Question 1 Comparison of TF scoring and TF-IDF scoring

Find 10 short text-items (10-20 words in each) E.g. Tweets; all dealing with common topic of interest

10 sample tweets relating with Rafael Nadal's victories around French open grand slam are selected.

[1] earlier today during the Balearic Golf Championship He was winning Roland Garros for a 13th time less than two weeks ago
[2] faced seven players with two handed backhand en route to his 13th French Open title
[3] Rafael Nadal just survived a real life Hunger Games at the French Open says The Ashfordian Perspective
[4] just two weeks back Rafael Nadal won his 13th french open title He is now playing Balearic Golf Championship and tied 10th in a field on day one
[5] Recently the French Open 2020 witnessed Rafael Nadal win a record extending 13th French Open title and level with his arch rival Roger Federer with 20 Grand Slams
[6] This is like Rafael Nadal complaining that the French Open is too predictable
[7] Nobody knows what Rafa Nadal has been through says Carlos Moya following French Open triumph
[8] Rafael Nadal beats Novak Djokovic in French Open final to tie Roger Federer with 20 Grand Slam titles
[9] French Open champion Rafael Nadal to take part in Paris Masters next month said Sports Mole
[10] Russians up and coming star Andrey Rublev has paid tribute to Rafael #Nadal after the Spaniard records equaling 20th Grand Slam success at Roland Garros hattrick

1(a) Remove the standard stop-words from them using some standard list, use nltk, so that you now have the remaining words (the R-words) for each short-text-item

Tweets are filtered for punctuations and converted to lowercase manually. **After stop word removal, 96 words are left in corpus for TF-IDF analysis.** Text left is:

earlier today Balearic Golf Championship He winning Roland Garros 13th time less two weeks ago faced seven players two handed backhand en route 13th French Open title Rafael Nadal survived real life Hunger Games French Open says The Ashfordian Perspective two weeks back Rafael Nadal 13th french open title He playing Balearic Golf Championship tied 10th field day one Recently French Open 2020 witnessed Rafael Nadal win record extending 13th French Open title level arch rival Roger Federer 20 Grand Slams This like Rafael Nadal complaining French Open predictable Nobody knows Rafa Nadal says Carlos Moya following French Open triumph Rafael Nadal beats Novak Djokovic French Open final tie Roger Federer 20 Grand Slam titles French Open champion Rafael Nadal take part Paris Masters next month said Sports Mole Russians coming star Andrey Rublev paid tribute Rafael # Nadal Spaniard records equaling 20th Grand Slam success Roland Garros hattrick

1(b) Compute the TF scores for these R-words across all the text items and use R to show the wordcloud for these words. Provide the matrix of TF scores and the wordcloud.

Term Frequency (TF) refers to frequency of a preprocessed word(t) in a single document(d). It is formulated as $tf(t,d) = f(t,d)$ Following screenshot of **TF score matrix** represents frequency of few terms [*in columns*] in each of 10 documents [in rows]. We can note that, some words occurs in single document [e.g. Winning, tied etc], but words like Rafael occur in multiple documents with varied frequency.

rafael	hattrick	russians	french	faced	nadal	masters	federer	golf	take	beats	spaniard
0	0	0	0	0	0	0	0	0.071429	0	0	0
0	0	0	0.083333	0.083333	0	0	0	0	0	0	0
0.083333	0	0	0.083333	0	0.083333	0	0	0	0	0	0
0.055556	0	0	0.055556	0	0.055556	0	0	0.055556	0	0	0
0.045455	0	0	0.090909	0	0.045455	0	0.045455	0	0	0	0
0.142857	0	0	0.142857	0	0.142857	0	0	0	0	0	0
0	0	0	0.090909	0	0.090909	0	0	0	0	0	0
0.066667	0	0	0.066667	0	0.066667	0	0.066667	0	0	0.066667	0
0.071429	0	0	0.071429	0	0.071429	0.071429	0	0	0.071429	0	0
0.05	0.05	0.05	0	0	0.05	0	0	0	0	0	0.05

Figure 1 TF matrix

Cumulative TF is calculated in context entire corpus[*Set of Documents*]. It is obtained by dividing number of occurrences of a term by total number of words in all documents. Partial set of top cumulative TF scores are listed below:
{'open': 0.06206896551724138, 'french': 0.06206896551724138, 'nadal': 0.05517241379310345, 'rafael': 0.04827586206896552, '13th': 0.027586206896551724, 'title': 0.020689655172413793, 'two': 0.020689655172413793, 'grand': 0.020689655172413793, 'roland': 0.013793103448275862, 'championship': 0.013793103448275862,…… 'level': 0.006896551724137931, 'star': 0.006896551724137931, 'following': 0.006896551724137931, 'mole': 0.006896551724137931}

The above results are confirmed by wordcloud. Wordcloud is created using frequencies recorded in the TF Scores matrix. TF score is measure of Frequency of a word in relation to corpus; hence importance of words like “Nadal”, “Open”, “French”, “Rafael”, “Grand” is noticeable.

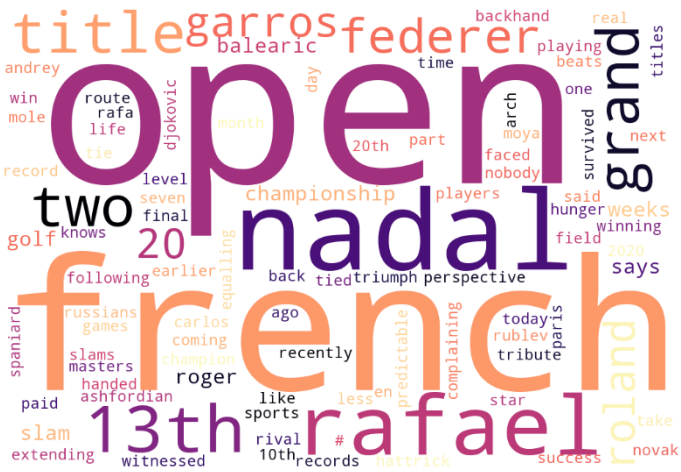

Figure 2 TF wordcloud

1 (c) compute the TF-IDF scores for these R-words across all the text-items. Also, provide the matrix of TF-IDF scores. Generating TF-IDF Scores for the same set of words.

Inverse Document frequency (IDF) is measure of rarity of a term. Since DF of a term is count of documents containing the term divided by total documents in set; IDF is defined as:

$$\mathrm{IDF}(t, \mathrm{D}) = \mathrm{N}/\,\mathrm{DF}(t, \mathrm{D}) = \log\left(\frac{\mathrm{N}}{\mathrm{DF}(t, \mathrm{D})}\right)$$

Thus, IDF MAXIMISES the value for words which seldom occur in corpus D. As Observed in partial set of IDFs below, words like “Nadal”,” French” which occur in many documents [refer figure 1] are penalized irrespective of their importance in each of those documents [i.e. TF].
{'Nadal': 0.09691001300805642, 'Open': 0.1549019599857432, 'French': 0.1549019599857432, 'Rafael': 0.1549019599857432,…… 'ago': 1.0, 'win': 1.0, 'Slams': 1.0, 'field': 1.0, 'Mole': 1.0}

TF-IDF is literally multiplication of TF for a term IN A PARTICULAR DOCUMENT (d) and IDF score of that term IN CONTEXT OF ENTIRE CORPUS OF DOCUMENTS (D).

rafael	back	french	next	tribute	nadal	federer	open	hunger	russians	garros
0	0	0	0	0	0	0	0	0	0	0.049926
0	0	0.008076	0	0	0	0	0.008076	0	0	0
0.012908	0	0.008076	0	0	0.008076	0	0.008076	0.083333	0	0
0.008606	0.055556	0.005384	0	0	0.005384	0	0.005384	0	0	0
0.007041	0	0.00881	0	0	0.004405	0.031771	0.00881	0	0	0
0.022129	0	0.013844	0	0	0.013844	0	0.013844	0	0	0
0	0	0.00881	0	0	0.00881	0	0.00881	0	0	0
0.010327	0	0.006461	0	0	0.006461	0.046598	0.006461	0	0	0
0.011064	0	0.006922	0.071429	0	0.006922	0	0.006922	0	0	0
0.007745	0	0	0	0.05	0.004846	0	0	0	0.05	0.034949

Figure 3 TF - IDF matrix

From **cumulative TF-IDF score,** we can see that **TF-IDF maximizes RARITY.** Terms with rare occurrence across multiple documents are maximized, while those occurring in multiple documents are minimized.
{'13th': 0.010977655411642416, 'title': 0.010818180936834572, 'two': 0.010818180936834572, 'grand': 0.010818180936834572, 'roland': 0.009640965577048535, 'championship': 0.009640965577048535, 'golf': 0.009640965577048535,…… 'mole': 0.006896551724137931, 'open': 0.006015104255672467, 'french': 0.006015104255672467, 'nadal': 0.005346759338375527}

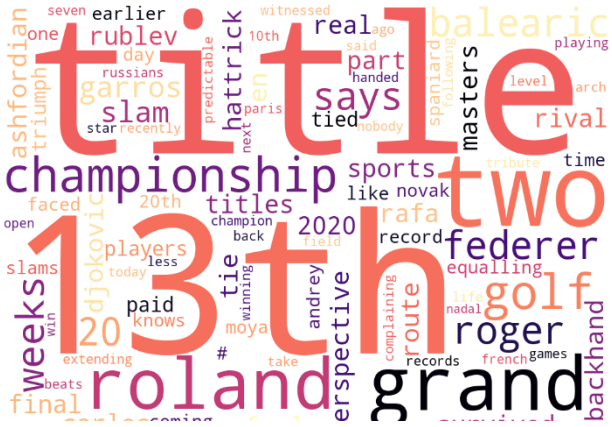

Figure 4 TF-IDF matrix

1(d) Discuss the changes that occur in the relative ranking of the top ranked words between the TF scoring and the TF-IDF scoring and explain why these changes occur (if any)

Words “nadal”, “open”, “french” occur in 8-9 documents each and words and are top ranking words in TF score. Whereas, word “13th”, “title” occurs in 4 documents each with moderate TF score. Yet, TF-IDF score of “13th”, “title” is top-ranking, while “nadal”, “open”, “french” are last in ranking.

Partial set of TF-IDF matrix shown in figure 3 confirms the fact that, tf–idf has higher weight when word has high TF (within document d) and a low DF (within entire corpus D). That is why, word "nadal" and "Federer" have comparable term frequencies in 5th document [tf = 0.045455] [*refer figure 1*]. But, in TF-IDF frequency for document 5, word "nadal" has very low frequency that "federer" [0.004<<0.03] [*refer figure 3*].

Question 2 Significance of Pointwise mutual information (PMI) score

Using Python or R, compute the PMI scores for all adjacent pairs of words in your 10-doc corpus List the top-5 pairs based on the PMI scores found each pair. Do the results make sense? If not, then introduce a minimal cut-off frequency and re-compute the top-5 until they seem sensible.

PMI Score: Pointwise mutual information (PMI) score dictates possibility of cooccurrence of set off words in relative order. Hence, we need probability of cooccurrence of words; and probability of their independent occurrence to comment about existence of a word given another. We can say that, PMI is manifestation of condition probability that maximises entropy (Information content) of presence of word(X) at position i given that, word(Y) is also present at position i+1. When PMI is calculated for bigrams (set of pairs of words cooccurring in stopword-removed corpus of Question 1), top 5 pairs by PMI value we obtained are:

	bigram	PMI
0	(10th, field)	7.179909
24	(russians, coming)	7.179909
26	(nobody, knows)	7.179909
27	(novak, djokovic)	7.179909
28	(one, recently)	7.179909

Even though these PMI are correct, they add little information while answering question WHAT the dataset is about. PMI index for cooccurrence of word (novak, djokovic) is high because it has single occurrence in single document [8t tweet] within entire corpus. So naturally, Novak → Djokovic

To find necessary PMI score, frequency filter is applied on the data by using minimal cut-off frequency of 2. Hence, a word pair is considered a bigram if and only if it has 2 or more cooccurrence in corpus. Obtained bigrams with :

f_{cut} = 2

	bigram	PMI
0	(Balearic, Golf)	6.179909
1	(Federer, 20)	6.179909
2	(Golf, Championship)	6.179909
3	(Roger, Federer)	6.179909
4	(Roland, Garros)	6.179909

f_{cut} = 3

	bigram	PMI
0	(french, open)	4.009984
1	(open, title)	4.009984
2	(rafael, nadal)	3.957517
3	(13th, french)	3.594947

Above PMI pairs indicate frequent bigrams in data. It should be noted that, only 4 bigrams with 3 or more cooccurrences can be located in the corpus.

Question 3 Calculation of Entropy

Entropy is used to determine whether tweet set is interesting (contains variety) or repetitive (spam).

3(a) Create two sets of 10 made-up tweets each:

a. spam-set: where tweets are very similar containing advert for product like a spam tweet
b. random-set: where tweets are all different, chosen at random from Twitter

Spam Tweets set: Regarding law suit against Google

[1] Today's lawsuit by the Department of Agriculture is deeply flawed. People do not use Google because they cannot.
[2] The Defence Department has filed an antitrust lawsuit against Google for allegedly abusing its money and good looks over smaller rivals.
[3] It is a far-left wiki for conspiracy theorists that makes libellous claims about conservatives. It is also promoted by Google.
[4] Absolute depressing of a growth by #google in the last month
[5] I have been calling for this for years. So glad to see it finally happening.
[6] Yesterday, Ken Buck broke from the DOJ telling PTV "breaking up Google may be the right answer."
[7] #NEW The DOD files an trust lawsuit against Google under Section 12 of the Patriot Act.
[8] Google given 20m aid by Department of defence for spreading American happiness across world
[9] Really want to know nothing? Read lawsuit Hunky-dory organization v. Google, Amazon, Facebook
[10] Google responded that the lawsuit would deprive world of "naturally prop up high-quality search alternatives"

Genuine Tweet set:

[1] Did you know it's is free to send mail to residents in Care Facilities across Ireland until 31st January 2021?
[2] Businesses and workers impacted by #Level5 restrictions are being supported.
[3] We've passed 100,000 signatures, OVER ONE HUNDRED THOUSAND SIGNATURES IN 36 HOURS to #RepealTheSeal and #OpentheArchive.
[4] Khabib Nurmagomedov has announced his retirement. He promised his mother this would be his final fight
[5] Fact: Sinn Féin is the RICHEST party in Ireland. Watch to find out how they exploited legal loopholes
[6] I love Hyderabad Am contributing Rs. 5 lakhs to support the flood affected people of Hyderabad.
[7] South Africa's only high-speed rail network is drawing up a multibillion rand plan to expand outside Johannesburg and Pretoria
[8] Morgan Stanley to Acquire Eaton Vance; Creates a Leading Asset Manager Positioned for Growth
[9] Be it in Arunachal Pradesh, Chandigarh, Madhya Pradesh, Gujarat and several other places, there are many who are promoting reading.
[10] Europe's cases are exploding. US is entering its third wave. But interestingly Asia is fending off another wave.

3(b) Now, find a Python/R program or package that computes entropy and find the entropy values for (i) spam-set, (ii) random-set, (iii) the two sets combined. Report the program you used and its source, the tweet data and the entropy values found.

Entropy is defined as degree to which information is surprising to the system. That is, with spam tweets talking about google law suit, essentially talk much more about "google", "law" and "suit". Hence information obtained cumulatively from point of view of computer is lesser than genuine tweets discussing multiple topics [sources of information]. Program to calculate Entropy is given as :

```
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)
```

Source: Natural Language Processing with Python, Ch. 6, by Steven Bird, Ewan Klein and Edward Loper

Entropy for Spam Tweet corpus: 6.277790641185837

Entropy for Genuine Tweet corpus: 6.774012911227603

Entropy for Combined List: 7.397926625110604

Naturally, information from spam[repetitive, ingenuine content] is lesser than standard tweets. And, when both corpuses are collectively treated, information is increased hence entropy increases. Also, document frequency for spam words is inorganically reduced.