Projet Industriel

Amadou-Bély GUISSE, Thomas HALOUIS, Chaïmae LAHRAIRI
Tristan MICHEL, Rémy PERRODIN, Aissam RABHI

Polytech Sorbonne

Présentation du projet

- **Capgemini** est une entreprise de services du numérique française. Il s'agit du cabinet de conseil qui a le plus gros chiffre d'affaires du pays et elle figure parmi les dix plus grosses du secteur mondialement.
- L'objectif du projet **ANBLIC** est d'industrialiser pour la première fois des technologies d'amélioration de la confidentialité qui sont à la limite de la théorie et de la pratique. En effet, ces technologies s'appuient sur des techniques cryptographiques telles que le chiffrement entièrement homomorphe et l'écrasement fonctionnel.

Classifications non-supervisées

Pour regrouper les textes du corpus en blocs similaires, nous utilisons des méthodes dites de classifications non-supervisées ou clustering. Mesurer la similitude de textes vectorisés est l'opération qui permet de les regrouper de manière pertinente. Nous avons testé 3 méthodes de classifications différentes et essayer de déterminer laquelle était la plus pertinente par rapport à notre cas d'usage.

- K-moyennes
  C'est la méthode de classification la plus connue et la plus simple. Après avoir initialisé ses centroïdes en prenant des données au hasard dans le jeu de données, l'algorithme des K-moyennes alterne plusieurs fois les deux étapes suivantes pour optimiser les centroïdes et leurs groupes :
  - 1. Regrouper chaque objet autour du centroïde le plus proche.
  - 2. Replacer chaque centroïde selon la moyenne des éléments de son groupe.
- Classification hiérarchique
  Nous avons utilisé la classification hiérarchique agglomérative. C'est à dire qu'au départ, nous considérons que tous les textes appartiennent à des groupes différents puis à chaque itération l'algorithme va rassembler deux groupes similaires jusqu'à ce qu'il ne reste qu'un seul groupe. Cette approche différente des K-moyennes nous a permis de créer des groupes de textes équilibrés en terme de taille.
- Classification à niveaux multiples
  Nous avons repris les deux classifications précédentes afin de mettre en place la classification en arbre. Cette méthode consiste à appliquer une première classification sur le corpus entier afin de créer des groupes de textes puis à effectuer de nouvelles classifications sur chacun des groupes. Il est possible de rajouter encore un niveau en reproduisant l'étape précédente sur les nouveaux groupes.

Optimisation des classifications

Pour accélérer notre algorithme de recherche, nous avons utilisé 2 méthodes:

- La première pour accélérer la recherche du groupe le plus proche de notre requête est la définition de centroïdes fictifs.
- La seconde permet d'accélérer la recherche des textes au sein du groupe. Pour cela, on utilise la méthode des K plus proches voisins qui nous renvoie les textes les plus proches de la requête.

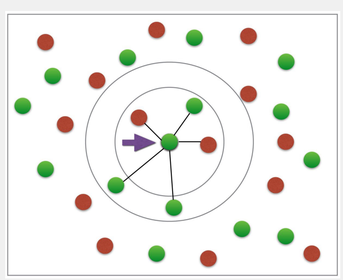


Schéma K plus proches voisins

Présentation de notre jeu de données

Pour effectuer notre projet nous nous sommes appuyés sur une base de données du *CDC*. Ce corpus de textes regroupe plusieurs *dizaine de milliers de textes* autour du Covid-19. La première étape du projet a été de **traiter** ces textes pour permettre aux ordinateurs de les utiliser.

On parle de **vectorisation** des textes. Vectoriser un texte s'effectue en 3 étapes: la tokenisation, la lemmatisation et le calcul des poids.

Regardons comment cela se traduit sur le texte T_1 inclus dans le corpus $C = \{T_1, T_2\}$

T_1 = *"This3 is An ExampL31, 0.1 and 0.2 are two floats? I am a stopword but mathematics is not a stopword. I lovingly love mathematic. mathematician@test.fr"*

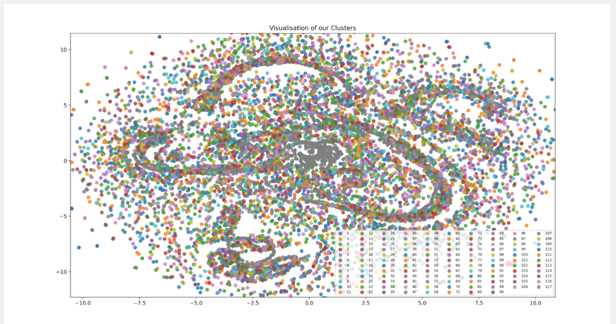
T_2 = *This is the second text of the mathematic corpus.*

- Tokenisation: **Suppression des mails, chiffres, caractères et mots inutiles**
  $T_{1_{tokenize}}$ = *['example', 'two', 'floats', 'stopword', 'mathematics', 'stopword', 'lovingly', 'love', 'mathematic']*
- Lemmatisation: **Racinisation des mots**
  $T_{1_{lemmatise}}$ = *['exampl', 'two', 'float', 'stopword', 'mathemat', 'stopword', 'love', 'love', 'mathemat']*
- Calcul des poids: **Méthode Tf-Idf**
  $T_{1_{vectorise}} = [0.34, 0.24, 0.34, 0.34, 0.68, 0.34]$

Une fois la transformation de T_1 terminée, il est possible de manipuler ce texte comme un vecteur classique et donc utiliser l'ensemble des opérations vectorielles usuelles. Cela nous permet de comparer les textes entre eux et ainsi effectuer de la classification non-supervisée entre autres choses.

Visualisation T-SNE

Nos classifications s'effectuant sur des jeux de données de taille importante, il a été nécessaire de mettre en place une méthode visuelle pour *regarder* et *vérifier* nos regroupements. T-SNE est une méthode de réduction de dimensions permettant de visualiser des données de taille trop imposante.

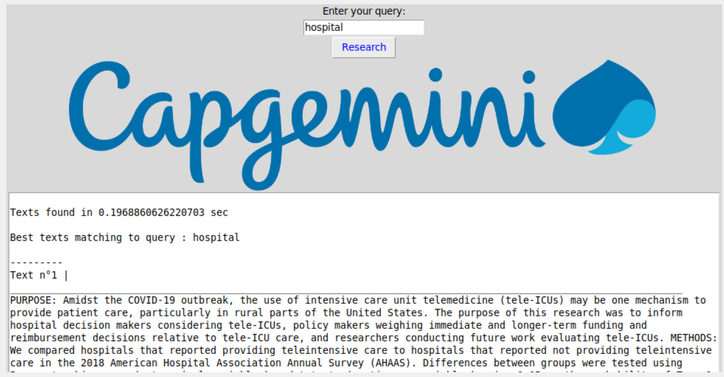


Visualisation d'une classification non-supervisée hiérarchique

Ci-dessus, nous observons la visualisation T-SNE d'un clustering hiérarchique de 14.000 textes. Chaque texte est défini par un point et une couleur. La couleur d'un point défini l'appartenance à un regroupement. Ce type de visualisation sert à vérifier si les textes sont bien répartis et que certains clusters n'ont pas une taille disproportionnée.

Interface graphique

Le moteur de recherche renvoie le texte le plus proche de la requête entrée par l'utilisateur.



Résultats des tests d'exécutions

Tableau comparatif des méthodes de classification

Mesures pour 14000 textes	Temps moyen pour trouver le meilleur score	Temps moyen pour trouver le meilleur groupe	Score moyen du meilleur texte	Score moyen des 5 meilleurs textes
K-Mean	0.11 s	0.0023 s	0.43	0.32
HAC	0.04s	0.0024s	0.39	0.32
Double-HAC	0.01s	0.0024s	0.30	0.29
Meilleure méthode	**Double HAC**	**Equivalents**	**K-Means**	**K-Means et HAC**

- L'équilibre temps d'exécution/score des textes pertinents ainsi que la faible variance de la taille des groupes de la classification hiérarchique nous a convaincu de l'efficacité de ce dernier.

Remerciements

Xavier TANNIER
Tuteur académique

Yassine ABBAR
Ingénieur

Zied BECHA
Ingénieur

Bernard DELMAS
Responsable relations

Mallick SAYANTA
Ingénieur R&D