Soutenance lecture d'article

Learning to Extract Local Events from the Web

ASQUIN Paul
AUBIN Victor
LAGATTU Mickaël

Motivations et méthode

Personnalisation
expérience utilisateur
lieu/préférence

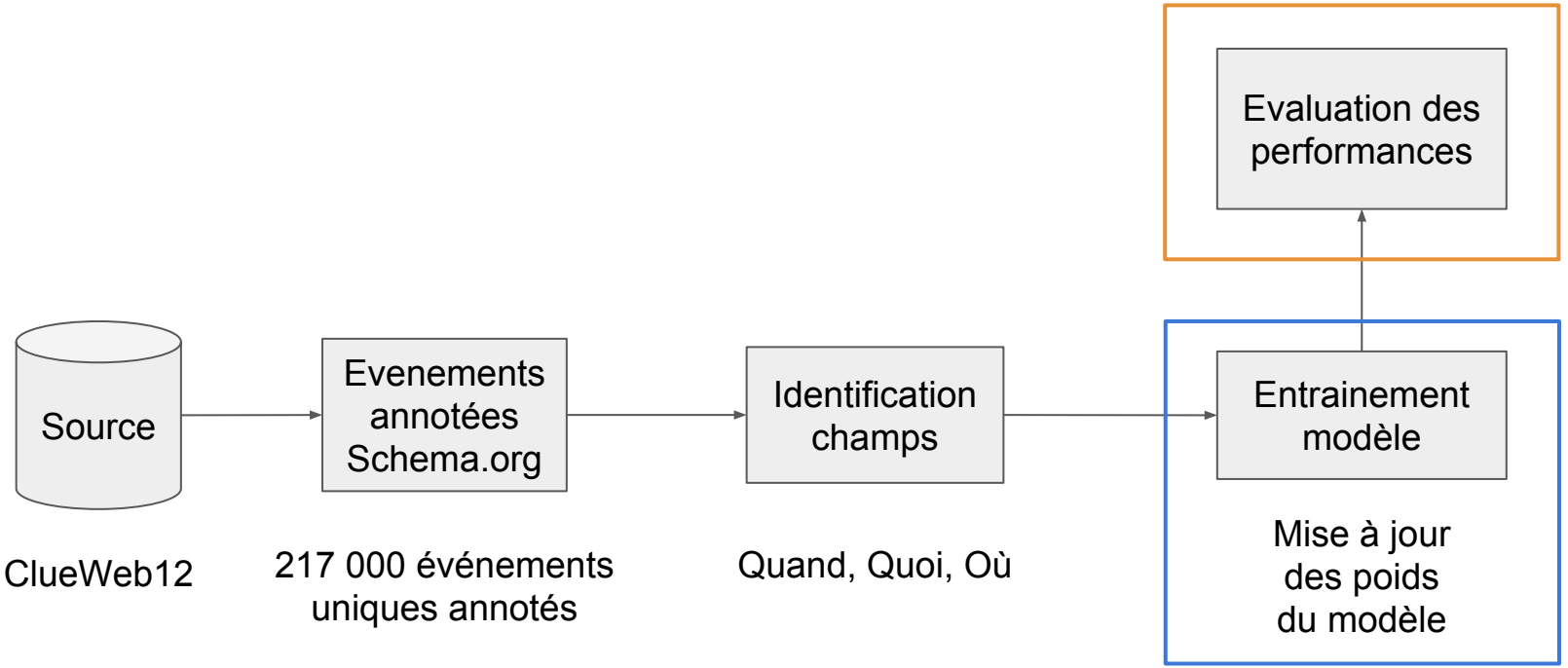
?

Analyse de pages
web non annotés
→Quoi, Quand, Où ?

!

Proposition d’
événements locaux

Modèle de classification d’événements



Schema.org

```
<div vocab="http://schema.org/" typeof="TouristAttraction">
        <h1><span property="name">Musée Marmottan Monet</span></h1>
        <div>
                <span property="description">It's a museum of Impressionism and fre
        </div>
        <div property="event" typeof="Event">It is hosting the
                <span property="about">Hodler</span>'s
                <span property="about">Monet</span>'s
                <span property="about">Munch</span>'s exibit:
                <span property="name">"Peindre l'impossible"</span>.
                <meta property="startDate" content="2016-10-01" />Start date: Septe
                <meta property="endDate" content="2017-02-05" />End date: Genuary 2
        </div>
</div>
```

Modèle d'extraction d'événements

De quoi s'agit ce champ dans cette région et dans ce document…?

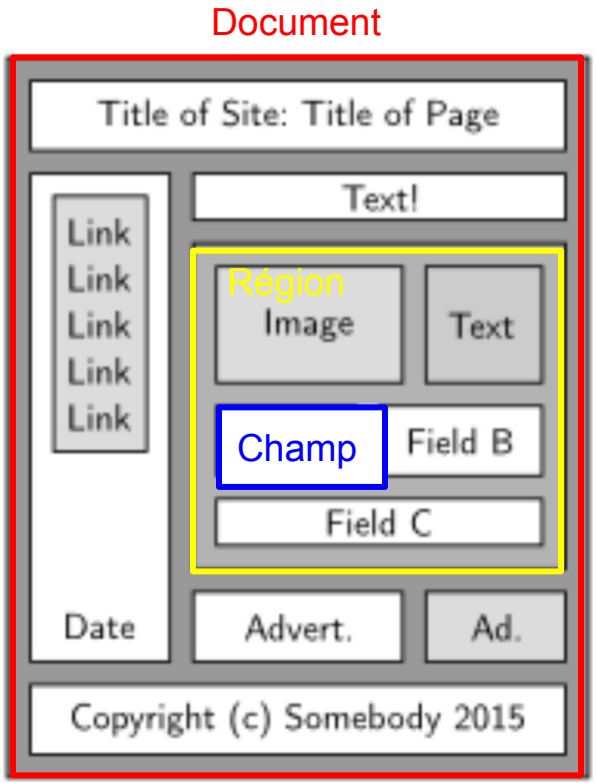
Score global : $\phi(\mathcal{F}, \mathcal{R}, \mathcal{D}) = \alpha(\mathcal{D})\beta(\mathcal{R})\gamma(\mathcal{F})$

Score de document : $\alpha(\mathcal{D})$

Score de la région : $\beta(\mathcal{R})$

Score du set de champs : $\gamma(\mathcal{F})$

Recherche du lieu de la date et du contenu d'un événement (Quand, Quoi, Où)



Score de document

Problème de classification => Naive-Bayes et Langage modeling framework (Ponte et Croft)

Définition du score de document : $$\alpha(\boldsymbol{\mathcal{D}}) = \begin{cases} 1 & logP(E|\boldsymbol{\mathcal{D}}) - logP(\overline{E}|\boldsymbol{\mathcal{D}}) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Décomposition des probabilité : $$P(E|D) = \prod_{w \in \boldsymbol{\mathcal{D}}} \lambda P(w \in E) + (1 - \lambda) P(w \in C)$$

Fréquence d'un mot dans le corpus: $$P(\boldsymbol{\mathcal{D}} \in X) = \prod_{w \in \boldsymbol{\mathcal{D}}} \frac{tf(w, X)}{tf(*, X)} \qquad \boldsymbol{X \in \{E, \overline{E}\}}$$

Score de région

Filtrage selon la taille de la région : $$\beta(\boldsymbol{\mathcal{R}}) = \begin{cases} 1 & |\boldsymbol{\mathcal{R}}| < \tau \\ 0 & \text{otherwise} \end{cases} \qquad \tau = 2^{12}$$

Tau est assigné de façon empirique

Les approches de probabilité de distribution étaient plus complexes et pas plus efficaces

Score de champ

Problème de classification avec 4 classes

$$\mathcal{K} = [\text{'}What\text{'}, \text{'}When\text{'}, \text{'}Where\text{'}, \text{`}Other\text{'}],$$

Fonction discriminantes

$$\delta_{\text{What}}(f) = \vec{W}^T_{\text{What}} \cdot \vec{X}_f$$
$$\delta_{\text{Where}}(f) = \text{matches}(f, \text{Address}) \cdot \vec{W}^T_{\text{Where}} \cdot \vec{X}_f$$
$$\delta_{\text{When}}(f) = \text{matches}(f, \text{Date/Time}) \cdot \vec{W}^T_{\text{When}} \cdot \vec{X}_f$$
$$\delta_{Other}(f) = \vec{W}_{Other} \cdot \vec{X}_f$$

Features Textuelles, NLP et structurelles

$$\vec{X}_f$$

Attribution de la classe

$$\text{PREDICTEDTYPE}(f) = \underset{k \in \boldsymbol{\mathcal{F}}^R}{\text{argmax}}\, \delta_k(f)$$

LIBLINEAR => apprend les poids

$$\vec{W}$$

Score final (nul si toutes les classes ne sont pas représentées)

$$\gamma_S(\boldsymbol{\mathcal{F}}) = \prod_{f \in \boldsymbol{\mathcal{F}}} \max_{k \in \boldsymbol{\mathcal{F}}^R} \delta_k(f)$$

Evaluation

Besoin de générer des événements factices pour éviter les biais

Evaluation de la précision et du rappel

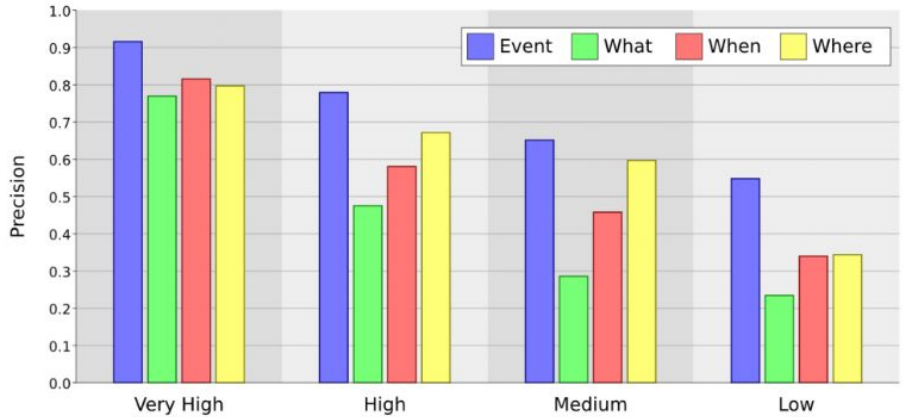


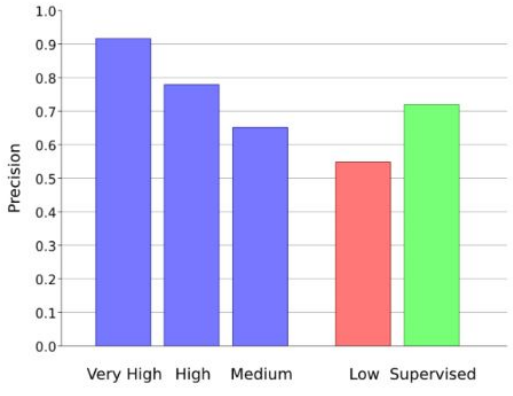
Figure 6: Precision Evaluated at Recall Levels

Conséquences

Rappel doublé pour 85% de précision, quadruplé pour 65% de précision

Il reste beaucoup de faux positifs quand on augmente le rappel

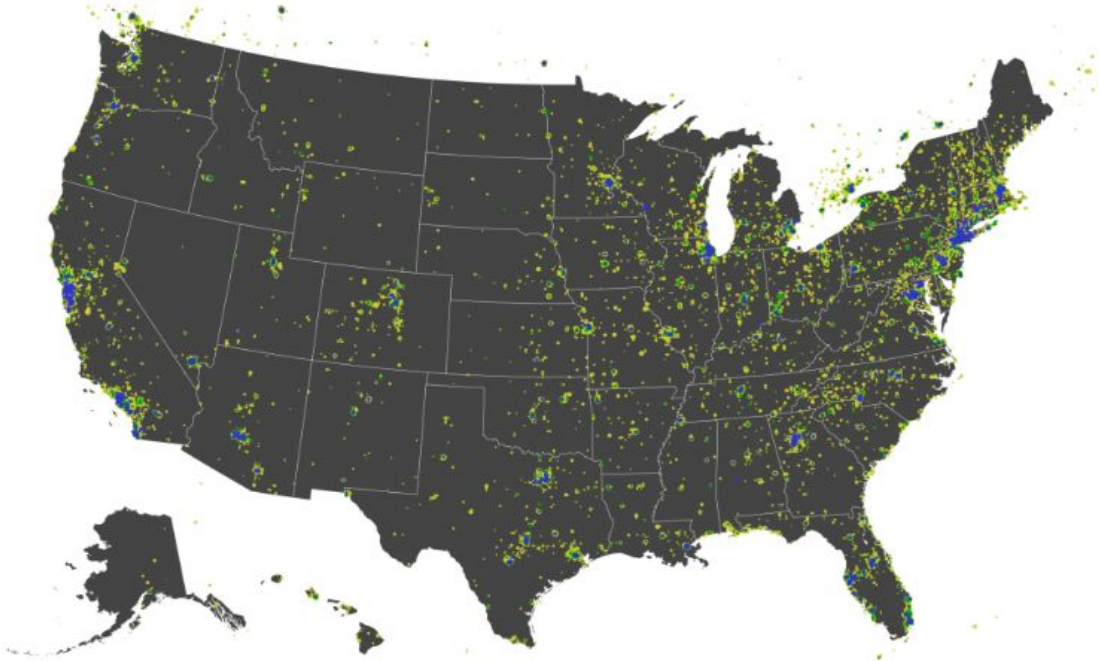
➢ Supervision par des humains : 30% de précision en plus



amazon beta mechanical turk

Etude géographique

Dataset fortement biaisé : concentration des événements aux Etats-Unis

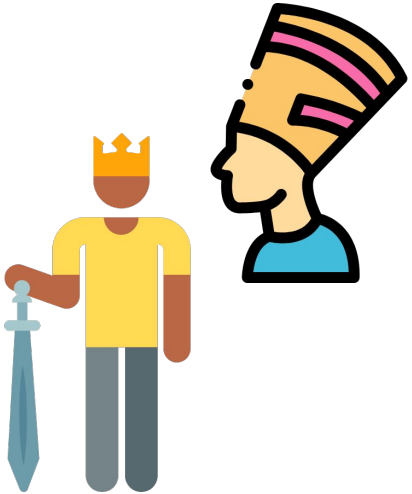


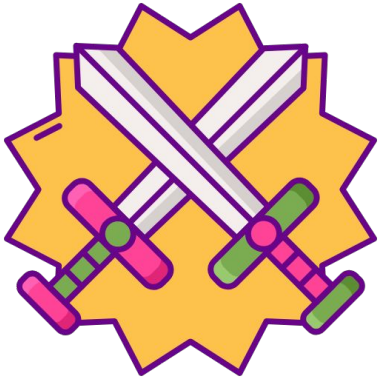
Analyse critique

Ne fonctionne qu’en Anglais

- Ne prouve pas que le Bootstrap sampling va réduire le biais
- Catégories de précision et de rappel qui peuvent être choisies de manière avantageuse pour présenter leurs résultats

Travaux similaires - Recherches historiques





Comprendre l’histoire malgré des données incomplètes

Projet d’ouverture

Mise en relation par centres d’intérêts sur des événements locaux
Qui, Quoi, Quand, Où ?...

Éléments à la volée

- Il existe un extracteur d’événement assez pratique : https://schema.org/Event

Méthode

- Procédé proposé :
  → ciblage par Nom event, date, heure, lieu.
  → groupages des événements.
- Évaluation de la méthode avec extractions depuis dataset
- Comment donner du poids à une information
  → Document scoring : probabilité d'une page de parler d'un événement
  → Region scoring : probabilité que la région du document ait l'information
- Problèmes d'avoir plusieurs événements, dates et lieux nommés sur une même page, ou des faux positifs (date de copyright du site par ex)
  → Si une date revient trop de fois, certainement faux positif, si une date dans le footer, certainement faux positif

Éléments à la volée

- RELATED WORK
  Analyse d'événements / lieux
  → Analyser les événements d'Histoire et détecter de mauvaises affectations lieux/dates.
  → Proposer des événements à des utilisateurs connus et leur demande si cela les intéresse
  → Détection d'événements dans twitter : catastrophe naturelle, attentat etc.
  → Analyse des unes de journaux
  + Beaucoup de travaux sur l'extraction de données