ALGO DE CO APPRENTISSAGE

=> Algo itératif

=> classification semi supervisée

<https://aclanthology.org/2009.jeptalnrecital-recital.1.pdf>

section 3 : **co-training** => pour des données déjà étiquetées et beaucoup non étiquetées. => semi supervisée

L’algorithme de Blum et Mitchell

“Le co-training consiste à entraîner plusieurs classifieurs, chacun basé sur une vue du corpus, puis à les améliorer entre eux à l’aide d’une masse importante de données non annotées ; les classifieurs plus à même de classer un exemple donné jouant le rôle de "professeurs" pour les autres.”

(corpus = données organisées en classes)

“Le pool d’exemples non annotés (U) est séparé en autant de partitions (Ui) que l’on souhaite d’itérations de co-training, et à chaque itération i, on soumet la partition Ui aux deux classifieurs. Pour chaque exemple, si un classifieur annonce un score de confiance supérieur à un certain seuil, il y apposera son label et ajoutera cet exemple au corpus d’apprentissage de l’autre classifieur.”

Ils divisent leur corpus en 2; selon les différentes classes (articles appartenant aux classes sport, télé .. d’un côté et international, société de l’autre)

dans l’article: travaillent avec 2 classifieurs : icsiboost et LIBLINEAR

font en premier leur apprentissage sur le même corpus puis auront leur propre corpus constitué par l’autre classifieur.

pour icsiboost : 1000 itérations

<https://www.mdpi.com/2073-8994/12/1/8/htm>

=> Stratégie de sélection d'échantillons pour co-training

sélectionner des échantillons ayant un niveau de confiance élevé, puis de les ajouter au tour de formation suivant:

- probabilité a posteriori du classificateur comme métrique de confiance pour l'étiquetage

- mesuré la confiance de l'étiquetage en utilisant la validation croisée à dix reprises sur l'ensemble étiqueté original

ce qu’ils font:

stratégie de double vérification pour la sélection d'échantillons, cad :

1. Un échantillon avec la même étiquette prédite par les deux classificateurs est ajouté au premier ensemble de candidats
2. Un échantillon avec une grande similarité avec l'ensemble d'entraînement dans une classe est ajouté au deuxième ensemble de candidats. Les échantillons finaux sont déterminés en fonction de l'intersection des deux ensembles de candidats.

algo bénéfique que si les ensembles de données sont indépendants, (ie si l'un des classificateurs étiquette correctement un point de données que l'autre classificateur a mal classé auparavant.) Si les classificateurs sont d'accord sur toutes les données non étiquetées, (donc dépendants), l'étiquetage des données ne crée pas de nouvelles informations => mauvais résultats.

<https://link.springer.com/chapter/10.1007/978-3-540-74958-5_42>

faut 2 vues des données (2 façons différentes de décrire nos données; ie selon différentes caractéristiques)

Il faut aussi 2 classificateurs et chacun va étiqueter des données non étiquetées pour l’autre.

après un certain nombre d’itération, on ne peut plus améliorer => s’arrêter

Combien d’itération?

<https://www.researchgate.net/publication/350194568_A_review_of_research_on_co-training>