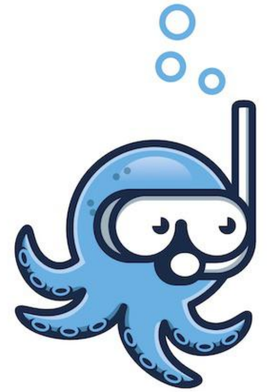
OPENCLASSROOMS

VEILLE THEMATIQUE : SNORKEL

Projet 8



snorkel

Azim Makboulhoussen
10 Août 2018

Sommaire

- Introduction
- Présentation de la solution Snorkel
- Les étapes de mise en œuvre de Snorkel
- Application sur un prototype
- Evaluation des résultats
- Conclusion

Introduction

Objectif du projet

- Monter en compétence sur une nouvelle thématique dans le domaine du machine learning
- S’appuyer sur une recherche récente dans le choix de sa thématique
- Prototypage de la solution sur un cas concrêt

Présentation de Snorkel

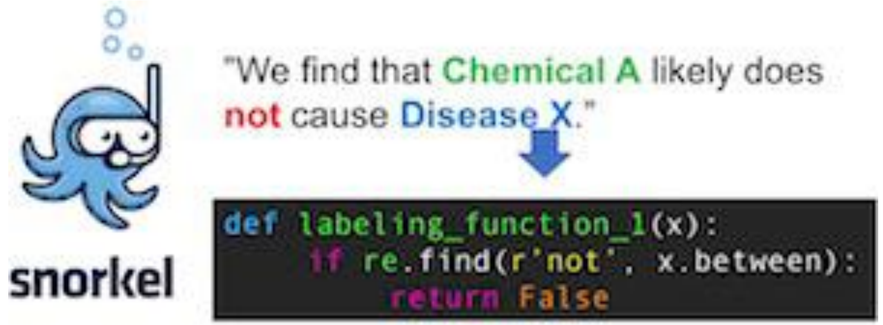
https://hazyresearch.github.io/snorkel/

La problématique

- ✓ Fort succès du Machine Learning et Deep Learning
- ✓ Disponibilité de librairies ➔ facilite l'étape Apprentissage
- ✓ Automatisation des Features Extraction avec le Deep Learning

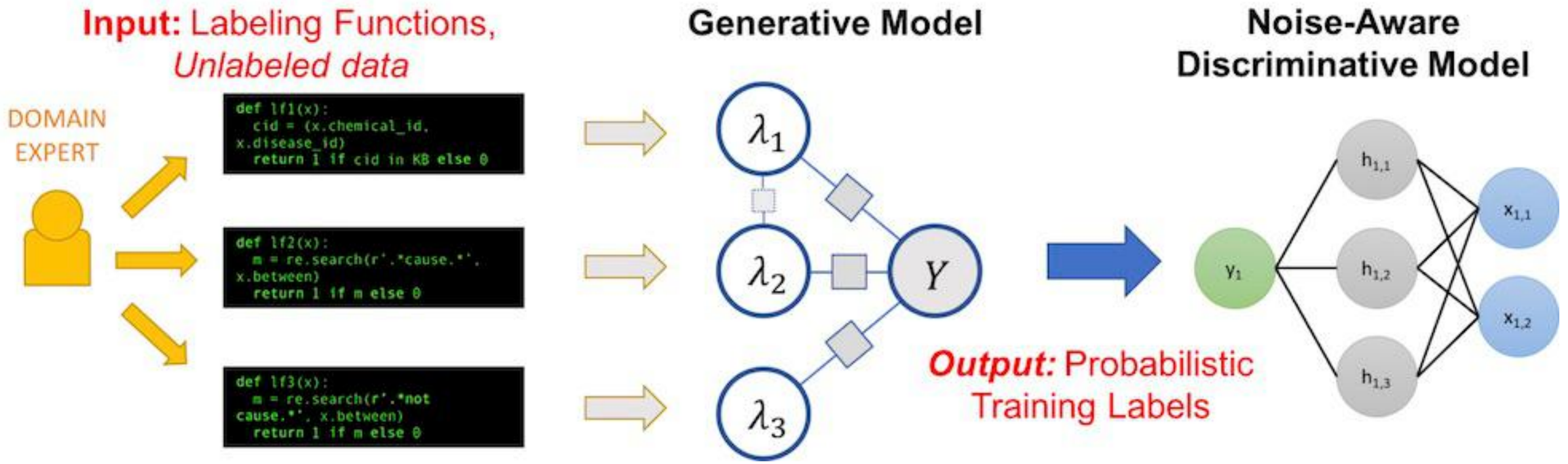
- ❑ Important besoin de données d'entrainement
- ❑ Beaucoup de données non lisibles directement par les machines
- ❑ Intervention humaine et expertise nécessaires pour les annoter

Une solution : Snorkel



- Comment avoir plus de données d'entrainement et sans complexité ?
- Solution développée par Stanford
- Présentée à la prestigieuse conférence VLDB en 2018
- Plus besoin d'annotation manuelles mais sur le Data Programming
- Utilisateur programme des fonctions de labélisation à la place
- Utilisé dans de nombreux projets de recherche scientifiques mais également par des entreprises comme la Nasa, le groupe Alibaba ou Intel.

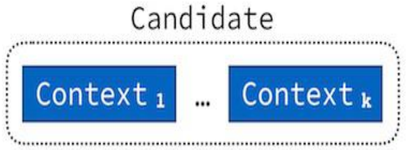
Le workflow Snorkel

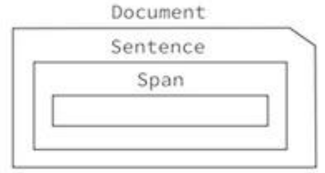


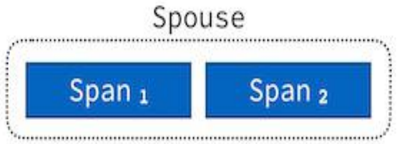
Etapes d’un projet Snorkel

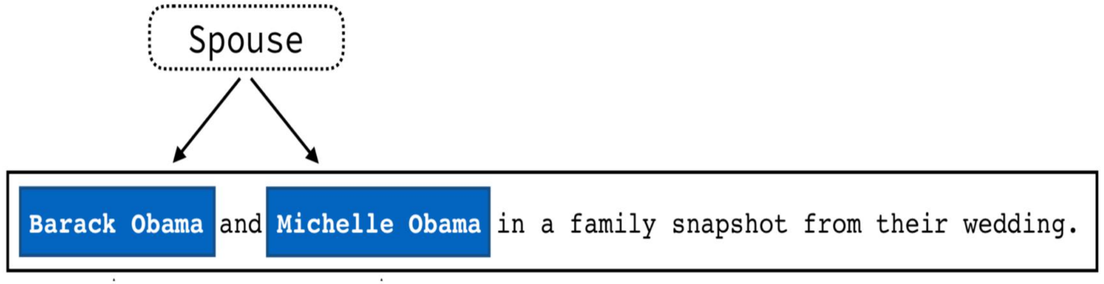
Définition et extraction des candidats

- Déterminer l'information que l'on veut extraire et quel sera son utilité
- Définir le schéma : définition formelle de ce que l'on veut extraire
- Implémenter le schéma à l'aide de l'objet **Candidate** et **Contexte**
- Utiliser les fonctions de Snorkel pour extraire les candidats dans l'ensemble de notre corpus d'entrainement et les stocker en base de données









Ecriture des fonctions de labélisation (LFs)

- Ce sont simplement des fonctions écrites par l'utilisateur qui implémentent des règles métiers
- Elles seront appliqués à l'ensemble des candidats de nos données d'entrainement et vont servir à les annoter (labéliser)
- Elles doivent prédire à la fois les labels positifs et négatifs
- Elles peuvent être de 2 types : Pattern based LF ou Distant Supervision LF

```
def LF_same_last_name(c):
    """
    Label as positive if both
    """
    p1_last_name = last_name(c.person1.get_span())
    p2_last_name = last_name(c.person2.get_span())
    if p1_last_name and p2_last_name and p1_last_name == p2_last_name:
        if c.person1.get_span() != c.person2.get_span():
            return 1
    return 0
```

```
def LF_dating(c):
    dating = {'boyfriend', 'girlfriend'}
    return -1 if len(dating.intersection(get_between_tokens(c))) > 0 else 0
```

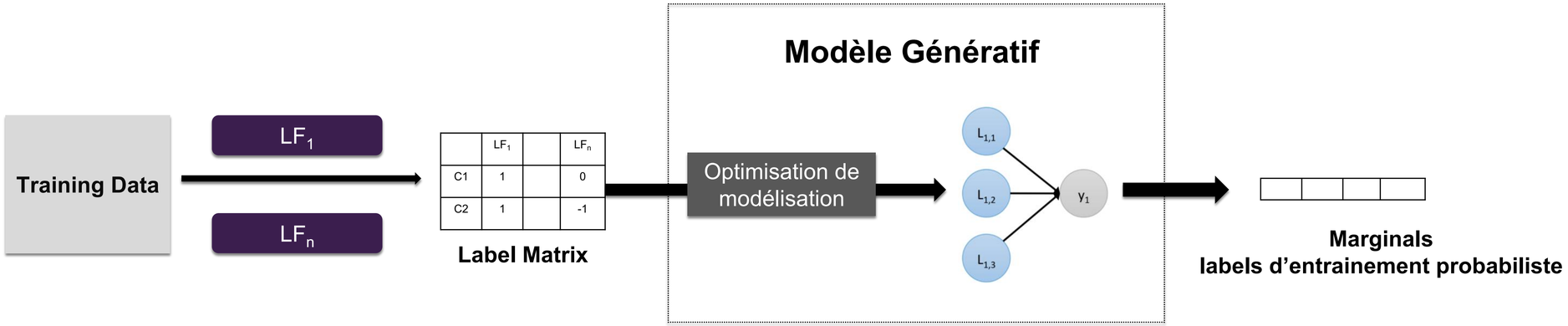
```
def known_spouse(x):
  pair = (x.person1_id, x.person2_id)
  return 1 if pair in KB else 0
```

Former U.S. president **Barack Obama** and
first lady **Michelle Obama** arrive to talk …

Knowledge Base (KB)
CONTAINS (A B)

→ **Label** = True

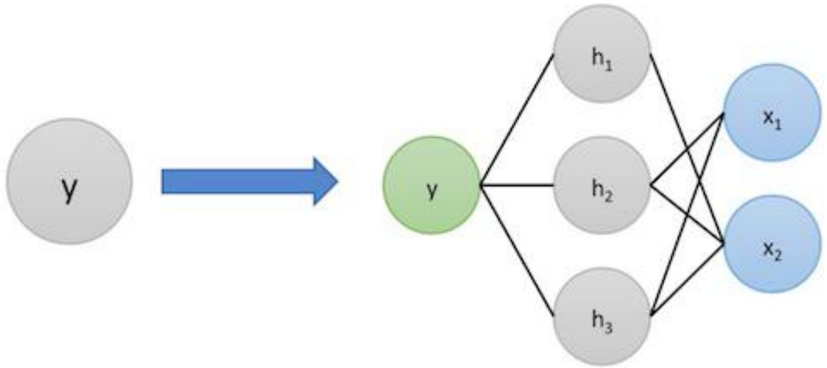
Apprentissage avec un modèle génératif



- Application des fonctions de labélisation sur les données d'entrainement
- Le modèle génératif va apprendre les précisions des LFs et les uniformiser
- Exploitation des chevauchements et conflits entre LFs pour apprendre et supprimer le bruit
- Le modèle génératif va donner la distribution des marginales

Prédiction avec un modèle discriminatif

- Les prédictions marginales du modèle précédents sont utilisés comme labels d'entrainement d'un modèle discriminatif
- Snorkel dispose de liens vers les modèles TensorFlow et implémente directement les modèles LSTM et de Régression logistiques.
- LSTM (Long Short-Term Memory) est un type de réseau de neurone récurrent très efficace notamment pour la classification textuelle.
- Ce dernier modèle va faire notre prédiction finale



Application sur un prototype

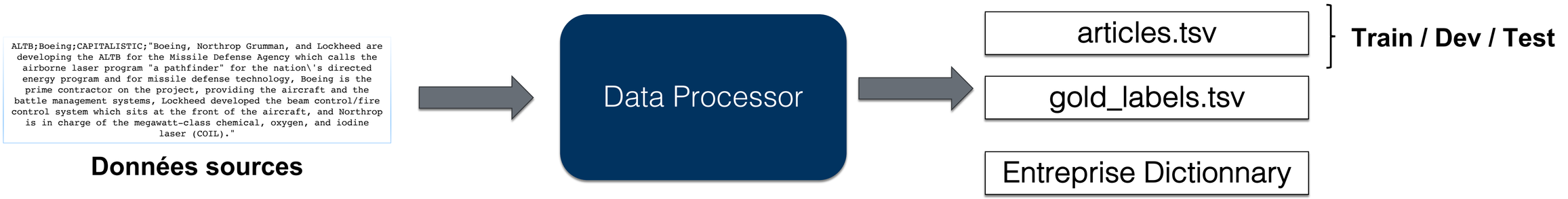
Notre mise en application

- Utilisation de Snorkel pour extraire des partenariat entre acteurs économiques
- Sources de données : 645 articles de presse concernant plus de 400 entreprises.
- Données annotés fournies par la société GeoTrend (Business Discovery)
- GeoTrend est intéressé par Snorkel pour sa solution
- Le format du fichier de données :

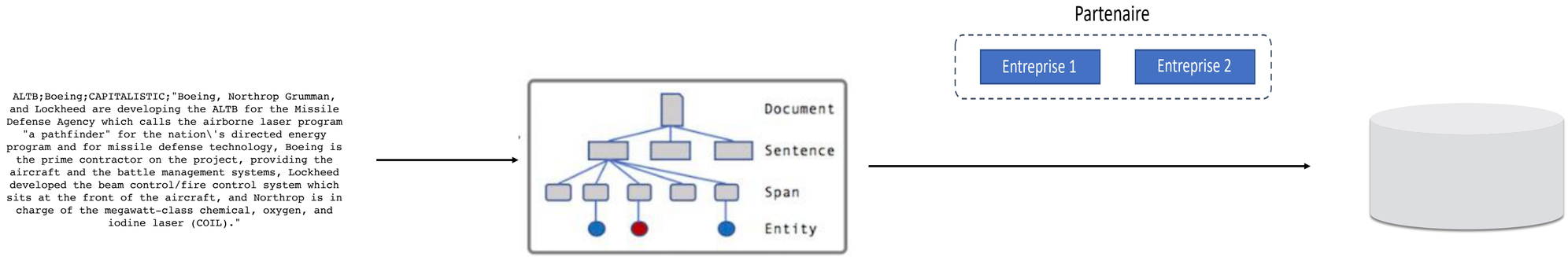
```
actor1; actor2; type; sentence
actor2; actor1; type; sentence
…
```

Préparation des données

- Préparation des données pour être utilisable dans Snorkel
- Utilisation des données annotés pour notre jeu de développement et test
- Mise en forme des données pour être compatible avec Snorkel
- C’est l’étape qui nous a pris le plus de temps.



Extraction des candidats



- Définition du schéma de nos candidats
- Utilisation d'un dictionnaire pour identifier les entreprise (abandon de l'OrganisationMatcher de Spacy)
- Découpage des données en jeu d'entrainement (80%), développement (10%) et test (10%)
- Chargement des articles et des données annotés (appelés gold labels) dans la base de données Snorkel

Ecriture des Labeling Functions (LFs)

- Utilisation de SentenceNGramViewer pour aide à l'écriture des LFs
- Une douzaine de fonctions implémentées
- Utilisation des regexps et de helper fonctions de Snorkel
- Les métriques précision / recall / F1 score ont été utilisés pour optimiser les LFs
- LFs pour prédictions positives et négatives

{Negative, Abstain, Positive}

```
def LF_clients(c):
    clients = { 'by', 'from', 'include'}
    return -1 if len(clients.intersection(get_left_tokens(c[1], window=3))) > 0 else 0


labeled = coverage(session, LF_clients, split=0)
tp, fp, tn, fn = error_analysis(session, LF_clients, split=1, gold=L_gold_dev)
```

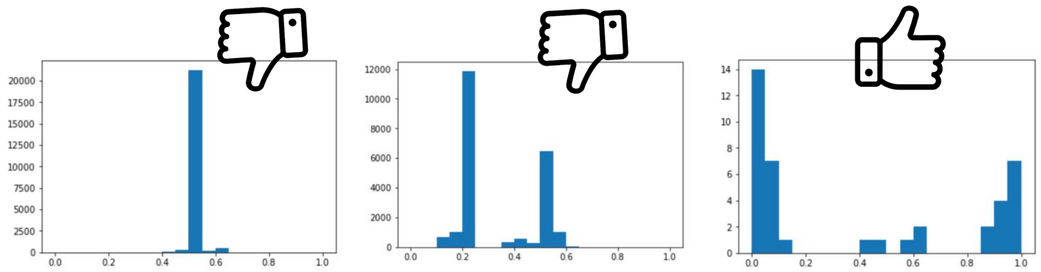
```
def LF_with_relations(c):
    return 1 if re.search(r'{{A}}.+with.{0,40}{{B}}',get_tagged_text(c) , re.I) else 0
```

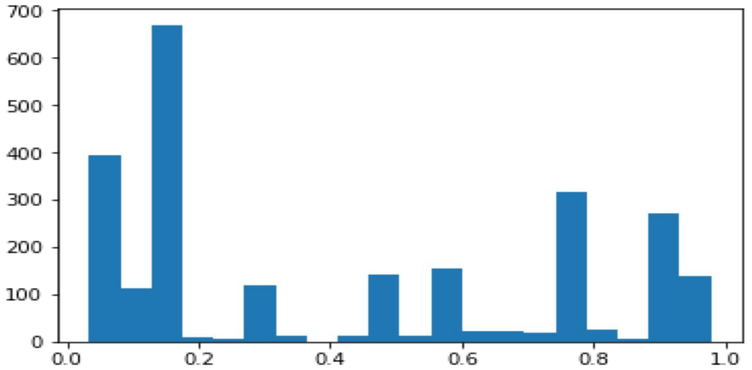
```
against = ['alternative(s)?', 'vs', 'rival(s)?', 'versus', 'surpass(es|ed)?' , 'competitor(s)?', 'leaders ']
def LF_against(c):
    return -1 if re.search(r'{{A}}.{0,50} ' + ltp(against) + '.{0,50}{{B}}',get_tagged_text(c) , re.I) else 0
```

Métriques évaluation des LFs

- Jeu de développement pour évaluer la qualité des LFs
- Amélioration de la fonction en se basant sur les nombres TP, FP, TN, FN
- Utilisation des métriques **précision** / **rappel** et **F1 score** optimiser chaque LF
- Pour évaluer l'application de l'ensemble des LFs :
  - **Accuracy** : pourcentage de candidats que la fonction labellise correctement.
  - **Coverage** : pourcentage de candidats labélisés par une ou plusieurs LFs.
  - **Conflit** : pourcentage de candidats avec des résultats en désaccord entre LFs

Entrainement des modèles





- Application du modèle génératifs sur l'ensemble des LFs écrites
- Utilisation de la distribution des marginales pour évaluer la qualité de nos LFs
- Evaluation du modèle à l'aide du jeu de développement
- Utilisation d'un réseau LSTM pour le modèle discriminatif
- Multiple itérations du pipeline Snorkel pour avoir améliorer les résultats
- Une recherche sur grille aurait pu être utiliser pour optimiser les hyper-paramètres

Les résultats

Scores obtenus

Jeu de développement :

	Précision	Rappel	F1-Score
Modèle génératif	37,9 %	58,1 %	45,9 %
LSTM	39,7%	62,8%	48,6%

Jeu de test :

	Précision	Rappel	F1-Score
Modèle génératif	44,8 %	61,5 %	51,9 %
LSTM	38,6 %	83,5%	52,8 %

Conclusion

Conclusion

- Utilisation d'une solution issue des dernières recherches de Stanford
- Montée en compétence dans le domaine du Data Programing non vu en cours
- Quelques difficultés pour la prise en main au début
- Pas de baseline pour comparaison mais gain de temps indéniable
- Amélioration possible des résultats par :
  - Augmentation de la volumétrie des données,
  - Travail sur les fonctions de labélisation
  - Optimisation des algorithmes
  - Tester la prédiction multi-classes
- Scores obtenus par équipes Snorkel meilleurs que ceux avec méthodes classiques
- Recherche et amélioration en cours de Snorkel

Merci à mon mentor Amine Abdaoui pour ses conseils, son expertise et sa capacité à toujours guider vers les meilleures solutions