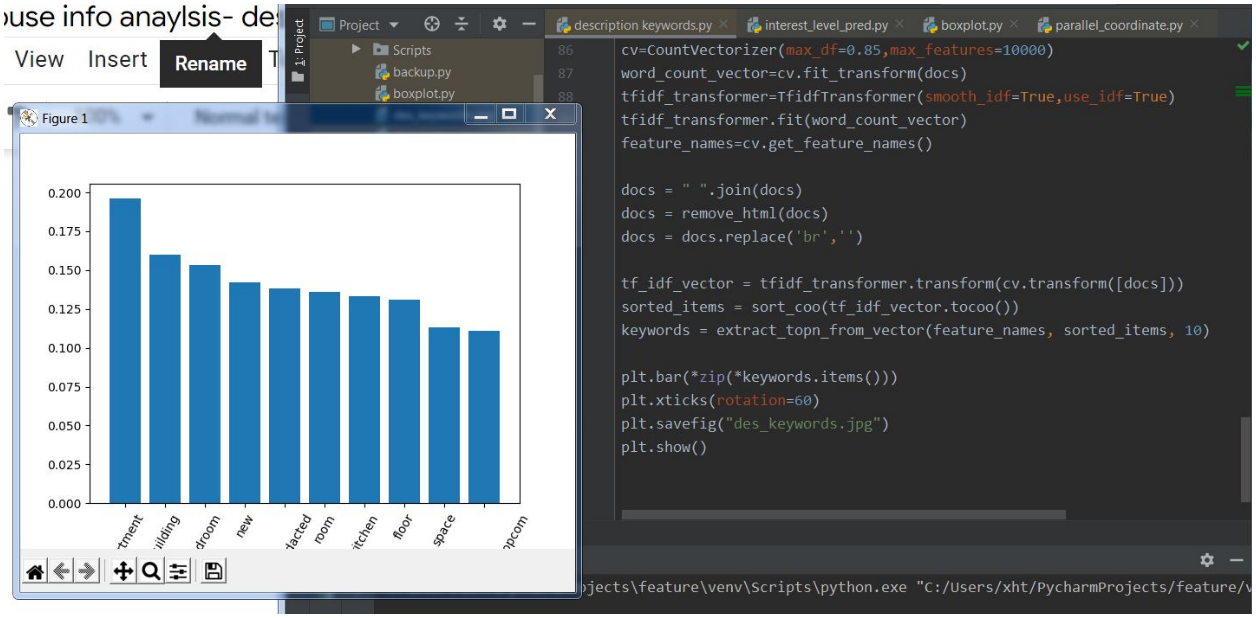
- Developer: Sophia Xiao @ 2020/12/18
    - Edit @ 2021/1/2
- Dataset downloaded from Tianchi

  https://tianchi.aliyun.com/dataset/dataDetail?dataId=83994
- 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.



```
cv=CountVectorizer(max_df=0.85,max_features=10000)
word_count_vector=cv.fit_transform(docs)
tfidf_transformer=TfidfTransformer(smooth_idf=True,use_idf=True)
tfidf_transformer.fit(word_count_vector)
feature_names=cv.get_feature_names()

docs = " ".join(docs)
docs = remove_html(docs)
docs = docs.replace('br','')

tf_idf_vector = tfidf_transformer.transform(cv.transform([docs]))
sorted_items = sort_coo(tf_idf_vector.tocoo())
keywords = extract_topn_from_vector(feature_names, sorted_items, 10)

plt.bar(*zip(*keywords.items()))
plt.xticks(rotation=60)
plt.savefig("des_keywords.jpg")
plt.show()
```

```
ojects\feature\venv\Scripts\python.exe "C:/Users/xht/PycharmProjects/feature/v
```