Q1: TF-IDF

1. **Select texts.**

Recently, there has been a rent crisis in Ireland, especially in Dublin. I am curious what the local people think about the rent crisis on Twitter. Therefore, I intercepted ten tweets and completed the normalization of them. Lowercase the text, remove punctuations and convert numbers to English words using nltk.corpus.stopwords.words. Here are the normalized tweets.

```
a = ['god', 'bless', 'politicians', 'oppose', 'homes', 'built', 'complain
', 'housing', 'rent', 'crisis', 'even', 'think', 'could', 'troll', 'level
', 'one', 'hundredk', 'year', 'make', 'lads']

b = ['protests', 'dublin', 'sharp', 'rise', 'costs', 'living', 'huge', 'ac
comodation', 'crisis', 'places', 'rent', 'people', 'afford', 'buy', 'anyth
ing']

c = ['entitled', 'opinion', 'young', 'person', 'dublin', 'anywhere', 'irel
and', 'living', 'crisis', 'impossible', 'save', 'monthly', 'rent', 'half',
 'wage', 'bills', 'extortionate', 'prices', 'basic', 'costs', 'rising']

d = ['dublin', 'student', 'accommodation', 'crisis', 'would', 'appear', 'b
ad', 'students', 'knocking', 'random', 'doors', 'asking', 'people', 'spare
', 'room', 'rent', 'happened', 'us', 'hypothetical', 'effective', 'governm
ent', 'would', 'probably', 'something']

e = ['issue', 'high', 'cost', 'lack', 'supply', 'nowhere', 'even', 'rent',
 'dublin', 'talk', 'people', 'twentys', 'annoys', 'many', 'people', 'like
', 'dan', 'touch', 'regards', 'rent', 'crisis', 'tragic', 'reality']

f = ['yes', 'many', 'people', 'putting', 'major', 'milestones', 'lives', '
tight', 'money', 'last', 'years', 'cost', 'living', 'crisis', 'going', 'co
st', 'rent', 'dublin']

g = ['rent', 'crisis', 'wild', 'like', 'seriously', 'feels', 'like', 'neve
r', 'find', 'flat']

h = ['doubt', 'someone', 'try', 'spin', 'vacant', 'houses', 'rural', 'part
s', 'west', 'reason', 'rent', 'crisis', 'dublin']

i = ['airbnbs', 'huge', 'factor', 'housing', 'crisis', 'dublin', 'rental',
 'prices', 'skyrocketed', 'amount', 'properties', 'actually', 'available',
 'rent', 'time', 'low', 'landlords', 'instead', 'airbnb', 'extortionate',
'prices', 'need', 'address']

j = ['housing', 'crisis', 'reached', 'tipping', 'point', 'pretty', 'much',
 'rented', 'property', 'available', 'outside', 'dublin', 'people', 'buy',
'due', 'cbi', 'rules', 'whose', 'landlords', 'selling', 'due', 'rent', 'co
ntrols', 'vulnerable', 'position']
```

Let min_freq = 1, here is the plot in WordCloud.

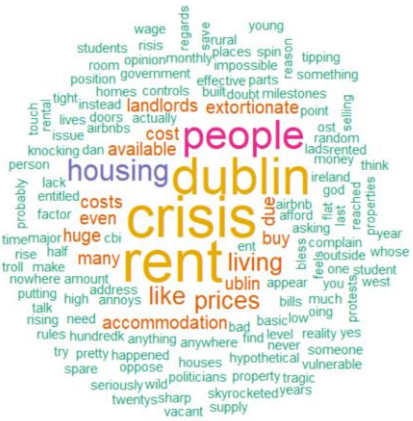

Figure 1: Repeated words

2. Compute TF and TF-IDF

The results obtained after the pre-processing are as follows. I have selected the TF scores for the first t en and last ten examples.

	0
god	1
someone	1
doubt	1
flat	1
find	1
never	1
feels	1
seriously	1
wild	1
going	1

140 rows × 1 columns Open in new tab

	0
many	2
like	3
prices	3
living	3
cost	3
housing	4
people	6
crisis	10
rent	10
dublin	10

140 rows × 1 columns Open in new tab

	god	bless	dublin	politicians	oppose	homes	built	complain	housing
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
6	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
10	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0

10 rows × 140 columns Open in new tab

Figure 2: TF scores and matrix

The result of TF-IDF:

	0
money	0.046078
risingdublin	0.046078
rise	0.046078
rented	0.046078
rental	0.046078
...	...
housing	0.138233
people	0.276465
dublin	0.322543
rent	0.414698
crisis	0.460776

136 rows × 1 columns Open in new tab

	accomodation	actually	address	afford	airbnb	airbnbs	amount	annoys
0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
1	0.266090	0.000000	0.000000	0.313013	0.000000	0.000000	0.000000	0.0000
2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
3	0.179273	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.2347
5	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
8	0.000000	0.225713	0.225713	0.000000	0.225713	0.225713	0.225713	0.0000
9	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000

10 rows × 140 columns Open in new tab

Figure 3: TF-IDF scores and matrix

There is little change in word ranking between TF and TF-IDF. The ranking of the last five is still in the last five. Furthermore, There are some internal changes in the ranking of the last three words, possibly due to changes in word combinations throughout the corpus.

Q2: PMI

Here are the top five PMI scores calculated by using nltk.collocations.

```
{('god', 'bless'): 7.584962500721156,
 ('bless', 'politicians'): 7.584962500721156,
 ('politicians', 'oppose'): 7.584962500721156,
 ('oppose', 'homes'): 7.584962500721156,
 ('homes', 'built'): 7.584962500721156}
```

The results do not make sense. Many words appear only one time but achieve a high PMI val ue, like `'homes'` and `'built'`, which does not match the expected results. In this example, t he expected value would be `'rent'` and `'crisis'` to have a maximum PMI value. Set this m inimal cut-off frequency to 3 and calculate it again.

```
('rent', 'crisis') = 2.803602787196496
```

Q3: Entropy

I chose ten tweets from Tesla as the spam_set, all about its electric cars product. Apart from that, I als o chose some other ten tweets as the random_set from the trend of Twitter which covers almost all asp ects of life, including news, sports, entertainment, etc and has little correlation between them. Here ar e the set.

```
spam_set = ['Your Tesla now shows energy consumed vs projected & gives ran
ge tips', 'The Tesla Model Y is officially the best selling vehicle in Ger
many, accounting for just over 4.5% of the car market.','Tesla Immersive,
our multichannel audio upmixer, enables stereo content to be remixed in re
al time, optimizing the listening experience for our vehicles specifically
', 'The brain of your Tesla: Neural networks with 1 billion parameters, co
mpleting 144 trillion operations per second', 'Enter your destination & yo
ur Tesla will automatically include Supercharging stops in your route', 'W
e are launching the Tesla Shareholder Platform — join the program to parti
cipate in Tesla events and hear more updates soon ', 'Tesla will ask share
holders to vote at this year’s annual meeting to authorize additional shar
es in order to enable a stock split.', 'Non-Tesla vehicles can now charge
at select Superchargers in Denmark, Finland, Germany, Luxembourg and Switz
erland via the Tesla app. ', 'Tesla navigation will now take predicted cro
sswind, headwind, humidity & temperature into account for calculating batt
ery % on arrival', 'Tesla Vehicle Production & Deliveries and Date for Fin
ancial Results & Webcast for First Quarter 2022']
```

```
random_set = [ 'There are two groups who I will simply not accept any mora
l judgement or admonition from: Irish people feverishly supporting Ukraine
 but saying the girls singing Up The Ra is wrong+ anyone supporting a part
y who wanted to commemorate the Black and Tans. Hypocrisy is not for me.',
 'On February 24, Russia launched an invasion of Ukraine by land, air and
sea after months of tensions between Moscow and Kyiv. The attack triggered
 a chain of events over the next six months including unprecedented sancti
```

```
ons against Russia and the expansion of NATO.', 'Philly McMahon has hit ou
t at Sky Sports News presenter Rob Wotton following his controversial ques
tioning of Republic of Ireland star Chloe Mustaki.', 'Thoroughly enjoying
Rob Wooton getting (deservedly) shat on from Irish Twitter. Typical Englis
h ignorance with the usual superiority complex. Lbu they’d do your head i
n.', 'Was told by my lecture today at 11.11am as we all walked out that th
e protest was a load of bullshit, he’s not the one struggling to pay for a
ccommodation, college fees, food, parking he’s not on €10.50 an hour. The
arrogance of some people.', 'Join dynamic alumni Nicai de Guzman (Wolfgang
 Digital), Emmet Daniel (Hubspot), and expert Dr. Linda Yang (Intercultura
l Development Programme, UCD) as they offer advice on thriving in multicul
tural work environments', 'Brendan Fraser hadn’t played the lead in a majo
r movie in 12 years, a gap on his résumé that’s been attributed to persona
l issues, health problems and an assault allegation against the former hea
d of the HFPA.', '58 years on the road in 2022, proud to have brought the
story of Ireland around the world since 1964! Let The People Sing', 'Here’
s Xavi celebrating a group stage goal to avoid Europa league and here’s An
celotti reacting to the greatest come back of all time to make the UCL fin
al. Levels to this game.', 'Venus Williams on sisterhood, securing equal p
ay for women in sport and building an empire off the court']
```

Here is the entropy method found in the previous PowerPoint. Before calculating the entropy value, I have done some normalization in the pre-precessing section using the pre_precessing() method.

```
import nltk
import math
def entropy(labels):
    freqdist = nltk.FreqDist(labels)
    probs = [freqdist.freq(l) for l in freqdist]
    return -sum(p * math.log(p,2) for p in probs)

print("Entropy of the spam_set: {}".format(entropy(pre_precessing(spam_se
t))))
print("Entropy of the random_set: {}".format(entropy(pre_precessing(random
_set))))
print("Entropy of the span_set and random_set: {}".format(entropy(pre_prec
essing(spam_set + random_set))))
```

```
Test (1)
D:\Python\TextAnalytics\Scripts\python.exe D:/Python/TextAnalytics/Assignment4/Test.py
Entropy of the spam_set: 6.610308719803565
Entropy of the random_set: 7.4797027681764625
Entropy of the span_set and random_set: 8.036545353326574

Process finished with exit code 0
```

Figure 4: Results of Entropy

Here it computes the results of three different entropy, corresponding to three cases. Entropy values symbolize a high or low state of disorder, even in text analytics, which can also reflect whether it is a disorder or not. The spam_set is primarily based on Tesla’s product, which has lower entropy than the random_set. Moreover, the random_set is chosen entirely at random from the Twitter trends. It has a higher degree of disorder which is reflected in the result. At last, combining the two corpus sets has a higher disorder which gets the highest entropy value.