CMPE256 - Assignment
ML Based Spam filter.

3rd Aim. Apply TF and IDF and develop Spam Filter mode for documents.

Span Dictionary:
free, click here, visit, open attachment, call this number, money, out, extra, offer, available, Pension, opportunity, chance, investment, Pension.

Solution: t_1 = free
t_2 = click here
t_3 = visit
t_4 = open attachment
t_5 = call this number
t_6 = money
t_7 = out
t_8 = extra
t_9 = offer
t_{10} = available
t_{11} = pension
t_{12} = opportunity
t_{13} = chance
t_{14} = Investment

Terms \ Documents.
free
click here
visit
open attachment.

document / term.

	Terms	d_1	d_2	d_3	d_4	d_5	d_6
t_1	free	1	1	1	2	0	0
t_2	clickhere	0	0	0	1	0	0
t_3	visit	1	1	0	0	0	0
t_4	openattachment	0	0	0	0	0	0
t_5	call this number	0	0	0	0	0	0
t_6	money	0	0	0	0	1	0
t_7	out	0	0	1	0	0	0
t_8	extra	0	0	0	0	2	0
t_9	offer	1	1	0	0	2	4
t_{10}	available	1	1	0	0	0	0
t_{11}	pension	0	0	2	0	0	0
t_{12}	opportunity	0	0	0	0	1	0
t_{13}	chance	0	0	0	1	1	0
t_{14}	Investment.	0	0	0	0	0	1

Term frequency table:

	d_1	d_2	d_3	d_4	d_5	d_6
t_1	1	1	1	1.11	0	0
t_2	0	0	0	1	0	0
t_3	1	1	0	0	0	0
t_4	0	0	0	0	0	0
t_5	0	0	0	0	0	0
t_6	0	0	0	0	1	0
t_7	0	0	1	0	0	0
t_8	0	0	0	0	1.11	0
t_9	1	1	0	0	1.11	1.204
t_{10}	1	1	0	0	0	0
t_{11}	0	0	1.11	0	0	0
t_{12}	0	0	0	0	1	0
t_{13}	0	0	0	1	1	0
t_{14}	0	0	0	0	0	1

IDF calculations below.

IDF: $t_1 = 0.2430$
$t_2 = 0.845$
$t_3 = 0.5440$
$t_4 = 0$
$t_5 = 0$
$t_6 = 0.845$
$t_7 = 0.845$
$t_8 = 0.845$
$t_9 = 0.243$
$t_{10} = 0.5440$
$t_{11} = 0.845$
$t_{12} = 0.845$
$t_{13} = 0.5440$
$t_{14} = 0.845$

* Tf-IDF calculations below for spam words in all documents.

	d_1	d_2	d_3	d_4	d_5	d_6
t_1	0.2430	0.2430	0.2430	0.9379	0	0
t_2	0	0	0	0.845	0	0
t_3	0.5440	0.5440	0	0	0	0
t_4	0	0	0	0	0	0
t_5	0	0	0	0	0	0
t_6	0	0	0	0	0.845	0
t_7	0	0	0.845	0	0	0
t_8	0	0	0	0	0.9379	0
t_9	0.2430	0.2430	0	0	0.9379	0.2925
t_{10}	0.5440	0.5440	0	0	0	0
t_{11}	0	0	0.9379	0	0	0
t_{12}	0	0	0	0	0.845	0
t_{13}	0	0	0	0.5440	0.5440	0
t_{14}	0	0	0	0	0	0.845