**CA 6**

Due 3/17 11:59 pm

You'll work with the UN-Debate dataset for this CA.

1. Read the dataset as in TextInsight.ipynb discussed in Module 2. Extract the data for the most recent 10 years into a new dataframe. Clean the data and generate, for each article, lemmas, nounds, and noun-phrases.
2. Compute TF-IDF values for lemmas, excluding spaCy stopwords.
3. Compute TF-IDF values for noun-phrases, excluding spaCy stopwords.
4. Create a wordcloud image for the top 80 lemmas in the entire corpus. Define the score for a lemma in the corpus to be the summation of the tf-idf values of the lemmas with respect to individual documents in the corpus. Given below is such a wordcloud image. Note that your output may show a different layout, but the top 80 lemmas should be the same.Text

   Description automatically generated
5. Create a wordcloud for the top 50 nouns in the entire corpus. Given below is such a wordcloud image. Your output may show a different layout, but the top 50 nouns should be the same. Text

   Description automatically generated
6. Create a bar graph for the top 20 noun-phrases for the entire corpus. Define the score for a noun-phrase to be the summation of the tf-idf values of the nouns contained in the noun-phrase, where a noun-phrase may contain adjectives. **Note:** Some words in in a noun-phrase may belong to the stopword list that have been dropped. For such words, assign a small value of 0.0001 as its tf-idf value. Given below is such a bar plot.

Chart, histogram

Description automatically generated

   
      Note: to generate a bar plot with well-aligned keys on the x-axis, you may use the following module:

import matplotlib.pyplot as plt

plt.rcdefaults()  
fig, ax = plt.subplots()

plt.bar(range(len(np\_top)), np\_top.values(), tick\_label=list(np\_top.keys()))  
fig.autofmt\_xdate(rotation=45)   
plt.show()

To serve as references, the following are the top 80 lemmas and the top 50 nouns (based on clean and normalized text):

[('united', 157.52153009249693),

('country', 148.79042594121998),

('international', 142.72790594557364),

('nations', 142.4950682828524),

('development', 131.50496476255412),

('world', 110.55881431742782),

('people', 97.41535999400209),

('security', 97.25706189385761),

('peace', 88.76750235318616),

('global', 72.22783688000361),

('year', 69.17332646347471),

('support', 67.00915107153914),

('human', 64.91958201645909),

('right', 63.744016379892656),

('effort', 62.41677353036296),

('community', 60.36255308284536),

('general', 60.18195935747829),

('change', 59.97154546526611),

('economic', 59.49073840839019),

('need', 58.52374746475446),

('council', 57.41633071020268),

('new', 55.764257664672854),

('states', 55.64086303924497),

('climate', 54.97238450468376),

('state', 54.85122501557446),

('government', 53.83751832310836),

('continue', 53.633667560827796),

('assembly', 53.595978960468784),

('conflict', 53.27234314315861),

('develop', 52.88981815825049),

('challenge', 51.50652815338624),

('organization', 49.32190812620009),

('work', 49.078544409515196),

('crisis', 47.898046998383755),

('african', 45.72099970133691),

('political', 45.57222692614106),

('nation', 44.68731339635627),

('region', 44.1260042728391),

('member', 43.83336308858346),

('issue', 43.247725387249695),

('africa', 42.90401059374601),

('time', 42.77359052633918),

('sustainable', 42.34508528322077),

('achieve', 41.881049082819445),

('reform', 40.94515224440497),

('republic', 40.78085262242604),

('cooperation', 40.62944695220699),

('commitment', 39.838958290071595),

('national', 39.82228518309732),

('nuclear', 39.529108578888426),

('like', 39.45519276904182),

('great', 39.35889453359088),

('action', 38.49021254376047),

('today', 38.41515009421385),

('session', 37.86548492707909),

('process', 37.54717944183682),

('law', 37.13868779532062),

('president', 37.10739029399666),

('terrorism', 36.330097802158804),

('ensure', 35.78582108189829),

('resolution', 35.43818719425585),

('respect', 34.78168626182056),

('order', 34.66405019134354),

('social', 34.5133059047348),

('agenda', 34.49119394790438),

('address', 34.34047498342529),

('remain', 34.3201622562329),

('solution', 34.17931492237972),

('small', 33.87691134181654),

('war', 33.64719967935829),

('include', 33.07832392564047),

('hope', 32.93387335985484),

('situation', 32.886232571196075),

('poverty', 32.72198534796623),

('important', 32.71396501929874),

('principle', 32.60796773581254),

('responsibility', 32.54606689205432),

('problem', 32.4603112345899),

('future', 32.34761998950234),

('progress', 32.24654845334519)]  
  
[('country', 148.79042594121998),

('development', 131.50496476255412),

('world', 110.55881431742782),

('people', 97.41535999400209),

('security', 97.25706189385761),

('peace', 88.76750235318616),

('year', 69.17332646347471),

('support', 67.00915107153914),

('right', 63.744016379892656),

('change', 59.97154546526611),

('state', 54.85122501557446),

('conflict', 53.27234314315861),

('challenge', 51.50652815338624),

('organization', 49.32190812620009),

('nation', 44.68731339635627),

('region', 44.1260042728391),

('member', 43.83336308858346),

('issue', 43.247725387249695),

('time', 42.77359052633918),

('reform', 40.94515224440497),

('cooperation', 40.62944695220699),

('commitment', 39.838958290071595),

('action', 38.49021254376047),

('session', 37.86548492707909),

('process', 37.54717944183682),

('terrorism', 36.330097802158804),

('respect', 34.78168626182056),

('order', 34.66405019134354),

('agenda', 34.49119394790438),

('solution', 34.17931492237972),

('war', 33.64719967935829),

('situation', 32.886232571196075),

('poverty', 32.72198534796623),

('principle', 32.60796773581254),

('responsibility', 32.54606689205432),

('future', 32.34761998950234),

('progress', 32.24654845334519),

('role', 30.814794834568872),

('regard', 30.218534983565334),

('resource', 29.993243194434523),

('democracy', 29.63902135589767),

('woman', 29.574565710854596),

('life', 29.474745372425456),

('end', 28.995215120060685),

('dialogue', 28.775567592985972),

('opportunity', 28.740323306991574),

('threat', 28.223989329946466),

('area', 28.04851900067583),

('policy', 28.04291751955034),

('trade', 27.84289504631387)]