7CCSMDM1 Data Mining – Coursework 2

Chan, Hsin-Ju (1920253)

1 WORKING WITH A TEXT CORPUS – NLP

Q1. The TF-IDF formula is from Russell & Norvig, which is $IDF(T) = \log\frac{M-DF(T)+0.5}{DF(T)+0.5}$

	polarity	subjectivity	word counts	most_freq_term	normalised freq	tf	idf	tf_idf
benjamin-bunny.txt	0.086364	0.387471	1373	bunny	0.0095	0.0165	1.098612	0.018127
ginger-pickles.txt	0.085720	0.423212	1420	shop	0.0092	0.0157	1.098612	0.017248
jeremy-fisher.txt	0.090068	0.414693	1305	fisher	0.0115	0.0194	1.558145	0.030228
jungle-book.txt	0.040444	0.451593	20082	jungle	0.0065	0.0271	1.558145	0.042226
just-so-stories.txt	0.155150	0.489808	11226	wild	0.0093	0.0332	0.223144	0.007408
kim.txt	0.081631	0.471863	40320	sahib	0.0056	0.0222	1.098612	0.024389
man-who-would-be-king.txt	0.068823	0.428368	5892	king	0.0073	0.0160	0.223144	0.003570
peter-rabbit.txt	0.088911	0.401538	1326	rabbit	0.0068	0.0114	0.223144	0.002544
puck-of-pooks-hill.txt	0.089953	0.458353	22656	puck	0.0055	0.0206	2.302585	0.047433
squirrel-nutkin.txt	0.086166	0.372316	1466	squirrels	0.0082	0.0142	2.302585	0.032697

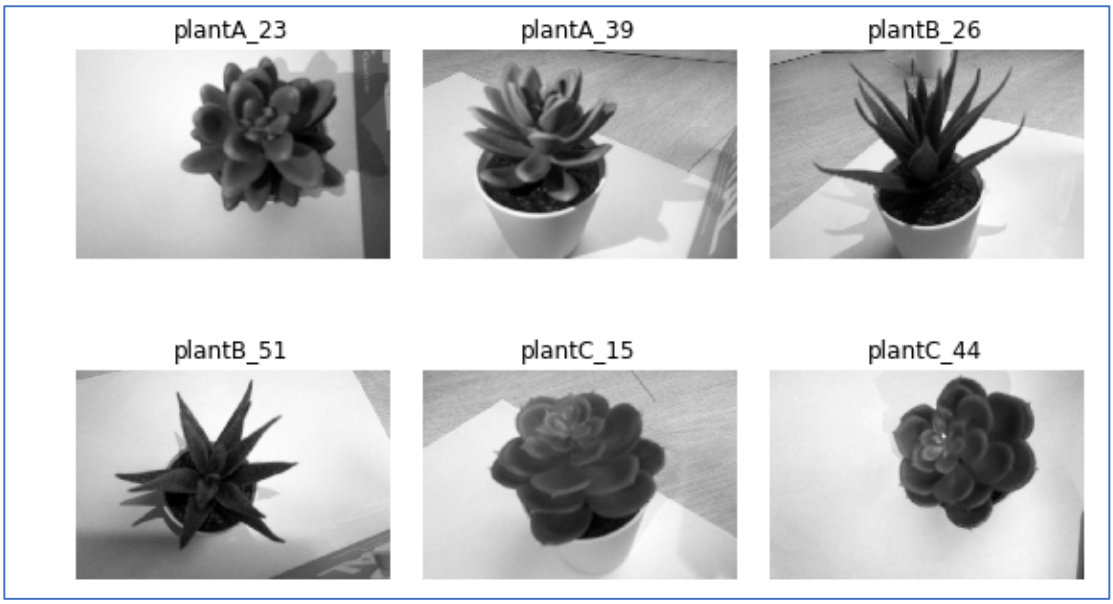
Q2.

Since polarity, subjectivity, word counts and most frequent term are directly related to the document itself, and the normalized frequent, tf, idf and tf-idf are related to the most frequent term found from the document; therefore, I chose the first four statistics as the feature. With the first four statistics, document which has lower word counts is labelled by author Potter, others which has over 2000 words are from Kipling. Also, the most frequent term that are related to animals are from Potter, others are from Kipling.

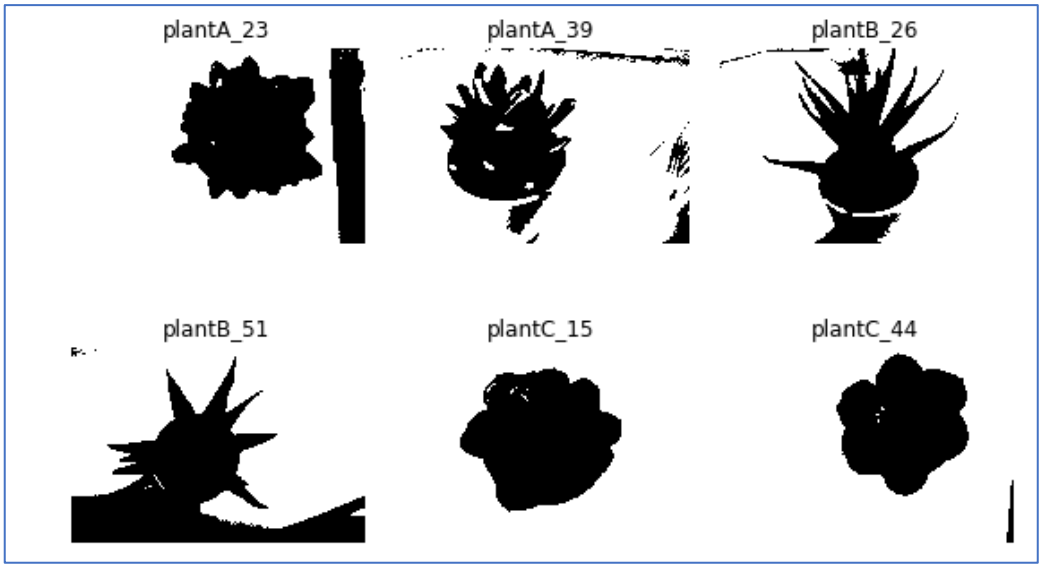
After read in the documents, I split data into training and test dataset. The size of data sets can separate into two: verses and labels. For the former one, it has 5 training dataset and 5 test sets, for the latter one, it has also 5 training dataset and 5 test sets. By using Naïve Bayes classifier in scikit-learn, the true positive I've got is 3 and the true negative I've got is 2, which has no error prediction.

2 WORKING WITH AN IMAGE DATA SET – IMAGE PROCESSING

(a) Convert the image object from RGB format to greyscale.

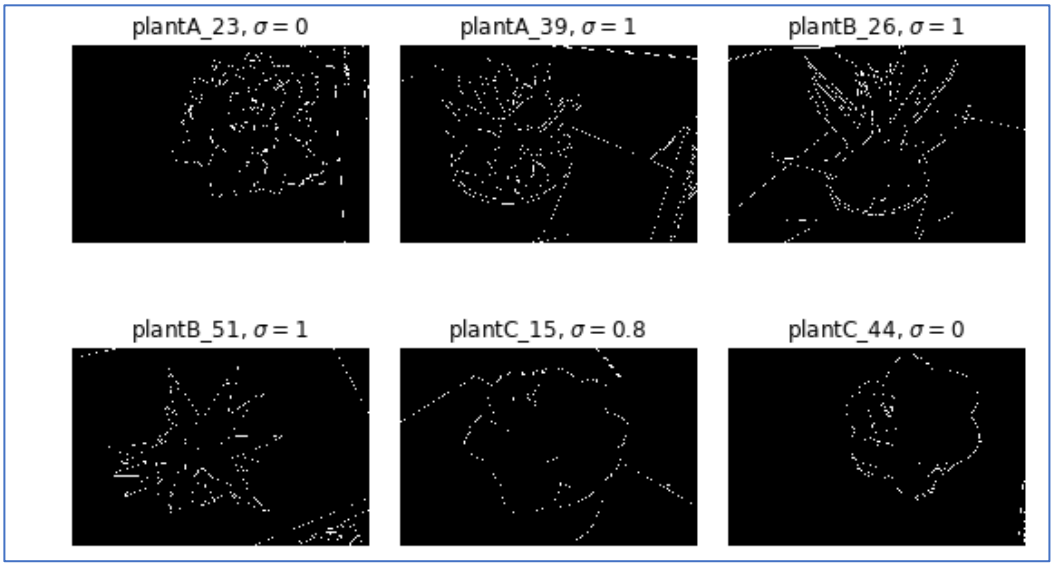


(b) Convert the image from greyscale to Black and White version, using Otsu threshold

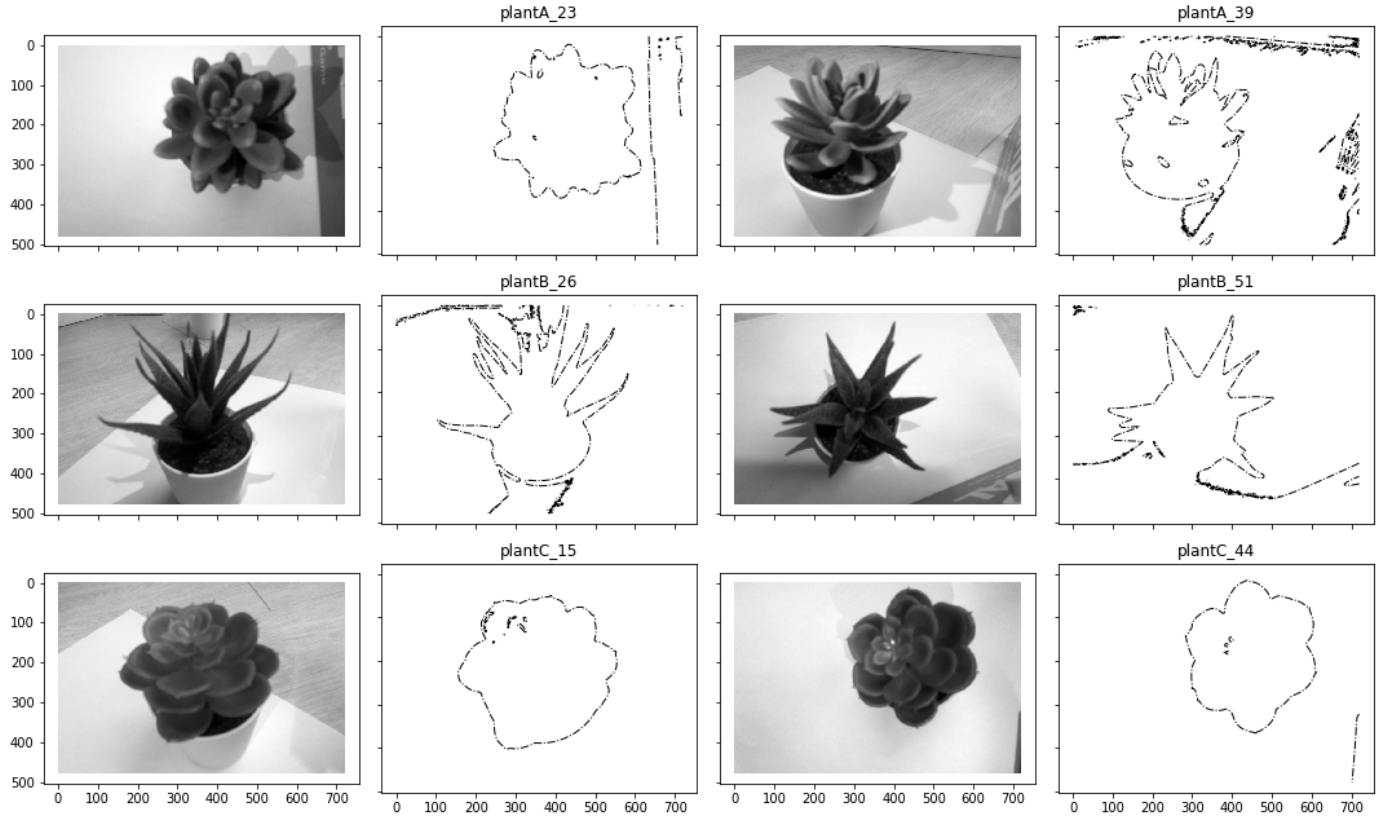


(c) Detect edges using Canny algorithm for each image.

Since different sigma shows different for each image, I choose the best one for these images, which can obviously show the edges of the plants.



(d) Detect contours in each image.



(e) Detect the green in each image

(f) Detect straight lines in each image using the probabilistic Hough transformation.

If I have a long line length, the picture will only show several lines but can not represent the plants, so I set the line length to 5 and line gap to 3, which can obviously show the plants.