EDA

September 21, 2021

[1]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[2]:
```
import os
import pathlib
from pathlib import Path
os.chdir("/content/drive/My Drive/Classroom/projects/BERT")
!ls -l
```

```
total 1247
drwx------ 2 root root   4096 Sep 12 15:56 Data
-rw------- 1 root root 737225 Sep 21 15:49 EDA.ipynb
-rw------- 1 root root 531326 Sep 20 05:02 EDA.pdf
drwx------ 2 root root   4096 Sep 12 15:53 papers
```

[3]:
```
pip install fuzzywuzzy
```

```
Collecting fuzzywuzzy
  Downloading fuzzywuzzy-0.18.0-py2.py3-none-any.whl (18 kB)
Installing collected packages: fuzzywuzzy
Successfully installed fuzzywuzzy-0.18.0
```

[4]:
```
# util lib import
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import xml.etree.ElementTree as et
import os
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pathlib
from pathlib import Path
import csv
```

```
from tqdm.notebook import tqdm
from fuzzywuzzy import fuzz
from wordcloud import WordCloud, STOPWORDS
```

[5]:
```
data = 'Data//aesw2016(v1.2)_train.xml'
#data = 'Data//New Text Document.xml'
csvfile = 'Data//data.csv'
```

[6]:
```
etree = et.parse(data)
```

[7]:
```
myroot = etree.getroot()
```

[ ]:
```
print(myroot)
```

```
<Element 'aesw' at 0x7f040cf975f0>
```

[ ]:
```
type(myroot)
```

[ ]:
```
xml.etree.ElementTree.Element
```

[ ]:
```
p = set()
for x in myroot.iter():
  p.add(x.tag)
```

[ ]:
```
p
```

[ ]:
```
{'aesw',
 'affiliation',
 'attribution',
 'copyright',
 'del',
 'email',
 'header',
 'ins',
 'license',
 'licenseText',
 'par',
 'sentence',
 'training',
 'year'}
```

[ ]:
```
p = set()
for x in myroot[1].iter():
  p.add(x.tag)
```

[ ]:
```
p
```

[ ]:
```
{'del', 'ins', 'par', 'sentence', 'training'}
```

[ ]:
```
p = set()
for x in myroot[1][0].iter():
  p.add(x.tag)
```

[ ]:
```
p
```

[ ]:
```
{'del', 'ins', 'par', 'sentence'}
```

[ ]:
```
p = set()
for x in myroot[1][0][0].iter():
  p.add(x.tag)
p
```

[ ]:
```
{'del', 'ins', 'sentence'}
```

[ ]:
```
myroot[1][0][0].text
```

[ ]:
```
'To facilitate an easier notation throughout the paper we define Heaviside and
Dirac functions for complex'
```

[111]:
```
'''
functions to read strigs before editing(SBE) and strings after editing(SAE)
'''
##https://github.com/samuelstevens/sentence-editing-interpretability/blob/main/
 ↪paper/aesw_to_sentences.py
## code snippet taked\n from the above github repo, an implementation of Bert␣
 ↪editing paper


def SBE(sent_elem) -> str:
  ''' SBE : String Before Editing'''
  assert sent_elem.tag == "sentence"
  string_builder = [str(sent_elem.text) if sent_elem.text else ""]


  del_word= ''
  for del_ins in sent_elem:
      if del_ins.tag == "del" and del_ins.text:
          string_builder.append(str(del_ins.text))
          del_word = del_ins.text
      if del_ins.tail:
          string_builder.append(str(del_ins.tail))

  return "".join(string_builder),del_word




##https://github.com/samuelstevens/sentence-editing-interpretability/blob/main/
 ↪paper/aesw_to_sentences.py
## code snippet taked\n from the above github repo, an implementation of Bert␣
 ↪editing paper
def SAE(sent_elem) -> str:
  ''' SAE : String After Editing'''
```

```
  assert sent_elem.tag == "sentence"
  string_builder = [str(sent_elem.text) if sent_elem.text else ""]


  ins_word=''

  for del_ins in sent_elem:
      if del_ins.tag == "ins" and del_ins.text:
          string_builder.append(str(del_ins.text))
          ins_word = del_ins.text

      if del_ins.tail:
          string_builder.append(str(del_ins.tail))

  return "".join(string_builder),ins_word
```

[112]:
```
if os.path.isfile(csvfile):
  os.remove(csvfile)

with open(csvfile, "w") as file:
  writer = csv.writer(file)
  writer.writerow(['SID','Domain','SBE','SAE','del_word','ins_word','Label'])

  for para in tqdm(myroot.iter('par')):
    domain = para.attrib['domain']

    for sent in para.iter('sentence'):
      l = 1 if len(list(sent)) > 0 else 0
      sid = str(sent.attrib["sid"])
      sbe,del_word = SBE(sent)
      sae,ins_word = SAE(sent)

      writer.writerow([sid, domain, sbe, sae, del_word, ins_word, l])
```

```
0it [00:00, ?it/s]
```

[117]:
```
df = pd.read_csv(csvfile)
```

[118]:
```
df
```

[118]:

	SID	Domain	...	ins_word	Label
0	1.0	Physics	...	,	1
1	1.1	Physics	...	NaN	0
2	1.2	Physics	...	NaN	0
3	1.3	Physics	...	NaN	0
4	2.0	Mathematics	...	NaN	0
...	...	...	...	...	...
1189407	254143.4	Computer Science	...	hand-held	1

```
1189408  254143.5  Computer Science  ...          ,     1
1189409  254144.0       Mathematics  ...        NaN     0
1189410  254144.1       Mathematics  ...          ,     1
1189411  254144.2       Mathematics  ...        NaN     0

[1189412 rows x 7 columns]
```

[119]:
```
print("Number of data points:",df.shape[0])
```

```
Number of data points: 1189412
```

[120]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1189412 entries, 0 to 1189411
Data columns (total 7 columns):
 #   Column    Non-Null Count    Dtype
---  ------    --------------    -----
 0   SID       1189412 non-null  float64
 1   Domain    1189412 non-null  object
 2   SBE       1189321 non-null  object
 3   SAE       1189315 non-null  object
 4   del_word  331932 non-null   object
 5   ins_word  411244 non-null   object
 6   Label     1189412 non-null  int64
dtypes: float64(1), int64(1), object(5)
memory usage: 63.5+ MB
```

[121]:
```
df.isnull().sum()
```

[121]:
```
SID              0
Domain           0
SBE             91
SAE             97
del_word    857480
ins_word    778168
Label            0
dtype: int64
```

[122]:
```
df[df['SAE'].isnull()]
```

[122]:
```
              SID            Domain  ... ins_word Label
12356      2694.0       Engineering  ...      NaN     1
23774      5167.0       Mathematics  ...      NaN     1
30674      6667.0       Mathematics  ...      NaN     1
41501      8985.1       Mathematics  ...      NaN     1
45142      9754.0       Engineering  ...      NaN     1
...           ...               ...  ...      ...   ...
1113929  238119.0  Computer Science  ...      NaN     1
```

```
1136308  242920.2       Mathematics  ...       NaN     1
1153671  246576.0  Computer Science  ...       NaN     1
1176268  251399.0       Engineering  ...       NaN     1
1179687  252091.0           Physics  ...       NaN     1

[97 rows x 7 columns]
```

[123]:
```
df[df['SBE'].isnull()]
```

[123]:
```
              SID            Domain  ...                          ins_word Label
6045       1315.0       Engineering  ...  Methods to remove redundancy.      1
22220      4820.0       Mathematics  ...                              2.      1
41504      8985.4       Mathematics  ...               Let _MATHDISP_.      1
41505      8985.5       Mathematics  ...              Then _MATHDISP_.      1
47023     10158.0       Mathematics  ...                              2.      1
...           ...               ...  ...                             ...   ...
1085085  231936.0       Mathematics  ...                               .      1
1129431  241447.0       Mathematics  ...                               .      1
1162047  248363.0       Mathematics  ...                                .     1
1179447  252032.0       Mathematics  ...                             1.      1
1180893  252354.0  Computer Science  ...              Configuration 3.      1

[91 rows x 7 columns]
```

[124]:
```
df[df['Label']==1][df['ins_word'].isnull()]
```

[124]:
```
              SID            Domain  ... ins_word Label
8             3.0       Engineering  ...      NaN     1
10            3.2       Engineering  ...      NaN     1
16            5.1  Computer Science  ...      NaN     1
124          33.8         Chemistry  ...      NaN     1
130          35.1       Engineering  ...      NaN     1
...           ...               ...  ...      ...   ...
1189343  254129.0       Mathematics  ...      NaN     1
1189355  254131.1       Engineering  ...      NaN     1
1189366  254134.0           Physics  ...      NaN     1
1189374  254135.0       Mathematics  ...      NaN     1
1189393  254141.0       Engineering  ...      NaN     1

[55428 rows x 7 columns]
```

[125]:
```
df.fillna('',inplace= True)
```

[126]:
```
df[df['Label']==1][df['ins_word']=='']
```

[126]:
```
              SID            Domain  ... ins_word Label
8             3.0       Engineering  ...              1
10            3.2       Engineering  ...              1
16            5.1  Computer Science  ...              1
124          33.8         Chemistry  ...              1
```

```
130          35.1       Engineering  ...              1
...           ...               ...  ...      ...   ...
1189343  254129.0       Mathematics  ...              1
1189355  254131.1       Engineering  ...              1
1189366  254134.0           Physics  ...              1
1189374  254135.0       Mathematics  ...              1
1189393  254141.0       Engineering  ...              1

[55428 rows x 7 columns]
```

[127]:
```
df[df['Label']==1][df['del_word']=='']
```

[127]:
```
              SID            Domain  ... ins_word Label
13            4.2       Engineering  ...      and     1
17            6.0       Engineering  ...        ,     1
20            7.1  Computer Science  ...        ,     1
22            7.3  Computer Science  ...        ,     1
35           11.2  Computer Science  ...     the      1
...           ...               ...  ...      ...   ...
1189386  254139.3       Mathematics  ...        ,     1
1189392  254140.0       Mathematics  ...        ,     1
1189398  254142.2       Mathematics  ...        ,     1
1189402  254142.6       Mathematics  ...        ,     1
1189410  254144.1       Mathematics  ...        ,     1

[134740 rows x 7 columns]
```

[128]:
```
def list_to_dict(lst):
  freq = {}
  for val in lst:
    if val in freq.keys():
      freq[val]+=1
    else:
      freq[val]=1
  return freq
```

[148]:
```
vocab = []
rows = df[df['Label']==1]
for _,row in rows.iterrows():
  try:
    vocab.extend(row['del_word'].split(' '))
  except:
    print(row['ins_word'])
    break
  #print(row['del_word'])
#wrds = [wrd for wrd in vocab if wrd != '']
```

[149]:
```
freq = list_to_dict(vocab)
```

[150]:
```
len(freq.keys())
```

[150]:
```
26732
```

[151]:
```
frq2=sorted(freq.items(), key=lambda x: x[1], reverse=True)
frq2
```

[151]:
```
[('', 244273),
 (',', 46806),
 ('-', 17618),
 ('the', 14057),
 (':', 13348),
 ('.', 6388),
 ('a', 4721),
 ('that', 4586),
 ('is', 4131),
 ('to', 4108),
 ('in', 3539),
 ('of', 3489),
 ('as', 3208),
 ('are', 2399),
 ('which', 2182),
 (';', 2081),
 ('section', 1990),
 ("'s", 1842),
 ('it', 1673),
 ('and', 1596),
 ('an', 1467),
 ('be', 1442),
 ('equation', 1365),
 ('have', 1363),
 ('for', 1292),
 ('on', 1205),
 ("n't", 1153),
 ('then', 1119),
 ('The', 1104),
 ("'", 1065),
 ('towards', 1062),
 ('with', 1050),
 ('2', 912),
 ('3', 875),
 ('has', 874),
 ('by', 868),
 ('can', 855),
 ('we', 808),
 ('figure', 773),
 ('"', 762),
 ('one', 742),
 ('will', 720),
 ('well', 656),
```

```
('this', 634),
('Theorem', 633),
('non-linear', 627),
('4', 625),
('non-zero', 612),
('when', 589),
('i.e.', 588),
('from', 584),
('such', 569),
('holds', 569),
('theorem', 543),
('above', 536),
('following', 536),
('In', 530),
('eq', 526),
('And', 517),
('was', 515),
('non-negative', 506),
('been', 505),
('Section', 499),
('being', 499),
('order', 491),
('while', 485),
('1', 483),
('so', 468),
('nonlinear', 467),
('exists', 442),
('also', 441),
('Lemma', 427),
('5', 408),
('_CITE_', 403),
('about', 400),
('It', 394),
('at', 390),
('fig', 385),
('Appendix', 371),
('two', 371),
('lemma', 358),
('if', 347),
('non-trivial', 343),
('table', 331),
('equations', 325),
('6', 319),
('due', 318),
('Figure', 315),
('below', 314),
('Ref.', 313),
```

```
('notations', 313),
('use', 313),
('were', 300),
('case', 298),
('Sec', 297),
("'ll", 296),
('Then', 296),
('follows', 295),
('there', 295),
('not', 295),
('respectively', 289),
('Model', 285),
('non-empty', 278),
('get', 278),
('condition', 269),
('appendix', 269),
('A', 266),
('into', 266),
('than', 264),
('its', 264),
('exist', 264),
('8', 263),
('where', 262),
('obtain', 261),
('show', 257),
('could', 255),
('percent', 255),
('e.g.', 254),
('parametrization', 249),
('As', 241),
('more', 236),
('paper', 236),
('other', 235),
('out', 234),
('cancellation', 232),
('us', 221),
('here', 214),
('Where', 213),
('function', 212),
('system', 210),
('nontrivial', 207),
('see', 207),
('do', 205),
('less', 202),
('We', 199),
('problem', 199),
('time', 197),
```

```
('See', 194),
('present', 193),
('model', 193),
('Eq.', 189),
('true', 187),
('they', 184),
('only', 184),
('find', 183),
('s', 179),
('both', 177),
('last', 176),
('what', 174),
('This', 174),
('_REF_', 173),
('each', 172),
('chapter', 172),
('shown', 172),
('Problem', 171),
('using', 170),
('Similar', 169),
('subsection', 167),
('works', 166),
('all', 166),
('While', 165),
('Let', 164),
('these', 164),
('or', 164),
('shows', 163),
('does', 162),
("can't", 162),
('those', 160),
('would', 160),
('number', 160),
('systems', 160),
('example', 159),
('Algorithm', 159),
('Fig', 158),
('Chapter', 157),
('non-decreasing', 157),
('their', 157),
('cf.', 156),
('given', 155),
('details', 154),
('From', 154),
('7', 153),
('So', 153),
('For', 150),
```

```
('Equation', 150),
('consider', 148),
('MATH', 147),
('On', 146),
('give', 144),
('them', 144),
('used', 143),
('coexistence', 142),
('results', 142),
('But', 141),
('some', 139),
('first', 138),
('gaussian', 137),
('value', 136),
('up', 134),
('That', 133),
('nonzero', 133),
('need', 133),
('non-linearity', 133),
('hand', 132),
('step', 132),
('but', 132),
('Because', 132),
('multi-core', 131),
('obtained', 131),
('follow', 130),
('Fig.', 130),
('like', 130),
('three', 130),
('denote', 130),
('non-increasing', 129),
('result', 129),
('too', 129),
('Subsection', 128),
("won't", 128),
("'ve", 127),
('/', 125),
('between', 125),
('since', 123),
('pre-computation', 123),
('on-line', 123),
('lead', 123),
('non-degenerate', 122),
('should', 121),
('form', 121),
('non-uniform', 120),
('I', 119),
```

```
('satisfies', 119),
('much', 118),
('later', 118),
('Firstly', 117),
('System', 117),
('every', 115),
('noncommutative', 115),
('nonnegative', 115),
('over', 114),
('make', 114),
('similar', 114),
('If', 112),
('method', 112),
('let', 112),
('gives', 112),
('c.f', 111),
('under', 111),
('study', 110),
('however,', 110),
('non-singular', 110),
('five', 110),
('Eqs', 109),
('hold', 109),
('presented', 109),
('found', 108),
('become', 108),
('shall', 108),
('set', 107),
('Pomeron', 107),
('further', 107),
('follows:', 107),
('enough', 106),
('non-thermal', 106),
('so-called', 105),
('yields', 104),
('part', 104),
('At', 103),
('take', 103),
('therefore', 102),
('ones', 102),
('done', 102),
('Condition', 101),
('may', 101),
('algorithm', 101),
('sections', 100),
('proposition', 98),
('large', 98),
```

```
('cut-off', 98),
('U.S.', 97),
('models', 97),
('By', 97),
('type', 97),
('because', 97),
('Such', 96),
('say', 96),
('increase', 96),
('formula', 96),
('denotes', 95),
('Cloud', 95),
('..', 94),
('fact', 94),
('conditions', 93),
("'re", 92),
('noted', 92),
('any', 92),
('eqs', 91),
('9', 91),
('larger', 90),
('Figs', 90),
('satisfy', 90),
('%', 89),
('however', 89),
('work', 89),
('no', 88),
('thus', 88),
('provide', 88),
('another', 88),
('Table', 88),
('imply', 87),
('Method', 87),
('depends', 86),
('state', 86),
('line', 85),
('1000', 85),
('having', 85),
('via', 85),
('way', 85),
('process', 85),
('considered', 84),
('iid', 83),
('different', 83),
('already', 82),
('criteria', 82),
('indexes', 82),
```

```
('assumption', 82),
('definition', 82),
('ie', 81),
('multi-dimensional', 81),
("it's", 81),
('pre-processing', 81),
('Secondly', 80),
('firstly', 80),
('analogue', 80),
('parameters', 80),
('term', 80),
('through', 80),
('next', 80),
('know', 80),
('inputs', 79),
('that:', 79),
('seen', 79),
('Proposition', 79),
('estimate', 79),
('define', 79),
('ten', 79),
('10', 79),
('2nd', 78),
('allow', 78),
('compare', 78),
('dimension', 78),
('assume', 78),
('detail', 78),
('set-up', 78),
('becomes', 78),
('Inequality', 77),
('There', 77),
('_CITE_,', 77),
('now', 77),
('pre-defined', 76),
('effect', 76),
('provides', 75),
('point', 75),
('sun', 75),
('functions', 75),
('sake', 74),
('direction', 74),
('means', 74),
('nonsingular', 74),
('pseudorandom', 73),
('Principle', 73),
('four', 73),
```

```
('had', 73),
('literatures', 73),
('solution', 73),
('Series', 73),
('Figures', 73),
('happen', 72),
('S', 72),
('our', 72),
('setup', 71),
('note', 71),
('methods', 71),
('second', 71),
('Refs.', 71),
('small', 70),
('non-dimensional', 70),
('seconds', 70),
('non-vanishing', 70),
('though', 70),
('etc', 70),
('values', 69),
('implies', 69),
('namely', 69),
('leads', 69),
('region', 69),
('upon', 69),
('sec', 69),
('correspond', 69),
('spacetime', 68),
('non-overlapping', 68),
('gets', 68),
('proved', 67),
('al.', 67),
('figures', 67),
('contains', 67),
('represent', 67),
('solar', 67),
('Since', 67),
('eg', 67),
('3rd', 66),
('distribution', 66),
('yield', 66),
('cutoff', 66),
('zero', 65),
('proof', 65),
('hundred', 65),
('occurence', 65),
('thanks', 65),
```

```
('dynamic', 65),
('below:', 64),
('able', 64),
('hours', 64),
('non-stationary', 64),
('principle', 64),
('Case', 64),
('reference', 63),
('solutions', 63),
('fibre', 63),
('phenomena', 63),
('smaller', 63),
('d', 63),
('he', 63),
('twenty', 62),
('web', 62),
('virtue', 62),
('online', 62),
('network', 62),
('nonperturbative', 62),
('prove', 62),
('describe', 61),
('times', 61),
('determine', 61),
('Eqns', 61),
('parameter', 61),
('measure', 60),
('introduction', 60),
('minutes', 60),
('size', 60),
('reduce', 60),
('straight-forward', 60),
('non-smooth', 60),
('data', 60),
('as:', 60),
('hypothesis', 59),
('discussed', 59),
('occuring', 59),
('pre-impact', 59),
('analysis', 59),
('disc', 59),
('twelve', 59),
('kind', 59),
('non-coding', 58),
('degree', 58),
('how', 58),
('inequality', 58),
```

```
('expression', 58),
('might', 58),
('cases', 58),
('either', 57),
('Following', 57),
('near', 57),
('They', 57),
('extend', 57),
('compared', 57),
('together', 57),
('iff', 57),
('general', 57),
('item', 56),
('Grid', 56),
('made', 56),
('problems', 56),
('To', 56),
('represents', 56),
('MATHDISP', 56),
('Further', 56),
('Universe', 56),
('maximum', 56),
('4th', 55),
('multi-scale', 55),
('explicitely', 55),
('wind', 55),
('needs', 55),
('An', 55),
('per', 54),
('higgs', 54),
('change', 54),
('mass', 54),
('steps', 54),
('resp.', 54),
('occur', 54),
('proposed', 54),
('cycle', 54),
('others', 53),
('baker', 53),
('nonempty', 53),
('showed', 53),
('nor', 53),
('resulting', 53),
('apply', 53),
('matrix', 53),
('semi-definite', 53),
('Theory', 53),
```

```
('studied', 53),
('non-magnetic', 53),
('described', 53),
('satisfying', 52),
('spatio-temporal', 52),
('nm', 52),
('Sections', 52),
('grey', 52),
('same', 52),
('multi-agent', 52),
('min', 52),
('corresponds', 52),
('field', 52),
('agreement', 51),
('down', 51),
('consist', 51),
('hasto', 51),
('depend', 51),
('amount', 51),
('eleven', 51),
('nonrelativistic', 51),
('E.g.', 50),
('lines', 50),
('Eq', 50),
('most', 50),
('high', 50),
('Logic', 50),
('non-equilibrium', 50),
('sub-graph', 50),
('effects', 50),
('Example', 50),
('non-local', 50),
('regions', 50),
('density', 50),
('happens', 50),
('sub-section', 49),
('toward', 49),
('Type', 49),
('consists', 49),
('satisfied', 49),
('so-', 49),
('Equations', 49),
('CMEs', 49),
('non-secretors', 49),
('compute', 48),
('hence', 48),
('1st', 48),
```

```
('theorems', 48),
('P.L.', 48),
('mean', 48),
('whilst', 48),
('non-congruent', 48),
('max', 48),
('With', 47),
('pre-existing', 47),
('seems', 47),
('period', 47),
('non-negligible', 47),
('examples', 47),
('increases', 47),
('mentioned', 47),
('hamiltonian', 47),
('networks', 47),
('allows', 47),
('suggest', 46),
('cent', 46),
('?', 46),
('control', 46),
('transfered', 46),
('non-convex', 46),
('build', 46),
('correspondingly', 46),
('space-time', 46),
('non-random', 46),
('forms', 46),
('application', 45),
('comparing', 45),
('purpose', 45),
('axis', 45),
('nonparametric', 45),
('One', 45),
('corollary', 45),
('close', 45),
('known', 45),
('list', 45),
('retarget', 45),
('decrease', 45),
('takes', 45),
('paring', 45),
('assure', 45),
('whereas', 45),
('rate', 45),
('makes', 44),
('3D', 44),
```

```
('Notice', 44),
('focused', 44),
('Physics', 44),
('solve', 44),
('upwards', 44),
('choose', 44),
('bigger', 44),
('states', 44),
('contrary', 44),
('Column', 44),
('space', 44),
('Denote', 44),
('non-loop', 44),
('notice', 44),
('Analysis', 43),
('pre-specified', 43),
('taking', 43),
('before', 43),
('optic', 43),
('player', 43),
('i.e.,', 43),
('non-conservative', 43),
('place', 42),
('parametrizations', 42),
('th', 42),
('power', 42),
('among', 42),
('outputs', 42),
('side', 42),
('square', 42),
('independent', 42),
('equality', 42),
('Skorohod', 42),
('Semi', 42),
('mash-up', 42),
('easily', 42),
('assures', 42),
('taken', 41),
('non-parametric', 41),
('spectral', 41),
('occured', 41),
('uses', 41),
('First', 41),
('Alexandrov', 41),
('indicate', 41),
('Via', 41),
('multi-grid', 41),
```

```
('These', 41),
('re-written', 41),
('propose', 41),
('anti-phase', 41),
('active', 41),
('observed', 41),
('till', 41),
('informations', 41),
('mode', 41),
('structure', 40),
('Corollary', 40),
('remain', 40),
('defined', 40),
('that,', 40),
('multi-portfolio', 40),
('amongst', 40),
('corresponding', 40),
('estimation', 40),
('hyper-chaotic', 40),
('appear', 40),
('appears', 40),
('discuss', 39),
('tradeoff', 39),
('points', 39),
('West', 39),
('big', 39),
('test', 39),
('investigate', 39),
('occurs', 39),
('very', 39),
('include', 39),
('statement', 39),
('Thanks', 39),
('Queue', 39),
('presents', 39),
('J.', 39),
('structures', 39),
('require', 39),
('non', 39),
('word', 38),
('3-', 38),
('backwards', 38),
('specially', 38),
('based', 38),
('evidences', 38),
('non-rigid', 38),
('Then,', 38),
```

```
('usually', 38),
('Note', 38),
('separation', 38),
('yet', 38),
('followings', 38),
('Standard', 38),
('non-positive', 38),
('assured', 38),
('produce', 37),
('possible', 37),
('still', 37),
('off', 37),
('must', 37),
('Remark', 37),
('identified', 37),
('vs.', 37),
('non-standard', 37),
('Thm.', 37),
('appeared', 37),
('A.', 37),
('x', 37),
('low', 37),
('many', 37),
('Spline', 37),
('converge', 37),
('just', 37),
('describes', 37),
('quasi-periodic', 37),
('co-ordinates', 37),
('tunnelling', 37),
('Assumption', 37),
('non-constant', 37),
('brownian', 37),
('0', 37),
('goes', 37),
('ROI', 37),
('length', 36),
('techniques', 36),
('afterwards', 36),
('rather', 36),
('parenthesis', 36),
('needsto', 36),
('changes', 36),
('considering', 36),
('M.', 36),
('worth', 36),
('pixels', 36),
```

```
('Formula', 36),
('cartesian', 36),
('QPP', 36),
('variable', 36),
('generate', 36),
('Which', 36),
('occurred', 36),
('Transform', 36),
('image', 36),
('sub-surface', 36),
('equals', 36),
('South', 36),
('path-wise', 36),
('algorithms', 35),
('Thus', 35),
('relationship', 35),
('transform', 35),
('far', 35),
('level', 35),
('thirty', 35),
('flatfield', 35),
('tail', 35),
('Whereas', 35),
('relation', 35),
('!', 35),
('shortly', 35),
('understand', 35),
('respectively,', 35),
('includes', 35),
('sub-optimal', 35),
('nonlinearity', 35),
('fermi', 35),
('come', 35),
('needed', 35),
('S.', 35),
('behaviors', 35),
('arbitrary', 35),
('going', 35),
('pair', 34),
('et', 34),
('Lemmas', 34),
('perform', 34),
('pre-computed', 34),
("RGE's", 34),
('monotone', 34),
('conclusion', 34),
('disk', 34),
```

```
('non-contractible', 34),
('indicates', 34),
('Solar', 34),
('Though', 34),
('constant', 34),
("ODE's", 34),
('Results', 34),
('requires', 34),
('greater', 34),
('non-linearities', 34),
('Odderon', 34),
('automaton', 34),
('previous', 33),
('even', 33),
('approach', 33),
('Because,', 33),
('Using', 33),
('finally', 33),
('Notations', 33),
('remind', 33),
('inequalities', 33),
('increasing', 33),
('Law', 33),
('non-symmetric', 33),
('simulatability', 33),
('during', 33),
('introduce', 33),
('spacial', 33),
('non-normal', 33),
('Rule', 33),
('signalling', 33),
('Step', 33),
(']', 33),
('multi-level', 33),
('dependent', 33),
('called', 33),
('improve', 33),
('hardcore', 33),
('PhD', 33),
('constraints', 32),
('including', 32),
('put', 32),
('numbers', 32),
('General', 32),
('multi-sided', 32),
('comparable', 32),
('Paragraph', 32),
```

```
('necessary', 32),
('Process', 32),
('computation', 32),
('nondecreasing', 32),
('multi-path', 32),
('calculate', 32),
('belongs', 32),
('originated', 32),
('assumptions', 32),
('subspace', 32),
('non-cooperative', 32),
('spectrum', 32),
('difference', 32),
('limit', 32),
('trans-equatorial', 32),
('applications', 31),
('six', 31),
('parametrisation', 31),
('remains', 31),
('investigated', 31),
('operator', 31),
('Also', 31),
('nineties', 31),
('property', 31),
('alone', 31),
('tables', 31),
('non-homogeneous', 31),
('variables', 31),
('Test', 31),
('nodes', 31),
('component', 31),
('vs', 31),
('t', 31),
('contain', 31),
('internet', 31),
('ref.', 31),
('aircraft', 30),
('5th', 30),
('instance', 30),
('right', 30),
('When', 30),
('construction', 30),
('degenerated', 30),
('non-potential', 30),
('account', 30),
('enable', 30),
('fifteen', 30),
```

```
('pre-determined', 30),
('non-autonomous', 30),
('papers', 30),
('few', 30),
('filed', 30),
('Towards', 30),
('sufficient', 30),
('anti-top', 30),
('nondimensional', 30),
('then,', 30),
('processes', 30),
('bi-directional', 30),
('theatre', 30),
('zeroes', 30),
('sixteen', 30),
('Especially', 30),
('Eqs.', 30),
('applying', 30),
('electromagnetic', 30),
('derive', 30),
('anti-symmetric', 30),
('attentions', 30),
('multi-wavelength', 30),
('Approximation', 30),
('Function', 29),
('whole', 29),
('indeed', 29),
('reach', 29),
('focus', 29),
('course', 29),
('phylogenetic', 29),
('hand,', 29),
('looses', 29),
('position', 29),
('base', 29),
('subsections', 29),
('non-contiguous', 29),
('extracellular', 29),
('chose', 29),
('higher', 29),
('experiment', 29),
('observe', 29),
('good', 29),
('Schwartz', 29),
('resulted', 29),
('non-stationarity', 29),
('Equ', 29),
```

```
('got', 29),
('observations', 29),
('against', 29),
('non-ideal', 29),
('dependance', 29),
('non-parallel', 28),
('unit', 28),
('non-inferior', 28),
('constraint', 28),
('Afterwards', 28),
('downwards', 28),
('MJLs', 28),
('wave', 28),
('consequence', 28),
('criterion', 28),
('De', 28),
('reason', 28),
('Theorems', 28),
('Due', 28),
('temperature', 28),
('micro-scale', 28),
('s.t.', 28),
('conclude', 28),
('nonholonomic', 28),
('converges', 28),
('important', 28),
('suggests', 28),
('formulae', 28),
('plane', 28),
('auto-correlation', 28),
('non-supported', 28),
('fifty', 28),
('phase', 28),
('II', 28),
('lagrangian', 28),
('loose', 28),
('identify', 28),
('non-diagonal', 28),
('comparison', 28),
('basis', 28),
('regards', 28),
('eigen-skeleton', 28),
('range', 27),
('anti-synchronization', 27),
('short', 27),
('denoted', 27),
('Conjecture', 27),
```

```
 ('parametrize', 27),
 ('coexist', 27),
 ('ref', 27),
 ('west', 27),
 ('group', 27),
 ('coefficients', 27),
 ('comes', 27),
 ('re-write', 27),
 ('Different', 27),
 ('belong', 27),
 ('P.', 27),
 ('again', 27),
 ('earlier', 27),
 ('fixed', 27),
 ('introduced', 27),
 ('non-commutative', 27),
 ('noting', 27),
 ...]
```

[152]:
```
wc = WordCloud(background_color="white", max_words=2000)
wc.generate_from_frequencies(freq)
print ("Word Cloud for Deleted Words")
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

```
Word Cloud for Deleted Words
```

[134]:
```
vocab = []
rows = df[df['Label']==1]
for _,row in rows.iterrows():
```

```
  try:
    vocab.extend(row['ins_word'].split(' '))
  except:
    print(row['ins_word'])
    break
  #print(row['del_word'])
#wrds = [wrd for wrd in vocab if wrd != '']
```

[135]:
```
freq = list_to_dict(vocab)
```

[136]:
```
len(freq.keys())
```

[136]:
```
19803
```

[137]:
```
frq2=sorted(freq.items(), key=lambda x: x[1], reverse=True)
frq2
```

[137]:
```
[('', 148864),
 (',', 106377),
 ('-', 34339),
 ('the', 27676),
 ('.', 14912),
 ('a', 9799),
 (':', 9536),
 ('that', 7633),
 ('and', 5788),
 (';', 5584),
 ('of', 4478),
 ('is', 4209),
 ('an', 3051),
 ('to', 2753),
 ('in', 2636),
 ('are', 2429),
 ('have', 2335),
 ('The', 2307),
 ('Section', 2146),
 ('for', 1957),
 ('we', 1879),
 ('which', 1870),
 ('by', 1710),
 ('be', 1609),
 ('as', 1539),
 ('not', 1460),
 ('it', 1228),
 ('with', 1212),
 ('on', 1149),
 ('one', 1081),
 ('see', 1029),
 ('toward', 1015),
```

```
('th', 861),
('us', 849),
('this', 837),
("'", 812),
('Equation', 809),
('was', 779),
('where', 773),
('two', 760),
('then', 751),
('three', 751),
('Figure', 745),
('has', 721),
('were', 668),
('from', 663),
('Eq', 633),
('will', 628),
('following', 604),
('Theorem', 599),
('nonzero', 599),
('nonlinear', 596),
('four', 561),
('We', 546),
('theorem', 535),
('at', 526),
('also', 499),
('This', 495),
('nonnegative', 487),
('section', 472),
('A', 465),
('so', 456),
('can', 453),
('Fig', 436),
('exist', 434),
('non-linear', 429),
('these', 418),
('been', 414),
('follows', 412),
('into', 411),
("'s", 409),
('"', 385),
('Table', 373),
('appendix', 359),
('In', 355),
('such', 350),
('model', 348),
('_REF_', 347),
('five', 345),
```

```
('nontrivial', 339),
('being', 337),
('than', 329),
('Lemma', 327),
('lemma', 327),
('all', 325),
('For', 324),
('find', 314),
('system', 310),
('they', 308),
('used', 304),
('if', 302),
('notation', 299),
('s', 294),
('equation', 287),
('let', 286),
('Appendix', 286),
('result', 284),
('Eq.', 283),
('exists', 283),
('Sect', 280),
('nonempty', 279),
('2', 279),
('six', 272),
('whereas', 269),
('shows', 265),
('%', 262),
('1', 262),
('or', 262),
('is,', 260),
('only', 259),
('obtain', 258),
('because', 253),
('following.', 245),
('may', 244),
('parameterization', 239),
('those', 235),
('show', 234),
('If', 232),
('By', 232),
('eight', 231),
('cancelation', 229),
('Figs', 228),
('Eqs', 228),
('shall', 226),
('since', 225),
('_CITE_', 221),
```

```
('Equations', 220),
('example', 219),
('using', 218),
('problem', 217),
('first', 216),
('Since', 216),
('algorithm', 215),
('Fig.', 203),
('use', 200),
('Figures', 199),
('their', 197),
('non-trivial', 197),
('cannot', 195),
('functions', 194),
('case', 193),
('see,', 192),
('there', 187),
('detail', 187),
('its', 186),
('To', 186),
('ii', 183),
('respectively', 182),
('must', 181),
('other', 178),
('method', 178),
('owing', 175),
('It', 172),
('values', 172),
('thus', 171),
('given', 171),
('found', 169),
('obtained', 169),
('work', 168),
('following:', 168),
('An', 164),
('results', 164),
('However,', 163),
('condition', 157),
('equations', 154),
('example,', 150),
('type', 150),
('cf', 150),
('more', 150),
('here', 150),
('3', 149),
('Then', 148),
('well', 145),
```

```
('get', 145),
('Subsection', 145),
('MATHDISP', 145),
('Condition', 143),
('Gaussian', 142),
('co-existence', 140),
('shown', 140),
('consider', 138),
('nondecreasing', 138),
('do', 137),
('Note', 137),
('Sections', 137),
('inequality', 137),
('second', 136),
('does', 135),
('5', 135),
('al.', 134),
('becomes', 134),
('chapter', 133),
('denote', 133),
('latter', 132),
('non-zero', 132),
('nonlinearity', 132),
('i', 132),
('when', 131),
('multicore', 131),
('These', 129),
('nondegenerate', 129),
('Ref.', 129),
('but', 128),
('hold', 127),
('satisfies', 127),
('systems', 127),
('article', 126),
('function', 126),
('seven', 126),
('Here', 126),
('follows:', 126),
('while', 125),
('no', 125),
('Chapter', 125),
('higher', 124),
('nonuniform', 123),
('similar', 123),
('precomputation', 122),
('nonincreasing', 121),
('holds', 120),
```

```
('would', 120),
('First', 120),
('conditions', 119),
('Although', 119),
('them', 119),
('presented', 119),
('However', 118),
('present', 118),
('considered', 118),
('occur', 117),
('Similarly', 116),
('time', 115),
('4', 114),
('numbers', 113),
('similarly', 111),
('made', 110),
('nonsingular', 110),
('non-commutative', 110),
('online', 109),
('satisfy', 109),
('gives', 108),
('lower', 108),
('now', 108),
('Let', 107),
('non-negative', 106),
('US', 105),
('0', 105),
('implies', 105),
('Thus', 105),
('et', 104),
('pomeron', 104),
('become', 103),
('Step', 102),
('nonthermal', 102),
('give', 102),
('Problem', 101),
('follows.', 99),
('occurs', 99),
('showed', 99),
('literature', 98),
('iii', 98),
('analysis', 97),
('solutions', 97),
('cloud', 97),
('terms', 97),
(']', 95),
('form', 95),
```

```
('.,', 95),
('cutoff', 95),
('our', 94),
('Proposition', 94),
('number', 93),
('Sun', 93),
('authors', 93),
('parameter', 91),
('criterion', 91),
('leads', 91),
('indices', 91),
('figure', 91),
('Theorems', 90),
('times', 89),
('due', 89),
('both', 88),
('note', 88),
('There', 88),
('cases', 88),
('whether', 88),
('order', 87),
('above', 87),
('solution', 86),
('models', 86),
('?', 86),
('allows', 86),
('need', 86),
('having', 86),
('input', 85),
('point', 85),
('process', 85),
('series', 85),
('fact', 84),
('about', 84),
('proposed', 84),
('fewer', 84),
('Sect.', 83),
('compared', 83),
('respect', 83),
('state', 83),
('provides', 83),
('1,000', 82),
('analog', 82),
('earlier', 82),
('yields', 82),
('many', 81),
('studied', 81),
```

```
('among', 81),
('straightforward', 81),
('value', 81),
('described', 80),
('subsection', 80),
('As', 80),
('occurring', 80),
('Eqs.', 80),
('parameters', 79),
('corresponds', 79),
('versus', 79),
('fourth', 78),
('principle', 78),
('dimensional', 78),
('lead', 78),
('ensure', 78),
('occurrence', 77),
('means', 77),
('multidimensional', 76),
('define', 76),
('assumed', 76),
('preprocessing', 76),
('take', 76),
('processes', 75),
('nine', 75),
('out', 75),
('setup', 75),
('large', 75),
('denotes', 75),
('information', 75),
('predefined', 74),
('problems', 74),
('increases', 74),
('below', 74),
('sets', 73),
('through', 73),
('etc', 73),
('assume', 73),
('corresponding', 73),
('Model', 73),
('lines', 72),
('From', 72),
('previous', 72),
('depend', 72),
('nonvanishing', 72),
('recall', 72),
('should', 72),
```

```
('non-singular', 72),
('pseudo-random', 71),
('discussed', 71),
('12', 71),
('therefore', 71),
('min', 71),
('theorem.', 70),
('i.e.', 70),
('Therefore', 70),
('h', 70),
('Definition', 70),
('set-up', 69),
('e.g.', 69),
('third', 69),
('proposition', 69),
('coordinates', 69),
('had', 69),
('make', 68),
('who', 68),
('known', 68),
('10', 68),
('Second', 67),
('data', 67),
('set', 67),
('methods', 67),
('needs', 67),
('could', 67),
('Example', 67),
('field', 67),
('increase', 67),
('regarding', 67),
('matrix', 66),
('Introduction', 66),
('prove', 66),
('phenomenon', 65),
('close', 65),
('depends', 65),
('estimate', 65),
('types', 65),
('space-time', 64),
('Web', 64),
('nondimensional', 64),
('might', 64),
('uses', 64),
('especially', 64),
('step', 64),
('larger', 64),
```

```
('way', 64),
('theory', 64),
('describe', 64),
('cut-off', 64),
('points', 63),
('defined', 63),
('satisfying', 63),
('Lemmas', 63),
('ten', 63),
('proof', 62),
('preimpact', 62),
('states', 62),
('explicitly', 62),
('represent', 62),
('nonoverlapping', 61),
('fiber', 61),
('determine', 61),
('on-line', 61),
('20', 60),
('noncoding', 60),
('yield', 60),
('approach', 60),
('nonsmooth', 60),
('Assumption', 60),
('i.i.d.', 60),
('D', 60),
('ensures', 60),
('directions', 59),
('Tables', 59),
('non-perturbative', 59),
('correspond', 59),
('greater', 59),
('preceding', 59),
('On', 59),
('line', 58),
('provide', 58),
('respectively,', 58),
('sections', 58),
('ones', 58),
('Kapitsa', 58),
('disk', 58),
('Corollary', 58),
('seen', 58),
('smaller', 57),
('hypotheses', 57),
('various', 57),
('nonstationary', 57),
```

```
('Higgs', 57),
('output', 57),
('contrast', 57),
('When', 57),
('behavior', 57),
('forms', 57),
('sufficient', 56),
('constraints', 56),
('each', 56),
('occurred', 56),
('consists', 56),
('side', 56),
('Refs.', 56),
('regards', 56),
('represents', 56),
('approximation', 56),
('continuous', 56),
('whilst', 55),
('study', 55),
('parts', 55),
('different', 55),
('Furthermore', 55),
('semidefinite', 55),
('Case', 55),
('Lagrangian', 55),
('universe', 54),
('spatiotemporal', 54),
('general', 54),
('denoted', 54),
('gray', 54),
('degrees', 54),
('called', 54),
('nonmagnetic', 54),
('coefficients', 54),
('requires', 54),
('another', 53),
('Item', 53),
('makes', 53),
('multiscale', 53),
('very', 53),
('effect', 53),
('approximately', 53),
('between', 53),
('dynamics', 53),
('contain', 53),
('Using', 53),
('multiagent', 53),
```

```
('what', 53),
('grid', 52),
('either', 52),
('physics', 52),
('logic', 52),
('subgraph', 52),
('SW', 52),
('11', 52),
('follow', 52),
('distributions', 52),
('non-relativistic', 52),
('highest', 52),
('works', 52),
('i.e.,', 51),
('easily', 51),
('Conditions', 51),
('possible', 51),
('dependence', 51),
('spacetime', 51),
('detailed', 51),
('mentioned', 51),
('any', 51),
('Hamiltonian', 51),
('non-empty', 50),
('matrices', 50),
('discuss', 50),
('details', 50),
('observed', 50),
('equal', 50),
('applied', 50),
('although', 50),
('proved', 49),
('same', 49),
('over', 49),
('network', 49),
('constant', 49),
('Algorithm', 49),
('presents', 49),
('interest', 49),
('comparison', 49),
('nonsecretors', 49),
('6', 48),
('part', 48),
('transferred', 48),
('change', 48),
('performed', 48),
('equilibrium', 48),
```

```
('nonequilibrium', 48),
('some', 48),
('dimensions', 48),
('appears', 48),
('column', 48),
('transform', 48),
('preexisting', 47),
('generally', 47),
('Baker', 47),
('based', 47),
('100', 47),
('principal', 47),
('I', 47),
('noncongruent', 47),
('appear', 47),
('-hand', 47),
('still', 47),
('structure', 47),
('nonlocal', 47),
('contains', 47),
('hand', 47),
('MATH', 47),
('Lyapunov', 46),
('during', 46),
('further', 46),
('towards', 46),
('most', 46),
('backward', 46),
('increasing', 46),
('introduced', 46),
('derive', 46),
('evidence', 46),
('under', 46),
('They', 46),
('next', 46),
('previously', 46),
('test', 46),
('pairing', 46),
('inequalities', 46),
('allow', 46),
('region', 45),
('until', 45),
('nonconvex', 45),
('expression', 45),
('rewritten', 45),
('provided', 45),
('independent', 45),
```

```
('taking', 45),
('irreducible', 44),
('parameterizations', 44),
('components', 44),
('multiportfolio', 44),
('upward', 44),
('takes', 44),
('chosen', 44),
('choose', 44),
('Consider', 44),
('technique', 44),
('regions', 44),
('rule', 44),
('coefficient', 44),
('nonloop', 44),
('Moreover', 43),
('finding', 43),
('nonnegligible', 43),
('non-parametric', 43),
('Aleksandrov', 43),
('developed', 43),
('like', 43),
('optical', 43),
('rewrite', 43),
('remaining', 43),
('high', 42),
('indicates', 42),
('mean', 42),
('Skorokhod', 42),
('ref.', 42),
('semi', 42),
('nonrandom', 42),
('re-target', 42),
('minutes', 42),
('led', 42),
('mashup', 42),
('prespecified', 41),
('formula', 41),
('west', 41),
('decreases', 41),
('imply', 41),
('Therefore,', 41),
('determined', 41),
('Finally', 41),
('kinds', 41),
('obtains', 41),
('hyperchaotic', 41),
```

```
('require', 41),
('zero', 41),
('Inequality', 41),
('include', 41),
('low', 41),
('queue', 41),
('much', 41),
('discussion', 40),
('even', 40),
('variables', 40),
('did', 40),
('rather', 40),
('multigrid', 40),
('range', 40),
('e.g.,', 40),
('addition', 40),
('pixel', 40),
('law', 40),
('up', 40),
('zeros', 40),
('applications', 40),
('nonconservative', 40),
('ODEs', 40),
('write', 39),
('spline', 39),
('probability', 39),
('antiphase', 39),
('Also', 39),
('Fermi', 39),
('assumption', 39),
('m', 39),
('Cycle', 39),
('table', 39),
('trade-off', 38),
('references', 38),
('otherwise', 38),
('however', 38),
('suppose', 38),
('RGEs', 38),
('corollary', 38),
('nonrigid', 38),
('paper', 38),
('nonpositive', 38),
('transequatorial', 38),
('converges', 38),
('resulting', 38),
('fields', 38),
```

```
('subsurface', 37),
('One', 37),
('perform', 37),
('gets', 37),
('networks', 37),
('With', 37),
('briefly', 37),
('calculated', 37),
('Assumptions', 37),
('Cartesian', 37),
('Thus,', 37),
('particular', 37),
('pathwise', 37),
('ensured', 37),
('Because', 37),
('positive', 37),
('associated', 36),
('remains', 36),
('spatial', 36),
('nonparametric', 36),
('reader', 36),
('describes', 36),
('parentheses', 36),
('consist', 36),
('Recall', 36),
('existence', 36),
('nonstandard', 36),
('Chap', 36),
('affect', 36),
('relation', 36),
('suboptimal', 36),
('namely', 36),
('necessary', 36),
('determining', 36),
('nonlinearities', 36),
('Brownian', 36),
('main', 36),
('seem', 35),
('variable', 35),
('corollary.', 35),
('hypothesis', 35),
('subsystem', 35),
('distribution', 35),
('analyzed', 35),
('measurements', 35),
('assumptions', 35),
('algorithms', 35),
```

```
('Separation-List', 35),
('done', 35),
('QPPs', 35),
('neither', 35),
('focussed', 35),
('properties', 35),
('referred', 35),
('fifth', 34),
('lemmas', 34),
('three-dimensional', 34),
('built', 34),
('features', 34),
('nonhomogeneous', 34),
('performance', 34),
('remain', 34),
('afterward', 34),
('axes', 34),
('degenerate', 34),
('precomputed', 34),
('term', 34),
('changes', 34),
('several', 34),
('rate', 34),
('taken', 34),
('nonsymmetric', 34),
('conclusions', 34),
('noncontractible', 34),
('operator', 34),
('antitop', 34),
('investigated', 34),
('Player', 34),
('columns', 34),
('Max', 34),
('decrease', 34),
('achieved', 34),
('System', 34),
('structures', 34),
('past', 33),
('words', 33),
('parameterisation', 33),
('de', 33),
('Method', 33),
('after', 33),
('analyses', 33),
('wavelengths', 33),
('factor', 33),
('15', 33),
```

```
('considering', 33),
('Notation', 33),
('conclude', 33),
('30', 33),
('flat-field', 33),
('Propositions', 33),
('Part', 33),
('Simulatability', 33),
('At', 33),
('significant', 33),
('group', 33),
('apply', 33),
('noncooperative', 33),
('signaling', 33),
('satisfied', 33),
('lose', 33),
('eigenvalues', 33),
('indicate', 33),
('nonconstant', 33),
('reduces', 33),
('attention', 33),
('ROIs', 33),
('needed', 33),
('application', 32),
('1990', 32),
('Hence', 32),
('quasiperiodic', 32),
('1990s', 32),
('complete', 32),
('later', 32),
('maximum', 32),
('expressions', 32),
('noise', 32),
('X', 32),
('non-linearity', 32),
('nonnormal', 32),
('space', 32),
('vertices', 32),
('suggested', 32),
('Internet', 32),
('error', 32),
('16', 32),
('necessarily', 32),
('difficult', 32),
('chain', 32),
('hard-core', 32),
('representations', 32),
```

```
('longer', 31),
('characteristic', 31),
('solving', 31),
('derivatives', 31),
('far', 31),
('length', 31),
('enables', 31),
('according', 31),
('observations', 31),
('basis', 31),
('techniques', 31),
('non-decreasing', 31),
('Toward', 31),
('studying', 31),
('Statement', 31),
('sizes', 31),
('minimum', 31),
('October', 31),
('whose', 31),
('standard-model', 31),
('multilevel', 31),
('loses', 31),
('convergence', 31),
('research', 31),
('constants', 31),
('antisymmetric', 31),
('Ph.D.', 31),
('aircrafts', 30),
('analogous', 30),
('ratio', 30),
('approaches', 30),
('suffices', 30),
('know', 30),
('estimates', 30),
('before', 30),
('multisided', 30),
('steps', 30),
('analyze', 30),
('transforms', 30),
('p', 30),
('required', 30),
('understanding', 30),
('iv', 30),
('independently', 30),
('off', 30),
('submillimeter', 30),
('Schwarz', 30),
```

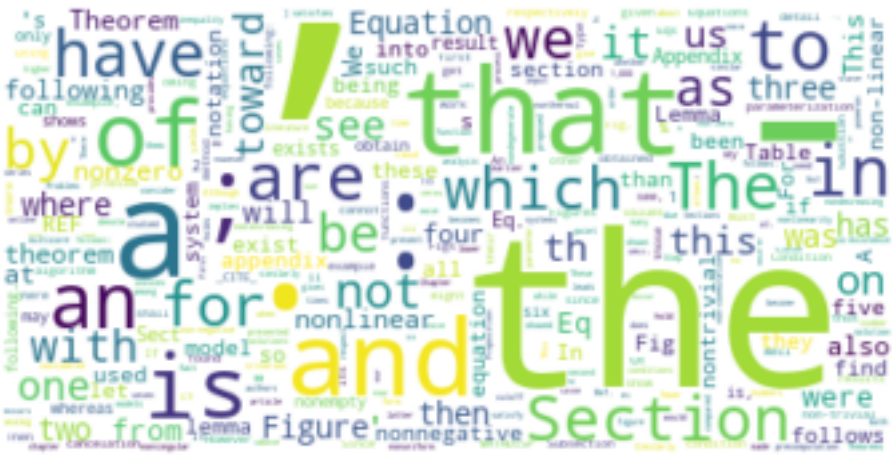
```
('question', 30),
('yet', 30),
('spectrum', 30),
('significantly', 30),
('definition', 30),
('tunneling', 30),
('odderon', 30),
('appearing', 30),
('g.', 30),
('figures', 30),
('constraint', 29),
('making', 29),
('differences', 29),
('That', 29),
('criteria', 29),
('phylo-genetic', 29),
('energy', 29),
('fixed', 29),
('loss', 29),
('material', 29),
('2.', 29),
('Suppose', 29),
('programming', 29),
('extent', 29),
('clearly', 29),
('arbitrary', 29),
('multipath', 29),
('noncontiguous', 29),
('vector', 29),
('arises', 29),
('considerably', 29),
('effects', 29),
('autocorrelation', 29),
('propose', 29),
('pairwise', 29),
('property', 29),
('Formula', 29),
('theater', 29),
('al', 29),
('Goldanskii', 29),
('8', 29),
('What', 29),
('better', 29),
('nondiagonal', 29),
('electro-magnetic', 29),
('produces', 28),
('therein', 28),
```

```
('pairs', 28),
('noninferior', 28),
('computed', 28),
('outward', 28),
('conjecture', 28),
('How', 28),
('source', 28),
('nonpotential', 28),
('arbitrarily', 28),
('seconds', 28),
('predetermined', 28),
('mmass-', 28),
('Based', 28),
('Steps', 28),
('microscale', 28),
('motion', 28),
('non-holonomic', 28),
('introduce', 28),
('extra-cellular', 28),
('element', 28),
('image', 28),
('sufficiently', 28),
('appropriate', 28),
('recent', 28),
('nonsupported', 28),
('Its', 28),
('multiwavelength', 28),
('onto', 28),
('measures', 28),
('See', 28),
('dependent', 28),
('south', 28),
('nonideal', 28),
('eigenskeleton', 28),
('nonparallel', 27),
('continuously', 27),
('downward', 27),
('orders', 27),
('periods', 27),
('purposes', 27),
('squares', 27),
('integers', 27),
('loop', 27),
('sees', 27),
('parameterize', 27),
('identifiable', 27),
('nonautonomous', 27),
```

```
 ('lemma.', 27),
 ('operators', 27),
 ('proven', 27),
 ('belong', 27),
 ('he', 27),
 ('wave', 27),
 ('All', 27),
 ('non-dimensional', 27),
 ('play', 27),
 ('Figs.', 27),
 ('matter', 27),
 ('non-equilibrium', 27),
 ('area', 27),
 ('bidirectional', 27),
 ('nonstationarity', 27),
 ('panel', 27),
 ('noncommutative', 27),
 ('Type', 27),
 ('however,', 27),
 ('non-linearly', 27),
 ('improve', 27),
 ('important', 27),
 ('right-hand', 27),
 ('hr', 27),
 ('examples', 27),
 ('dynamical', 27),
 ('left', 27),
 ('level', 27),
 ...]
```

[138]:
```
wc = WordCloud(background_color="white", max_words=2000)
wc.generate_from_frequencies(freq)
print ("Word Cloud for Instered Words")
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

```
Word Cloud for Instered Words
```



[153]:
```
vocab = []
rows = df[df['Label']==1]
for _,row in rows.iterrows():
  try:
    delw = row['del_word']
    insw = row['ins_word']
    both = delw + '' + insw
    both.strip()
    vocab.append(both)
  except:
    print(row['ins_word'])
    break
  #print(row['del_word'])
#wrds = [wrd for wrd in vocab if wrd != '']
```

[154]:
```
freq = list_to_dict(vocab)
```

[155]:
```
len(freq.keys())
```

[155]: 70665

[156]:
```
frq2=sorted(freq.items(), key=lambda x: x[1], reverse=True)
frq2
```

[156]:
```
[(',', 105367),
 (' -', 32446),
 ('the ', 20597),
 ('- ', 12854),
 (':', 11949),
 ('.', 8745),
 ('a ', 5932),
 (',,', 5817),
 ('that ', 4606),
 (',;', 3913),
```

```
('-', 2693),
(':.', 2323),
('and ', 1806),
(':,', 1794),
('sectionSection', 1704),
(',:', 1441),
('.,', 1401),
('whichthat', 1299),
('isare', 1281),
(' ,', 1276),
('aan', 1201),
(',and ', 1192),
('.:', 1191),
('an ', 1166),
('of ', 1133),
("n'tnot", 1113),
('athe', 1103),
(',.', 1099),
('thatwhich', 1083),
('thea', 1082),
("'s", 1076),
('areis', 1025),
('towardstoward', 976),
(';,', 954),
('the ,', 949),
(',the ', 922),
('it ', 901),
('to ', 836),
("'", 799),
('..', 795),
('-,', 776),
('is ', 757),
('th', 717),
('3three', 666),
('figureFigure', 617),
('ana', 595),
('2two', 594),
('non-linearnonlinear', 578),
('non-zerononzero', 574),
('in ', 545),
('as ', 541),
('equationEquation', 530),
('4four', 511),
('then ', 488),
(', ', 484),
(';', 478),
('for ', 471),
```

```
('eqEq', 459),
('Theoremtheorem', 453),
('non-negativenonnegative', 452),
('theoremTheorem', 448),
('one ', 442),
(',-', 442),
('by ', 438),
('as well asand', 426),
(';.', 422),
('nonlinearnon-linear', 421),
('the the ', 419),
(',that ', 415),
('.;', 398),
('inof', 395),
('theThe', 393),
('us ', 389),
('hashave', 381),
('"\'', 375),
('Sectionsection', 373),
('be ', 359),
(' -,', 345),
('existsexist', 344),
('figFig', 342),
('inon', 339),
('The ', 323),
('towith', 323),
('i.e.that is', 321),
('5five', 320),
('non-trivialnontrivial', 316),
('. ', 306),
('we ', 303),
('havehas', 303),
('iswas', 302),
(' and,', 299),
("'sus ", 297),
('whenwhere', 293),
('tableTable', 286),
('lemmaLemma', 286),
('Lemmalemma', 283),
(' the ', 281),
('Appendixappendix', 274),
('equation ', 272),
("'llwill ", 263),
(' .', 262),
('6six', 261),
('onin', 261),
('can ', 260),
```

```
('non-emptynonempty', 258),
('thathave', 253),
('SecSect', 253),
('anthe', 247),
('notationsnotation', 246),
("'sis ", 241),
('1one', 236),
('be be ', 236),
('toof', 233),
('onof', 229),
('ininto', 229),
('are ', 228),
('arewere', 228),
('Ref. ', 225),
('thean', 224),
(',a ', 224),
('cancellationcancelation', 224),
('Modelmodel', 222),
('parametrizationparameterization', 220),
('8eight', 220),
(': ', 220),
('suchso', 218),
('whilewhereas', 216),
('following ', 216),
('\'"', 213),
('appendixAppendix', 209),
('with ', 208),
('. : ', 208),
('two2', 202),
('Thethe', 200),
('have ', 200),
('i.e.that is,', 198),
('"', 196),
('to be ', 195),
('existexists', 193),
('nontrivialnon-trivial', 191),
(' percent%', 188),
('ofin', 187),
('one1', 187),
('holdshave', 186),
(';:', 185),
('-the ', 182),
('will ', 180),
('waswere', 178),
('a,', 175),
('2ii', 175),
('we have ', 174),
```

```
('willshall', 173),
('Seesee', 167),
(':the ', 158),
('in order ', 158),
('havesee', 158),
('also also ', 155),
("can'tcannot", 154),
('ifIf', 150),
('a ,', 148),
('let ', 148),
('couldcan', 147),
('withto', 145),
('the equationEq.', 144),
('Problemproblem', 143),
('dueowing', 143),
('Algorithmalgorithm', 143),
(' :', 141),
('coexistenceco-existence', 140),
('on ', 140),
('inat', 139),
('equationsEquations', 139),
('the,', 138),
('weWe', 137),
('detailsdetail', 136),
('workswork', 135),
(', . ', 131),
('the a ', 129),
('In order toTo', 128),
('withby', 127),
("won'twill not", 127),
('multi-coremulticore', 127),
('FigFigs', 126),
('gaussianGaussian', 126),
('nonzeronon-zero', 126),
("'vehave ", 126),
('cf.see', 124),
('tofrom', 124),
(', the', 124),
('forFor', 123),
('fromby', 123),
('offor', 122),
('7seven', 122),
('non-linearitynonlinearity', 122),
('in,', 122),
(':;', 121),
('thisthese', 120),
('non-decreasingnondecreasing', 120),
```

```
('canmay', 120),
('pre-computationprecomputation', 119),
(':-', 118),
('1i', 118),
(' . ', 117),
('MATHMATHDISP', 117),
('see ', 116),
('s', 116),
('non-degeneratenondegenerate', 116),
("'s,", 115),
('Chapterchapter', 115),
('that ,', 115),
('ofon', 115),
('out ', 114),
('thenthan', 114),
('non-uniformnonuniform', 113),
("'sthe ", 111),
('non-increasingnonincreasing', 111),
('byin', 111),
('has beenwas', 110),
('noncommutativenon-commutative', 109),
('c.fcf', 108),
('useused', 108),
('Andand', 107),
('on-lineonline', 107),
('werewas', 106),
('e.g.for example', 106),
('showshows', 105),
('isbe', 105),
('the -', 104),
('Equation ', 104),
('they ', 102),
('asby', 102),
('chapterChapter', 102),
('sosuch', 101),
('nonnegativenon-negative', 101),
(',the', 100),
('non-singularnonsingular', 99),
('assince', 98),
('being ', 98),
('aboutof', 98),
('Pomeronpomeron', 97),
('subsectionSubsection', 97),
('paperarticle', 96),
('forof', 96),
('SimilarSimilarly', 96),
('havefind', 96),
```

```
('thenwe have', 96),
('systemssystem', 96),
('only only ', 96),
('showsshow', 96),
('byof', 94),
('of ,', 94),
('non-thermalnonthermal', 94),
('U.S.US', 93),
('see, ', 93),
('thethe ', 93),
('Cloudcloud', 93),
('functionfunctions', 93),
('stepStep', 93),
('holdshold', 93),
('systemsystems', 93),
('FromBy', 92),
('forto', 91),
('inby', 91),
('3iii', 91),
('Systemsystem', 91),
('from ', 91),
(' a ', 91),
(' that,', 91),
("'reare ", 90),
('section,', 89),
('cut-offcutoff', 89),
('numbernumbers', 89),
('e.g.for example,', 89),
('the :', 89),
('intoin', 88),
('which,', 88),
('tofor', 87),
('FirstlyFirst', 87),
('Thenthen', 86),
('Wewe', 86),
('aA', 86),
('thatthose', 86),
('three3', 86),
('BecauseSince', 86),
('SubsectionSection', 85),
("',", 85),
(' -:', 85),
('is,', 84),
('conditionCondition', 84),
(' then,', 84),
('itstheir', 83),
('toin', 83),
```

```
('AsSince', 83),
('- and ', 83),
('belowin the following', 83),
('five5', 83),
('tointo', 82),
('of the ', 82),
('fromof', 80),
('-:', 80),
('Conditioncondition', 80),
(', respectively', 80),
(' as:', 80),
('Wherewhere', 80),
('Forfor', 79),
('whenas', 79),
(', -', 79),
('problemProblem', 79),
('are,', 79),
('10001,000', 78),
('so ', 78),
('that,', 78),
('similarsimilarly', 77),
('. .', 77),
('Methodmethod', 77),
('. , ', 76),
('implyimplies', 76),
('onewe', 76),
('inIn', 75),
('eachall', 75),
('inputsinput', 75),
('Appendixthe appendix', 75),
('and ,', 75),
('itIt', 75),
('the .', 75),
('itthis', 75),
('propositionProposition', 75),
(',is ', 75),
('doesdo', 74),
('at ', 74),
('ItThis', 74),
(',of ', 74),
('valuevalues', 74),
("it'sits", 74),
('followfollows', 73),
('laterlatter', 73),
('the result ', 73),
('EqsEq', 73),
('pre-definedpredefined', 73),
```

```
('to,', 73),
('-.', 73),
('where ', 72),
('multi-dimensionalmultidimensional', 72),
('asbecause', 72),
('pre-processingpreprocessing', 72),
('set-upsetup', 72),
('nonsingularnon-singular', 72),
('pseudorandompseudo-random', 71),
('.?', 71),
('; .', 71),
('athe ', 71),
('Seriesseries', 71),
('it ,', 71),
(' and;', 70),
('whatwhich', 70),
('eqsEqs', 70),
(':- ', 70),
('also ', 70),
('9nine', 70),
(', and', 69),
('setupset-up', 69),
('criteriacriterion', 69),
('FigureFigures', 69),
('indexesindices', 69),
('conditionconditions', 69),
('Refs. ', 69),
('%', 69),
('see,', 69),
('The equationEquation', 69),
('ofthe', 68),
('2ndsecond', 67),
('analogueanalog', 67),
('obtainfind', 67),
('non-vanishingnonvanishing', 67),
('lessfewer', 67),
('We have ', 67),
('to ,', 66),
('thisThis', 66),
('it is ', 66),
('Eq. ', 66),
('beingis', 65),
('.  ', 65),
('so-called ', 65),
('; . ', 65),
("''", 65),
('Figurefigure', 64),
```

```
('being being ', 64),
('Ifif', 63),
('beis', 63),
('isbeing', 63),
('findfound', 63),
('of,', 63),
('.; ', 63),
('everyall', 63),
('spacetimespace-time', 62),
('whileand', 62),
('asto', 62),
('followingfollows', 62),
('occurenceoccurrence', 62),
('equationequations', 62),
('Principleprinciple', 62),
('becomebecomes', 62),
('0', 62),
('cutoffcut-off', 62),
('SecondlySecond', 61),
('which ', 61),
('AAn', 61),
('sunSun', 61),
('resultresults', 61),
('resultsresult', 61),
('termterms', 61),
('true ', 61),
(' _CITE__CITE_ ', 61),
(',an ', 60),
('isthe ', 60),
('sectionsSections', 60),
('fibrefiber', 60),
('infor', 59),
('non-dimensionalnondimensional', 59),
('firstlyfirst', 59),
('10ten', 59),
('denotesdenote', 59),
('ButHowever,', 59),
('onto', 59),
('AThe', 58),
(' the,', 58),
('non-overlappingnonoverlapping', 58),
(', and;', 58),
('type ', 58),
('timetimes', 58),
('a functionfunctions', 58),
('onlineon-line', 58),
('suchthis', 58),
```

```
('pre-impactpreimpact', 58),
('exampleExample', 58),
('ten10', 58),
(':that ', 58),
(': . ', 58),
('modelsmodel', 57),
(', and ', 57),
('as,', 57),
('onat', 57),
('abouton', 57),
('withof', 57),
('iidi.i.d.', 57),
('twelve12', 57),
('nonperturbativenon-perturbative', 57),
('non-smoothnonsmooth', 57),
('as ,', 57),
('have beenwere', 57),
('modelModel', 57),
('non-codingnoncoding', 56),
('FigsFig', 56),
('atin', 56),
(';, and', 56),
('webWeb', 56),
('introductionIntroduction', 56),
('Propositionproposition', 56),
('non-stationarynonstationary', 56),
('ishas been', 56),
(' and ', 56),
('whilewhilst', 55),
('toat', 55),
(' :.', 55),
('phenomenaphenomenon', 55),
('ofto', 55),
(' that', 55),
('literaturesliterature', 55),
('givesgive', 55),
('getfind', 55),
('givegives', 55),
('been been ', 54),
('straight-forwardstraightforward', 54),
('And theThe', 54),
('Let usWe', 54),
('aboveearlier', 54),
('WhileAlthough', 54),
('wasis', 54),
('.and ', 53),
('hoursh', 53),
```

```
('Itit', 53),
('occuringoccurring', 53),
('parametersparameter', 53),
('thenThen', 53),
('discdisk', 53),
('such ', 53),
('into', 53),
('solutionsolutions', 53),
('thatthan', 53),
('or ', 53),
(':and ', 53),
('3rdthird', 53),
(',by ', 52),
('linelines', 52),
('setsets', 52),
('happenoccur', 52),
('for,', 52),
('was ', 52),
(': the following.', 52),
('higgsHiggs', 52),
('Section ', 52),
('inwith', 52),
('semi-definitesemidefinite', 52),
('is is ', 52),
('equation,', 52),
('twenty20', 51),
('the sake of ', 51),
('directiondirections', 51),
('dodoes', 51),
(', ,', 51),
('a the ', 51),
('this ', 51),
('statestates', 51),
('explicitelyexplicitly', 51),
('toby', 51),
('leadleads', 51),
('satisfiessatisfy', 51),
('thatthis', 51),
('secs', 51),
('. ; ', 51),
('four4', 51),
('Casecase', 51),
('-a ', 51),
('that the ', 50),
('thereThere', 50),
('holds ', 50),
('follows: ', 50),
```

```
('sub-graphsubgraph', 50),
('",', 50),
('due tofrom', 50),
('On the other handHowever', 49),
('multi-scalemultiscale', 49),
('nonemptynon-empty', 49),
('spatio-temporalspatiotemporal', 49),
('greygray', 49),
('toon', 49),
('obtainsee', 49),
('asto be', 49),
('foras', 49),
('non-magneticnonmagnetic', 49),
('EqnsEqs', 49),
('Ref.', 49),
('non-secretorsnonsecretors', 49),
('Gridgrid', 48),
('4thfourth', 48),
('Thatthat', 48),
('dependsdepend', 48),
('suchthese', 48),
('will beis', 48),
('viaby', 48),
('Logiclogic', 48),
('dD', 48),
('modelmodels', 48),
('eleven11', 48),
('nonrelativisticnon-relativistic', 48),
('up ', 48),
('multi-agentmultiagent', 48),
('itemItem', 47),
('iei.e.', 47),
('asand', 47),
('getsee', 47),
('leadslead', 47),
('non-equilibriumnonequilibrium', 47),
('definitionDefinition', 47),
('about ', 47),
('following ,', 47),
('functionsfunction', 47),
('withas', 47),
('non-congruentnoncongruent', 47),
('typetypes', 47),
('Equationequation', 47),
('minutesmin', 47),
('arehave been', 47),
(' the following', 47),
```

```
('sincebecause', 46),
('pre-existingpreexisting', 46),
('thisthe', 46),
('bakerBaker', 46),
('which is ', 46),
('.the ', 46),
(' -;', 46),
('processprocesses', 46),
('aboutregarding', 46),
('Universeuniverse', 46),
('non-localnonlocal', 46),
('ref. ', 46),
('MATHDISPMATH', 46),
('solutionssolution', 45),
('a a ', 45),
('for all ', 45),
(',in ', 45),
('partparts', 45),
('thanfrom', 45),
('Aa', 45),
('an,', 45),
('shall ', 45),
('mightmay', 45),
('correspondcorresponds', 45),
(',and', 45),
('iffif and only if', 45),
('orand', 45),
('satisfysatisfies', 45),
('arebe', 45),
('uponon', 45),
('TheA', 45),
('increaseincreases', 45),
('largerhigher', 44),
('P.L.Kapitsa', 44),
(',that', 44),
('thatthe ', 44),
('the abovethis', 44),
('isto be', 44),
('holdholds', 44),
('casecases', 44),
('differentvarious', 44),
('is ,', 44),
('SSection ', 44),
('WhereHere', 44),
('paringpairing', 44),
(',then ', 44),
('Figure Figures', 44),
```

```
('non-negligiblenonnegligible', 43),
('transferedtransferred', 43),
('non-convexnonconvex', 43),
('thosethese', 43),
('nonparametricnon-parametric', 43),
('on,', 43),
('hastomust', 43),
(' one', 43),
('in ,', 43),
('-, ', 43),
('provideprovides', 43),
('space-timespacetime', 43),
('whilstwhile', 43),
('noror', 43),
('when,', 43),
('itsthe', 42),
('allowallows', 42),
(',we have ', 42),
('considerconsidered', 42),
('asin', 42),
('the fact ', 42),
('have have ', 42),
('Fig. Figs', 42),
('tooalso', 42),
('retargetre-target', 42),
('caseCase', 42),
('mash-upmashup', 42),
('bywith', 42),
(' the following:', 42),
('non-loopnonloop', 42),
('showshowed', 42),
('hamiltonianHamiltonian', 42),
('denotedenotes', 42),
('enoughsufficient', 41),
('pre-specifiedprespecified', 41),
('thethis', 41),
('Inin', 41),
('corollaryCorollary', 41),
('upwardsupward', 41),
('problemproblems', 41),
('Fig. ', 41),
('needneeds', 41),
('andor', 41),
('Theorytheory', 41),
('informationsinformation', 41),
('formforms', 41),
('towardtowards', 40),
```

```
('AlexandrovAleksandrov', 40),
('degreedegrees', 40),
('multi-gridmultigrid', 40),
('has ', 40),
('non-randomnonrandom', 40),
('assureensure', 40),
('assuresensures', 40),
(': .', 40),
(',a', 40),
('saysee', 40),
('NoticeNote', 39),
('parametrizationsparameterizations', 39),
('maycan', 39),
('Typetype', 39),
('methodmethods', 39),
('1stfirst', 39),
('consistconsists', 39),
('multi-portfoliomultiportfolio', 39),
('the abovethese', 39),
('kindkinds', 39),
('notno', 39),
('tothe ', 39),
('Semisemi', 39),
('distributiondistributions', 39),
('yieldsyield', 39),
('solar windSW', 39),
('thanas', 39),
('Queuequeue', 39),
('arethe ', 39),
('secondss', 39),
('principleprincipal', 39),
('non-conservativenonconservative', 39),
('parameterparameters', 39),
('then ,', 38),
(',to ', 38),
('that of ', 38),
('lemma,', 38),
('occuredoccurred', 38),
('Physicsphysics', 38),
('algorithmAlgorithm', 38),
('hyper-chaotichyperchaotic', 38),
('non-rigidnonrigid', 38),
('ishas', 38),
('getobtain', 38),
('that: ', 38),
('thea ', 37),
('aton', 37),
```

```
('Inequalityinequality', 37),
('backwardsbackward', 37),
('below: ', 37),
('wethen', 37),
('outputsoutput', 37),
('offrom', 37),
('intoto', 37),
('FollowingThe following', 37),
('the a', 37),
('ofas', 37),
('theoremsTheorems', 37),
('knowknown', 37),
('non-positivenonpositive', 37),
('more ', 37),
('there ', 37),
('already ', 37),
(' to -', 37),
('tradeofftrade-off', 36),
('non-parametricnonparametric', 36),
('the equationEquation', 36),
('figuresFigures', 36),
('lesslower', 36),
('wherein which', 36),
('will beare', 36),
('/-', 36),
(', see', 36),
('1I', 36),
('onfor', 36),
('needsneed', 36),
('becomesbecome', 36),
('to the ', 36),
(',for ', 36),
('minminutes', 36),
('cartesianCartesian', 36),
('thosethe', 36),
('Columncolumn', 36),
('and the ', 36),
('been ', 36),
('thethat', 36),
('such athis', 36),
('path-wisepathwise', 36),
('assuredensured', 36),
(', _CITE_', 36),
('herewhere', 35),
('butand', 35),
('EqEqs', 35),
('conditionsConditions', 35),
```

```
('amongstamong', 35),
('dynamicdynamics', 35),
('non-standardnonstandard', 35),
('yieldyields', 35),
('representsrepresent', 35),
('need tomust', 35),
('whichand', 35),
('forin', 35),
('fermiFermi', 35),
('Splinespline', 35),
('thatthe', 35),
('donemade', 35),
('containscontain', 35),
(' as,', 35),
('2,', 35),
('both both ', 35),
('sub-surfacesubsurface', 35),
('in the ', 34),
('Subsectionsubsection', 34),
('afterwardsafterward', 34),
('cancould', 34),
('::', 34),
('an ,', 34),
("RGE'sRGEs", 34),
('smallerlower', 34),
('re-writtenrewritten', 34),
('SimilarA similar', 34),
('non-contractiblenoncontractible', 34),
('SkorohodSkorokhod', 34),
('wouldwill', 34),
('anti-phaseantiphase', 34),
('-hand', 34),
('playerPlayer', 34),
('focusedfocussed', 34),
('representrepresents', 34),
('becausesince', 34),
('non-linearitiesnonlinearities', 34),
('brownianBrownian', 34),
('workworks', 34),
('Analysisanalysis', 33),
('thesethis', 33),
(', we have', 33),
('providesprovide', 33),
('showedshown', 33),
('nearclose', 33),
('impliesimply', 33),
('wereare', 33),
```

```
(' below', 33),
('whenWhen', 33),
('flatfieldflat-field', 33),
('equationsequation', 33),
('simulatabilitySimulatability', 33),
(' as follows', 33),
('betweenamong', 33),
(']', 33),
('signallingsignaling', 33),
('tilluntil', 33),
('ifwhether', 33),
('followingsfollowing', 33),
('every,', 33),
('maxMax', 33),
('Then thethen ', 33),
('anand', 33),
('opticoptical', 33),
('Therethere', 33),
('cycleCycle', 33),
('Tabletable', 33),
('zero0', 32),
('needstomust', 32),
('pointpoints', 32),
('pre-computedprecomputed', 32),
('byas', 32),
(' true', 32),
('the following:  ', 32),
('as the ', 32),
(',exist', 32),
('Thm.Theorem', 32),
('sub-optimalsuboptimal', 32),
('nonlinearitynon-linearity', 32),
('non-cooperativenoncooperative', 32),
(';and ', 32),
('QPPQPPs', 32),
('Exampleexample', 32),
('likesuch as', 32),
('formfrom', 32),
('behaviorsbehavior', 32),
(', it', 32),
('the equationsEqs.', 32),
('co-ordinatescoordinates', 32),
('happensoccurs', 32),
('has has ', 32),
("ODE'sODEs", 32),
('trans-equatorialtransequatorial', 32),
('non-constantnonconstant', 32),
```

```
('boththe two', 32),
('hardcorehard-core', 32),
('vs.versus', 32),
('if ', 31),
('parametrisationparameterisation', 31),
(':th', 31),
(',- ', 31),
('parenthesisparentheses', 31),
('speciallyespecially', 31),
('NotationsNotation', 31),
('our ', 31),
('the and ', 31),
(' the following.', 31),
('thanksdue', 31),
('correspondinglyrespectively', 31),
('evidencesevidence', 31),
('thirty30', 31),
('non-symmetricnonsymmetric', 31),
('non-homogeneousnonhomogeneous', 31),
('OnIn', 31),
('nondecreasingnon-decreasing', 31),
('3,', 31),
('detaildetails', 31),
(',an', 31),
('separation listSeparation-List', 31),
('assumptionAssumption', 31),
('xX', 31),
('non-normalnonnormal', 31),
('ofthe ', 31),
('toas', 31),
('tothe', 31),
('ThoughAlthough', 31),
('virtue of ', 31),
('multi-levelmultilevel', 31),
('quasi-periodicquasiperiodic', 31),
('ThatThis', 31),
('letLet', 31),
('networksnetwork', 31),
('ROIROIs', 31),
(' and', 30),
('hypothesishypotheses', 30),
('do ', 30),
('. ,', 30),
('axisaxes', 30),
('TheoremTheorems', 30),
('non-linear,', 30),
('methodMethod', 30),
```

```
(',as ', 30),
('but ', 30),
('fifteen15', 30),
('the figureFig.', 30),
('two ', 30),
('Corollarycorollary', 30),
('problemsproblem', 30),
('eq,', 30),
('aloneonly', 30),
('both ', 30),
('LemmaLemmas', 30),
('TowardsToward', 30),
('sub-sectionsubsection', 30),
('pixelspixel', 30),
('Lawlaw', 30),
('anti-topantitop', 30),
('makemakes', 30),
("'llshall ", 30),
('..etc', 30),
('.-', 30),
('sixteen16', 30),
('internetInternet', 30),
('tunnellingtunneling', 30),
('the  ', 30),
('Odderonodderon', 30),
('the presentthis', 29),
('?.', 29),
('phylogeneticphylo-genetic', 29),
('thatas', 29),
('multi-sidedmultisided', 29),
('Westwest', 29),
('Processprocess', 29),
('beare', 29),
('proposeproposed', 29),
('abovepreceding', 29),
('4iv', 29),
('Testtest', 29),
('aa ', 29),
('spacialspatial', 29),
('Rulerule', 29),
(',th', 29),
('subsectionSection', 29),
('anotherthe other', 29),
('Standard Modelstandard-model', 29),
('zeroeszeros', 29),
('anotherother', 29),
('Transformtransform', 29),
```

```
('the following.  ', 29),
('anti-symmetricantisymmetric', 29),
('aircraftaircrafts', 28),
('non-inferiornoninferior', 28),
('3Dthree-dimensional', 28),
('ButHowever', 28),
('non-decreasing,', 28),
('-and ', 28),
('has beenis', 28),
('would ', 28),
('degenerateddegenerate', 28),
('non-negative,', 28),
('non-potentialnonpotential', 28),
('ofby', 28),
('theirthe', 28),
('analysisanalyses', 28),
('theorem,', 28),
('indicateindicates', 28),
('At lastFinally', 28),
('Eq.', 28),
('the following ', 28),
('its ', 28),
('has,', 28),
('pre-determinedpredetermined', 28),
('mass mmass-', 28),
('micro-scalemicroscale', 28),
('remindrecall', 28),
('thatwhen', 28),
('nonholonomicnon-holonomic', 28),
('multi-pathmultipath', 28),
('non-contiguousnoncontiguous', 28),
('extracellularextra-cellular', 28),
('sizesizes', 28),
('non-supportednonsupported', 28),
('nm ', 28),
('vsversus', 28),
('those ', 28),
('theatretheater', 28),
('formsform', 28),
('correspondscorrespond', 28),
('by,', 28),
('electromagneticelectro-magnetic', 28),
('toand', 28),
('PhDPh.D.', 28),
('dependancedependence', 28),
(',for all ', 28),
('Approximationapproximation', 28),
```

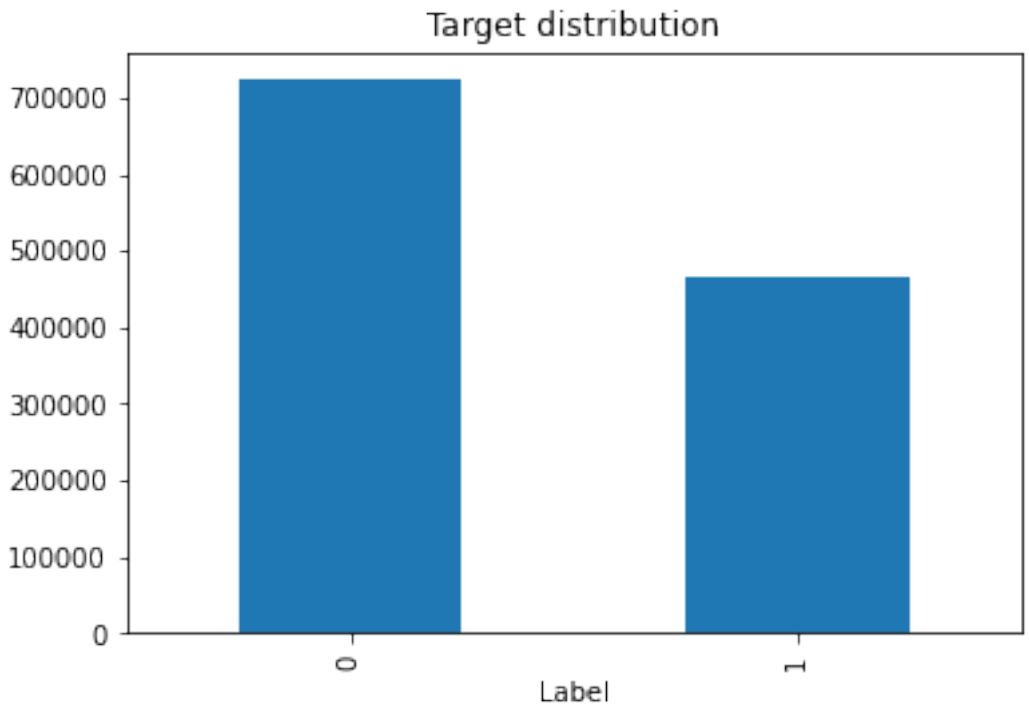
```
 ('muchmany', 27),
 ('non-parallelnonparallel', 27),
 ('5thfifth', 27),
 ...]
```

[157]:
```
wc = WordCloud(background_color="white", max_words=2000)
wc.generate_from_frequencies(freq)
print ("Word Cloud for del ins Words together")
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```

```
Word Cloud for Instered Words
```

[ ]:
```
plt.title('Target distribution')
df.groupby("Label")['SID'].count().plot.bar()
```

[ ]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7b6fe8b790>
```

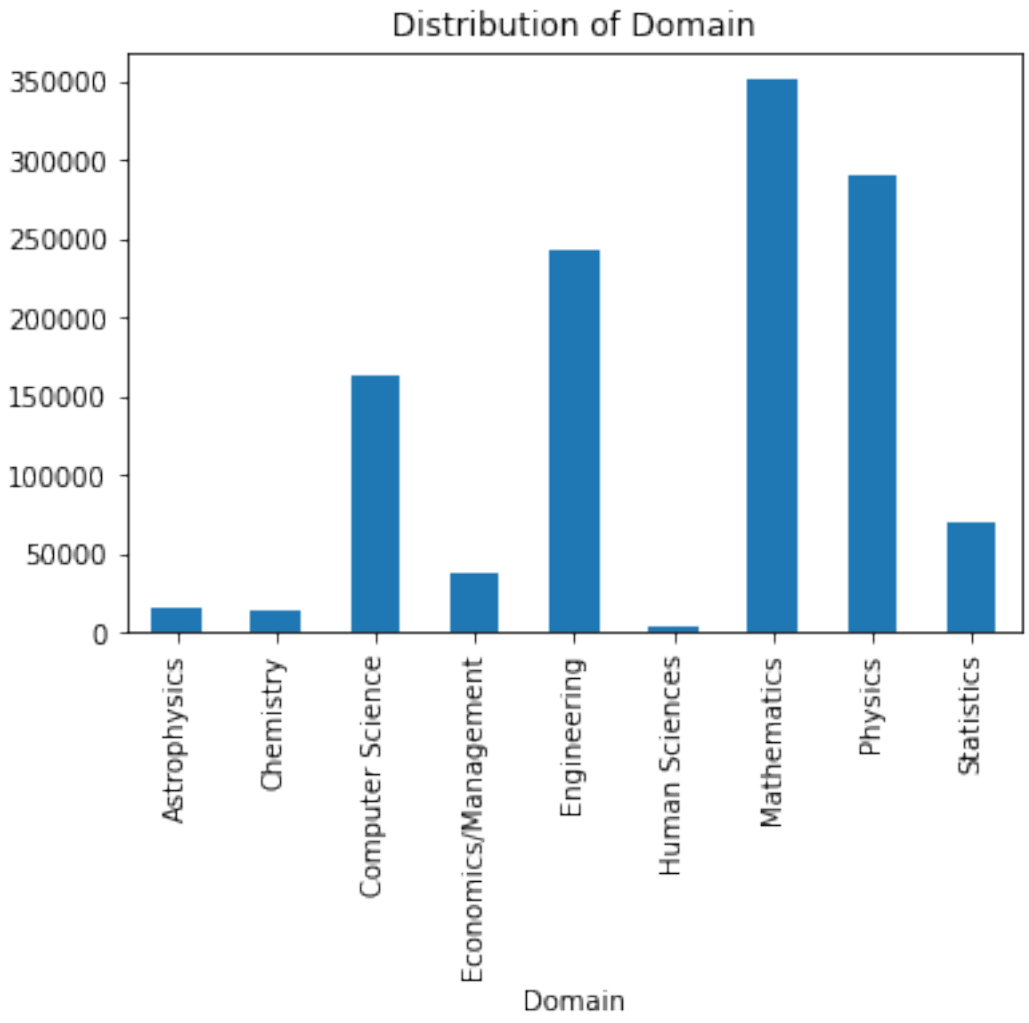


[ ]:
```
print('edited sentences percentage:   {}%'.format(round((df['Label'].
 ↪mean())*100, 2)))
print('not edited sentences percentage:   {}%'.format(100 - round((df['Label'].
 ↪mean())*100, 2)))
```

```
edited sentences percentage:   39.24%
not edited sentences percentage:   60.76%
```

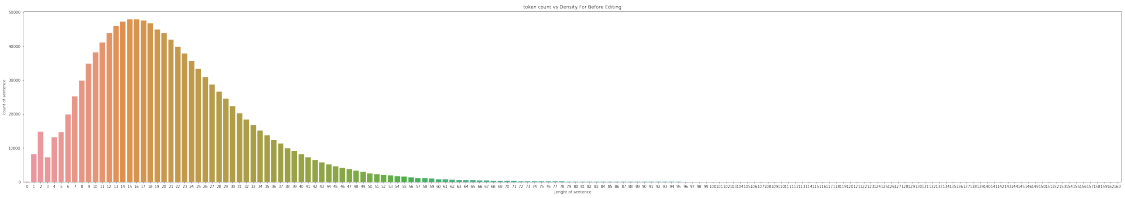
[ ]:
```
plt.title('Distribution of Domain')
df.groupby("Domain")['SID'].count().plot.bar()
```

[ ]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7b6f8c5e50>
```



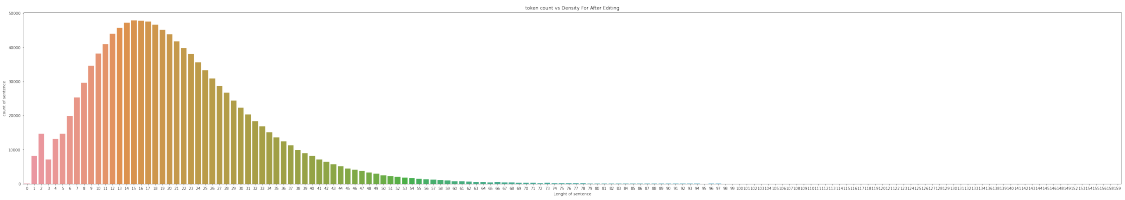
[ ]:
```
plt.figure(figsize=(50, 8))

SBE = df['SBE'].str.split().apply(len).value_counts().sort_index()

SBE = SBE[SBE.index<np.percentile(SBE.index,90)]

plt.title('token count vs Density For Before Editing')
sns.barplot(x=SBE.index ,y=SBE.values,order = SBE.index).set(xlabel='Lenght of␣
 ↪sentence', ylabel='count of sentence')
```

[ ]:
```
[Text(0, 0.5, 'count of sentence'), Text(0.5, 0, 'Lenght of sentence')]
```



[ ]:
```
plt.figure(figsize=(50, 8))

SBE = df['SAE'].str.split().apply(len).value_counts().sort_index()

SBE = SBE[SBE.index<np.percentile(SBE.index,90)]

plt.title('token count vs Density For After Editing')
sns.barplot(x=SBE.index ,y=SBE.values,order = SBE.index).set(xlabel='Lenght of␣
 ↪sentence', ylabel='count of sentence')
```

[ ]:
```
[Text(0, 0.5, 'count of sentence'), Text(0.5, 0, 'Lenght of sentence')]
```



[158]:
```
df['SBE_len'] = df['SBE'].str.len()
df['SAE_len'] = df['SAE'].str.len()
df['SAE_n_words'] = df['SAE'].astype('str').apply(lambda row: len(row.split("␣
 ↪")))
df['SBE_n_words'] = df['SBE'].astype('str').apply(lambda row: len(row.split("␣
 ↪")))
```

[159]:
```
df
```

[159]:

	SID	Domain	...	SAE_n_words	SBE_n_words
0	1.0	Physics	...	35	36
1	1.1	Physics	...	21	21
2	1.2	Physics	...	22	22
3	1.3	Physics	...	11	11
4	2.0	Mathematics	...	8	8
...	...	...	...	...	...
1189407	254143.4	Computer Science	...	24	24
1189408	254143.5	Computer Science	...	37	37
1189409	254144.0	Mathematics	...	23	23
1189410	254144.1	Mathematics	...	33	33

```
1189411  254144.2       Mathematics  ...          15          15

[1189412 rows x 11 columns]
```

[31]:
```
df.shape
```

[31]:
```
(1189412, 11)
```

[160]:
```
# More features form quora question pair similarity
def more_features_from_QQPS(row:str):

  '''
  featurs = WordCommon, WordShare,WordTotal
  '''
  w1 = set(map(lambda word: word.lower().strip(), row['SAE'].split(" ")))
  w2 = set(map(lambda word: word.lower().strip(), row['SBE'].split(" ")))
  common = 1.0 * len(w1 & w2)
  total = 1.0 * (len(w1) + len(w2))
  share = 1.0 * common/total

  return common,total,share
```

[161]:
```
df['word_Common'],df['word_Total'],df['word_share'] = zip(*df.
 ↪apply(more_features_from_QQPS,axis =1))
```

[34]:
```
df.head()
```

[34]:
```
   SID       Domain  ... word_Total word_share
0  1.0      Physics  ...       57.0   0.438596
1  1.1      Physics  ...       40.0   0.500000
2  1.2      Physics  ...       38.0   0.500000
3  1.3      Physics  ...       22.0   0.500000
4  2.0  Mathematics  ...       16.0   0.500000

[5 rows x 14 columns]
```

[36]:
```
df.shape
```

[36]:
```
(1189412, 14)
```

[35]:
```
df.columns
```

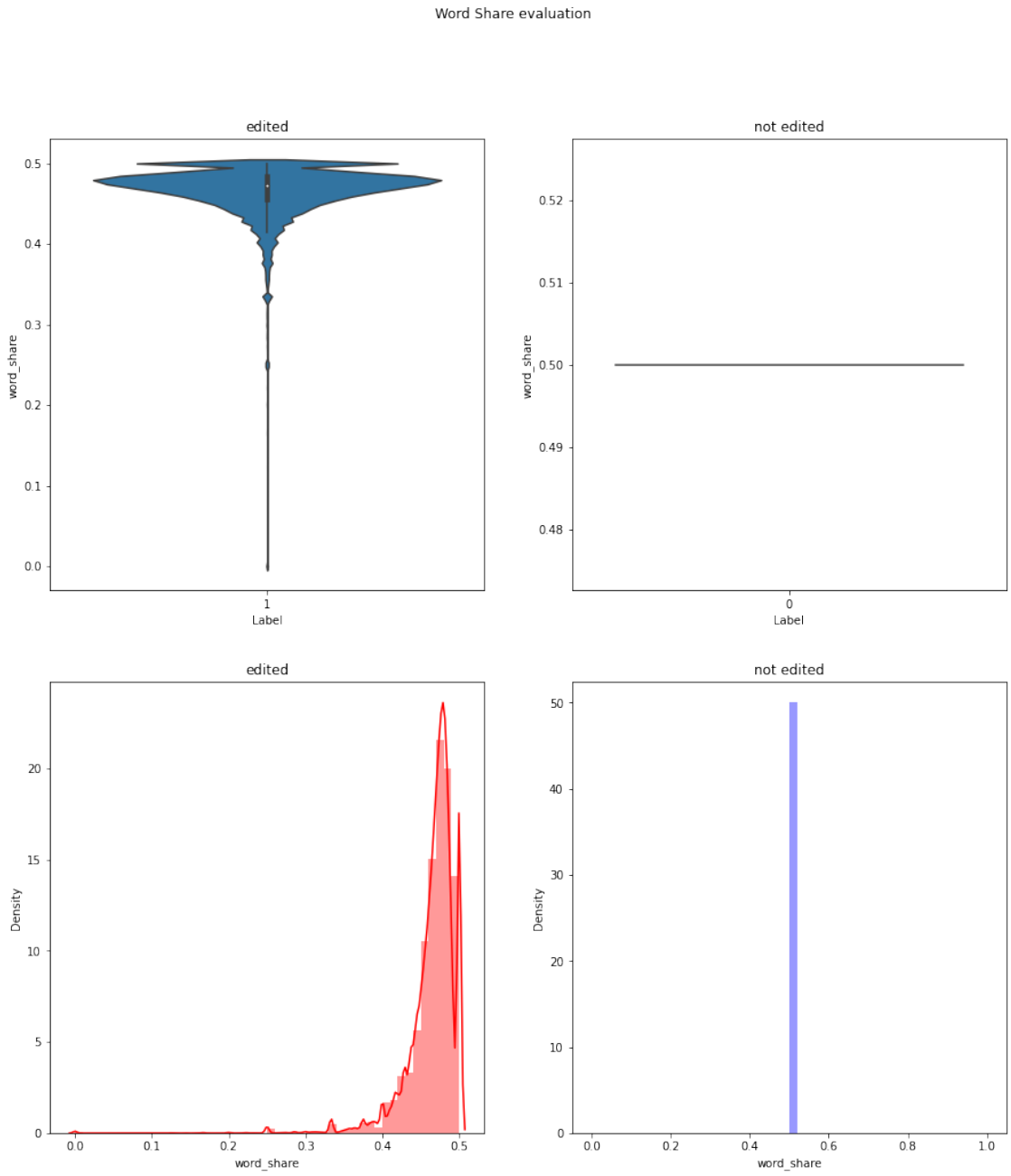
[35]:
```
Index(['SID', 'Domain', 'SBE', 'SAE', 'del_word', 'ins_word', 'Label',
       'SBE_len', 'SAE_len', 'SAE_n_words', 'SBE_n_words', 'word_Common',
       'word_Total', 'word_share'],
      dtype='object')
```

[37]:
```
print ("Minimum length of the SBE : " , min(df['SBE_n_words']))

print ("Minimum length of the questions in SAE : " , min(df['SAE_n_words']))
```

```
print ("Number of Questions with minimum length SBE :", df[df['SBE_n_words']==␣
 ↪min(df['SBE_n_words'])].shape[0])
print ("Number of Questions with minimum length SAE :", df[df['SAE_n_words']==␣
 ↪min(df['SAE_n_words'])].shape[0])
```

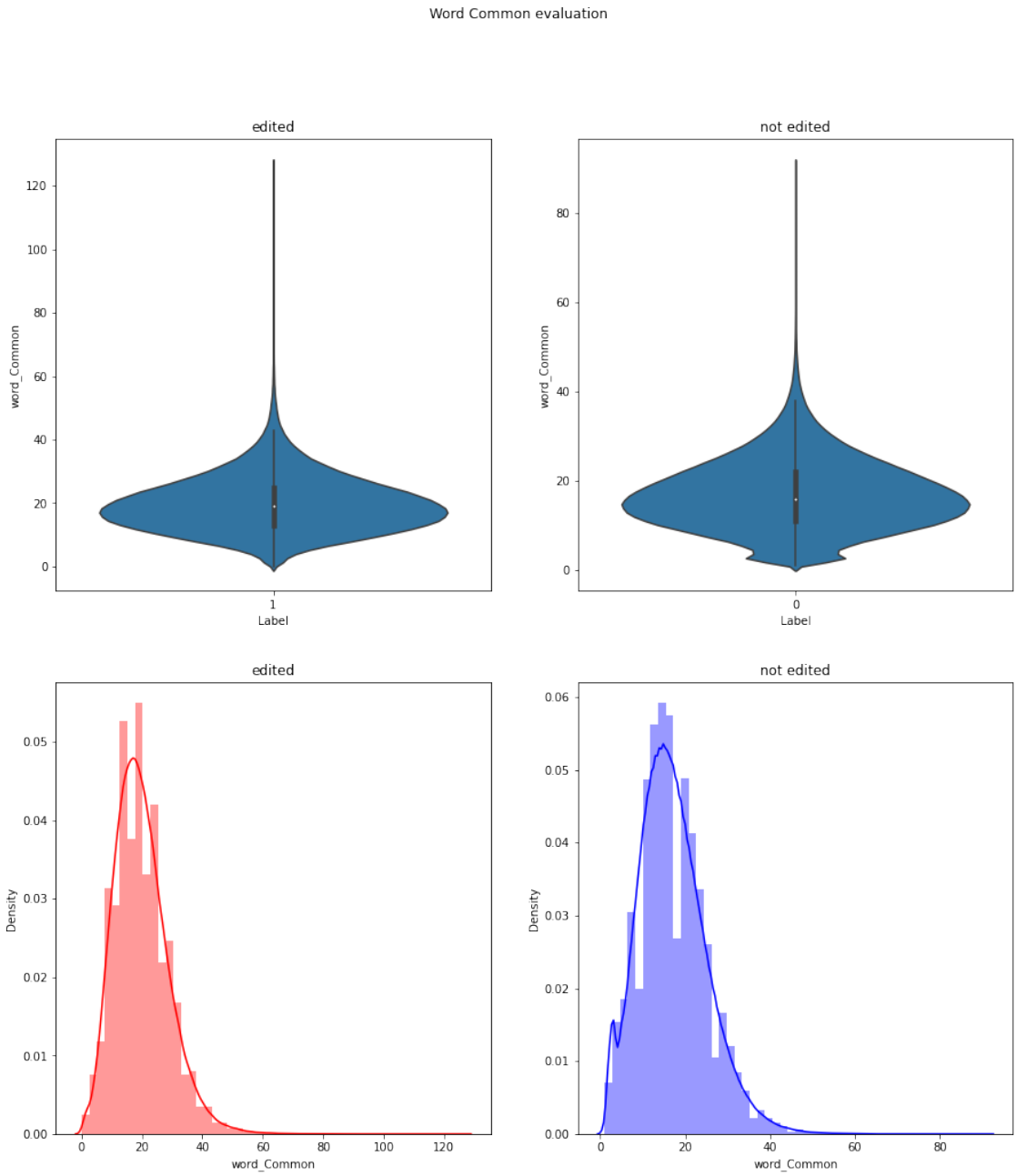
```
Minimum length of the SBE :  1
Minimum length of the questions in SAE :  1
Number of Questions with minimum length SBE : 1213
Number of Questions with minimum length SAE : 1183
```

[38]:
```
plt.figure(figsize=(15, 16))
plt.suptitle('Word Share evaluation')
plt.subplot(2,2,1)
plt.title('edited')
sns.violinplot(x = 'Label', y = 'word_share', data = df[df['Label'] == 1])

plt.subplot(2,2,2)
plt.title('not edited')
sns.violinplot(x = 'Label', y = 'word_share', data = df[df['Label'] == 0])

plt.subplot(2,2,3)
plt.title('edited')
sns.distplot(df[df['Label'] == 1.0]['word_share'][0:] , label = "1", color =␣
 ↪'red')

plt.subplot(2,2,4)
plt.title('not edited')
sns.distplot(df[df['Label'] == 0.0]['word_share'][0:] , label = "0" , color =␣
 ↪'blue' )
plt.show()
```



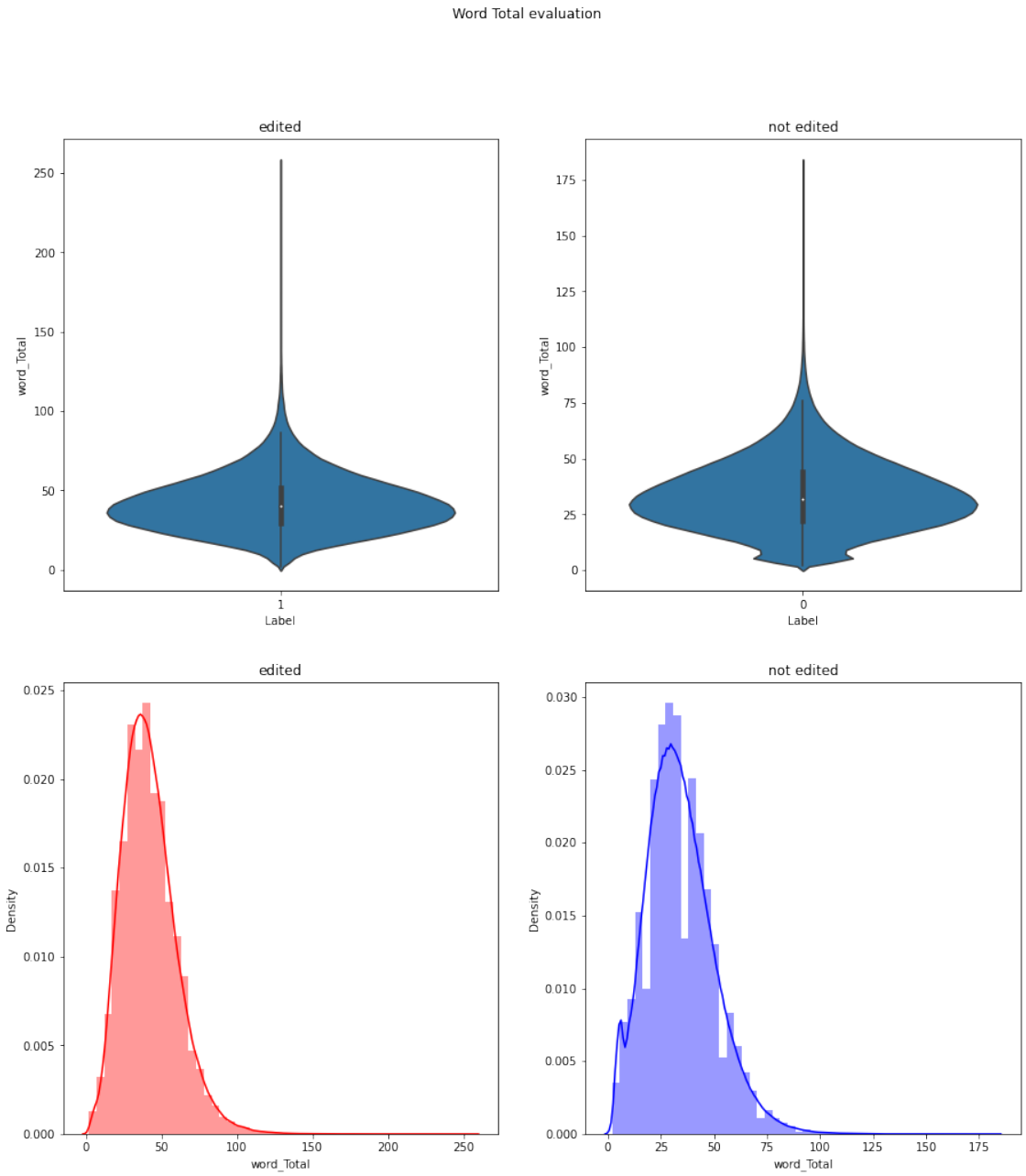
[ ]:
```
plt.figure(figsize=(15, 16))
plt.suptitle('Word Common evaluation')

plt.subplot(2,2,1)
plt.title('edited')
sns.violinplot(x = 'Label', y = 'word_Common', data = df[df['Label'] == 1])

plt.subplot(2,2,2)
```

```
plt.title('not edited')
sns.violinplot(x = 'Label', y = 'word_Common', data = df[df['Label'] == 0])

plt.subplot(2,2,3)
plt.title('edited')
sns.distplot(df[df['Label'] == 1.0]['word_Common'][0:] , label = "1", color =␣
 ↪'red')

plt.subplot(2,2,4)
plt.title('not edited')
sns.distplot(df[df['Label'] == 0.0]['word_Common'][0:] , label = "0" , color =␣
 ↪'blue' )
plt.show()
```



[ ]:
```
plt.figure(figsize=(15, 16))
plt.suptitle('Word Total evaluation')

plt.subplot(2,2,1)
plt.title('edited')
sns.violinplot(x = 'Label', y = 'word_Total', data = df[df['Label'] == 1])

plt.subplot(2,2,2)
plt.title('not edited')
```

```
sns.violinplot(x = 'Label', y = 'word_Total', data = df[df['Label'] == 0])

plt.subplot(2,2,3)
plt.title('edited')
sns.distplot(df[df['Label'] == 1.0]['word_Total'][0:] , label = "1", color =␣
 ↪'red')

plt.subplot(2,2,4)
plt.title('not edited')
sns.distplot(df[df['Label'] == 0.0]['word_Total'][0:] , label = "0" , color =␣
 ↪'blue' )
plt.show()
```



0.0.1 Using Fuzzywuzzy library for more EDA

[162]:
```
df['str_sim'] = df.apply(lambda row:  fuzz.ratio(row['SBE'],row['SAE']),axis=1)
```

[163]:
```
df['par_str_sim'] = df.apply(lambda row:  fuzz.
 ↪partial_ratio(row['SBE'],row['SAE']),axis=1)
```

[164]:
```
df['tok_set_ratio'] = df.apply(lambda row:  fuzz.
 ↪token_set_ratio(row['SBE'],row['SAE']),axis=1)
```

[165]:
```
df['tok_sort_ratio'] = df.apply(lambda row:  fuzz.
↪token_sort_ratio(row['SBE'],row['SAE']),axis=1)
```

[43]:
```
df.shape
```

[43]:
```
(1189412, 18)
```

[44]:
```
df.columns
```

[44]:
```
Index(['SID', 'Domain', 'SBE', 'SAE', 'del_word', 'ins_word', 'Label',
       'SBE_len', 'SAE_len', 'SAE_n_words', 'SBE_n_words', 'word_Common',
       'word_Total', 'word_share', 'str_sim', 'par_str_sim', 'tok_set_ratio',
       'tok_sort_ratio'],
      dtype='object')
```

[166]:
```
df.to_csv(csvfile,index=False)
```

[167]:
```
df = pd.read_csv(csvfile)
```

[168]:
```
df.head()
```

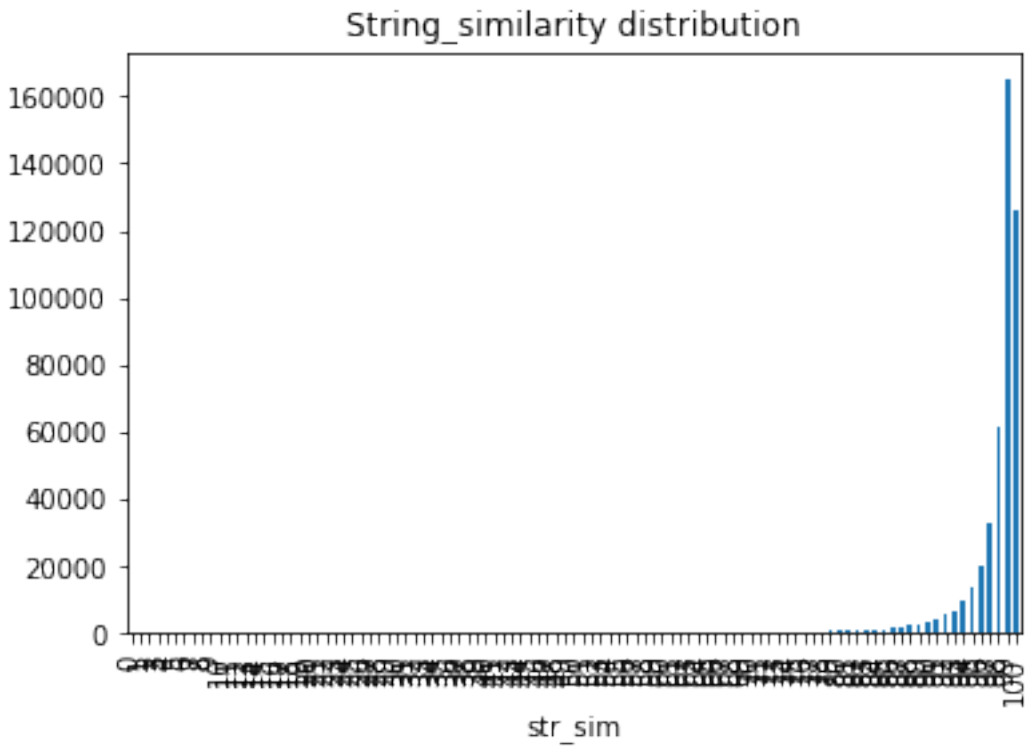
[168]:

	SID	Domain	...	tok_set_ratio	tok_sort_ratio
0	1.0	Physics	...	100	100
1	1.1	Physics	...	100	100
2	1.2	Physics	...	100	100
3	1.3	Physics	...	100	100
4	2.0	Mathematics	...	100	100

[5 rows x 18 columns]

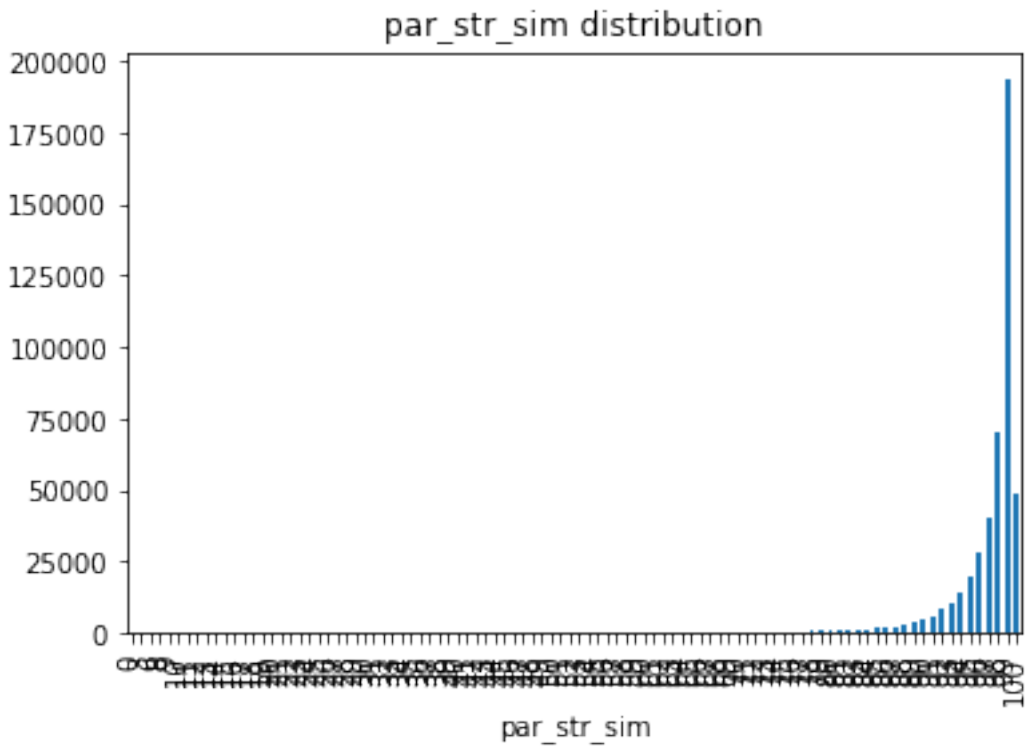
[169]:
```
plt.title('String_similarity distribution')
df[df['Label']==1].groupby('str_sim')['SID'].count().plot.bar()
```

[169]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fea16d1e910>
```



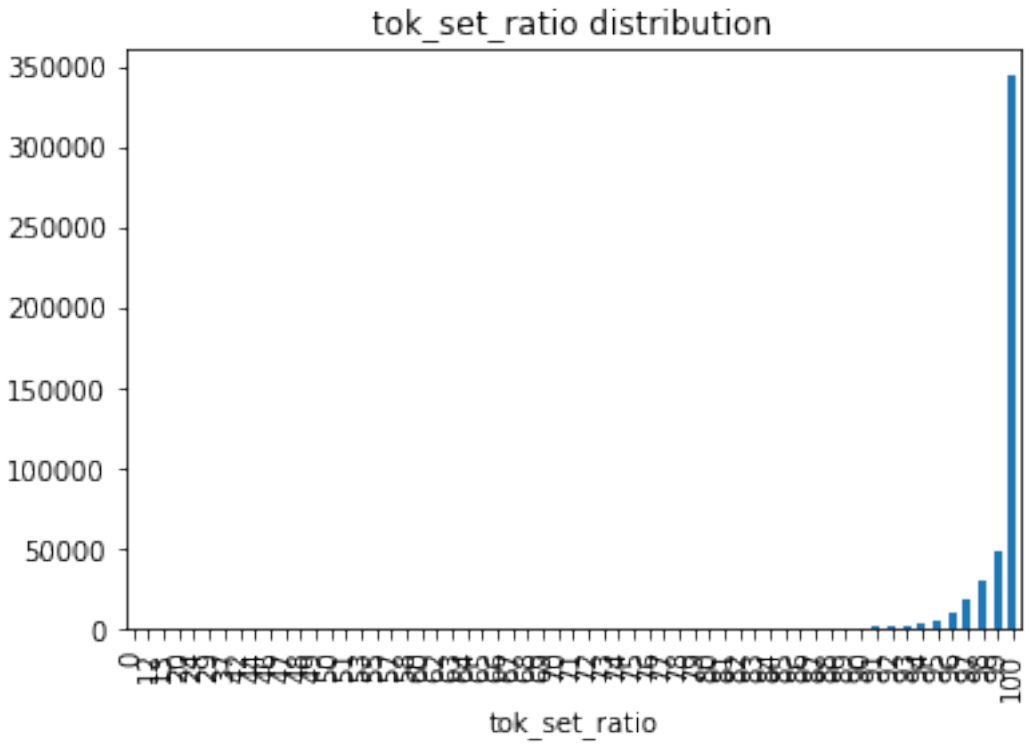
[170]:
```
plt.title('par_str_sim distribution')
df[df['Label']==1].groupby('par_str_sim')['SID'].count().plot.bar()
```

[170]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe9cdc48950>
```



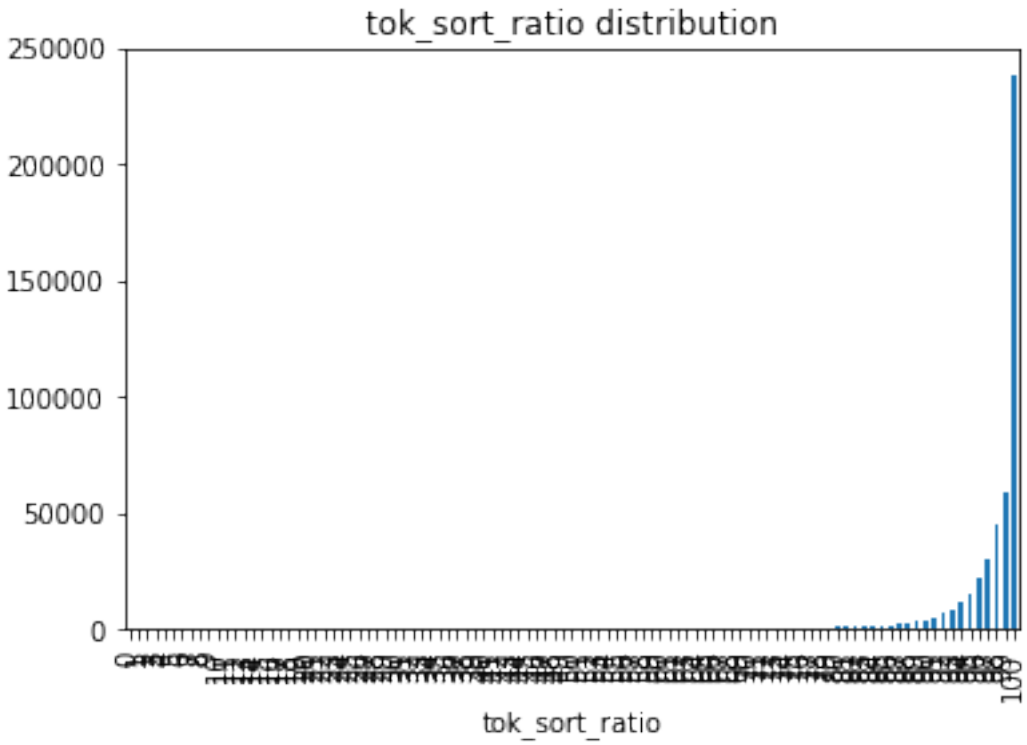
[171]:
```
plt.title('tok_set_ratio distribution')
df[df['Label']==1].groupby('tok_set_ratio')['SID'].count().plot.bar()
```

[171]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe9cda229d0>
```



[172]:
```
plt.title('tok_sort_ratio distribution')
df[df['Label']==1].groupby('tok_sort_ratio')['SID'].count().plot.bar()
```

[172]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe9cda31990>
```


tok_sort_ratio distribution
250000
200000
150000
100000
50000
0
tok_sort_ratio