Retour d'expérience :
Apprentissage automatique et web sémantique

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

Pascal Cuxac pascal.cuxac@inist.fr

INIST-CNRS
Vandoeuvre lès Nancy

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.

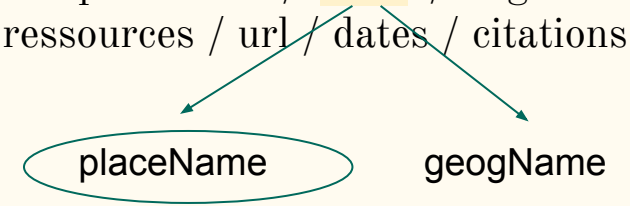
4 axes de travail :

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

Contexte

ISTEX-RD :

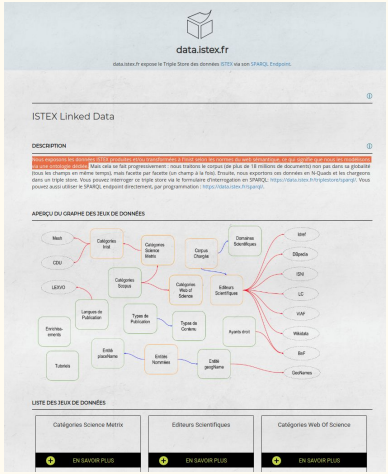
- reconnaissance d'entités nommées :
  - en collaboration avec le LI de Tours
  - avec l'outil Unitex/CasSys
  - extraction de 8 types d'EN : nom de personnes / lieux / organisations / projets financés / organismes hébergeur de ressources / url / dates / citations



Contexte

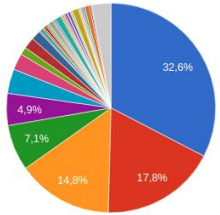
DATA-ISTEX : exposition des données ISTEX produites et/ou transformées à l'Inist selon les normes du web sémantique, modélisées via une ontologie dédiée.

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/

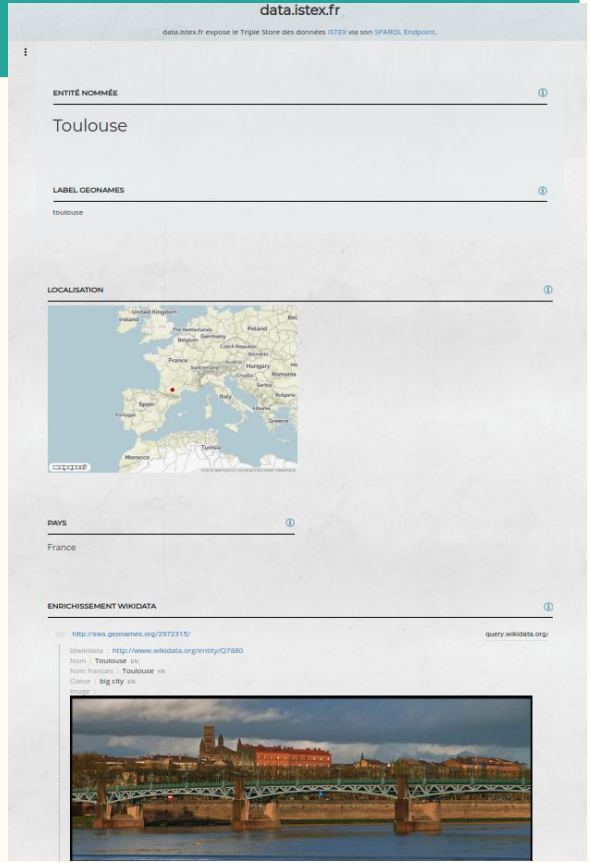
```
SELECT ?libellecatinist (count(?libellecatinist) as ?totallibellecatinist) (count (?libellecatinist)*100/?totalcontentTypePDR as ?percent)
where
{
  {
select ?libellecatinist
  where
  {
?docistex <https://data.istex.fr/ontology/istex#contentType> ?conttype.
   ?conttype<http://www.w3.org/2004/02/skos/core#altLabel> "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

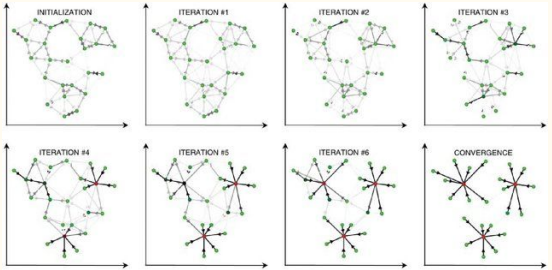
DATA-ISTEX

- 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



Projet ISTEX-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, Guénaël. (2010). Two-level Unsupervised Clustering driven by neighborhood and density.

Projet ISTEX-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 grand 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.

Mezard, Marc. (2007). Computer science. Where are the exemplars?. Science (New York, N.Y.). 315. 949-51. 10.1126/science.1139678.

Projet ISTEX-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, Vandoeuvreits-Nancy, Vandoeuvres-LèS-Nancy, Vandoevre-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

nom du cluster

composition du cluster

n° cluster

Projet ISTEX-LOD

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

Projet ISTEX-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 ISTEX-LOD

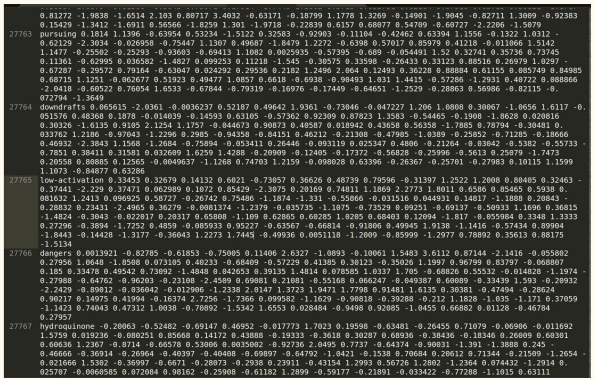
➢ une même entité peut correspondre à plusieurs lieux différents

• 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 ISTEX-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 ?



Ben c’est ça !

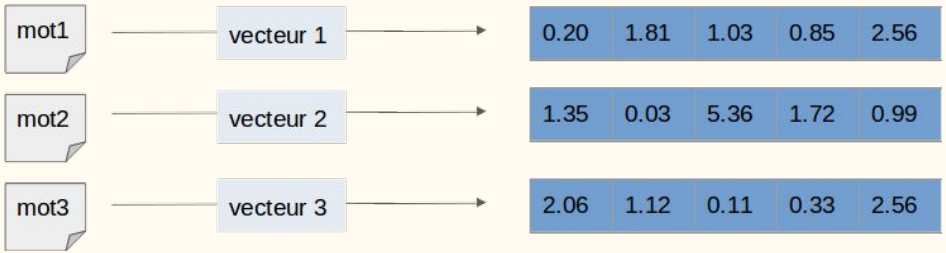
Ah…. ok…!

Projet ISTEX-LOD

➢ Word2Vec : == auto-encodeur

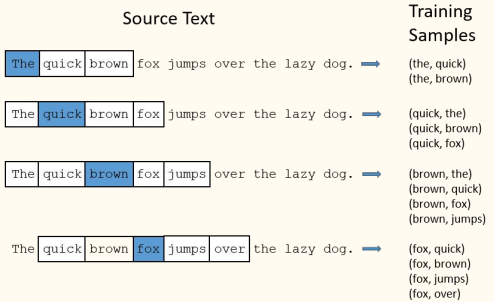
Efficient Implementation of Word Representations in Vector Space,
T. Mikolov, K. Chen, G. Corrado, and J. Dean, 2013.
http://arxiv.org/pdf/1301.3781.pdf

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 !!

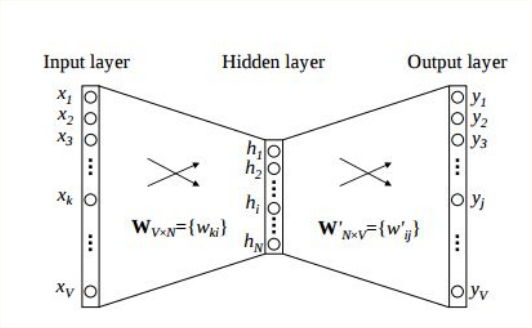


Projet ISTEX-LOD

- Fixer une fenêtre de prise en compte du contexte

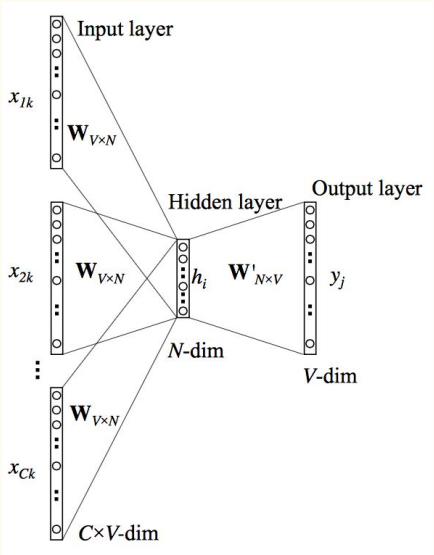


- Fixer la dimension de la couche cachée (dimension des vecteurs de sortie)



Projet ISTEX-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 ISTEX-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 [vMadrid—vSpain+vFrance] est vParis,
  - 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 ISTEX-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.

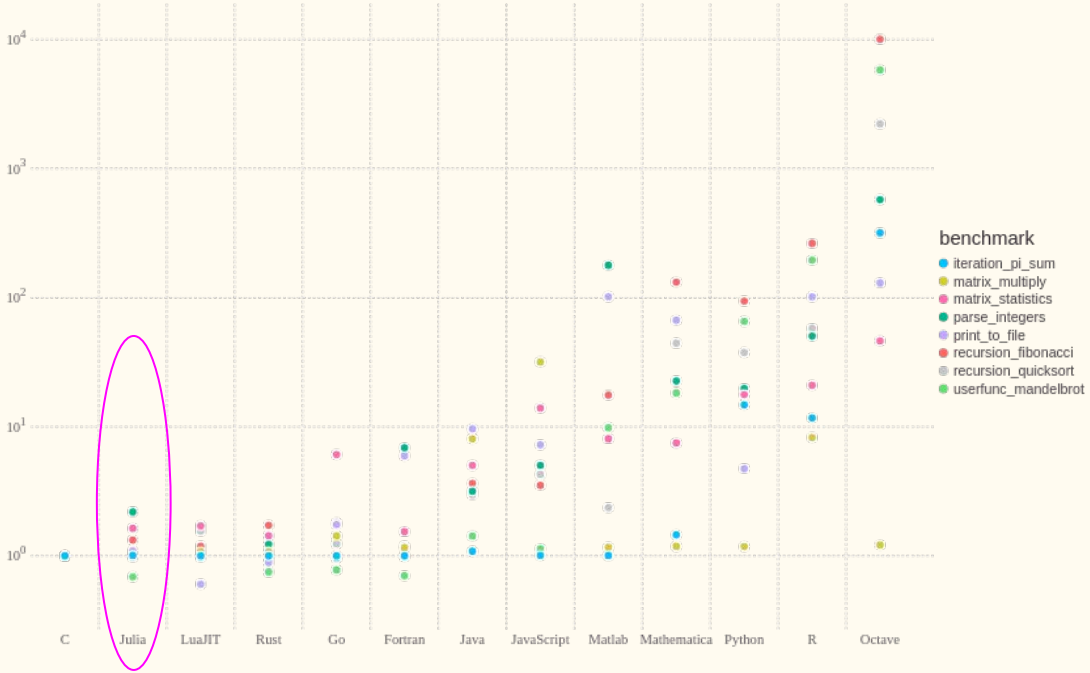
Méthodologie

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

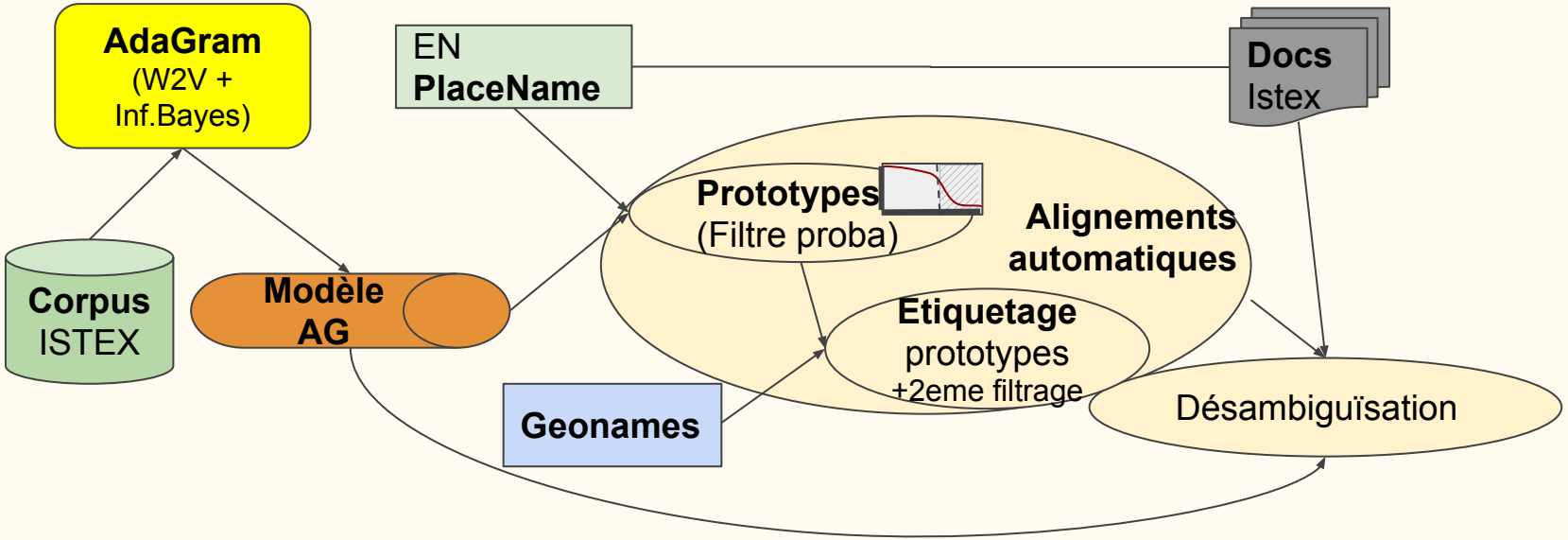
AdaGram - Programme en Julia : https://github.com/sbos/AdaGram.jl



Méthodologie

Alignement et Désambiguïsation :

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



Méthodologie

Exemple : Paris

➢ - Corpus d'apprentissage (full text ISTEX) → 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, quatar 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
  - - Paris, USA geonames.org/4717560

Méthodologie

Exemple : Paris

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

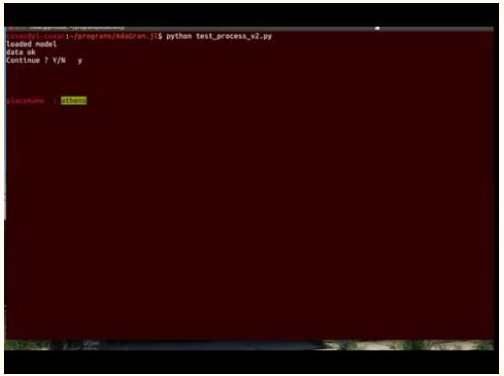
- ➢ Pour une document donné Doc ISTEX12345678 (full text)
  - Probabilité de chaque prototype :
    - -p1 proba 0.90
    - -p2 proba 0.10

- ➢ => dans ce document Doc ISTEX12345678
  - Paris = Paris, France (prob =0.9) → geonames.org/2988507

Méthodologie

Exemple D'application :

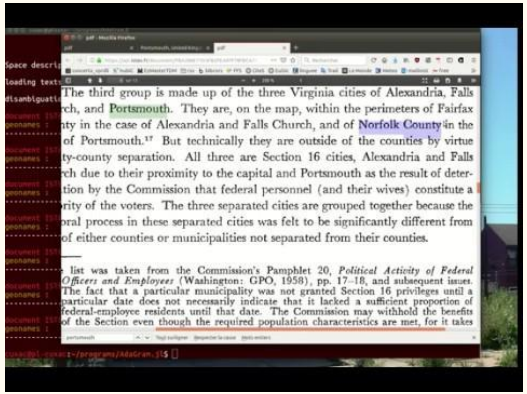
- ➢ corpus de 400 000 documents (textes pleins, format txt) issus d'ISTEX
- ➢ 5 prototypes demandés
- ➢ liste de 7 placeName (athens, carcassonne, montreal, annapolis, abbeville, lafayette, portsmouth)
- ➢ Apprentissage, sélection et labellisation des prototypes :
  - https://drive.google.com/open?id=1JzqZYlR4mKiCAAPRGn83RkMZeEl4ZLsE



Méthodologie

Exemple D'application :

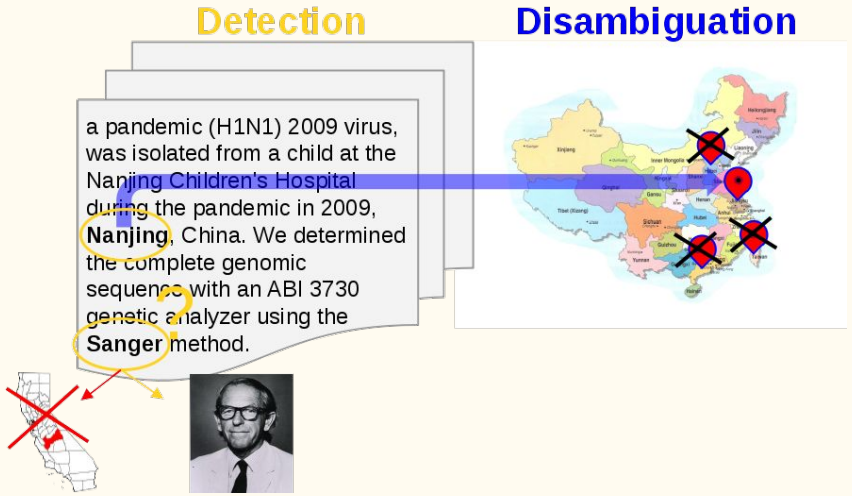
➢ Désambiguïsation dans 7 documents (textes pleins, format txt) issus d'ISTEX :
- https://drive.google.com/open?id=1HwrHyG27-zPICuT174GOYX5Sa0l80usD



Evaluation

Campagne SemEval 2019 :

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

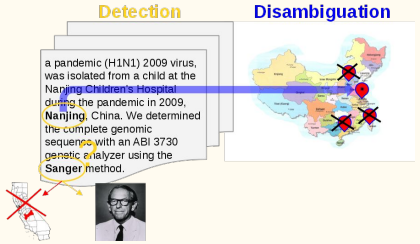


➢ Task 12 SemEval'19 :
Toponym Resolution in Scientific Papers
(https://competitions.codalab.org/competitions/19948)

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

Evaluation

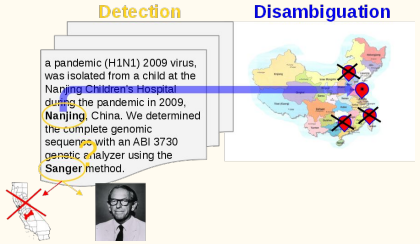
Campagne SemEval 2019 Pourquoi ?



- ➢ 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