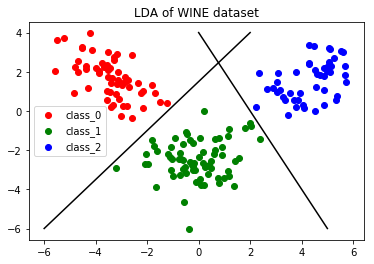
LDA « est un modèle génératif probabiliste permettant d'expliquer des ensembles d'observations, par le moyen de groupes non observés, eux-mêmes définis par des similarités de données. »

Critère à rentrer en paramètres :  Nombre de sujets que l’on veut extraire du sujet.

Le but du LDA est de réussir à extraire un nombre de thématiques traitées dans un document et de mettre en avant les mots les plus présents qui ont permis de déduire ce sujet.

  
Exemple de séparation des données en utilisant du LDA

Objectif de base dans notre cas : Mettre un grand nombre de données pour plusieurs journées (post, tweets,...). Noter pour chaque jour les thèmes principaux et l’évolution du bitcoin (positive ou négative). Ensuite, rechercher pour un nouveau jour, les thèmes qui ressortent et prédire de ce fait le cours du Bitcoin en fonction de l’historique en utilisant un modèle de prédiction (type Random Forest, Xgboost, Keras,…).

Résultat obtenu :

Après avoir travaillé sur le LDA, nous avons découvert les limites du LDA pour notre cas. D’abord, le LDA est très dur à paramétrer et il affiche assez rapidement, s’il n’est pas bien paramétré, pleins de fois les mêmes mots.

Cependant, même si le LDA est bien paramétré, comme c’est un algorithme de synthèse des sujets d’un corpus de texte, l’utilisation du LDA diminue grandement la quantité de données. De ce fait, nous nous sommes rendu compte que les résultats étaient beaucoup moins bons en utilisant le LDA car l’algorithme avait du mal à faire le lien entre les jours. Dernière chose, le LDA a comme but de synthétiser en quelques sujets, la pensée des utilisateurs. Malheureusement, la pensée des utilisateurs, parlant sur les subreddit autour du bitcoin, n’est pas facile à résumer car le subreddit combine du sarcasme, des blagues, des informations sérieuses,… Donc, le LDA avait du mal à sortir des sujets cohérents en associant la totalité de ces posts.