Compte rendu Text-Mining

Dans cette partie, on commence par récupérer les données de l'api Foursquare que l'on stocke dans des data frames. Nous avons dans ces données, les données locales et les données des autres lieux aux alentours qui vont nous servir de comparaison par la suite.

Une fois les données récupérées, on duplique les lignes grâce à la colonne priorité de la data frame ce qui va donner plus de poids aux mots des premières recherches. En dupliquant les lignes, on obtient une liste de mots que l'on va ensuite nettoyer, c'est-à-dire enlever la ponctuation, les caractères spéciaux, les mots communs...

Ensuite on récupère les mots les plus fréquents en distinguant les mots locaux et les mots des autres lieux. C'est donc ici que l'on va confronter et comparer les mots locaux avec les mots des autres lieux. Pour faire cela, on a utilisé la méthode tf_idf du module nltk. Cette méthode permet de calculer le score de proximité de chaque mot local avec les mots des autres lieux. C'est à dire que si le score du mot local est élevé, alors ce mot est proche des autres mots, il a une certaine proximité, une ressemblance avec les mots des autres lieux. Et si le score du mot local est faible, alors ce mot n'a aucune proximité et ressemblance avec les mots des autres lieux. C'est donc un mot unique et spécifique à mon lieu. Après avoir confronté les mots, on récupère les mots qui ont le score de proximité les plus faibles.

Ces mots sont très utiles car ils sont uniques au lieu et ils vont donc nous servir de requête pour récupérer des pages web qui correspondent au lieu. On utilise ces mots avec l'API de moteur de recherche qwant en spécifiant que l'on veut juste récupérer les url. Une fois les sites web récupérés, le but est d'identifier un très large panel de mots qui correspondent au lieu. Alors il faut faire du Webscrapping afin de récupérer le contenu de ces sites et d'en extraire les mots clés.

Enfin il faut refaire une requête avec l'API de moteur de recherche qwant en croisant les mots clés récemment obtenu avec les mots spécifiques au lieu. Avec cette requête, on aura un site web spécifique au lieu.