CSC 5741: Lecture #05—Exploratory Data Analysis

Lighton Phiri
<lighton.phiri@unza.zm>

April 23 2019

Contents

Introduction

During these "hands-on" activities, we briefly look at examplers of Exploratory Data Analysis (EDA).

In all instances, you are encouraged to make reference to online Python documentation and documentation for specific libraries. You are also encouraged to look up and explore other libraries, especially as you work towards the Mini Projects.

Importing Libraries and Modules

[1]:

```
# Import all libraries and modules for use during lecture session code walkthrough
import pandas as pd
import re
import string

from collections import Counter
from IPython.core.interactiveshell import InteractiveShell
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfVectorizer
import matplotlib.pyplot as plt

InteractiveShell.ast_node_interactivity = "all"
pd.set_option('display.latex.repr', True)
pd.set_option('display.latex.longtable', True)

var_stemmer = PorterStemmer()
```

Implementing Core Functions

[2]:

```
def fxn_case_folding(var_input):
    """
    Case Folding
    """
    return var_input.lower()

def fxn_punctuation(var_input_text):
    """
    Punctuation
    """
    var_output_text = re.sub("[%s]" % re.escape(string.punctuation), " ",␣
 ↪var_input_text)
    var_output_text = re.sub("[%s]" % re.escape(string.punctuation), " ",␣
 ↪var_output_text)
    var_output_text = re.sub('\w*\d\w*', '', var_output_text) # HINT: lookup isalpha()␣
 ↪function
    return var_output_text

def fxn_stopwords(var_input_text):
    """
    Stopwords
    """
    var_etd_stop = " ".join([
        var_etd_word for var_etd_word in var_input_text.split()
        if var_etd_word not in stopwords.words('english')
    ])
    return var_etd_stop

def fxn_stem(var_input_text):
    """
    Stemming
    """
    var_output_text = " ".join([
        var_stemmer.stem(var_etd_word) for var_etd_word in var_input_text.split()
    ])
    return var_output_text

def fxn_normalise_ict1110_minors(var_input_minor):
    """
    Returns normalised ICT 1110 minor
    """
    var_ict1110_minors = ["Geography", "History", "Languages", "Mathematics", "Civic",␣
 ↪"Art", "Religious Studies"]
    if "civic" in var_input_minor.lower():
        var_output_minor = "Civic Education"
    elif "religious" in var_input_minor.lower() or "res" in var_input_minor.lower():
        var_output_minor = ""
    elif "history" in var_input_minor.lower():
        var_output_minor = "History"
    elif "art" in var_input_minor.lower():
        var_output_minor = "Art"
    elif "language" in var_input_minor.lower() or "french" in var_input_minor.lower():
        var_output_minor = "Languages"
    elif "geography" in var_input_minor.lower():
        var_output_minor = "Geography"
    elif "math" in var_input_minor.lower():
        var_output_minor = "Mathematics"
    elif "writing" in var_input_minor.lower():
        var_output_minor = "Writing Skill"
    else:
        var_output_minor = var_input_minor
    return var_output_minor.title()
```

Exploratory Data Analysis

Example 1: 2018/19 ICT 1110 Information Survey

Dataset

[3]:
```
# Explore 2018/19 ICT 1110 survey
!tail -n 2 db-unza19-ict1110_2018_19-preliminary_survey.csv
```

```
2019/04/08 4:53:14 AM GMT+2|Participant38|5f59b44f0c6a470fe410c0df68985274|Chamba valley,
Lusaka|Mathematics|I felt maths would combine well with ICTs.. |"Though i didn't choose
to do ict in the first place.. But then i thought to myself, "" since it is a new
program, why not go for it, as jobs will be readily available."" And that's how got to
the decision.."|No|No||1 to 2 years|Yes|Am a guitarist
2019/04/08 11:33:44 AM GMT+2|Participant39|605fc2f11ca271de28344e0e13459fd5|Airport,
Sowezi/NWP|Religious Education| Passionate for it|To learner more about
Technology|No|Yes|Basics of computer.|More than 5 years|Yes|Researching.
```

Data Cleaning

[4]:
```
# Create DataFrame of survey
var_ict1110_survey = pd.read_csv("db-unza19-ict1110_2018_19-preliminary_survey.csv",␣
 ↪sep="|")
var_ict1110_survey.columns

# Rename columns
var_ict1110_survey.rename(columns={"Full Names": "StudentName",
                                  "Student ID": "StudentID",
                                  "Hometown (surburb/town/province---e.g. Kabwata/
 ↪Lusaka/Lusaka)": "HomeTown",
                                  "What is your programme Minor (e.g. Mathematics,␣
 ↪Languages)": "MinorProgramme",
                                  "What made you decide on your programme minor?":␣
 ↪"MinorProgrammeMotivation",
                                  "Why did you decide to major pursue the B.ICTs Ed.␣
 ↪Programme?": "MajorProgrammeMotivation",
                                  "Did you study Computer Studies at secondary school?
 ↪": "DidComputerStudies",
                                  "Have you undergone any computer related training?":␣
 ↪"HasComputerTraining",
                                  "If your response to the question above is year,␣
 ↪please provide details of the type of course and/or training":␣
 ↪"ComputerTrainingType",
                                  "How many years experience do you have using␣
 ↪computers?": "ExperienceWithComputers",
                                  "Do you currently own a computer or have regular␣
 ↪access to one?": "HasComputerAccess",
                                   "List one interesting fact about yourself (e.g. I␣
 ↪cycle everyday!):": "AboutMe"}, inplace=True)
```

[4]:
```
Index(['Timestamp', 'Full Names', 'Student ID',
       'Hometown (surburb/town/province---e.g. Kabwata/Lusaka/Lusaka)',
       'What is your programme Minor (e.g. Mathematics, Languages)',
       'What made you decide on your programme minor?',
       'Why did you decide to major pursue the B.ICTs Ed. Programme?',
       'Did you study Computer Studies at secondary school?',
       'Have you undergone any computer related training?',
       'If your response to the question above is year, please provide details of the
type of course and/or training',
       'How many years experience do you have using computers?',
       'Do you currently own a computer or have regular access to one?',
       'List one interesting fact about yourself (e.g. I cycle everyday!):'],
      dtype='object')
```

[5]:
```
var_ict1110_minors = list(set(var_ict1110_survey["MinorProgramme"].to_list()))
```

[6]:
```
# 1. Case Folding
```

```
var_ict1110_minors = [var_minor.lower() for var_minor in var_ict1110_minors]
var_ict1110_minors
```

```
[6]: ['art',
      'civic education ',
      'res1010',
      'languages',
      'art and design',
      'french',
      'geography',
      'academic writing and study skills',
      'religious education',
      'religious studies',
      'french',
      'geography',
      'language',
      'religious studies ',
      'history',
      'data mining',
      'history ',
      'languages ',
      'religious studies',
      'mathematics ',
      'mathematics',
      'civic education',
      'mathematics',
      'religious studies',
      'languages 1220 and 1200']
```

```
[7]: # 2. Deduplication
     var_ict1110_minors = list(set(var_ict1110_minors))
```

```
[8]: # 3. Punctuation
     var_ict1110_minors_punct = [var_minor_trim.strip() for var_minor_trim in␣
      ↪var_ict1110_minors]
     var_ict1110_minors_punct = list(set(var_ict1110_minors_punct))
     len(var_ict1110_minors_punct)

     var_ict1110_minors_punct
```

```
[8]: 15
```

```
[8]: ['geography',
      'languages 1220 and 1200',
      'religious education',
      'religious studies',
      'language',
      'art',
      'academic writing and study skills',
      'art and design',
      'civic education',
      'data mining',
      'history',
      'res1010',
      'languages',
      'french',
      'mathematics']
```

[9]:

```
# 4. Remove stopwords
var_ict1110_minors_stop = [ " ".join([x for x in var_ict1110_minor.split() if x not in␣
 ↪stopwords.words('english')]) for var_ict1110_minor in var_ict1110_minors_punct]

var_ict1110_minors_stop
```

[9]:

```
['geography',
 'languages 1220 1200',
 'religious education',
 'religious studies',
 'language',
 'art',
 'academic writing study skills',
 'art design',
 'civic education',
 'data mining',
 'history',
 'res1010',
 'languages',
 'french',
 'mathematics']
```

[10]:

```
# 5. Stemming
var_ict1110_minors_stem = [var_stemmer.stem(var_minor) if len(var_minor.split())==1␣
 ↪else var_minor for var_minor in var_ict1110_minors_stop]
```

[11]:

```
var_ict1110_minors_stem = list(set(var_ict1110_minors_stem))
var_ict1110_minors_stem
```

[11]:

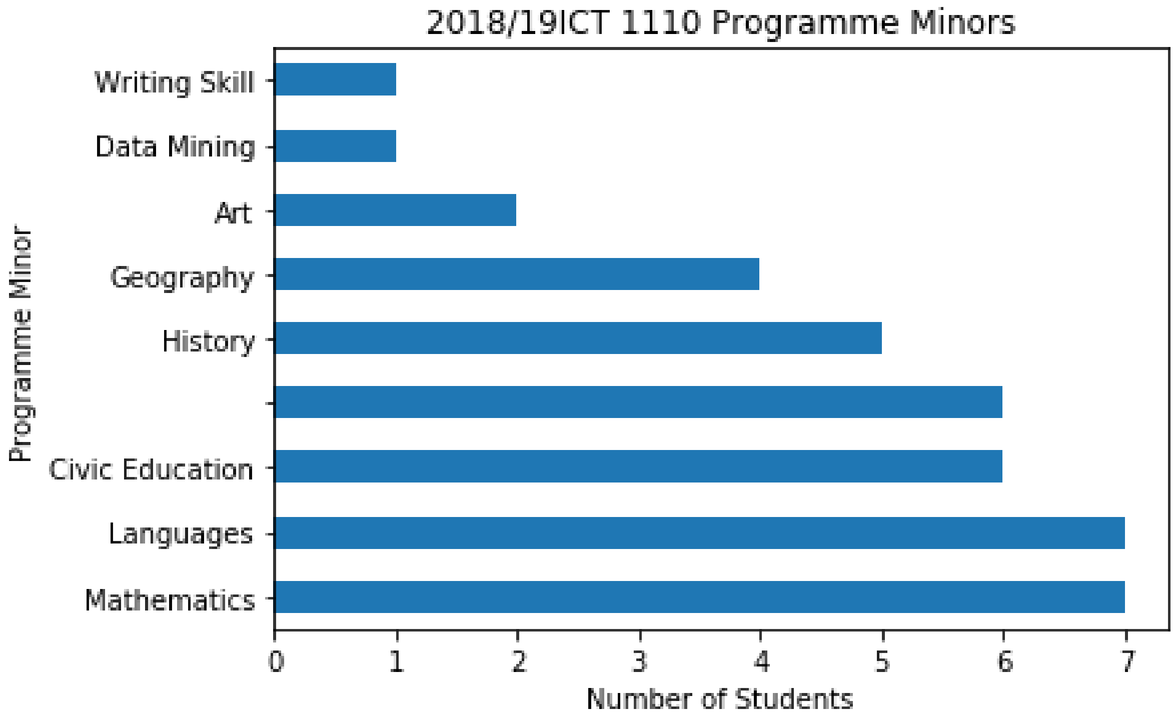
```
['histori',
 'languages 1220 1200',
 'religious education',
 'religious studies',
 'art',
 'academic writing study skills',
 'geographi',
 'civic education',
 'art design',
 'data mining',
 'languag',
 'res1010',
 'mathemat',
 'french']
```

Programme Minors

[12]:
```
var_ict1110_survey_eda = var_ict1110_survey

var_ict1110_survey_eda["MinorProgramme"] = var_ict1110_survey_eda["MinorProgramme"].
 ↪apply(fxn_case_folding)

var_ict1110_survey_eda["MinorProgramme"] = var_ict1110_survey_eda["MinorProgramme"].
 ↪apply(fxn_punctuation)

var_ict1110_survey_eda["MinorProgramme"] = var_ict1110_survey_eda["MinorProgramme"].
 ↪apply(fxn_normalise_ict1110_minors)

var_ict1110_survey_eda.columns
var_ict1110_survey_eda_plot1 = var_ict1110_survey_eda["MinorProgramme"].value_counts().
 ↪plot(kind="barh", title="2018/19ICT 1110 Programme Minors")
var_ict1110_survey_eda_plot1.set(xlabel="Number of Students", ylabel="Programme Minor")
```

[12]:
```
Index(['Timestamp', 'StudentName', 'StudentID', 'HomeTown', 'MinorProgramme',
       'MinorProgrammeMotivation', 'MajorProgrammeMotivation',
       'DidComputerStudies', 'HasComputerTraining', 'ComputerTrainingType',
       'ExperienceWithComputers', 'HasComputerAccess', 'AboutMe'],
      dtype='object')
```

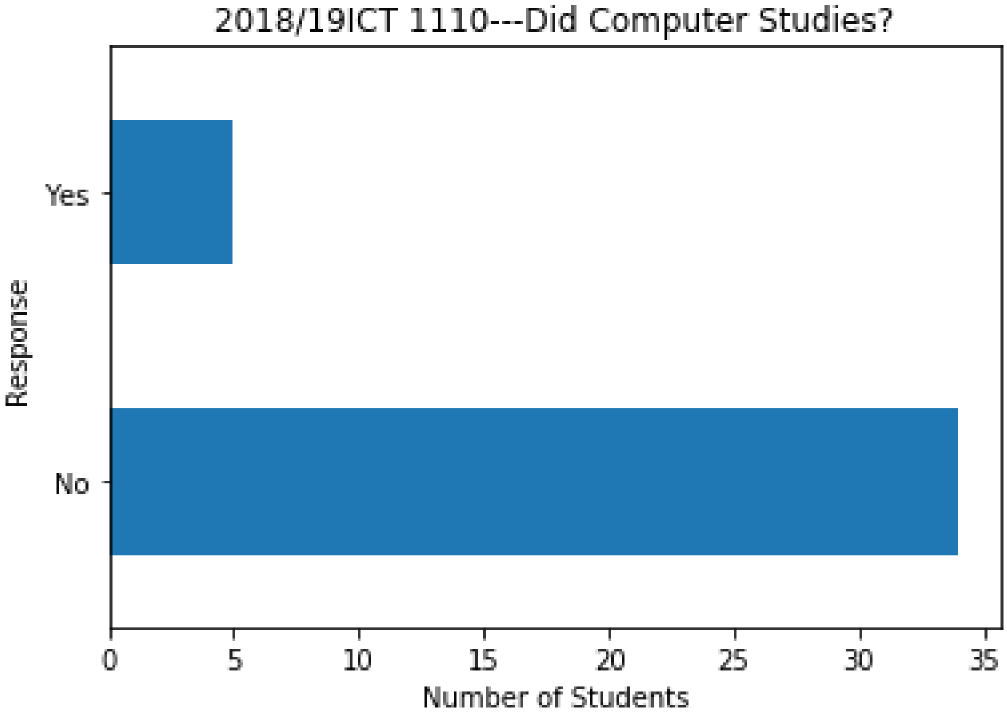
[12]:
```
[Text(0, 0.5, 'Programme Minor'), Text(0.5, 0, 'Number of Students')]
```



Computer Studies in High School

```
[13]: # 2. Computer Studies
var_ict1110_survey_eda_plot2 = var_ict1110_survey_eda["DidComputerStudies"].
 ↪value_counts().plot(kind="barh", title="2018/19ICT 1110---Did Computer Studies?")
var_ict1110_survey_eda_plot2.set(xlabel="Number of Students", ylabel="Response")
```

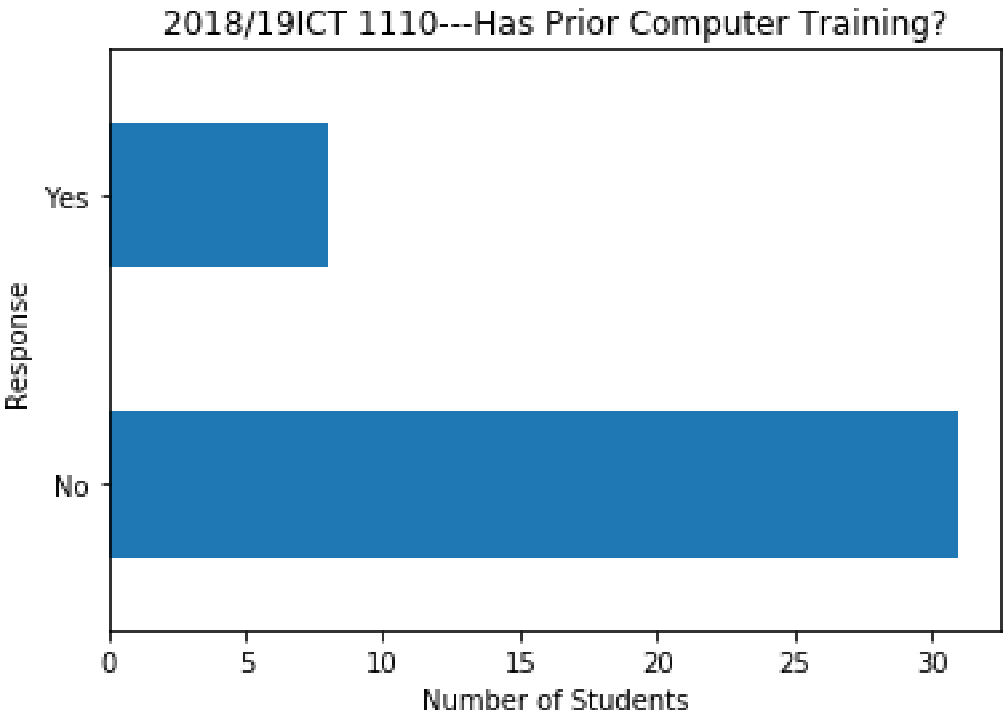
```
[13]: [Text(0, 0.5, 'Response'), Text(0.5, 0, 'Number of Students')]
```



Computer Training Done

```
[14]: # 3. Computer Training
var_ict1110_survey_eda_plot3 = var_ict1110_survey_eda["HasComputerTraining"].
 ↪value_counts().plot(kind="barh", title="2018/19ICT 1110---Has Prior Computer␣
 ↪Training?")
var_ict1110_survey_eda_plot3.set(xlabel="Number of Students", ylabel="Response")
```

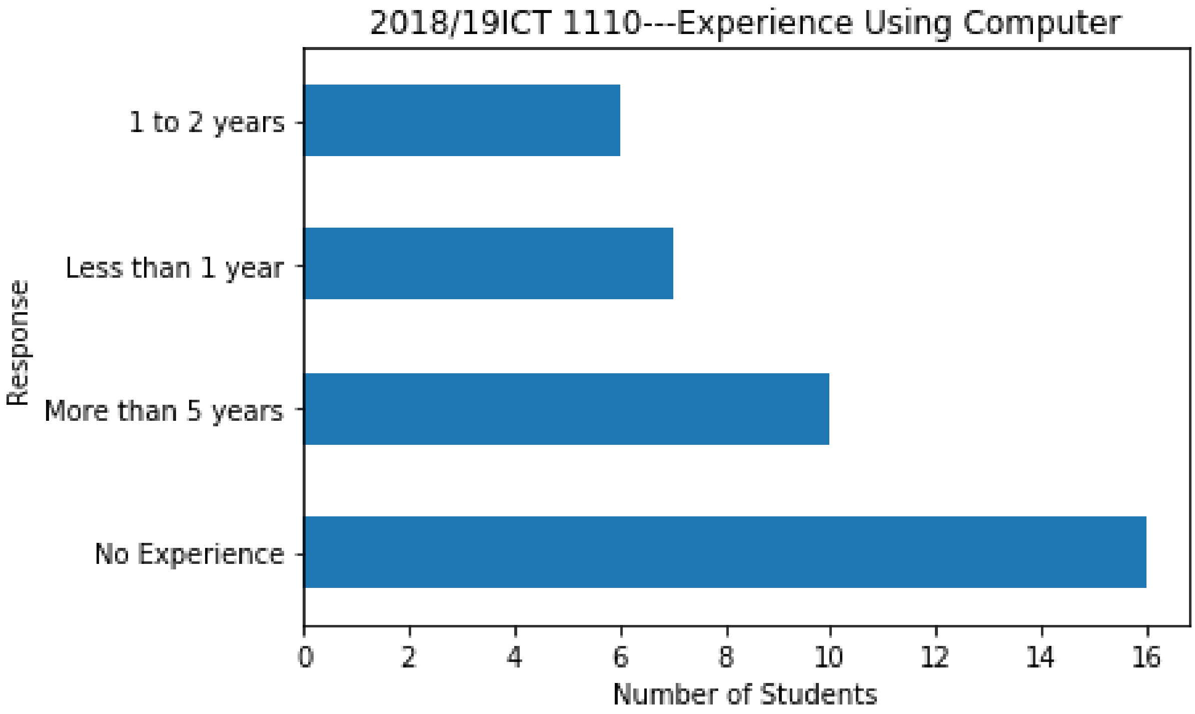
```
[14]: [Text(0, 0.5, 'Response'), Text(0.5, 0, 'Number of Students')]
```



Experience Using Computers

```
[15]: # 4. Experience with Computers
var_ict1110_survey_eda_plot4 = var_ict1110_survey_eda["ExperienceWithComputers"].
 ↪value_counts().plot(kind="barh", title="2018/19ICT 1110---Experience Using Computer")
var_ict1110_survey_eda_plot4.set(xlabel="Number of Students", ylabel="Response")
```

```
[15]: [Text(0, 0.5, 'Response'), Text(0.5, 0, 'Number of Students')]
```

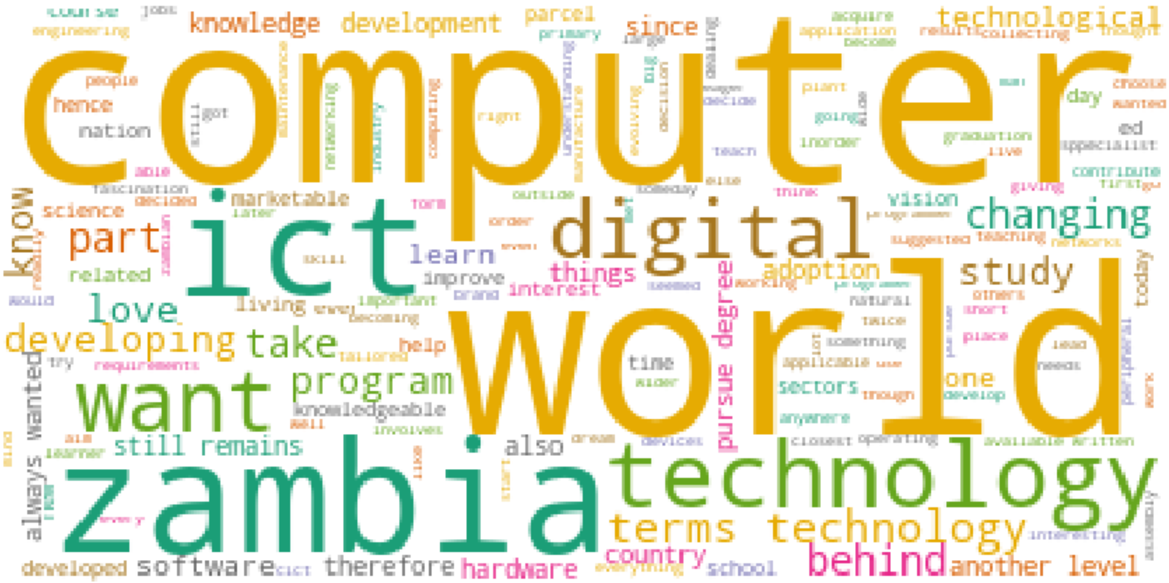


Motivation for Enrolling into Programme

[16]:

```
from wordcloud import WordCloud

var_ict1110_wordcloud = WordCloud(stopwords=stopwords.words("english"),␣
 ↪background_color="white", colormap="Dark2",
               max_font_size=150, random_state=42)

# fxn_punctuation
# fxn_stopwords
# 1. Missing values
var_ict1110_survey_eda["MajorProgrammeMotivation"].fillna("", inplace=True)
# 2. Case Folding
var_ict1110_survey_eda["MajorProgrammeMotivation"] =␣
 ↪var_ict1110_survey_eda["MajorProgrammeMotivation"].apply(fxn_case_folding)
# 3. Punctuations
var_ict1110_survey_eda["MajorProgrammeMotivation"] =␣
 ↪var_ict1110_survey_eda["MajorProgrammeMotivation"].apply(fxn_punctuation)
# 4. Stopwords
var_ict1110_survey_eda["MajorProgrammeMotivation"] =␣
 ↪var_ict1110_survey_eda["MajorProgrammeMotivation"].apply(fxn_stopwords)

var_ict1110_survey_eda_motivation = var_ict1110_survey_eda["MajorProgrammeMotivation"]
var_ict1110_wordcloud.generate(' '.join(var_ict1110_survey_eda_motivation))

plt.figure(figsize = (15,10))
plt.imshow(var_ict1110_wordcloud)
plt.axis("off")
```

```
[16]: <wordcloud.wordcloud.WordCloud at 0x7f9b6103eba8>

[16]: <Figure size 1080x720 with 0 Axes>

[16]: <matplotlib.image.AxesImage at 0x7f9b60be7ac8>

[16]: (-0.5, 399.5, 199.5, -0.5)
```



Example 2: University of Zambia ETD Abstracts

Dataset

```
[17]: # Explore dataset format
!head -n 2 db-unza19-dspace_unza_zm.csv
```

```
_identifier|_datestamp|_setSpec|title|creator|subject|description|date|type|identifier|la
nguage|format
oai:dspace.unza.zm:123456789/4153|2016-06-09T12:46:34Z|com_123456789_289=col_123456789_29
0|"Morphological characterisation of low and high oil sunflower(Hellanthus Annuus.
L.)Varieties for use in marker assisted selection"|"Chinyundo, Anthony"|"Helianthus
Annuus. L.=Sun flower oil=Cooking oil"|"Morphological characterization was done on three
sunflower varieties; CCA81, Milika and Record in order to see morphological differences
for possible use in marker assisted selection. The parameters that were looked at are
leaf size, leaf shape, colour of leaves, number of leaves per plant, hairiness at top of
stem, days to 50 % flowering and maturity, seed colour, presence of seed stripes, colour
of seed stripes, position of seed stripes, shape of seed, weight of 100 seeds, kernel and
oil percentages. Significant differences were noted in leaf size, plant height, days to
```

```
50 % flowering and maturity. Record had the highest oil percentage of 42.97, Milika 38.77
and CCA81 42.17. In the other parameters no significant differences were established.
Variation among these characteristics is important because it allows for development of
varieties adapted to specific environments or agro-climatic regions"|2015-11-11T13:39:13Z
=2015-11-11T13:39:13Z=2015-11-11|Other|http://hdl.handle.net/123456789/4153|en|applicatio
n/pdf
```

[18]:

```python
# Use pandas to pluck out abstracts
var_unza_etds = pd.read_csv("db-unza19-dspace_unza_zm.csv", sep="|")
# Change row index value to use IR object identifiers
# https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.
 ↪set_index.html
var_unza_etds.set_index("_identifier", inplace=True)
var_unza_etds.fillna(value={"description": ""}, inplace=True)
var_unza_etds.fillna(value={"title": ""}, inplace=True)
### var_unza_etds_setspecs[var_unza_etds_setspecs["SetSpec"].str.
 ↪contains("col_")==True]
print("Total objects in UNZA IR: ", len(var_unza_etds))
var_unza_etds = var_unza_etds[var_unza_etds["_setSpec"].str.
 ↪contains("com_123456789_18")==True] # Filter out ETDs only
print("Total ETDs in UNZA IR: ", len(var_unza_etds))
var_unza_etds.columns
```

```
Total objects in UNZA IR:  5440
Total ETDs in UNZA IR:  3356
```

[18]:

```
Index(['_datestamp', '_setSpec', 'title', 'creator', 'subject', 'description',
       'date', 'type', 'identifier', 'language', 'format'],
      dtype='object')
```

[19]:

```python
# 1. Data format
var_unza_etds.columns
var_unza_etds.head(1).T

var_unza_etds.describe().T
var_unza_etds.info()
```

[19]:

```
Index(['_datestamp', '_setSpec', 'title', 'creator', 'subject', 'description',
       'date', 'type', 'identifier', 'language', 'format'],
      dtype='object')
```

[19]:

_identifier	oai:dspace.unza.zm:123456789/3777
_datestamp	2016-06-09T10:16:03Z
_setSpec	com_123456789_18=col_123456789_76
title	Instruction based formative assessment in sele...
creator	Mwale, Fred M.
subject	Educational tests and measurement=Formative As...
description	The purpose of the study was to evaluate the u...
date	2015-04-13T07:36:13Z=2015-04-13T07:36:13Z=2015...

Continued on next page

_identifier	oai:dspace.unza.zm:123456789/3777
type	Thesis
identifier	http://hdl.handle.net/123456789/3777
language	en
format	application/pdf=application/pdf=application/pd...

[19]:

	count	unique	top	freq
_datestamp	3356	3347	2016-06-09T10:26:21Z	2
_setSpec	3356	13	com_123456789_18=col_123456789_83	772
title	3356	3273		20
creator	3332	3218	Muchemwa, Levy	3
subject	3316	3245	Agronomy	4
description	3356	3155		182
date	3336	3336	2017-11-23T10:16:07Z=2017-11-23T10:16:07Z=2016	1
type	3317	5	Thesis	3293
identifier	3336	3336	http://hdl.handle.net/123456789/1013	1
language	3317	3	en	3310
format	3336	18	application/pdf	2861

```
<class 'pandas.core.frame.DataFrame'>
Index: 3356 entries, oai:dspace.unza.zm:123456789/3777 to
oai:dspace.unza.zm:123456789/1149
Data columns (total 11 columns):
_datestamp     3356 non-null object
_setSpec       3356 non-null object
title          3356 non-null object
creator        3332 non-null object
subject        3316 non-null object
description    3356 non-null object
date           3336 non-null object
type           3317 non-null object
identifier     3336 non-null object
language       3317 non-null object
format         3336 non-null object
dtypes: object(11)
memory usage: 314.6+ KB
```

Data Cleaning

[20]:

```
var_unza_etds.head(1)
var_unza_etds_description = var_unza_etds[["description"]]
var_unza_etds_description.columns
var_unza_etds_description.fillna(value={"description": ""}, inplace=True)
var_unza_etds_dict = var_unza_etds_description
```

[20]:

	_datestamp	_setSpec	title
_identifier			
oai:dspace.unza.zm:123456789/3777	2016-06-09T10:16:03Z	com_123456789_18=col_123456789_76	Instruction b

[20]:
```
Index(['description'], dtype='object')

/home/lightonphiri/.local/lib/python3.6/site-packages/pandas/core/generic.py:6130:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/indexing.html#indexing-view-versus-copy
  self._update_inplace(new_data)
```

ETDs by Year

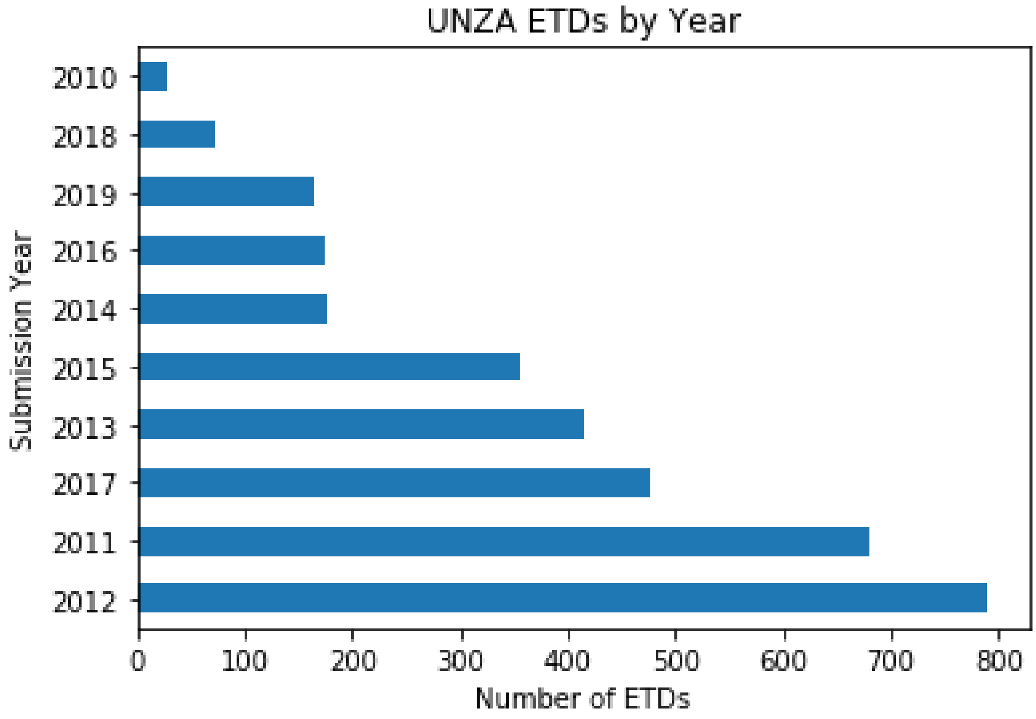
[21]:
```
var_unza_etds["date"].head(2)
```

[21]:

	date
_identifier	
oai:dspace.unza.zm:123456789/3777	2015-04-13T07:36:13Z=2015-04-13T07:36:13Z=2015...
oai:dspace.unza.zm:123456789/4729	2017-07-25T13:51:51Z=2017-07-25T13:51:51Z=2016

[22]:
```
var_unza_etds_eda_plot1 = var_unza_etds["date"].str[:4].sort_values().value_counts().
 ↪plot("barh", title="UNZA ETDs by Year")
var_unza_etds_eda_plot1.set(xlabel="Number of ETDs", ylabel="Submission Year")
```

[22]:
```
[Text(0, 0.5, 'Submission Year'), Text(0.5, 0, 'Number of ETDs')]
```


UNZA ETDs by Year
Submission Year
2010
2018
2019
2016
2014
2015
2013
2017
2011
2012
0
100
200
300
400
500
600
700
800
Number of ETDs