一、初步清洗

从excel获取内容，只读取文本，或者标题（标题是否要，取决于标题质量），其它一律不用

粗清洗

确保文本与“新疆棉花事件”相关，即确保文本出现相关词条，如：新疆棉花、疆棉、新疆、集中营、强迫劳动，以及新疆棉花事件中涉及的品牌名称如hm等（但一定要确保跟新疆棉花事件相关）

这个步骤完全可以通过正则或者很简单的判定包含关系的语句来完成，如:

if '疆棉' in text...

去重

文本去重，不要在数据集中包含两条相同的文本，直接用python的set()即可

二、精清洗

去除多余字符，如表情字符等各种杂符号串等：

text = "".join(ch for ch in text if unicodedata.category(ch)[0] != 'S') #去除表情等杂乱字符串

中英文符号的转换，英文符号转换成中文符号，具体符号可在数据中探索和补充，可参考以下中转英程序：

def EngPun\_2\_ChnPun(str):#中文符号或冗余符号转英文符号

EngPun=u',.!?[]()<>""""\'\'\';:---. . '

ChnPun=u'，。！？【】（）《》"“"”‘’’；：—-―.….  ️️'

table={ord(c):ord(e) for e,c in zip(EngPun,ChnPun)}

return str.translate(table)

规范表述，将关键名词的表述的统一起来，在中文中涉及的可能有点少，英文涉及更多，比如将h&m统一为HM；nike、耐克、Nike统一为nike等等。采用正则的方式完成，可参考twitter中：

text = re.sub(re.compile('(h&m)',re.IGNORECASE), 'HM', text) # 将HM替换成统一表述

# xjcottonPattern=re.compile('(xinjiang cotton)|(xinjiangcotton)|(xinjiang\'s cotton)',re.IGNORECASE)

text=re.sub(re.compile('(xinjiang)|(xin jiang)|(xingjiang)|(xiangjiang)',re.IGNORECASE),'Xinjiang',text)

text = re.sub(re.compile('(xinjiang cotton)|(xinjiangcotton)|(xinjiang\'s cotton)', re.IGNORECASE),

'Xinjiang Cotton', text) # 统一表述

text=re.sub(re.compile('(forcedlabor)|(forcedlabour)',re.IGNORECASE),'forced labor',text)

text = re.sub(re.compile('(nike)', re.IGNORECASE), 'NIKE', text)

text = re.sub(re.compile('(lining)|(li ning)|(li-ning)', re.IGNORECASE), 'LiNing', text)

text = re.sub(re.compile('(adidas)', re.IGNORECASE), 'Adidas', text)

text = re.sub(re.compile('(muji)', re.IGNORECASE), 'MUJI', text)

text = re.sub(re.compile('(uniqlo)', re.IGNORECASE), 'UNIQLO', text)

text=re.sub(re.compile('u\.s\.',re.IGNORECASE),'US',text)

冗余符号的处理，查看通过完成以上步骤的数据集，看看目前数据集还有哪些符号或者格式没处理统一，在这一步完成。

再次去重