

Module 1 : Python Basics

Projet Twitter Sentiment Classifier

Nous avons fourni des données Twitter synthétiques (fausses, générées de manière semi-aléatoire) dans un fichier csv nommé `project_twitter_data.csv` qui contient le texte d'un tweet, le nombre de retweets de ce tweet et le nombre de réponses à ce tweet. Nous avons également des mots qui expriment un sentiment positif et un sentiment négatif, dans les fichiers `positive_words.txt` et `negative_words.txt`.

Votre tâche consiste à créer un classificateur de sentiments, qui détectera le degré de positif ou de négatif de chaque tweet. Vous allez créer un fichier csv, qui contient des colonnes pour le nombre de retweets, le nombre de réponses, le score positif (qui est le nombre de mots heureux dans le tweet), le score négatif (qui est le nombre de mots en colère dans le tweet), et le score net pour chaque tweet.

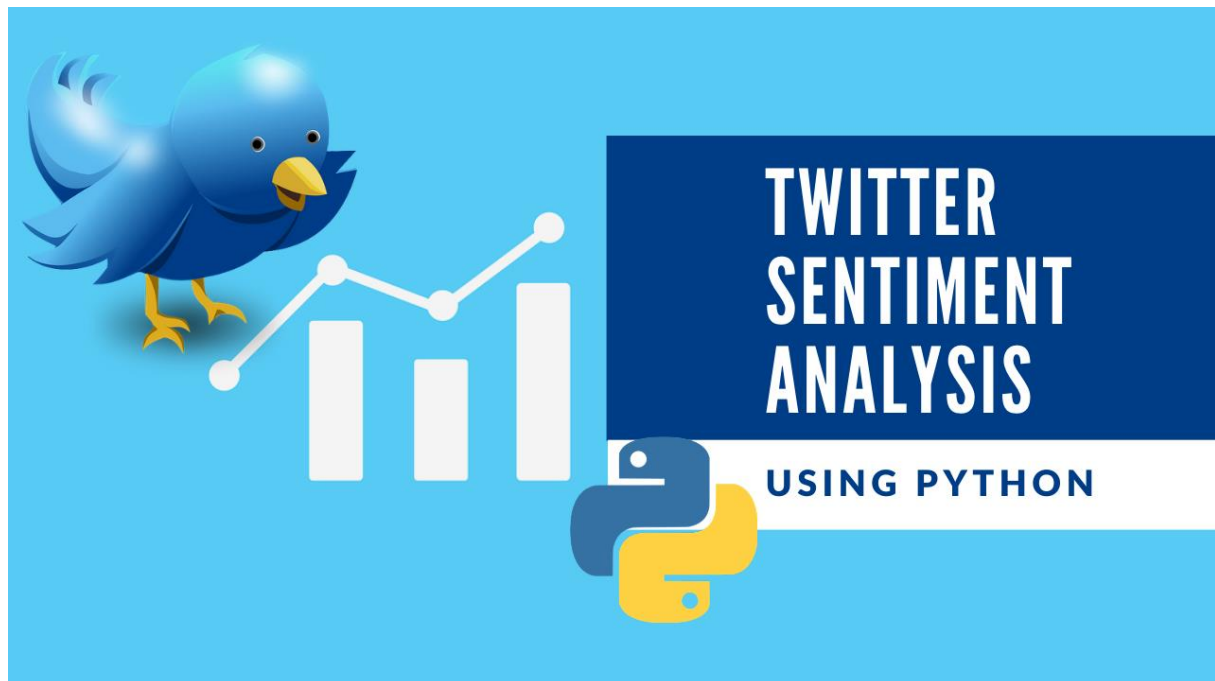
Pour commencer, définissez une fonction appelée `strip_punctuation` qui prend un paramètre, une chaîne qui représente un mot, et supprime les caractères considérés comme ponctuation de partout dans le mot. (Astuce: rappelez-vous la méthode `.replace()` pour les chaînes.)

Ensuite, copiez dans votre fonction `strip_punctuation` et définissez une fonction appelée `get_pos` qui prend un paramètre, une chaîne qui représente une ou plusieurs phrases, et calcule combien de mots dans la chaîne sont considérés comme des mots positifs. Utilisez la liste `positive_words` pour déterminer quels mots compteront comme positifs. La fonction doit renvoyer un entier positif - combien d'occurrences il y a de mots positifs dans le texte. Notez que tous les mots de `positive_words` sont en minuscules, vous devrez donc également convertir tous les mots de la chaîne d'entrée en minuscules.

Ensuite, copiez dans votre fonction `strip_punctuation` et définissez une fonction appelée `get_neg` qui prend un paramètre, une chaîne qui représente une ou plusieurs phrases, et calcule combien de mots dans la chaîne sont considérés comme des mots négatifs. Utilisez la liste, `negative_words` pour déterminer quels mots compteront comme négatifs. La fonction doit renvoyer un entier positif - combien d'occurrences il y a de mots négatifs dans le texte. Notez que tous les mots de `negative_words` sont en minuscules, vous devrez donc également convertir tous les mots de la chaîne d'entrée en minuscules.

Enfin, copiez vos fonctions précédentes et écrivez le code qui ouvre le fichier `project_twitter_data.csv` qui contient les fausses données Twitter générées (le texte d'un tweet, le nombre de retweets de ce tweet et le nombre de réponses à ce tweet). Votre tâche consiste à créer un classificateur de sentiments, qui détectera le degré de positif ou de négatif de chaque tweet. Copiez le code à partir des fenêtres de code ci-dessus et placez-le en haut de cette fenêtre de code. Maintenant, vous allez écrire du code pour créer un fichier csv appelé `result_data.csv`, qui contient

le nombre de retweets, le nombre de réponses, le score positif (qui est le nombre de mots heureux dans le tweet), le score négatif (qui est le nombre de courroux) les mots sont dans le tweet) et le score net (à quel point le texte est positif ou négatif) pour chaque tweet. Le fichier doit avoir ces en-têtes dans cet ordre. N'oubliez pas qu'il existe un autre composant dans ce projet. Vous allez télécharger le fichier csv sur Excel ou Google Sheets et produire un graphique du score net par rapport au nombre de retweets.



Format des tweets dans le fichier source : **the text of a tweet, the number of retweets of that tweet, and the number of replies to that tweet.**