Concept of Opinionator
Sentiment analysis is deciding tone behind a textual statement. Roughly, any human thought or action is:

- Positive: hopeful, encouraging, progressive, happy
- Negative: regressive, disregarding, biased, drowsy
- Neutral: Poised, balanced, moderate

To catch mood of a piece of text, Inter-rater reliability is used. Text itself is an OPINION and one who judges it to be positive negative or neutral is a RATER. Inter-rater reliability is the extent of agreement among raters. It gives a score of how much homogeneity, or consensus, there is in the ratings given by judges.

Expression of Consensus can be qualitative i.e. categorical (vote **1** or **0** for ONE OF THE candidate choices : Agree, Disagree, not sure) as well as quantitative i.e. numeric (vote degree of agreement 50%, 34% etc. for ALL OF THE candidate choices : Agree, Disagree, not sure).

Question 1 Human Ratings Task:
1 a) Get 3 classmates (opinion holders) to write three different opinions about their phone

Opinions gathered from 3 subject opinion holders:

1. "I am using this iPhone as I prefer small footprint. It has got good battery life, security and support and camera is excellent."
2. "My Nokia is performing consistently for 8 years and camera with ZEISS is good for city captures"
3. "I am Samsung user since it ran Symbian. It offers best cost to performance and I am satisfactory with my galaxy as it updates monthly."

1 b) Get 3 different people (raters) to rate these comments as positive, negative, neutral or can't-say
Raters provide 4 distinct feedbacks i.e. reviews about the opinion. As both neutral or can't-say feedbacks say that Raters are undecisive about sentiment conveyed in opinion; they are merged as "Neutral". Ratings are grouped into 3 categories : [Positive, Negative, Neutral]

Feedback rating matrix obtained from 3 independent reviewers:

Ratings: Positive = P, Neutral = O, Negative = N				
		Raters		
		R1	R2	R3
Opinion	O1	O	P	N
	O2	N	P	O
	O3	O	O	P

		Rater 1	Rater 2	Rater 3
Opinion 1	Positive	0	1	0
Opinion 2	Positive	0	1	0
Opinion 3	Positive	0	0	1
Opinion 1	Negative	0	0	1
Opinion 2	Negative	1	0	0
Opinion 3	Negative	0	0	0
Opinion 1	Neutral	1	0	0
Opinion 2	Neutral	0	0	1
Opinion 3	Neutral	1	1	0

1 c) Take this 3 x 3 matrix and find the inter-rater reliability between your 3 raters using Kappa

Cohen Kappa method finds agreement between 2 Raters who classify N items into C mutually exclusive categories. With 3 subject opinions and 3 categories of perceived sentiment, we get 3*3 matrix for each opinion for all raters. This method is used when Expression of Consensus are **qualitative and mutually exclusive**.

It is generally thought to be a more robust measure than simple percent agreement calculation, as kappa takes into account the possibility of the agreement occurring by chance.

$$k = \frac{P_o - P_e}{1 - P_e}$$ .... Po: Probability of observed Agreement; Pe: Probability of chance agreement

		Rater 2			
		Positive	Negative	Neutral	Total
Rater 1	Positive	0	0	0	0
	Negative	1	0	0	1
	Neutral	1	0	1	2
	Total	2	0	1	3

		Positive	Negative	Neutral	Total
	Agreement	0	0	1	1
	chance	0.666667	0	0.333333	0.666667
Kappa		0.142857			

		Rater 3			
		Positive	Negative	Neutral	Total
Rater 2	Positive	0	1	1	2
	Negative	0	0	0	0
	Neutral	1	0	0	1
	Total	1	1	1	3

		Positive	Negative	Neutral	Total
	Agreement	0	0	0	0
	chance	0.333333	0.333333	0.333333	1
Kappa		-0.5			

		Rater 3			
		Positive	Negative	Neutral	Total
Rater 1	Positive	0	0	0	0
	Negative	0	0	1	1
	Neutral	1	1	0	2
	Total	1	1	1	3

		Positive	Negative	Neutral	Total
	Agreement	0	0	0	0
	chance	0.333333	0.333333	0.333333	1
Kappa		-0.5			

Typically Kappa > 0.7 indicates high agreement, and values towards 0 successively indicate lower agreement. A negative kappa value indicates that predictive model performs even below baseline model with given dataset.

- As observed in above image, R1 and R2 agree about Opinion 3, and they both consider Opinion 3 as Neutral. Apart from that, there is no agreement hence small kappa score of 0.14 is seen. It is insufficient and poor agreement to unbiased Judgement of opinions does not provide enough evidence to classify any of opinions as Positive, negative or neutral.
- Pairs R1-R3 and R2-R3 do not agree on expressed in any opinion, hence kappa scores are 0 in both cases.

Kappa method is Expert rating system; it is not realistic for human raters to rate corpuses in size of millions. Also, kappa defines P_e "agreement by chance" as baseline function. This chance adjustment of kappa statistics supposes that, when not completely certain, raters simply guess; which is a very unrealistic scenario. It specifically fails with imbalanced classes; as it does not take into consideration agreement on rarity class.

1 d) If you wanted to get the correlation between raters (using Pearson's *rho*) what would you do? Then do this.

Pearson's coefficient method finds correlation between 2 Raters who classify N items into C categories; where categories need not be mutually exclusive. Hence, users may rate an opinion as P% positive, N% negative and O% neutral etc. This method is used when expression of Consensus is **quantitative and mutually correlated**.

Pearson's coefficient can be found using built in formula of MS excel: PEARSON(array1, array2). For example: TO calculate correlation between R1 and R2 on POSITIVE views, array 1=[20,90,0] & array 2=[25,80,10] are passed to function.

		Rater 1(R1)	Rater 2(R2)	Rater 3(R3)	Mean rating(Mean)
Opinion 1	Positive	20	25	0	15.00
Opinion 2	Positive	90	80	0	56.67
Opinion 3	Positive	0	10	10	6.67
Opinion 1	Negative	80	80	0	53.33
Opinion 2	Negative	0	10	0	3.33
Opinion 3	Negative	0	0	40	13.33
Opinion 1	Neutral	10	10	90	36.67
Opinion 2	Neutral	10	10	100	40.00
Opinion 3	Neutral	100	80	100	93.33

Correlations	Positive	Negative	Neutral
R1 * R2	1.00	0.99	1.00
R1 * R3	-0.67	-0.50	0.50
R2 * R3	-0.67	-0.60	0.50
R1 * Mean	1.00	0.98	1.00
R2 * Mean	1.00	0.95	1.00
R3 * Mean	-0.63	-0.33	0.54

It is observed that, R1 and R2 have good corelation with each other as well as with Mean expression about categories. Whereas, R3 do not concur with mean expression and naturally, it does not concur with R1 and R2. Though sample space is very limited, one can say that models (or mortal people) behind entity R1 and R2 can provide fair judgment about the sentiment in opinions.

Question 2 Do, some searches and find 3 sentiment lists that are commonly used in previous research. For 2 of these lists, select 10 positive and 10 negative words (randomly). Evaluate each word, discussing whether it is really positive/negative; for each one tries to find a sentential context in which it might be interpreted with the opposite valence.

Following word lists are referred which are commonly used for sentiment analysis:

- AFINN : Each word is assigned associative sentiment integer value ranging between -5 to 5. Though express definition is not provided from makers of dataset; values in range -1 to 1 are considered neutral.
- SentiWords: SentiWords is a high coverage resource containing roughly 155.000 English words associated with a sentiment score included between -1 and 1. Words in this resource are in the form lemma#PoS and are aligned with WordNet lists (that include adjectives, nouns, verbs and adverbs). Scores are learned from SentiWordNet
- SentiWordNet: It is lexical analysis resource based on WordNet The pair (POS,ID) . The values PosScore and NegScore are the positivity and negativity. It assigns positive score, negative score and objectivity [1 - (PosScore + NegScore)] to each sysnet of WordNet.

10 random positive and negative words are chosen for analysis from 2 of these lists each. Each word is followed by its score.

AFINN				SentiWords			
Positive		Negative		Positive		Negative	
aboard	1	censored	-2`	authorization	0.167	autocrat	-0.061
acquitted	2	deafening	-1	bangkok	0.000	bandage	-0.316
amusements	3	dysfunction	-2	ban	0.000	contaminated	-0.023
endorsement	2	idiotic	-3	classical	0.447	feasibility	-0.024
happiness	3	prosecuted	-2	drumbeat	0.066	negligible	-0.240
favourites	2	sarcastic	-2	nature	0.459	old	-0.456
medal	3	scapegoats	-2	noble	0.541	obesity	-0.571
thrilled	5	singleminded	-2	limitless	0.503	intrude	-0.522
rewards	2	unprofessional	-2	vacation	0.895	poison	-0.714
wonderful	4	verdict	-1	perk	0.608	unhappiness	-0.825

Positive words from AFINN (score > 0)

- aboard (Valence - Positive : 1) Typically refers to state of a traveler who is on a ship or train. Example in Opposite valence: "By the time, Watson realized that Sherlock was not aboard train to Brighten".
- acquitted (Valence - Positive : 2) Deals with announcing an accused free of charges. Example in Opposite valence: "Sentiment grew that chemical industrialist having environmental crimes should not be acquitted"
- amusements (Valence - Positive : 3) It is positive word expressing awe of delight. Example in Opposite valence: "We are not amused – said Queen Victoria"
- endorsement (Valence - Positive : 2) Act of promoting or supporting someone's ability. Example in Opposite valence: "Terrorist organization declared that Endorsement to more killings is their motto"
- happiness (Valence - Positive : 3) Happiness is positive feeling of joy. Example in Opposite valence: "After his war efforts, John never cherished his happiness"
- favorites (Valence - Positive : 2) Example in Opposite valence: "sandwich is not my favorite food"
- medal (Valence - Positive : 3) This is a noun referring to a souvenir presented to winners; mostly in games. It hardly has opposite valence usage.
- thrilled (Valence - Positive : 5) Superlative feeling of eminent enjoyment. Example in Opposite valence: "Harry was not thrilled watching Draco in his satin black robes."
- rewards (Valence - Positive : 2) Example in Opposite valence: "Caesar declared that Mutiny would be rewarded with death"
- wonderful (Valence - Positive : 4) Application in Opposite valence cannot be expressed. Rather, a negative word like disastrous can be applied.

Negative words from AFINN (score < 0)

- censored (Valence - Negative : -2) Opposite valence: "Democracy ensures that thoughts are not censored"
- deafening (Valence - Negative : -1) Opposite valence: "deafening crackers announced the arrival of a new year"
- dysfunction (Valence - Negative : -2) Opposite valence application could not be found
- idiotic (Valence - Negative : -3) Opposite valence: "An idiotic smile flashed on his smiling, rejuvenated face."
- prosecuted (Valence - Negative : -2) Opposite valence application could not be found
- sarcastic (Valence - Negative : -2) Opposite valence application could not be found
- scapegoats (Valence - Negative : -2) Opposite valence application could not be found
- singleminded (Valence - Negative : -2) This word should have a positive valence, as it is used to reflect committed and determinant nature of the person. Positive valence: "Edison single mindedly kept working in his train lab"
- unprofessional (Valence - Negative : -2) Opposite valence application could not be found
- verdict (Valence - Negative : -1)) This word should have a positive valence, as verdict is used to reflect decision or more often justice in case of court cases. Positive valence: "Verdict of Nierenberg trials was clean and fair"

Positive words from SentiWords(score > 0)

- authorization (Valence - Positive : 0.16695) Opposite valence: "Secretary must not be authorized to spend money"
- Bangkok (Valence - Positive : 0) This is noun and Opposite valence application could not be found
- Ban (Valence - Positive : 0) Ban is shown neutral/ positive. But the word typically deals with restrictions and prohibiting something. It is more naturally a negative word. Negative valence: "All kind of drugs must be banned from university campus at priority"
- Classical (Valence - Positive : 0.44662) A positive word used in context of music, art, culture. Opposite valence: "Vishy Anand made classical boardgame mistakes against his last game with Carlson"

- Drumbeat (Valence - Positive : 0.06646) This is noun and Opposite valence application could not be found
- Nature (Valence - Positive : 0.45878) Opposite valence: "It is in nature of Rudy the dinosaur to eat you, So always listen to buck"
- noble (Valence - Positive : 0.54113) Opposite valence: "Not a noble deal for England I am afraid – said Churchill"
- limitless (Valence - Positive : 0.50302) Opposite valence: "Limitless hunger for power took away Gandalf's beard"
- vacation (Valence - Positive : 0.89489) Opposite valence application could not be found
- perk (Valence - Positive : 0.60778) Opposite valence: "Do not abuse perks of a sales job, you are valuable to company"

Negative words from SentiWords (score < 0)

- autocrat (Valence - Negative : -0.06110) Opposite valence: "Pasha implemented autocracy for welfare of Turkey"
- bandage (Valence - Negative : -0.31643) This is noun and Opposite valence application could not be found
- contaminated (Valence - Negative : -0.02319) Contamination is associative with making a perfectly good thing impure. Opposite valence application could not be found. Also, a small negative score is surprising. The word should have larger negative score.
- feasibility (Valence - Negative : -0.02386) feasibility is degree of convenience. It should be given a positive score instead of negative one. Positive valence: "Project is feasible now that trade relations with Saudi are open"
- negligible (Valence - Negative : -0.24044) Used to undermine or disregard importance. Opposite valence: "Profound happiness gathered in small nature trails is not negligible"
- old (Valence - Negative : -0.45615) This word is often associated with concept of "ageing", "fraying". Opposite valence: "Sir Lee exclaimed – This is old intelligence! Come on Robert, grail is calling us"
- obesity (Valence - Negative : -0.57126) reference to overweight conditions and bad health. Opposite valence application could not be found.
- Intrude (Valence - Negative : -0.52168) Opposite valence application could not be found.
- poison (Valence - Negative : -0.71373) Opposite valence: "What is your poison today, Jack? – asked Antonio the bartender."
- Unhappiness (Valence - Negative : -0.82533) Opposite valence application could not be found.

Question 3 Bromberg's Sentiment Program:
Have a look at the simple program that does sentiment analysis. So, take a look at the program and see what is happening in the different variables, but adding print statements on its variables.
Andy Bromberg has implemented simple sentiment analysis classifier to predict positive and negative movie reviews. Dataset files rt-polarity.pos.txt and rt-polarity.neg.txt, contain 5331 positive and negative snippets respectively.
Code follows following steps:

1. Text blobs of Positive reviews and Negative reviews read from files are and split into individual review sentences using newline character and loaded in variables posSentences and negSentences
2. A for loop removes split sentences into list of words and punctuations using regex. Then 'pos' or 'neg' label are added to each wordlist and appended to lists posFeatures and negFeatures
3. Using these feature sets, train and test datasets are created with 3:1 split ratio. A naive bayes classifier is trained and metrices accuracy and precision, recall for positive and negative sentences is printed.

3 a) Now consider ways to improve the training. E.g. What might happen on removal of stop words from the inputs

Stop word removal is pre-processing technique used to retain words of significance in given text matter. In case of English stop words, mostly articles, conjunctions and 'wh' words are contained in list of stop words [find ref]. Filtering out stop words serves 2 purpose:

- Saves up valuable runtime space as Information retrieval engine concentrates on corpus of subject matter words
- In applications like search engines and find menus, text processing time is significantly reduced

Most modern web search engines do not remove stop words, but rather exploit language statistics for better understanding of phrase query [E.g. Grammarly].

In case of sentiment analysis, HOWEVER; stop word removal is likely to affect classifier performance. Because, sentiments are expressed as *Subtext* rather than *Context*. And stop words are important in training classifier on understanding the subtext.

Assuming a case with given opinions about cars:

Opinion	**Opinion after stop word removal**
My **chevy** is not **work**ing fine	My chevy working fine
Mercedes needs to **concentrate** on **electric market**	Mercedes needs concentrate electric market
I do not drive **French car**	I drive French car

- Stop word removal may help text analyser engine to concentrate on bold words
- BUT, opinion of text is completely changed for first and third snippet and this is misleading for classifier.

As predicted, classifier performance declines with stop word removal. Following classifier performance is observed using entire corpus of words[training on 7998 instances; test on 2666 instances]:

	maintain entire corpus	removing stop words using **nltk corpus**	removing stop words using **stopword.txt**
accuracy	0.7734	0.7678	0.7682
pos precision	0.7881	0.7721	0.7715
pos recall	0.7479	0.7599	0.7622
neg precision	0.7602	0.7637	0.7650
neg recall	0.7989	0.7757	0.7742

3 b) Implement another solution in the program and report what happens to the precision and recall of the classifier.

Recalling implementation of Naïve bayes algorithm in nltk, it is noticed that :

- Large number of unimportant features; which would have less P(f | label) are considered while prediction.
- Ultimately, unimportant features reduce numerator and add up to denominator; obscuring decision boundary.

$$P(label \mid feature) = \frac{P(label) * P(f1 \mid label) * \ldots . P(fn \mid label)}{sum(l) * [P(l) * P(f1 \mid l) \ldots . * P(fn \mid l)]}$$

Hence, if most informative words forming the feature sets are chosen, it is likely to improve overall accuracy of the algorithm. Andy Bromberg has provided implementation in python 2.7 [source]; which is converted for this problem into python 3 for consistency. This implementation:

- finds positive score and negative score for each word using chi squared test in BigramAssocMeasures package of nltk. These scores are between word frequency of word with respect to entire corpus as against within positive and negative sets respectively.
- Top 1000,5000,10000, 15000 and 20000 words are selected. It is noticed that best 5000 words provide best performance.

	maintain entire corpus	best 1000 word	best 5000 word	best 10000 word	best 15000 word	best 20000 word
accuracy	0.7734	0.7941	0.8507	0.8477	0.8477	0.7734
pos precision	0.7881	0.8151	0.8667	0.8699	0.8652	0.7877
pos recall	0.7479	0.7607	0.8290	0.8177	0.8237	0.7487
neg precision	0.7602	0.7757	0.8361	0.8280	0.8318	0.7605
neg recall	0.7989	0.8275	0.8725	0.8777	0.8717	0.7982

- Lower word corpus [< 1000] do not give significance precision and recall improvement as too little data is available. With larger word corpus [10000<] selected for forming features; unimportant features reduce precision and recall.

Other classifications techniques like SVM, Bernoulli Naïve bayes can also be attempted and optimal algorithm can be found using grid search.