Methods tried before arriving to Semi Supervised

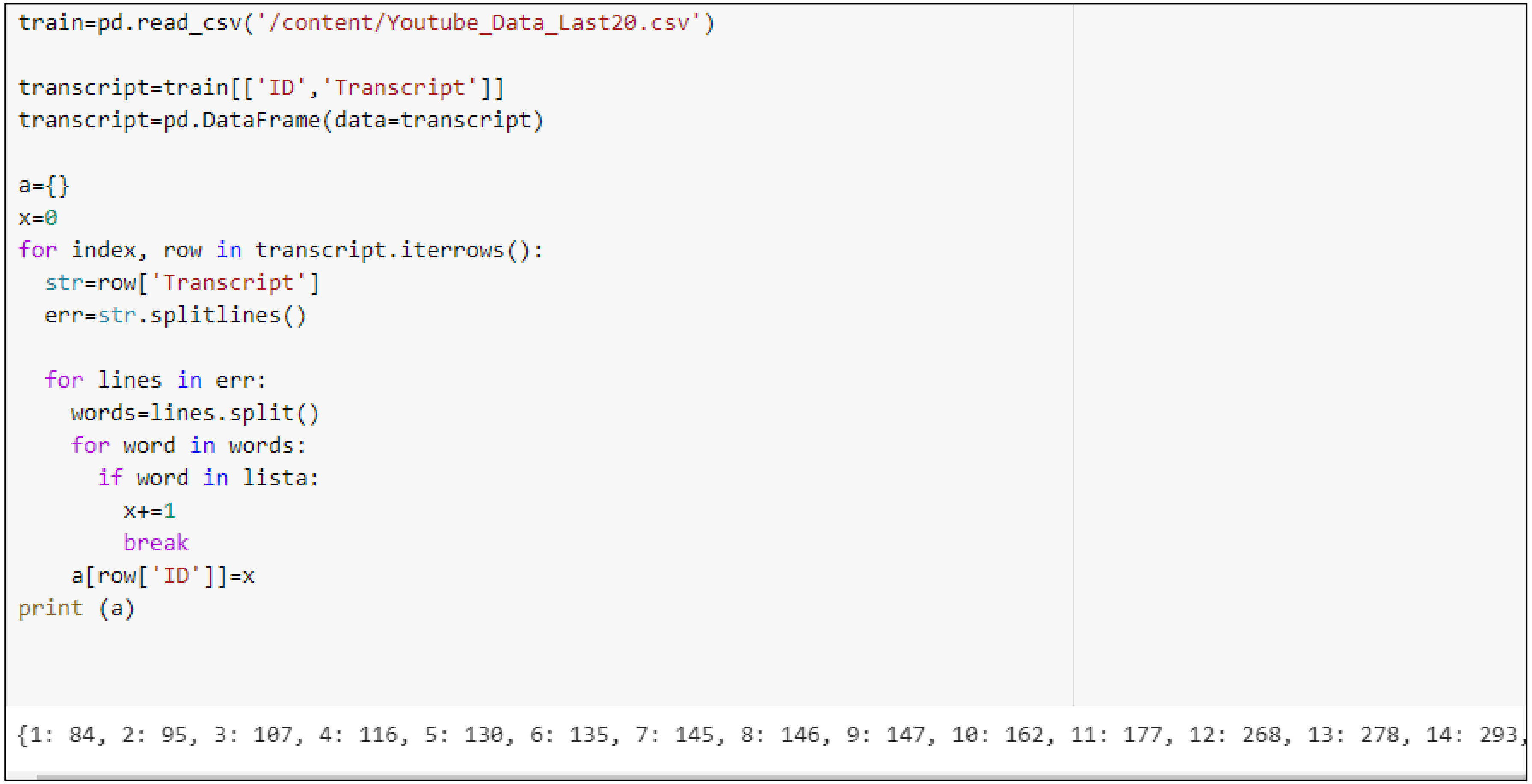
By:

Manthiramoorthy Cheranthian

Considering all the documents, here is the list of 20 most important words:

- Stage 1
- Stage 4
- Lymph
- Nodes
- Surgery
- cells
- treat
- hormonal
- invasive
- patients
- Treat
- Factors
- Risks
- Genes
- Affects
- Health-care
- Cancer
- Changes
- Time

Obtaining the Line counts having those 20 important words.



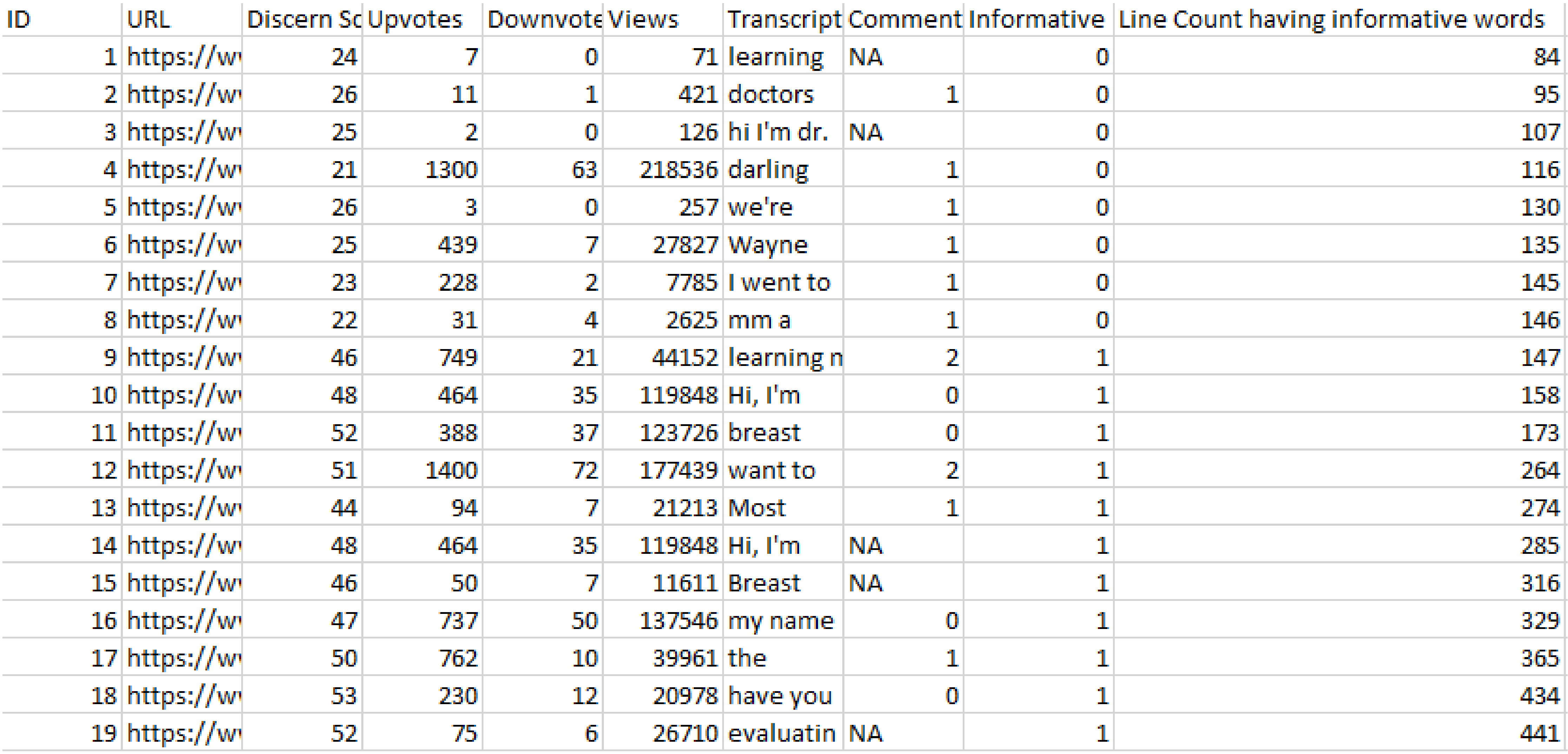
```
train=pd.read_csv('/content/Youtube_Data_Last20.csv')

transcript=train[['ID','Transcript']]
transcript=pd.DataFrame(data=transcript)

a={}
x=0
for index, row in transcript.iterrows():
  str=row['Transcript']
  err=str.splitlines()

  for lines in err:
    words=lines.split()
    for word in words:
      if word in lista:
        x+=1
        break
    a[row['ID']]=x
print (a)
```

```
{1: 84, 2: 95, 3: 107, 4: 116, 5: 130, 6: 135, 7: 145, 8: 146, 9: 147, 10: 162, 11: 177, 12: 268, 13: 278, 14: 293,
```

Classifying the videos based on, Line counts:

Checking for the number of lines from the video transcripts based on the 20-word reference identified from the previous step.

Classifying the video as Informative(1) or Misinformative(0) based on the number of count of informative lines present in the video



ID	URL	Discern Sc	Upvotes	Downvote	Views	Transcript	Comment	Informative	Line Count having informative words
1	https://w	24	7	0	71	learning	NA	0	84
2	https://w	26	11	1	421	doctors	1	0	95
3	https://w	25	2	0	126	hi I'm dr.	NA	0	107
4	https://w	21	1300	63	218536	darling	1	0	116
5	https://w	26	3	0	257	we're	1	0	130
6	https://w	25	439	7	27827	Wayne	1	0	135
7	https://w	23	228	2	7785	I went to	1	0	145
8	https://w	22	31	4	2625	mm a	1	0	146
9	https://w	46	749	21	44152	learning n	2	1	147
10	https://w	48	464	35	119848	Hi, I'm	0	1	158
11	https://w	52	388	37	123726	breast	0	1	173
12	https://w	51	1400	72	177439	want to	2	1	264
13	https://w	44	94	7	21213	Most	1	1	274
14	https://w	48	464	35	119848	Hi, I'm	NA	1	285
15	https://w	46	50	7	11611	Breast	NA	1	316
16	https://w	47	737	50	137546	my name	0	1	329
17	https://w	50	762	10	39961	the	1	1	365
18	https://w	53	230	12	20978	have you	0	1	434
19	https://w	52	75	6	26710	evaluatin	NA	1	441