ID:81127139

DATA SCIENCE - II Homework - 3

Q1)

	Name: Abhisher. Milind. Patwardhan	No. of the last
(3)	ID: 811271359	
	Data Science - II - HomeWork - 3	
7	Downwell 2: TE(Today) " 1/9	
1	Text Retrieval ()	
	The (mount) AT	
	· Today the Dawys Wor!	
	· Dawge nave Won today	
	· Dawys the Champion.	
	· The Dawys news today	
1	e () pane () Tr	
1)	Preprocessing (2014) AT	
	Today Dawgs Won (1)	
	2) Dawys have Won today 3) Dawys Champion	
	4) Dangs news Today 7 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
	There's round	
2)	Term-Dowment incidence Matrix	
	(2015) - 77.	
->	_ Downent -> DIA D2 P3 D4	
	Term to	
	Today (201) may south in much souther	4
	Daugs 1 1 1 1	
1	Wonthemass folds plant pa - 0 (1) 195	
	Champion 1 0	
	news 0 0 0	
	TDF (roday) = 109 (413) . 109 (413)	
3)	Implement + TF-11DF	
-	> term Frequency (TF) = No. of times teem t appears in	doe d
	7 Texm Frequency (TF) = No. of times teem t appears in Total No of items in doc of	1.
	3 DF (thempion) = 10g (4/1) = 10g(4)	
	Downen 1: TF(Today) = 1/3	
-0	TF (Dawgs) = 1/3 TF (Won) = 1/3	
	TF (100) = /3	
	TF (champion) = 0 TF (news) = 0	
	TF (news) = 0	ALCOHOL:

```
Dowment 2: TF(Today) = 1/3
TF(Dawqs) = 1/3
                TF (won)
                                = 1/3
                TF (News)
                                 - Toldy IN O Par
                TF (Champion) 2 000 1000 1
               TF (Today) = 0
Dowment 3:
                TF (Dawgs) = 1/2
TF (Won) = 0
                TF (News) _ 1000 more &
                TF (champion) = 1/2 / (e
                 hongrach some (e
 Dowment 4: TF (Today) = 1/3

TF (Dawgs) = 1/3

TF (Won) = 0
       TF (won) = 1/3

TF (champion = 0.
Invoise Dowment Frequency (IDF)
IDF(t) = log ( Total No. of documents with term't
IDF (Today) = log (4/3) = log (4/3)

IDF (Dawgs) = log (4/4) = 0

IDF (WON) = log (4/2) = log (2)

IDF (News) = log (4/1) = log (4)

IDF (Champion) = log (4/1) = log (4)
                  t (contempts) 31
                  ( stress ) 71
```

ID:81127139

	A TF- IDF	Mattix is	as follows:	probability of	
	Dowment -	Di samo	D ₂	.D3	04
	Team 1	11 1 11 1	(1)	13/13	1
	Today	1/3.109(4/3)	1/3.109(4/3)	0	1/3.109(4/3)
	Dawgs	4. 0	0	0	0
	Won	1/3. log (2)	1/3. 109(2)		0
	champion	0	0	1/2.109(4)	0
	News	0	0	0	1/3. 109(4)
			J. J. St.		
4)	Juverted In Joken Se		iven que		Won -
		1	00%		
	7	erm	Do coment I	D	
	Dawgs Won		POT 1	jeanoitia (
			1	13	
			Lineary	rivers Freeze	
	Dawqs		- 2	anye	
	Won		2	Fo mpion	
	Today		2		
	0 5-5		3 3	unte	
	enampion		3	mon	
		Dawgs	4		
		News	4	: weed !	4
		Today	4	1	
	A personal se		pout min	10) 1114 14	4
				nana aine	
			111111		
	11.5 - E				
0	11.5 - 5	- 1 - 1		Penerg!	

ID:81127139

		(3)
	2) Sorting by term	S: 27 KHOM 302 - 57
	0 0	
	Term ?	bocoment 10 : manual.
	- Champion	3
[818] pol. e]	Dawgs	(slaye)-st paper
	Dawgs	2 1
	Dawgs	(6)301 - eV Galve
7	Dawige	4 a migmans
(2) pool of	news	4) coels
	Today	1.
17 753	Today	2 continued (1)
	Today	4
	won	Ingress of the Control
	Won	2
	all amount	and leaves to the same of the
•	3) Dictionary of Posti	9:
	0	- Countil
	Texm Frequency	-> Posting Lists
	Dawgs 4	→ 1→2→3→4
	champion 1.	→ 3
	news 1	→ 4
	today 3 won 2	\rightarrow 1 \rightarrow 2 \rightarrow 4
	won ²	- 2
		Thereof.
	4) Query: "Dawy	75 Won 100
	0	The bolt with the same
-	y we will locate Dan	195 & "Won" Separately & nativere
	their respective pos	gs" & "Won" Separately & notrieve tings.
	Dawgs 4 ->	$1 \rightarrow 2 \rightarrow 3 \rightarrow 4$
	Inter 1 2	1 -> 2
	Won 2	
The second	AND DESCRIPTION OF THE PERSON	

		<u>(3)</u>
	0	
		Finding the Intersection of Document 10's for all terms in query: $[1,2,3,4] \cap [1,2] = [1,2].$
		Hence Document 1 & 2 are netziered documents for given query "Dawys won".
(5)	Explain advantages of inverted index over term-domment incidence matrices for organizing tents?
1	0	· Inverted index allows fast retrieval of documents Containing Specific terms because index directly
		maps terms to the documents they appear in. 2) Scalability:
		· Inverted indexes are generally more scalable than term-downent incidence matrices due to their distributed and parallel nature
		3) Dynamic Updation: . Inverted index can be easily updated when new documents are added. Whereas term-document incidence matrices may require recalculations or resiring operations.
	0	4) space Efficiency: . Inverted index stores only the vuique terms and their associated document references. Hence it is more
The Party of		space efficient than term-document incidence matrices.

Q 2)

```
PS C:\Users\abhis\OneDrive - University of Georgia\Desktop\Data Science\Iris_clf_mask_code> python -u "c:\User
Model: LogReg
Train Accuracy=0.9429 | F1=0.9444 | AUC=0.9812
Test Accuracy=0.9333 | F1=0.9231 | AUC=0.9864

Model: MLP
Train Accuracy=0.9857 | F1=0.9867 | AUC=0.9992
Test Accuracy=0.9667 | F1=0.9630 | AUC=0.9955

Model: NaiveBayesian
Train Accuracy=0.9286 | F1=0.9231 | AUC=0.9321

Train Accuracy=0.9333 | F1=0.9231 | AUC=0.9321

Model: DecisionTree
Train Accuracy=0.9667 | F1=0.9867 | AUC=0.9348
Test Accuracy=0.7667 | F1=0.7586 | AUC=0.9760

PS C:\Users\abhis\OneDrive - University of Georgia\Desktop\Data Science\Iris_clf_mask_code>
```

Conclusions:

The results indicate that all models perform well on the training set, with Logistic Regression, MLP, and Decision Tree achieving high accuracy, F1 score, and AUC. However, on the test set, MLP stands out as the best-performing model with the highest accuracy, F1 score, and AUC, indicating its superior generalization ability. Logistic Regression also performs reasonably well on the test set, showing robustness in its predictions. Naive Bayesian and Decision Tree models exhibit moderate performance, with Decision Tree showing signs of overfitting as evidenced by the significant gap between training and test accuracies. In conclusion, while MLP emerges as the top-performing model, Logistic Regression demonstrates competitive performance with good generalization capabilities, making it a reliable choice when interpretability and simplicity are valued.