Date / /	

	Date / /
	Assignment - 4
36	a strainer as to meet the contract of the cont
¥	- 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1 - 1
*	Title: - Stemming and feature extraction.
*	esoblem statement: Consider a suitable text data set. Remove stop words
7.2	apply stemming and feature extraction selection techniques to represent
	abcuments as vectors. Classify documents and evaluate precision
	and recall.
j .	the state of many in the state of the state
~ · *	Objective: > Implementation of the problem statement using Python.
	> Remove the woods apply stemming and feature selection.
2711 5	a described outside the formula to be trained in
h	autromes: - students will be able to: -
	> Implement the problem statement using python.
	> Remove stop words apply stemming and feature selection.
eg 11 m	the second of th
*	STW & H/W requirements:
-	=> Fedora 20 Window 10.
	= Julytes Notebook.
- Y	7-970/
₹	Theory:
	stop woods: In computing, stop woods are woods which are fittered out
25/26/1	before or after processing of text. Through there woods usually
	refex to the most common words in a language, there is no
1. e	single universal list of stop woods used by all natural.
- E4	language processing tools and indeed not all topls even use
	such a list - some tools specifically avoid removing those
	stop woxds to suppost phrase sparch.
4-11:57	
	and the state of t



<u>\$</u>	stemming - stemming is the process of reducing inflected for
	cometimes derived words to their word stem, base ox
	root from - generally a written word from. The stern
	need not be identical to the mosphological soot of the
	word, it is usually sufficient that related words map to
Trans	the same stem, even if this stem is not itself a valid
9	scot.
=>	Feature Extraction! In machine learning and statistics, feature
7 4	relection, also known as variable selection, attribute selection
	a variable subset relection, is the process of relecting a
	subset of relevant features (variables, foodictors) for use
	in model construction. Feature selection techniques are used
2	for foux reasons:
D	simplification of models to make them earies to interpret to use.
<u>a</u>)	shortex training times.
3)	To avoid the curse of dimensionality.
4	Enhanced generalization by reducing over fitting (reduction of
	vasiance).
	Postagged centences
	Raw Text (string) (Clist of tuples).
	Entity Detation.
3	sentence segmentation chunked sentences
<u> </u>	sentences (List of stoings)
The state of	Tokenization Relation Detertion 1
	Takenised string v
	Parts of speech Relations
ray Tanga	Tagging. (List of tuples).
	Fig. Feature Extraction Architecture





,	*	<u>Precision</u> : Precision mentions the proportions of the positive
		identifications that was actually correct. It means the percentage
		of your results that are relevant.
		U W
		Precision: True positive
		True Positive + False Positive
	女	Recall 1- Recall mentions the proportions of actual positive that were
		identified consectly · Recall refers to the percentage of total
- / -		relevant results correctly classified by your algorithm
		observed to the second of the
		Recall = True Positive
		True Positive + false Negative
		Ų
	=*	Conclusion. We have successfully removed stop words, applied stemming
		and feature selection techniques to represent documents as -
		vertoses and also calculated pserision & xerall.
/		
	10 H	
- 100	. 11 · .	
	() () () () () () () () () ()	
1	-	
	1 47	
	77	
5 7 40	- In	