Q1)
Tweets used:
Never underestimate what it's like in someone else's head. In this job we see people at breaking point everyday and that affects us too. Please remember that you are loved and important. Pick up the phone, talk and put the kettle on...you are not alone.WorldMentalHealthDay
...
As it's #WorldMentalHealthDay I encourage everyone who is struggling to try and book an appointment with a GP - get on that road! I know it can be scary, intimidating, embarrassing and/or cause anxiety - I cancelled two appointments before finally plucking up the courage...

On #WorldMentalHealthDay, let us not forget the conscious cruelty the Tories have inflicted on mental health sufferers: £105 million less funding now than in 2012. 30% fewer beds for mental health patients 40% of mental health trusts have experienced cuts

Today is #WorldMentalHealthDay. Around 70% of autistic children and young people have mental health problems - and many autistic adults do too. It's vital that they can get the support they need, from professionals who understand autism

It's #WorldMentalHealthDay today. Remember that even the happiest, smiliest and most confident people you know have days where they lose faith in themselves and don't feel comfortable in their own skin. Everyone is fighting their own battle, let's help each other out

Please dont be afraid to speak to someone if your struggling no matter what the reason is. You're family & friends will want you in their lives not just a memory of you. Its Ok not to be Ok. Stay strong & ask for help its a sign of great strength to do so. #WorldMentalHealthDay

In Peacehaven the Joff Youth Centre are working towards Takeover Day in November — a national initiative which will give young people a voice on issues and policies around mental health and emotional wellbeing #WorldMentalHealthDay

It's #WorldMentalHealthDay and 1 in 4 people suffer from mental health problems. We're proud to say we help hundreds of learners combat their personal barriers by motivating them to improve their skills and get back into work with our free training courses. DM us for more info!

You cannot recover from anxiety by staying calm. You can not recover from depression by just being positive. You cannot recover from anorexia nervosa by eating more. If mental illness were that simple we wouldn't be struggling in the first place. #WorldMentalHealthDay

Struggling with mental health was always a journey I kept to myself until recently, but opening up and seeking help got me to the good place I'm in today. It's hard to talk about it but it's even harder to go through it by yourself, lets break the stigma #WorldMentalHealthDay.

Without stopwords:
never underestimate 's like someone else 's head . job see people breaking point everyday affects us . please remember loved important . pick phone , talk put kettle ... alone.worldmentalhealthday ... 's # worldmentalhealthday encourage everyone struggling try book appointment gp - get road ! know scary , intimidating , embarrassing and/or cause anxiety - cancelled two appointments finally plucking courage ... # worldmentalhealthday , let us forget conscious cruelty tories inflicted mental health sufferers : £105 million less funding 2012 . 30 % fewer beds mental health patients 40 % mental health trusts experienced cuts today # worldmentalhealthday . around 70 % autistic children young people mental health problems - many autistic adults . ' vital get support need , professionals understand autism 's # worldmentalhealthday today . remember even happiest , smiliest confident people know days lose faith n't feel comfortable skin . everyone fighting battle , let 's help please dont afraid speak someone struggling matter reason . 're family & friends want lives memory . ok ok. stay strong & ask help sign great strength . # worldmentalhealthday peacehaven joff youth centre working towards takeover day november — national initiative give young people voice issues policies around mental health emotional wellbeing # worldmentalhealthday ' #

worldmentalhealthday 1 4 people suffer mental health problems . ' proud say help hundreds learners combat personal barriers motivating improve skills get back work free training courses . dm us info ! recover anxiety staying calm . recover depression positive . recover anorexia nervosa eating . mental illness simple ' struggling first place . # worldmentalhealthday struggling mental health always journey kept recently , opening seeking help got good place ' today . ' hard talk ' even harder go , lets break stigma # worldmentalhealthday .

TF for the first 3 tweets:

Doc 1
Word: i, TF: 0.11111
Word: as, TF: 0.03704
Word: worldmentalhealthday, TF: 0.03704
Word: encourage, TF: 0.03704
Word: everyone, TF: 0.03704
Word: struggling, TF: 0.03704
Word: try, TF: 0.03704
Word: book, TF: 0.03704
Word: appointment, TF: 0.03704
Word: gp, TF: 0.03704
Word: get, TF: 0.03704
Word: road, TF: 0.03704
Word: know, TF: 0.03704
Word: scary, TF: 0.03704
Word: intimidating, TF: 0.03704
Word: embarrassing, TF: 0.03704
Word: and/or, TF: 0.03704
Word: cause, TF: 0.03704
 Word: anxiety, TF: 0.03704
Word: cancelled, TF: 0.03704
Word: two, TF: 0.03704
Word: appointments, TF: 0.03704
Word: finally, TF: 0.03704
Word: plucking, TF: 0.03704
Word: courage, TF: 0.03704
Doc 2
Word: mental, TF: 0.10345
Word: health, TF: 0.10345
Word: on, TF: 0.03448
Word: worldmentalhealthday, TF: 0.03448
Word: let, TF: 0.03448
Word: us, TF: 0.03448
Word: forget, TF: 0.03448
Word: conscious, TF: 0.03448

Word: cruelty, TF: 0.03448
Word: tories, TF: 0.03448
Word: inflicted, TF: 0.03448
Word: sufferers, TF: 0.03448
Word: £105, TF: 0.03448
Word: million, TF: 0.0344
Word: less, TF: 0.03448
Word: funding, TF: 0.03448
Word: 2012, TF: 0.03448
Word: 30, TF: 0.03448
Word: fewer, TF: 0.03448
Word: beds, TF: 0.03448
Word: patients, TF: 0.03448
Word: 40, TF: 0.03448
Word: trusts, TF: 0.03448
Word: experienced, TF: 0.03448
Word: cuts, TF: 0.03448

Doc 3
Word: autistic, TF: 0.08
Word: today, TF: 0.04
Word: worldmentalhealthday, TF: 0.04
Word: around, TF: 0.04
Word: 70, TF: 0.04
Word: children, TF: 0.04
Word: young, TF: 0.04
Word: people, TF: 0.04
Word: mental, TF: 0.04
Word: health, TF: 0.04
Word: problems, TF: 0.04
Word: many, TF: 0.04
Word: adults, TF: 0.04
Word: too, TF: 0.04
Word: it, TF: 0.04

This technique produces a good big wordcloud but it adds more words that are repeated too much and don't provide a lot of value to the cloud and pollute it.Words like around,wordmentalhealth and lets are not that significant due to their frequency in each document. But TF wordcloud still emphasizes them.
By using Tf-idf we can make the wordcloud shorter and display more important words than before.
When we use tfidf it is clear that wordmentalhealth tag that was the most popular in the previous example now has a value of nearly 0 because it occurs in each document.

Words selected in first three docs based by tfidf cutoff

Doc1
Word: never, TF-IDF: 0.05748
Word: underestimate, TF-IDF: 0.05748
Word: like, TF-IDF: 0.05748
Word: elses, TF-IDF: 0.05748
Word: head, TF-IDF: 0.05748
Word: job, TF-IDF: 0.05748
Word: see, TF-IDF: 0.05748
Word: breaking, TF-IDF: 0.05748
Word: point, TF-IDF: 0.05748
Word: everyday, TF-IDF: 0.05748

Word: encourage, TF-IDF: 0.05961

Word: try, TF-IDF: 0.05961
Word: book, TF-IDF: 0.05961
Word: appointment, TF-IDF: 0.05961
Word: gp, TF-IDF: 0.05961
Word: road, TF-IDF: 0.05961
Word: scary, TF-IDF: 0.05961
Word: intimidating, TF-IDF: 0.05961

Doc2
Word: i, TF-IDF: 0.13377
Word: as, TF-IDF: 0.05961

Doc3
Word: let, TF-IDF: 0.0555
Word: forget, TF-IDF: 0.0555

Word: conscious, TF-IDF: 0.0555   Word: sufferers, TF-IDF: 0.0555
Word: cruelty, TF-IDF: 0.0555   Word: £105, TF-IDF: 0.0555
Word: tories, TF-IDF: 0.0555   Word: million, TF-IDF: 0.0555
Word: inflicted, TF-IDF: 0.0555   Word: less, TF-IDF: 0.0555

Hence we successfully weeded out the important words among tweets. The words in yellow are the important words in the tweets.

Q2

PMI is a measure of how much words that occur together relay some special information. For example Candy and cane are 2 very different words but together they relay very different information. This is useful to gather meaning from text

PMI of best 10 pairs are

(('Around', '70'), 8.988684686772165)
(('Its', 'Ok'), 8.988684686772165)
(('Joff', 'Youth'), 8.988684686772165)
(('Never', 'underestimate'), 8.988684686772165)
(('November', '—'), 8.988684686772165)
(('Ok.', 'Stay'), 8.988684686772165)
(('Stay', 'strong'), 8.988684686772165)
(('Takeover', 'Day'), 8.988684686772165)
(('Youth', 'Centre'), 8.988684686772165)
(('an', 'appointment'), 8.988684686772165)

The algorithm manages to find pairs properly like 'joff youth' and 'youth center' and "takeover day"

There are some wrong pairs like "and/or cause" and "ok stay" These are probably because of the small size of the dataset as well as the low number of pair words having inherent meaning.

To achieve this collocations package from nltk was used. We can also use this package for finding trigrams.

Q3

Entropy is a measure of disorder between different items. A high entropy usually suggests new and interesting data. While low entropy suggests that the data is repetitive .Very low entropy usually suggests spam.

It is the measured by simply checking how dissimilar items are among their tokens

In the data tested for the usual tweets the entropy was : 4.705634740381123

While for spam-set it was =4.505642995480449

It is apparent that there is less disorder in spam-set. While for normal data entropy is high. Once can set a threshold to filter spam tweets using this simple technique.

The entropy of the combined sets was =4.790216962849482.

This is obvious as the spam set and the random set are quite different from each other. Combining them results in a more disordered set than each of them. If they both were similar entropy would understandably lower than both of them.