc ISTE $X_{12345678}$ (full text)
 - Probabilité de chaque prototype :
 - -p1 proba 0.90
 - -p2 proba 0.10

- \Rightarrow dans ce document Doc ISTE $X_{12345678}$
 - Paris = Paris, France (prob =0.9) \rightarrow [geonames.org/2988507](https://www.geonames.org/2988507)

Méthodologie

Exemple D'application :

- corpus de 400 000 documents (textes pleins, format txt) issus d'ISTEX
- 5 prototypes demandés
- liste de 7 placeName (athens, carcassonne, montreal, annapolis, abbeville, lafayette, portsmouth)
- Apprentissage, sélection et labellisation des prototypes :
 - <https://drive.google.com/open?id=1JzqZY1R4mKiCAAPRGn83RkMZel4ZLsE>



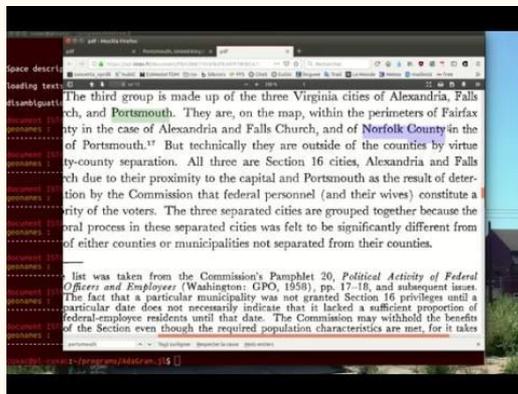
```
~/Documents/program/istex/15 python test_process_v2.py
Loading model
data ok
Continue ? y/n  y

Calculating
```

Méthodologie

Exemple D'application :

- Désambiguïsation dans 7 documents (textes pleins, format txt) issus d'ISTEX :
 - <https://drive.google.com/open?id=1HwrHyG27-zPICuT174GOYX5Sa0l80usD>



Evaluation

Campagne SemEval 2019 :

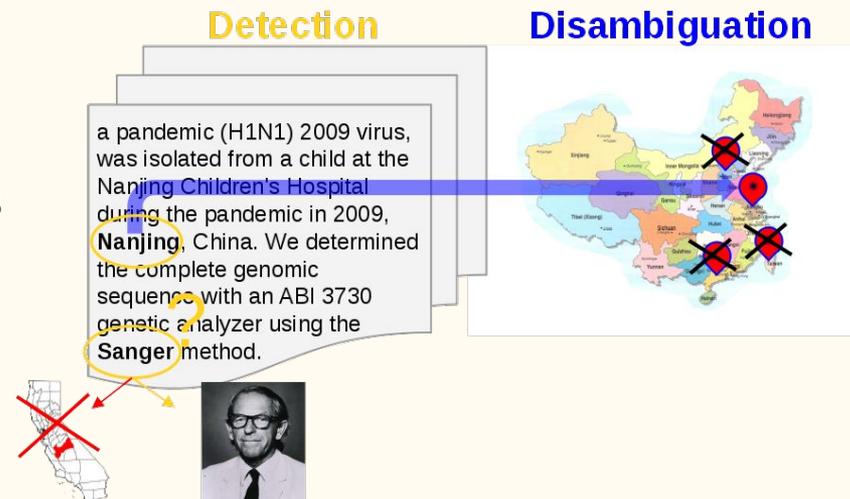
➤ SemEval (Évaluation sémantique) est une série continue de campagnes d'évaluations de systèmes d'analyse sémantique.

➤ Task 12 SemEval'19 :

Toponym Resolution in Scientific Papers

(<https://competitions.codalab.org/competitions/19948>)

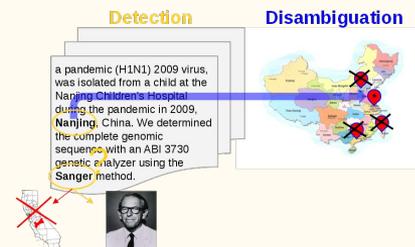
- Subtask 1: Toponym detection
- Subtask 2: Toponym disambiguation
- Subtask 3: end-to-end, toponym resolution



Evaluation

Campagne SemEval 2019 Pourquoi ?

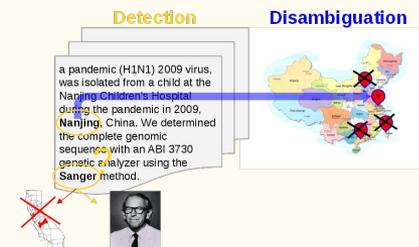
- Avoir un corpus étiqueté internationalement reconnu pour l'évaluation.
- Corriger et adapter l'approche proposée
- S'évaluer à partir de ce corpus
- Éventuellement se positionner par rapport aux participants
- *remarque : notre approche étant non supervisée les corpus d'apprentissage peuvent être utilisés pour évaluer/corriger la méthode.*



Evaluation

Campagne SemEval 2019 -Premiers résultats :

	Fmes
corpus échantillon	0.85
corpus “training”	0.83



Bilan

Quels sont les avantages ?

- Apprentissage non supervisé...pas de taggage,
- Application sur placeName mais potentiellement utilisable sur d'autres types de données,
- Peut désambiguïser dans un texte global mais également dans un paragraphe ou même une phrase,
- Premiers résultats encourageants,
- Parallélisation facile du code Julia.

Bilan

Désambiguïsation d'une EN dans une phrase

`disambiguate(vm,dict,"athens",split ("reflects a 70% budget allocation for the Athens center consulting services"))`

3-element Array{Float64,1}:

0.259825

0.702366 → "georgia"

0.0378085

Bilan

Quels sont les inconvénients ?

- Nécessité d'un corpus apprentissage volumineux (volume/représentativité)
- Capacité de calcul (+ CPU ... + rapidité)
- Capacité de stockage (corpus d'apprentissage, modèle...)

Merci !

Merci de votre attention...

Quelques liens :

Data.istex :

<https://data.istex.fr/> <https://data.istex.fr/sparql/>

<https://data.istex.fr/triplestore/sparql>

Julia

<https://julialang.org/>

AdaGram

<https://github.com/sbos/AdaGram.jl>

Word2vec

<https://github.com/tmikolov/word2vec>

