The tools used in the comment filtering module are python libraries centered around handling and manipulating the data obtained from the crawlers. The filtering model uses the pandas library to read the raw data from a csv form of the output json data from the crawlers. The filtering model uses the nltk library to clean our data and to tokenize the data and remove stopwords.

The outputs collected from the crawlers are all initially in json files. These are then converted into a uniformly structured csv type file. The main technologies used is a form of the bag of words model which is used it to analyse the importance of each generated token within the context of the extracted data.

In the implementation, note that input is the csv file and output is a cleaned and appended list of comments and replies. The comments and replies are first read into a dataframe following which the following cleaning methods are applied:

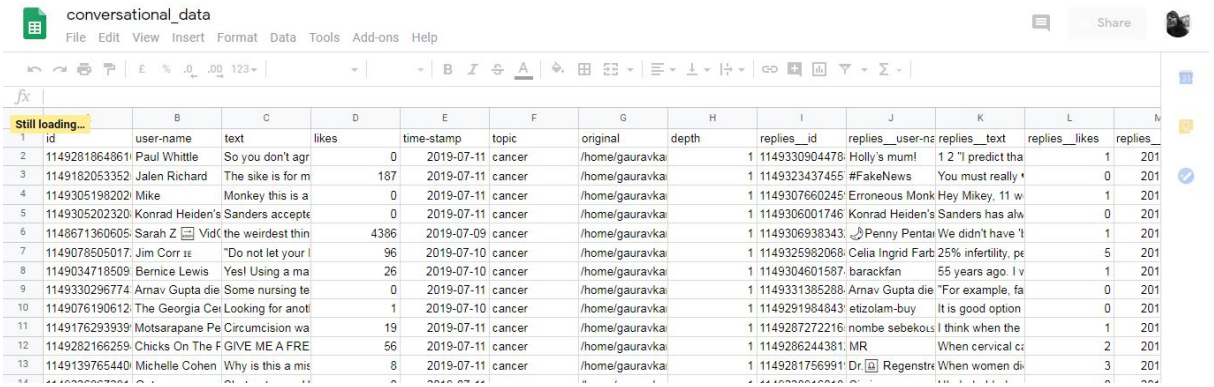
1. Convert to lowercase
2. "[/(){}\[\]\|@,;" symbols are replaced by a space
3. "^0-9a-z +_" symbols are removed
4. Stopwords are removed according the 'english' stopwords from the nltk stopwords library

Then the first 30 comments are analysed to generate a list of tokens and their frequencies are counted. Tokens with "#" are given high preference and a high frequency list of words is taken as a subset of the original list. Then all the comments and replies from the dataframe are cross-referenced. Entries which do not contain any of the words in the list of high frequency words are rejected. The remaining entries which have been filtered are the output.

For eg:

Consider a twitter crawler output for the search term "cancer".

1. Convert .json files to a standard .csv file.



	id	user-name	text	likes	time-stamp	topic	original	depth	replies__id	replies__user-na	replies__text	replies__likes	replies_
2	1149281864861	Paul Whittle	So you don't agr	0	2019-07-11	cancer	/home/gauravka	1	1149330904478	Holly's mum!	1 2 "I predict tha	1	201
3	1149182053352	Jalen Richard	The sike is for m	187	2019-07-11	cancer	/home/gauravka	1	1149323437455	#FakeNews	You must really ♥	0	201
4	1149305198202	Mike	Monkey this is a	0	2019-07-11	cancer	/home/gauravka	1	1149307660245	Erroneous Monk	Hey Mikey, 11 w	1	201
5	1149305202320	Konrad Heiden's	Sanders accepte	0	2019-07-11	cancer	/home/gauravka	1	1149306001746	Konrad Heiden's	Sanders has alw	0	201
6	1148671360605	Sarah Z	the weirdest thin	4386	2019-07-09	cancer	/home/gauravka	1	1149306938343	Penny Pentai	We didn't have 'b	1	201
7	1149078505017	Jim Corr	"Do not let your l	96	2019-07-10	cancer	/home/gauravka	1	1149325982068	Celia Ingrid Farb	25% infertility, pe	5	201
8	1149034718509	Bernice Lewis	Yes! Using a ma	26	2019-07-10	cancer	/home/gauravka	1	1149304601587	barackfan	55 years ago. I v	1	201
9	1149330296774	Arnav Gupta die	Some nursing te	0	2019-07-11	cancer	/home/gauravka	1	1149331385288	Arnav Gupta die	"For example, fa	0	201
10	1149076190612	The Georgia Cer	Looking for anot	1	2019-07-10	cancer	/home/gauravka	1	1149291984843	etizolam-buy	It is good option	0	201
11	1149176293939	Motsarapane Pe	Circumcision wa	19	2019-07-11	cancer	/home/gauravka	1	1149287272216	nombe sebekou	I think when the	1	201
12	1149282166259	Chicks On The F	GIVE ME A FRE	56	2019-07-11	cancer	/home/gauravka	1	1149286244381	MR	When cervical ca	2	201
13	1149139765440	Michelle Cohen	Why is this a mis	8	2019-07-11	cancer	/home/gauravka	1	1149281756991	Dr. Regenstre	When women di	3	201

2. Extract the comments and replies columns into a dataframe.

	text	replies__text
0	So you don't agree that the subgroup from Stu...	1 2 "I predict that Gardasil [vaccine] will b...
1	The sike is for my feeling being hurt ! My dau...	You must really ♥ her when Immunization prev...
2	Monkey this is a completely pointless tweet. ...	Hey Mikey, 11 women died yesterday from cervi...
3	Sanders accepted an endorsement from Wayne La...	Sanders has always placed class above both ge...
4	the weirdest thing about gender reveal parties...	We didn't have 'baby showers' in the UK until ...

```
<class 'pandas.core.frame.DataFrame'>
```

3. Clean the raw data.
4. Tokenize the entries in the dataframe and calculate the words frequency for each unique word.

```
let ------------- 2
cause ------------- 1
whether ------------- 1
tests ------------- 1
backed ------------- 1
gender ------------- 1
right ------------- 1
stick ------------- 1
biology ------------- 1
nobody ------------- 1
could ------------- 1
get ------------- 3
career ------------- 1
understanding ------------- 1
uk ------------- 3
diseases ------------- 2
parties ------------- 1
incidence ------------- 1
endorsement ------------- 1
available ------------- 2
health ------------- 1
seanad ------------- 1
hammer ------------- 1
```

5. Generate a high frequency word list as a subset of the former.

```
pap ------3
vaccinations ------3
years ------3
#hpvvaccine ------6
cancer ------9
give ------2
smear ------2
life ------3
evidence ------2
say ------2
take ------4
like ------3
risk ------3
#vaccine ------6
dey ------2
got ------2
change ------2
also ------2
go ------4
passed ------2
```

6. Cross refer the remaining entries and filter those that have at token in the high frequency word list.

```
Original number of comments and replies :1480
Extracted number of comments and replies :1212
Rejected number of comments and replies :268
```

7. Output is cleaned, filtered data.