Comparaison de deux courbes polygonales dans l'espace des embeddings FastText pour l'étude de la relation de traduction entre deux phrases

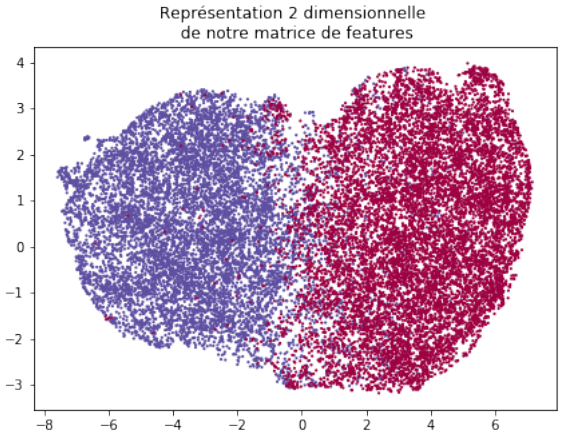
Adel Nabli

Résumé

Suite au TD précédent, nous avons avons utilisé la même méthode et avons obtenus une accuracy en légère baisse sur ce nouveau dataset : 95 % contre 97 % précédemment.

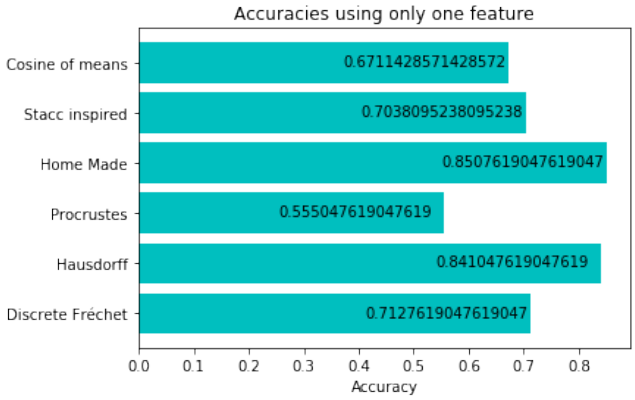
1 Résultats :

Notre méthode demandant le calcul de 6 distances pour chaque couple de phrase, il est long de traiter de gros datasets. Pour évaluer notre méthode, nous nous sommes donc restreints a l'étude de 21 000 phrases extraites du *test set*, 10 000 phrases en relation de traduction et 11 000 qui ne le sont pas *(nous avons extraits les lignes 0 :5000, 330270 :335270, 380000 :385000 et 727787 :733787 du test set)*. Le calcul de nos features a pris 35 minutes sur ce dataset.



On remarque de nouveau que les distances calculées arrivent tres bien a séparer notre dataset en 2 classes *(en utilisant UMAP pour le calcul de la représentation)*.

En utilisant qu'une seule distance la fois et la moitié du dataset extrait pour entrainer notre modele, on arrive aux résultats suivants avec un random forest :

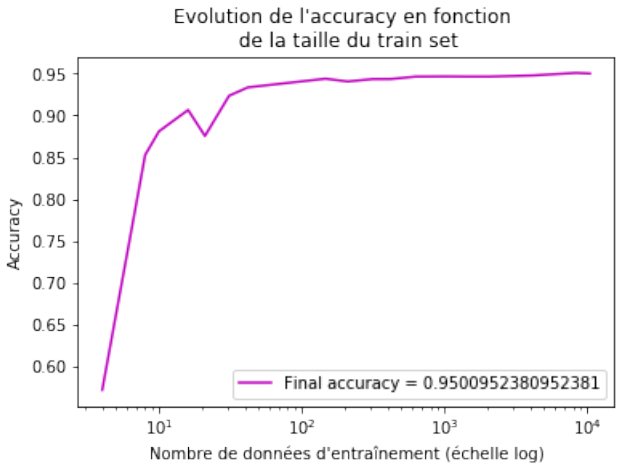


On remarque donc de nouveau que les deux distances générant les meilleurs résultats sont la distance "créé" et la distance de Hausdorff, avec chacune une accuracy de 85%.

En utilisant toutes les features, et encore une fois 50% du dataset extrait pour entrainement du random forest, on arrive a une accuracy de 95%, une précision de 96%, un recall de 94% et un F1-score de 95%. L'importance relative des features est donné dans le tableau suivant :

	Relative Importance of the features
Home Made	40.8 %
Hausdorff	31.1%
Discrete Fréchet	11.5%
Cosine of means	7.0 %
Stacc inspired	6.1 %
Procrustes	3.5 %

On remarque de nouveau que peut de données suffisent pour bien séparer le dataset, une centaine d'éxemples d'entrainements suffisant pour atteindre les 95% d'accuracy :



2 Discussion

Meme si un peu plus de données d'entrainement (100 contre 20) que sur l'ancien dataset sont nécessaires pour obtenir une accuracy similaire, finalement les résultats sur ces deux datasets sont comparables et certains commentaires faits précédemment tiennent encore. En effet, en inspectant les phrases mal classifiées, on se rend compte qu'en utilisant les données numériques présentes dans certains textes, on aurait pu améliorer la qualité de prédiction. On remarque cependant que nous avons perdu ici un biais qui était présent sur l'ancien dataset : 2 phrases longues ne sont plus systématiquement classifiées comme étant en traduction l'une de l'autre.

Toute choses prises en compte, on peut donc conclure que notre méthode semble etre un moyen efficace et ne nécessitant pas beaucoup de donnés d'entrainement en plus des embeddings utilisés en entré pour évaluer la relation de traduction entre 2 phrases anglais-francais.