IS20140 Final Project - Ecological Stressors Along the Shannon Catchment

Install all required files

In [4]:
```
#Import modules required to run 8-bit T-5 network
!pip install --quiet bitsandbytes
!pip install --quiet git+https://github.com/huggingface/transformers.git # Install lat
!pip install --quiet accelerate
!pip install --quiet sentencepiece
!pip install --quiet tokenizers
!pip install --quiet datasets
!pip install --quiet evaluate
!pip install --quiet torch
!pip install --quiet numpy
!pip install pandas==1.3.4 #Force version 1.3.4 as read_excel fails otherwise
!pip install --quiet sentence-transformers
!pip install --quiet sklearn
!pip uninstall xlrd -y
!pip install -quiet xlrd==1.2.0
!pip install --quiet networkx
!pip install --quiet pygraphviz
```

```
     |████████████████████████████████| 62.5 MB 1.1 MB/s
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
    Preparing wheel metadata ... done
     |████████████████████████████████| 182 kB 37.5 MB/s
     |████████████████████████████████| 7.6 MB 59.6 MB/s
  Building wheel for transformers (PEP 517) ... done
     |████████████████████████████████| 191 kB 36.1 MB/s
     |████████████████████████████████| 1.3 MB 18.6 MB/s
     |████████████████████████████████| 451 kB 26.7 MB/s
     |████████████████████████████████| 132 kB 76.1 MB/s
     |████████████████████████████████| 212 kB 79.9 MB/s
     |████████████████████████████████| 127 kB 72.2 MB/s
     |████████████████████████████████| 72 kB 1.5 MB/s
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/pu
blic/simple/
Collecting pandas==1.3.4
  Downloading pandas-1.3.4-cp38-cp38-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1
1.5 MB)
     |████████████████████████████████| 11.5 MB 25.8 MB/s
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.8/dist-packages
(from pandas==1.3.4) (2022.6)
Requirement already satisfied: python-dateutil>=2.7.3 in /usr/local/lib/python3.8/dist
-packages (from pandas==1.3.4) (2.8.2)
Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.8/dist-packages
(from pandas==1.3.4) (1.21.6)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.8/dist-packages (fro
m python-dateutil>=2.7.3->pandas==1.3.4) (1.15.0)
Installing collected packages: pandas
  Attempting uninstall: pandas
    Found existing installation: pandas 1.3.5
    Uninstalling pandas-1.3.5:
      Successfully uninstalled pandas-1.3.5
Successfully installed pandas-1.3.4
     |████████████████████████████████| 85 kB 5.1 MB/s
  Building wheel for sentence-transformers (setup.py) ... done
  Building wheel for sklearn (setup.py) ... done
Found existing installation: xlrd 1.1.0
Uninstalling xlrd-1.1.0:
  Successfully uninstalled xlrd-1.1.0

Usage:
  pip3 install [options] <requirement specifier> [package-index-options] ...
  pip3 install [options] -r <requirements file> [package-index-options] ...
  pip3 install [options] [-e] <vcs project url> ...
  pip3 install [options] [-e] <local project path> ...
  pip3 install [options] <archive url/path> ...

no such option: -u
     |████████████████████████████████| 120 kB 28.6 MB/s
  Building wheel for pygraphviz (setup.py) ... error
  ERROR: Failed building wheel for pygraphviz
    Running setup.py install for pygraphviz ... error
ERROR: Command errored out with exit status 1: /usr/bin/python3 -u -c 'import io, os,
 sys, setuptools, tokenize; sys.argv[0] = '"'"'/tmp/pip-install-hipm0vpq/pygraphviz_55
855ddbbdf64511a80713098d7a58de/setup.py'"'"'; __file__='"'"'/tmp/pip-install-hipm0vpq/
pygraphviz_55855ddbbdf64511a80713098d7a58de/setup.py'"'"';f = getattr(tokenize, '"'"'o
pen'"'"', open)(__file__) if os.path.exists(__file__) else io.StringIO('"'"'from setup
tools import setup; setup()'"'"');code = f.read().replace('"'"'\r\n'"'"',
 '"'"'\n'"'"');f.close();exec(compile(code, __file__, '"'"'exec'"'"'))' install --reco
rd /tmp/pip-record-gw2ifmn9/install-record.txt --single-version-externally-managed --c
ompile --install-headers /usr/local/include/python3.8/pygraphviz Check the logs for fu
ll command output.
```

Preprocess actual data

In [5]:
```
#Import and preprocess dataset
import pandas as pd
#dataset = pd.read_excel("/content/drive/MyDrive/College/IS20140/Final Project (1)/sa
dataset = pd.read_excel("/content/drive/Shareddrives/GEOG30370 Group Project/geography
dataset = dataset.dropna(axis=0, subset=["Abstract"]) #Drop rows with no known abstra
#print(dataset.head())
print(dataset)
```

```
   Column1 Publication Type  \
0       32                J
1       51                J
2       11                J
3       35                J
4       34                J
..     ...              ...
58      36                J
59      26                J
60      15                J
61      61                J
62      33                J

                                      Authors  Book Authors  \
0                          Khaliq, MN; Cunnane, C           NaN
1                                    DOYLE, P           NaN
2                          Sheehan, TL; Healy, MC           NaN
3   Gharbia, S; Riaz, K; Anton, I; Makrai, G; Gill...           NaN
4         Gharbia, SS; Gill, L; Johnston, P; Pilla, F           NaN
..                                           ...           ...
58        McGee, C; Brougham, C; Roche, J; Fogarty, A           NaN
59  Levesque, S; Reusch, K; Baker, I; O'Brien, J; ...           NaN
60                 Minchin, D; Penk, M; Igoe, F           NaN
61  Sun, MM; Luo, YM; Christie, P; Jia, ZJ; Li, ZG...           NaN
62  Giltrap, M; Ronan, J; Bignell, JP; Lyons, BP; ...           NaN

   Book Editors  Book Group Authors  \
0           NaN                 NaN
1           NaN                 NaN
2           NaN                 NaN
3           NaN                 NaN
4           NaN                 NaN
..          ...                 ...
58          NaN                 NaN
59          NaN                 NaN
60          NaN                 NaN
61          NaN                 NaN
62          NaN                 NaN

                                   Author Full Names  Book Author Full Names  \
0                          Khaliq, MN; Cunnane, C                     NaN
1                                    DOYLE, P                     NaN
2                        Sheehan, T. L.; Healy, M. C.                     NaN
3   Gharbia, Salem; Riaz, Khurram; Anton, Iulia; M...                     NaN
4   Gharbia, Salem S.; Gill, Laurence; Johnston, P...                     NaN
..                                               ...                     ...
58    McGee, C.; Brougham, C.; Roche, J.; Fogarty, A.                     NaN
59  Levesque, Stephanie; Reusch, Katharina; Baker,...                     NaN
60              Minchin, Dan; Penk, Marcin; Igoe, Fran                     NaN
61  Sun, Mingming; Luo, Yongming; Christie, Peter;...                     NaN
62  Giltrap, Michelle; Ronan, Jenny; Bignell, John...                     NaN

   Group Authors                                      Article Title  ...  \
0            NaN  Modelling point rainfall occurrences with the ...  ...
1            NaN  MODELING CATCHMENT EVAPORATION - AN OBJECTIVE ...  ...
2            NaN  Sub-recent changes in annual average water lev...  ...
3            NaN  Hybrid Data-Driven Models for Hydrological Sim...  ...
4            NaN  GEO-CWB: GIS-Based Algorithms for Parametrisin...  ...
..           ...                                                ...  ...
58           NaN  FIRST REPORT OF INTERSEX ROACH RESIDING IN IRI...  ...
59           NaN  PHOTO-IDENTIFICATION OF BOTTLENOSE DOLPHINS (T...  ...
60           NaN  EVIDENCE OF THE WINTER OCCURRENCE OF THE WHITE...  ...
61           NaN  Methyl-beta-cyclodextrin enhanced biodegradati...  ...
62           NaN  Integration of biological effects, fish histop...  ...

   IDS Number   Pubmed Id  Open Access Designations  Highly Cited Status  \
0         NaN         NaN                       NaN                  NaN
1         NaN         NaN                       NaN                  NaN
2         NaN         NaN                       NaN                  NaN
3         NaN         NaN                       NaN                  NaN
4         NaN         NaN                       NaN                  NaN
..        ...         ...                       ...                  ...
58        NaN         NaN                       NaN                  NaN
59        NaN         NaN                       NaN                  NaN
60        NaN         NaN                       NaN                  NaN
61        NaN  22893972.0                       NaN                  NaN
62        NaN  28501102.0                       NaN                  NaN

    Hot Paper Status Date of Export   UT (Unique WOS ID)  \
0                NaN            NaN  WOS:A1996UP76700006
1                NaN            NaN  WOS:A1990ET82200015
2                NaN            NaN  WOS:000202961300036
3                NaN            NaN  WOS:000781869400001
4                NaN            NaN  WOS:000581431700001
..               ...            ...                  ...
58               NaN            NaN  WOS:000303225800007
59               NaN            NaN  WOS:000390618000004
60               NaN            NaN  WOS:000401783300005
61               NaN            NaN  WOS:000304632700021
62               NaN            NaN  WOS:000407981500010

   Web of Science Record                                       Topics  \
0                      0              The rainstorms and thunderstorms.
1                      0          Environmental stewardship and respect
2                      0        Water level change and land reclamation
3                      0                     Water, sediment, and flood
4                      0                  Climate and land use changes
..                   ...                                            ...
58                     0                                   Intersex fish
59                     0  Multiple bottlenose dolphins are mentioned in ...
60                     0              Child exploitation and abandonment
61                     0      Environmental degradation and bioavailability
62                     0                Environmental risk and resource use

    Similarity
0     1.000000
1     0.798735
2     0.779308
3     0.770020
4     0.750733
..         ...
58    0.530397
59    0.528779
60    0.522999
61    0.518193
62    0.500254

[63 rows x 75 columns]
```

References:


```
@inproceedings{specter2020cohan,
  title={{SPECTER: Document-level Representation Learning using Citation-informed Transformers}},
  author={Arman Cohan and Sergey Feldman and Iz Beltagy and Doug Downey and Daniel S. Weld},
  booktitle={ACL},
  year={2020}
  }
```


Use models as they are:

In [6]:
```
#Guess topics using T5
# As it stands, we MIGHT be able to run a fixed question against each abstract and us
# It all feels a bit on the tenuous side. Heck, it all feels VERY on the tenuous side
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer
import torch

# Use a variant of T5 that has already been fine-tuned on question-answering tasks

tokenizer = AutoTokenizer.from_pretrained("MaRiOrOsSi/t5-base-finetuned-question-answe
model = AutoModelForSeq2SeqLM.from_pretrained("MaRiOrOsSi/t5-base-finetuned-question-a

def get_topics(input, model=model, tokenizer=tokenizer, max_output_length=50):
    question = "What are the main themes in this article?"
    context = input
    context = str(context)
    fully_formed_input = "question: "+question+" context: "+context
    fully_formed_input = str(fully_formed_input)
    input_ids = tokenizer(fully_formed_input, return_tensors="pt").input_ids
    input_ids = input_ids.to('cpu')
    outputs = model.generate(input_ids, max_new_tokens=max_output_length)
    output_string = tokenizer.decode(outputs[0], skip_special_tokens=True)
    return output_string

all_abstracts = dataset["Abstract"].to_numpy()
topics_list = list()
model = model
tokenizer = tokenizer
max_output_length = 50

for abstract in all_abstracts:
    topics = get_topics(abstract, model, tokenizer, max_output_length)
    topics_list.append(topics)

dataset["Topics"] = topics_list
print(dataset)
```

```
Token indices sequence length is longer than the specified maximum sequence length for
this model (621 > 512). Running this sequence through the model will result in indexin
g errors
   Column1 Publication Type  \
0       32                J
1       51                J
2       11                J
3       35                J
4       34                J
..     ...              ...
58      36                J
59      26                J
60      15                J
61      61                J
62      33                J

                                      Authors  Book Authors  \
0                          Khaliq, MN; Cunnane, C           NaN
1                                    DOYLE, P           NaN
2                          Sheehan, TL; Healy, MC           NaN
3   Gharbia, S; Riaz, K; Anton, I; Makrai, G; Gill...           NaN
4         Gharbia, SS; Gill, L; Johnston, P; Pilla, F           NaN
..                                           ...           ...
58        McGee, C; Brougham, C; Roche, J; Fogarty, A           NaN
59  Levesque, S; Reusch, K; Baker, I; O'Brien, J; ...           NaN
60                 Minchin, D; Penk, M; Igoe, F           NaN
61  Sun, MM; Luo, YM; Christie, P; Jia, ZJ; Li, ZG...           NaN
62  Giltrap, M; Ronan, J; Bignell, JP; Lyons, BP; ...           NaN

   Book Editors  Book Group Authors  \
0           NaN                 NaN
1           NaN                 NaN
2           NaN                 NaN
3           NaN                 NaN
4           NaN                 NaN
..          ...                 ...
58          NaN                 NaN
59          NaN                 NaN
60          NaN                 NaN
61          NaN                 NaN
62          NaN                 NaN

                                   Author Full Names  Book Author Full Names  \
0                          Khaliq, MN; Cunnane, C                     NaN
1                                    DOYLE, P                     NaN
2                        Sheehan, T. L.; Healy, M. C.                     NaN
3   Gharbia, Salem; Riaz, Khurram; Anton, Iulia; M...                     NaN
4   Gharbia, Salem S.; Gill, Laurence; Johnston, P...                     NaN
..                                               ...                     ...
58    McGee, C.; Brougham, C.; Roche, J.; Fogarty, A.                     NaN
59  Levesque, Stephanie; Reusch, Katharina; Baker,...                     NaN
60              Minchin, Dan; Penk, Marcin; Igoe, Fran                     NaN
61  Sun, Mingming; Luo, Yongming; Christie, Peter;...                     NaN
62  Giltrap, Michelle; Ronan, Jenny; Bignell, John...                     NaN

   Group Authors                                      Article Title  ...  \
0            NaN  Modelling point rainfall occurrences with the ...  ...
1            NaN  MODELING CATCHMENT EVAPORATION - AN OBJECTIVE ...  ...
2            NaN  Sub-recent changes in annual average water lev...  ...
3            NaN  Hybrid Data-Driven Models for Hydrological Sim...  ...
4            NaN  GEO-CWB: GIS-Based Algorithms for Parametrisin...  ...
..           ...                                                ...  ...
58           NaN  FIRST REPORT OF INTERSEX ROACH RESIDING IN IRI...  ...
59           NaN  PHOTO-IDENTIFICATION OF BOTTLENOSE DOLPHINS (T...  ...
60           NaN  EVIDENCE OF THE WINTER OCCURRENCE OF THE WHITE...  ...
61           NaN  Methyl-beta-cyclodextrin enhanced biodegradati...  ...
62           NaN  Integration of biological effects, fish histop...  ...

   IDS Number   Pubmed Id  Open Access Designations  Highly Cited Status  \
0         NaN         NaN                       NaN                  NaN
1         NaN         NaN                       NaN                  NaN
2         NaN         NaN                       NaN                  NaN
3         NaN         NaN                       NaN                  NaN
4         NaN         NaN                       NaN                  NaN
..        ...         ...                       ...                  ...
58        NaN         NaN                       NaN                  NaN
59        NaN         NaN                       NaN                  NaN
60        NaN         NaN                       NaN                  NaN
61        NaN  22893972.0                       NaN                  NaN
62        NaN  28501102.0                       NaN                  NaN

    Hot Paper Status Date of Export   UT (Unique WOS ID)  \
0                NaN            NaN  WOS:A1996UP76700006
1                NaN            NaN  WOS:A1990ET82200015
2                NaN            NaN  WOS:000202961300036
3                NaN            NaN  WOS:000781869400001
4                NaN            NaN  WOS:000581431700001
..               ...            ...                  ...
58               NaN            NaN  WOS:000303225800007
59               NaN            NaN  WOS:000390618000004
60               NaN            NaN  WOS:000401783300005
61               NaN            NaN  WOS:000304632700021
62               NaN            NaN  WOS:000407981500010

   Web of Science Record                                       Topics  \
0                      0              The rainstorms and thunderstorms.
1                      0          Environmental stewardship and respect
2                      0        Water level change and land reclamation
3                      0                     Water, sediment, and flood
4                      0                  Climate and land use changes
..                   ...                                            ...
58                     0                                   Intersex fish
59                     0  Multiple bottlenose dolphins are mentioned in ...
60                     0              Child exploitation and abandonment
61                     0      Environmental degradation and bioavailability
62                     0                Environmental risk and resource use

    Similarity
0     1.000000
1     0.798735
2     0.779308
3     0.770020
4     0.750733
..         ...
58    0.530397
59    0.528779
60    0.522999
61    0.518193
62    0.500254

[63 rows x 75 columns]
```

In [14]:
```
dataset[["Article Title", "Abstract", "Topics"]].to_csv("/content/drive/MyDrive/Colle
```

In [12]:
```
dataset
```

Out[12]:

	Column1	Publication Type	Authors	Book Authors	Book Editors	Book Group Authors	Author Full Names	Book Author Full Names	Group Authors	Article Title
0	32	J	Khaliq, MN; Cunnane, C	NaN	NaN	NaN	Khaliq, MN; Cunnane, C	NaN	NaN	Modelling point rainfall occurrences with the ...
1	51	J	DOYLE, P	NaN	NaN	NaN	DOYLE, P	NaN	NaN	MODELING CATCHMENT EVAPORATION - AN OBJECTIVE ...
2	11	J	Sheehan, TL; Healy, MC	NaN	NaN	NaN	Sheehan, T. L.; Healy, M. C.	NaN	NaN	Sub-recent changes in annual average water lev...
3	35	J	Gharbia, S; Riaz, K; Anton, I; Makrai, G; Gill...	NaN	NaN	NaN	Gharbia, Salem; Riaz, Khurram; Anton, Iulia; M...	NaN	NaN	Hybrid Data-Driven Models for Hydrological Sim...
4	34	J	Gharbia, SS; Gill, L; Johnston, P; Pilla, F	NaN	NaN	NaN	Gharbia, Salem S.; Gill, Laurence; Johnston, P...	NaN	NaN	GEO-CWB: GIS-Based Algorithms for Parametrisin...
...	...	...	...	...	...	...	...	...	...	...
58	36	J	McGee, C; Brougham, C; Roche, J; Fogarty, A	NaN	NaN	NaN	McGee, C.; Brougham, C.; Roche, J.; Fogarty, A.	NaN	NaN	FIRST REPORT OF INTERSEX ROACH RESIDING IN IRI...
59	26	J	Levesque, S; Reusch, K; Baker, I; O'Brien, J; ...	NaN	NaN	NaN	Levesque, Stephanie; Reusch, Katharina; Baker,...	NaN	NaN	PHOTO-IDENTIFICATION OF BOTTLENOSE DOLPHINS (T...
60	15	J	Minchin, D; Penk, M; Igoe, F	NaN	NaN	NaN	Minchin, Dan; Penk, Marcin; Igoe, Fran	NaN	NaN	EVIDENCE OF THE WINTER OCCURRENCE OF THE WHITE...
61	61	J	Sun, MM; Luo, YM; Christie, P; Jia, ZJ; Li, ZG...	NaN	NaN	NaN	Sun, Mingming; Luo, Yongming; Christie, Peter;...	NaN	NaN	Methyl-beta-cyclodextrin enhanced biodegradati...
62	33	J	Giltrap, M; Ronan, J; Bignell, JP; Lyons, BP; ...	NaN	NaN	NaN	Giltrap, Michelle; Ronan, Jenny; Bignell, John...	NaN	NaN	Integration of biological effects, fish histop...

63 rows × 75 columns

In [8]:
```
dataset["Abstract"]
```

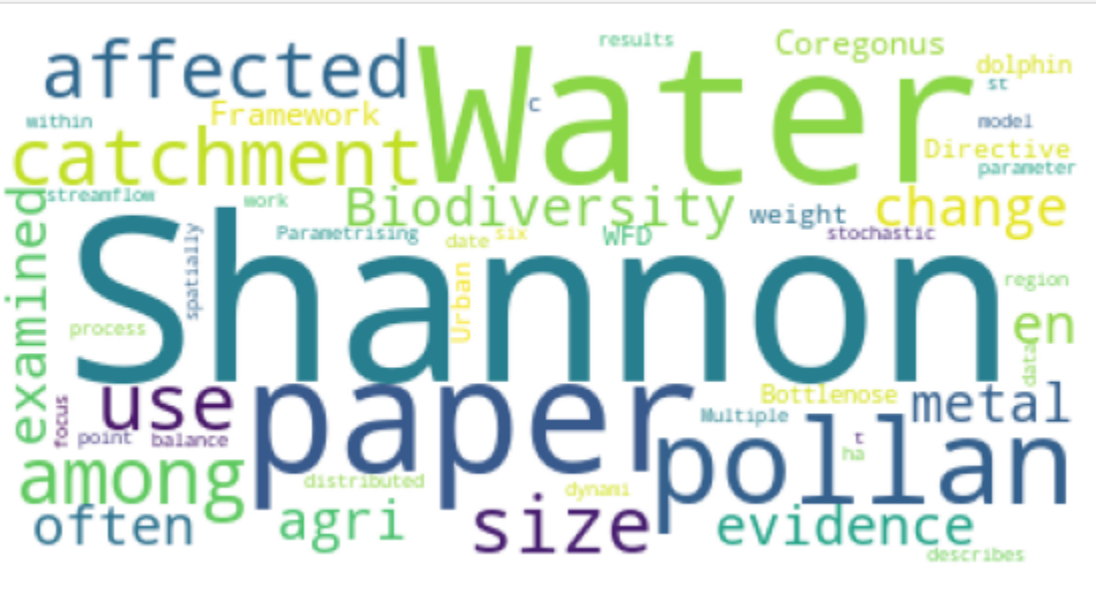
Out[8]:
```
0     A six parameter stochastic point process model...
1     Water-balance data from the Shannon catchment ...
2     This paper describes results to date on work t...
3     Changes in streamflow within catchments can ha...
4     Parametrising the spatially distributed dynami...
                            ...
58    Endocrine disrupting chemicals (EDCs) are chem...
59    The Lower River Shannon is a Special Area of C...
60    Coregonus pollan Thompson, 1835 is an Irish en...
61    The contamination of soils by polycyclic aroma...
62    This study investigates the use of a weight of...
Name: Abstract, Length: 63, dtype: object
```

In [9]:
```
torch.cuda.empty_cache()
```

In [10]:
```
## Create word cloud of the abstracts
from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt
from matplotlib import rcParams

abstracts = dataset["Abstract"]
abstracts = abstracts.to_string()
stopwords = set(STOPWORDS)
stopwords = ["river", "study", "Ireland", "Irish"] + list(stopwords)

abstracts_word_cloud = WordCloud(stopwords=stopwords, max_words=50, background_color='
rcParams["figure.figsize"] = (10,5)
plt.imshow(abstracts_word_cloud)
plt.axis("off")
plt.show()
```



In [11]:
```
topics = dataset["Topics"]
topics = topics.to_string()

topics_word_cloud = WordCloud(stopwords=stopwords, max_words=30, background_color="wh

rcParams["figure.figsize"] = (10,5)
plt.imshow(topics_word_cloud)
plt.axis("off")
plt.show()
```

References

https://huggingface.co/blog/hf-bitsandbytes-integration

https://discuss.huggingface.co/t/t5-finetuning-tips/684/5

https://colab.research.google.com/github/patil-suraj/exploring-T5/blob/master/T5_on_TPU.ipynb#scrollTo=QLGiFCDqvuil