- Developer: Sophia Xiao @ 2020/12/18@

- Dataset downloaded from Tianchi

- Goal: get keywords from rental house description using TF-IDF technique

- Website for study (cited):

  - Data preprocessing:

  - https://www.kaggle.com/sudalairajkumar/getting-started-with-text-preprocessing

  - TF- IDF technique tutorial

  - https://kavita-ganesan.com/extracting-keywords-from-text-tfidf/#.X9re3FUzbt9

- Steps:

  - Pip install modules needed for analysis

  - Read dataset

  - Preprocessing text.

    - All to lowercase

    - Remove punctuations, stop words

    - Lemmatization and remove any html tags that might left

    - Calculate TF-IDF, "sort" dictionary and list the words with top score

- Problems I had during the project:

  - Text preprocessing will be different if I the text language is not English

  - I don't know how IF I can apply cluster on keywords extraction

  - Although I cleaned the html tags before forming the keywords dictionary, I still got words like "br" in my result, don't know why this happens.

  - For the functions that sort scores, I used the example code. But the website only calculates one cell of a column for some reason. But I need the keywords out of

the whole column, so I try to combine the text in the column all together into a

single text to calculate TF-IDF, which works out fine.