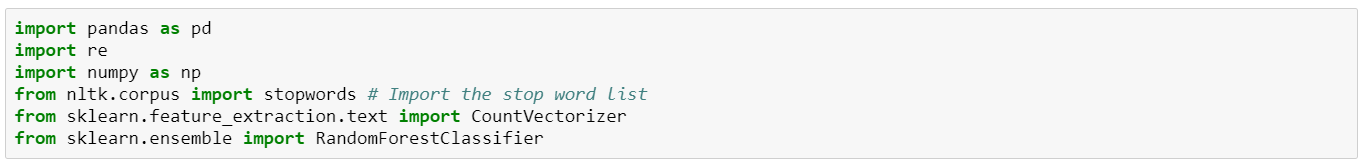
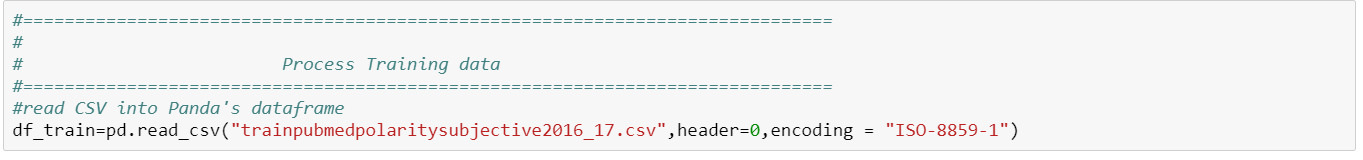
# **Cross Validation of Pubmed Abstracts using Bag of Words**

1. **Import modules**: Using pandas for dataframe processing



1. **Read CSV:** Unprocessed Training file with Abstracts classified by Pos, Neg, Neutral



1. **File Information:**
2. Number of records



# Get the number of abstracts

     2966

1. Column Names

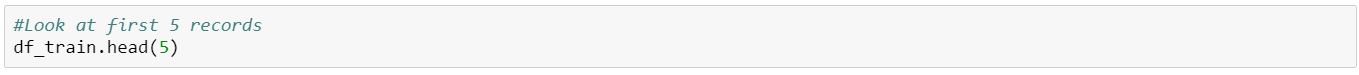


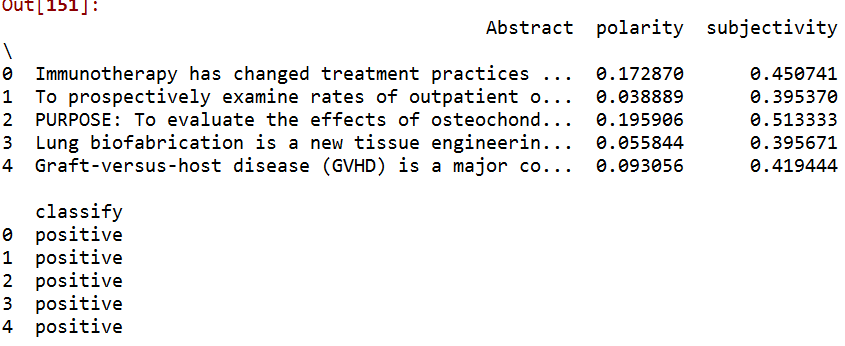
#get column names

df\_train.columns.values

#array(['Abstract', 'polarity', 'subjectivity', 'classify'], dtype=object)

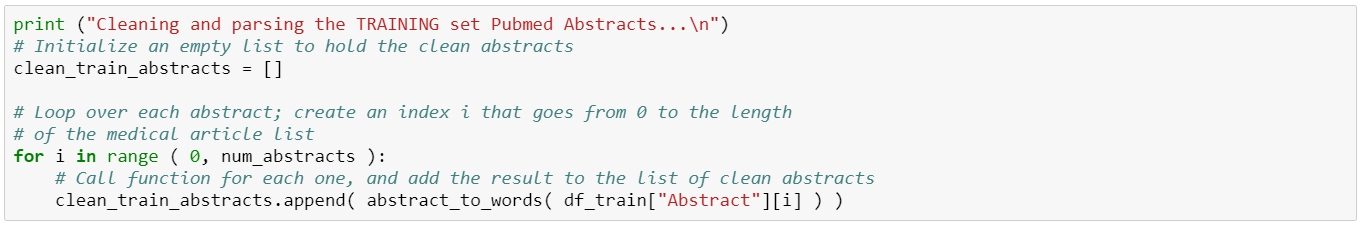
1. Look at first 5 rows



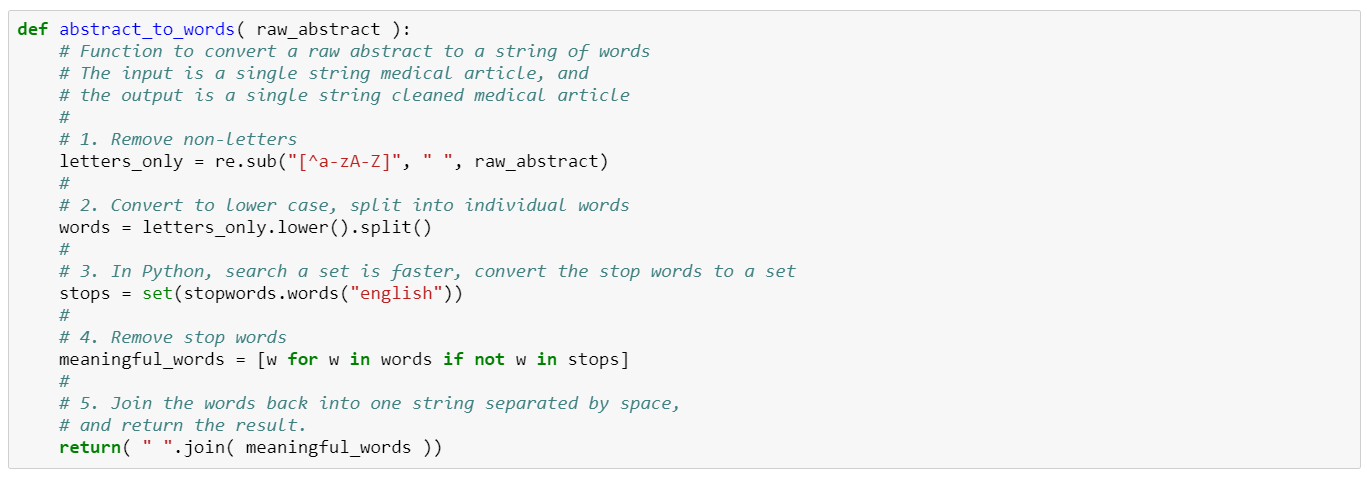


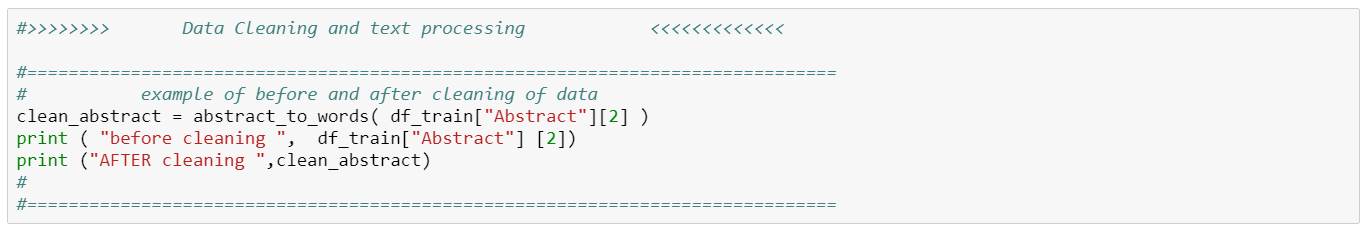
1. **Data Cleaning:**

Process each row



1. Remove none letters – use re (regular expression) to remove numbers and punctuation
2. Convert all letters to lower case
3. Search stop words
4. Remove word if found in stopwords



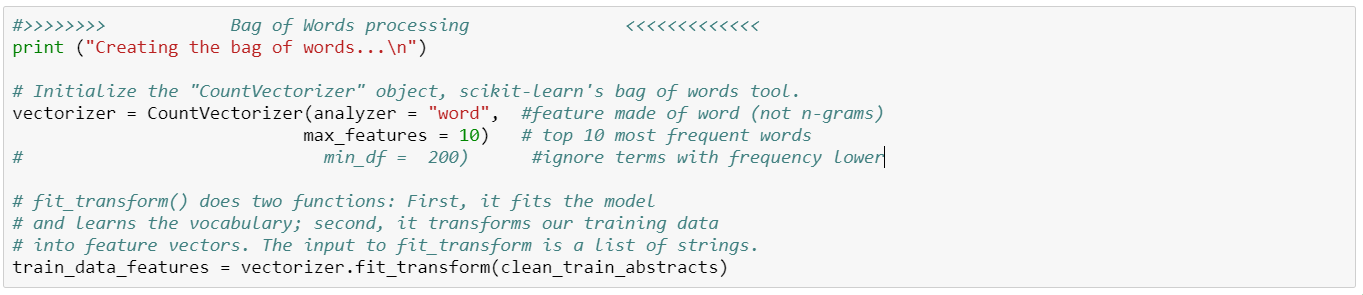


before cleaning PURPOSE: To evaluate the effects of osteochondral autograph transplantation (OAT) mosaicplasty as a concomitant procedure with opening-wedge valgus high tibial osteotomy (HTO) for spontaneous osteonecrosis of the medial femoral condyle (MFC) on clinical outcomes and cartilage status in comparison with bone marrow stimulation (BMS) by drilling and to assess the relation between lesion size and postoperative cartilage status.METHODS: Fifty-eight patients with spontaneous osteonecrosis of the MFC were treated with opening-wedge HTO and a concomitant procedure of BMS (28 patients) or OAT (30 patients)…………

AFTER cleaning purpose evaluate effects osteochondral autograft transplantation oat mosaicplasty concomitant procedure opening wedge valgus high tibial osteotomy hto spontaneous osteonecrosis medial femoral condyle mfc clinical outcomes cartilage status comparison bone marrow stimulation bms drilling assess relation lesion size postoperative cartilage status methods fifty eight patients spontaneous osteonecrosis mfc treated opening wedge hto concomitant procedure bms patients oat patients

5)**Portion of Bag of Words Processing:** use Scikit-learn CountVectorizer

a) max\_features use the top frequently used words



Creating the bag of words...

(2966, 10)

3589 bone

1924 cd

4691 cell

5448 cells

2087 disease

3231 marrow

6321 patients

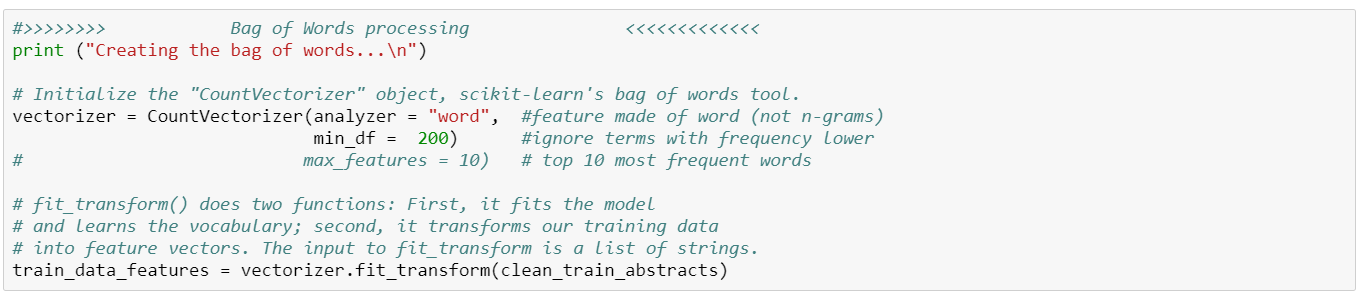
2583 stem

3068 transplantation

2086 treatment

**Bag of Words:** Fine tune for Random forest

b) min\_df ignores word frequency below this threshold



Creating the bag of words...

(2966, 252)

324 activation

389 activity

1098 acute

305 addition

361 adult

666 age

271 aim

882 allogeneic

800 also

379 although

379 among……..

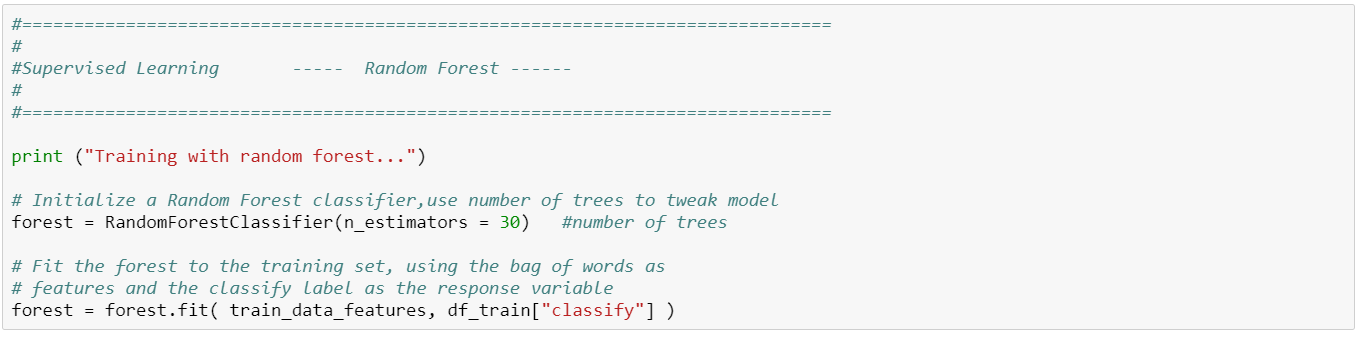
858 analysis

343 analyzed…….

6)**Full Bag of Words Processing:**



7)**Random Forest:** use Scikit-learn RandomForestClassifier. Fits a number of decision tree classifiers on sub-samples and averages response. Use bag of words to determine how to classify each abstract as Positive, Negative or Neutral in sentiment.

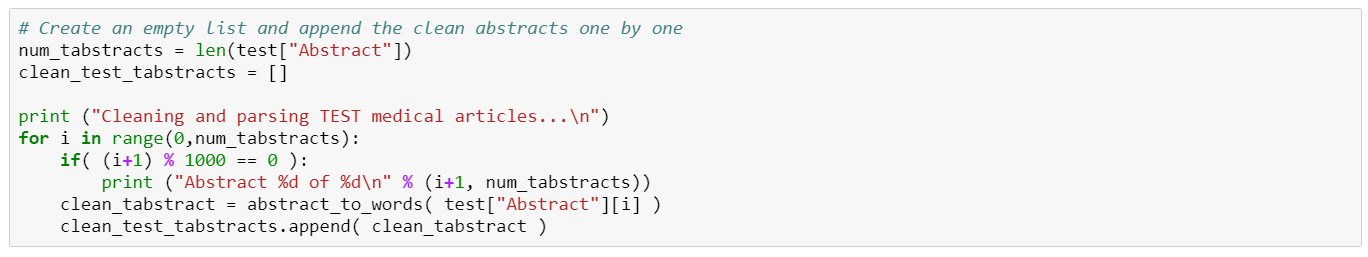


COMPLETED TRAINING

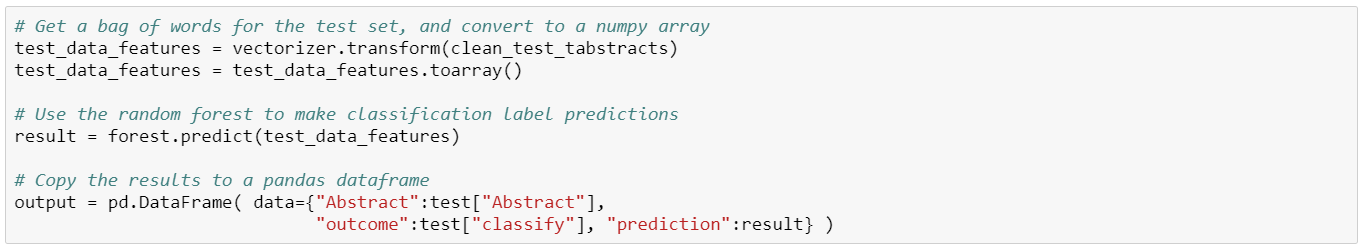
7)**Process Test data for Cross Validation**

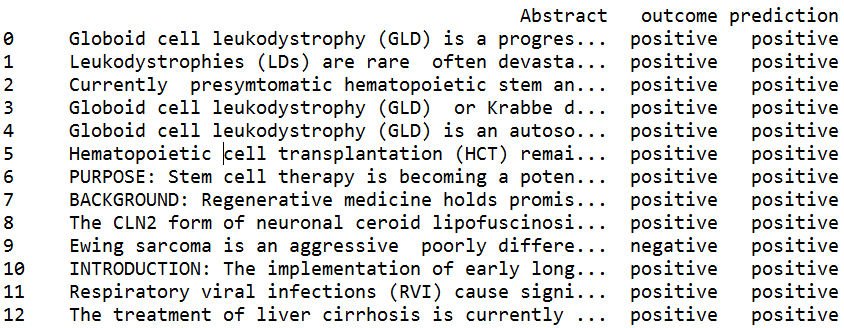


8)**Clean Test Data:** same process used for training data

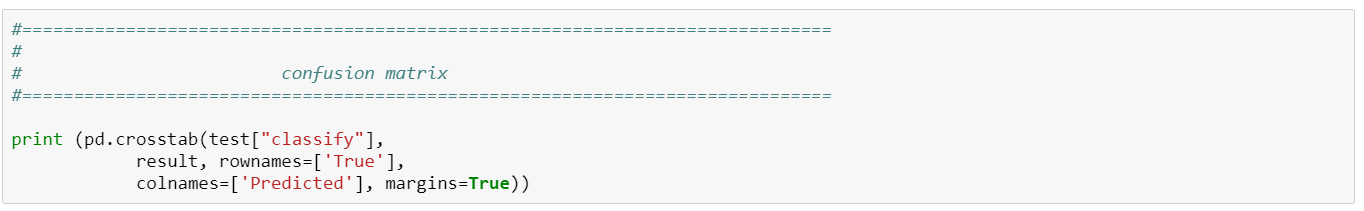


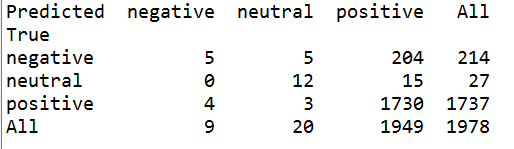
8)**Bag of Words processing and Random Forest**





9)**Assess outcome:** Confusion Matrix





10)**Write to CSV:**

