2020/21 CSC 5741: Data Mining and Warehousing
Jupyter Notebook—Data Cleaning and Preprocessing

Lighton Phiri
<lighton.phiri@unza.zm>

May 17 2021

Contents

Introduction

During these "hands-on" activities, we look at practical examples of how to clean data by implementing common pre-processing tasks and, additionally, focusing on text-specific pre-processing tasks. The motivation behind focusing on text is that it tends to require additional cleaning in comparison to other types of data. Specifically, we focus on the following pre-processing activities:

1. Case Folding
2. Stemming
3. Removing Stopwords
4. Removing Punctuations
5. Deduplication
6. Handling Missing Values

You will notice that the examples use native Python features as opposed to libraries such as Pandas. This is done to highlight the flexibility that Python provides. In cases were they are not used, you are encouraged to explore how Pandas and other libraries can be used.

In all instances, you are encouraged to make reference to online documentation for the various tools. Additionally, you can exploit tools like Zeal Offline Documentation Browser to download and search through offline documentation. You are also encouraged to look up and explore other libraries, especially as you work towards the Mini Projects.

General Notebook Configuration

[1]:
```
# Aesthetics for pandas cell outsput
import pandas as pd

pd.set_option('display.latex.repr', True)
pd.set_option('display.latex.longtable', True)
pd.set_option('max_colwidth', 30)

# Show all Jupyter Notebook cell output
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

Python Packages for Data Pre-processing

[2]:
```
# 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
```

Data Preprocessing

Dataset #1: 2018/19 ICT 1110 Information Survey

Link to dataset: http://bit.ly/3b5V6uc

Students at enrolled into the "ICT 1110: Computer Systems and Architecture" course, at The University of Zambia, respond to a preliminary survey aimed at collecting background information about them. This is done using Google Forms.

Dataset Description

This dataset comprises of 42 student responses for the 2018/19 cohort. The dataset has observations presented in CSV format, using "|" as the separator. In addition, each observation is associated with the following 13 data attributes: * 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

[3]:
```
# Use Bash to explore 2018/19 ICT 1110 survey
!tail -n 3 db-unza21-csc5741-ict1110_2018_19-preliminary_survey.csv | cat -n
```

```
     1  2019/04/05 9:01:53 AM GMT+2|Participant37|28814f84db09b59b95f863b9c143c3a9|8
miles / chibombo/ Central province |History |It's seemed like the best option |Computers
interest me|Yes|No||No Experience|No|I am ambidextrous
     2  2019/04/08 4:53:14 AM GMT+2|Participant38|94984a8c4896946d9bafd24959cb6181|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
     3  2019/04/08 11:33:44 AM
GMT+2|Participant39|8e4d9eeed250a9d065ac2bb8bdc67b30|Airport, Sowezi/NWP|Religious
Education| Passionate for it|To learner more about Technology|No|Yes|Basics of
computer.|More than 5 years|Yes|Researching.
```

[4]:
```
!cat db-unza21-csc5741-ict1110_2018_19-preliminary_survey.csv | wc -l
```

```
43
```

[5]:
```
pd.read_csv("db-unza21-csc5741-ict1110_2018_19-preliminary_survey.csv", sep="|").
↪head(2).T
```

[5]:

	0	1
Timestamp	2019/03/28 11:13:51 PM GMT+2	2019/03/28 11:55:27 PM GMT+2
Full Names	Participant1	Participant2
Student ID	NaN	742b8abe5776a6d942a92ce7dc...
Hometown (surburb/town/prov...	Chudleigh/Lusaka/Lusaka	Copperbelt,luanshya,Mpatamato
What is your programme Mino...	Data Mining	Mathematics
What made you decide on you...	I love data	I find it easy to study an...
Why did you decide to major...	I love computers	Wanted to acquire more kno...
Did you study Computer Stud...	No	No
Have you undergone any comp...	Yes	No
If your response to the que...	I have studied Computer Sc...	NaN
How many years experience d...	More than 5 years	1 to 2 years
Do you currently own a comp...	Yes	Yes
List one interesting fact a...	I cycle everyday!	A day doesn't pass by with...

[6]:
```
# Create a Pandas DataFrame of the survey responses
#
var_ict1110_survey = pd.read_csv("db-unza21-csc5741-ict1110_2018_19-preliminary_survey.
↪csv", sep="|")
var_ict1110_survey.columns

# Rename columns to ensure consistent naming format is used
```

```
#
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)

var_ict1110_survey.columns

# Inspect some of the records
#
var_ict1110_survey.tail(2).T
```

[6]:
```
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')
```

[6]:
```
Index(['Timestamp', 'StudentName', 'StudentID', 'HomeTown', 'MinorProgramme',
       'MinorProgrammeMotivation', 'MajorProgrammeMotivation',
       'DidComputerStudies', 'HasComputerTraining', 'ComputerTrainingType',
       'ExperienceWithComputers', 'HasComputerAccess', 'AboutMe'],
      dtype='object')
```

[6]:

	37	38
Timestamp	2019/04/08 4:53:14 AM GMT+2	2019/04/08 11:33:44 AM GMT+2
StudentName	Participant38	Participant39
StudentID	94984a8c4896946d9bafd24959...	8e4d9eeed250a9d065ac2bb8bd...
HomeTown	Chamba valley, Lusaka	Airport, Sowezi/NWP
MinorProgramme	Mathematics	Religious Education
MinorProgrammeMotivation	I felt maths would combine...	Passionate for it
MajorProgrammeMotivation	Though i didn't choose to ...	To learner more about Tech...
DidComputerStudies	No	No
HasComputerTraining	No	Yes
ComputerTrainingType	NaN	Basics of computer.
ExperienceWithComputers	1 to 2 years	More than 5 years
HasComputerAccess	Yes	Yes
AboutMe	Am a guitarist	Researching.

[7]:
```
type(var_ict1110_survey["MinorProgramme"])
type(var_ict1110_survey["MinorProgramme"].to_list())
```

[7]:
```
pandas.core.series.Series
```

[7]:
```
list
```

[8]:
```
# Explore Programme Minor entries
var_ict1110_survey["MinorProgramme"].tail(15)

# List unique Programme Minor entries
len(var_ict1110_survey["MinorProgramme"].to_list())

# Extract unique Programme Minor entries
list(set(var_ict1110_survey["MinorProgramme"].to_list()))

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

[8]:

	MinorProgramme
24	History
25	History
26	french
27	Mathematics
28	Academic writing and study...
29	MATHEMATICS
30	MATHEMATICS
31	French
32	Geography
33	Geography
34	Language
35	Geography
36	History

Continued on next page

	MinorProgramme
37	Mathematics
38	Religious Education

[8]:
```
39
```

[8]:
```
['Mathematics ',
 'Religious Studies',
 'French',
 'Data Mining',
 'Religious studies ',
 'Geography',
 'Languages ',
 'Languages 1220 and 1200',
 'Academic writing and study skills',
 'RES1010',
 'Language',
 'History ',
 'Civic education ',
 'Mathematics',
 'art and design',
 'french',
 'RELIGIOUS STUDIES',
 'History',
 'LANGUAGES',
 'GEOGRAPHY',
 'Art',
 'Religious studies',
 'Religious Education',
 'MATHEMATICS',
 'civic education']
```

Case Folding

[9]:
```
var_ict1110_minors
```

[9]:
```
['Mathematics ',
 'Religious Studies',
 'French',
 'Data Mining',
 'Religious studies ',
 'Geography',
 'Languages ',
 'Languages 1220 and 1200',
 'Academic writing and study skills',
 'RES1010',
 'Language',
 'History ',
 'Civic education ',
 'Mathematics',
```

```
 'art and design',
 'french',
 'RELIGIOUS STUDIES',
 'History',
 'LANGUAGES',
 'GEOGRAPHY',
 'Art',
 'Religious studies',
 'Religious Education',
 'MATHEMATICS',
 'civic education']
```

[10]:
```
var_z_result = []
for var_z in var_ict1110_minors:
    var_z_result.append(var_z.lower())
var_z_result
```

[10]:
```
['mathematics ',
 'religious studies',
 'french',
 'data mining',
 'religious studies ',
 'geography',
 'languages ',
 'languages 1220 and 1200',
 'academic writing and study skills',
 'res1010',
 'language',
 'history ',
 'civic education ',
 'mathematics',
 'art and design',
 'french',
 'religious studies',
 'history',
 'languages',
 'geography',
 'art',
 'religious studies',
 'religious education',
 'mathematics',
 'civic education']
```

[ ]:
```

```

[11]:
```
# 1. Case Folding

len(var_ict1110_minors)

# 1 (a) Use consistent casing
var_ict1110_minors = [var_minor.lower() for var_minor in var_ict1110_minors]
```

```
var_ict1110_minors
len(var_ict1110_minors)
```

[11]: 25

[11]:
```
['mathematics ',
 'religious studies',
 'french',
 'data mining',
 'religious studies ',
 'geography',
 'languages ',
 'languages 1220 and 1200',
 'academic writing and study skills',
 'res1010',
 'language',
 'history ',
 'civic education ',
 'mathematics',
 'art and design',
 'french',
 'religious studies',
 'history',
 'languages',
 'geography',
 'art',
 'religious studies',
 'religious education',
 'mathematics',
 'civic education']
```

[11]: 25

[12]:
```
[var_minor.lower() for var_minor in var_ict1110_minors]
```

[12]:
```
['mathematics ',
 'religious studies',
 'french',
 'data mining',
 'religious studies ',
 'geography',
 'languages ',
 'languages 1220 and 1200',
 'academic writing and study skills',
 'res1010',
 'language',
 'history ',
 'civic education ',
 'mathematics',
 'art and design',
 'french',
```

```
 'religious studies',
 'history',
 'languages',
 'geography',
 'art',
 'religious studies',
 'religious education',
 'mathematics',
 'civic education']
```

Deduplication

[13]:
```
# 2. Deduplication
var_ict1110_minors = list(set(var_ict1110_minors))
len(var_ict1110_minors)
var_ict1110_minors
```

[13]:
```
20
```

[13]:
```
['history ',
 'res1010',
 'data mining',
 'language',
 'french',
 'art and design',
 'religious education',
 'civic education ',
 'religious studies ',
 'mathematics',
 'languages ',
 'languages 1220 and 1200',
 'religious studies',
 'academic writing and study skills',
 'mathematics ',
 'history',
 'languages',
 'geography',
 'art',
 'civic education']
```

Punctuation

[14]:
```
# 3. Punctuation
# Remove training spaces
#
var_ict1110_minors
len(var_ict1110_minors)

#
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
```

[14]:
```
['history ',
 'res1010',
 'data mining',
 'language',
 'french',
 'art and design',
 'religious education',
 'civic education ',
 'religious studies ',
 'mathematics',
 'languages ',
 'languages 1220 and 1200',
 'religious studies',
 'academic writing and study skills',
 'mathematics ',
 'history',
 'languages',
 'geography',
 'art',
 'civic education']
```

[14]:
```
20
```

[14]:
```
15
```

[14]:
```
['data mining',
 'art',
 'language',
 'french',
 'art and design',
 'religious education',
 'mathematics',
 'languages 1220 and 1200',
 'religious studies',
 'res1010',
 'history',
 'languages',
 'geography',
 'academic writing and study skills',
 'civic education']
```

Stopwords

[15]:
```
len(stopwords.words('english'))
len(stopwords.words('portuguese'))

#
stopwords.words('english')[0:10]
```

[15]:
```
179
```

[15]:
```
203
```

[15]:
```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're"]
```

[16]:
```
# 4. Stopwords

# import stopwokds from nltk library
# from nltk.corpus import stopwords
stopwords.words('english')[0:20] # How about non-english languages? Lozi, IciBemba,␣
 ↪IciTonga???
```

[16]:
```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his']
```

[17]:
```
var_x_result = []
for var_x in var_ict1110_minors_punct:
    var_x_string = " "
    for var_y in var_x.split():
        if (not(var_y in stopwords.words('english'))):
            var_x_string += "".join(var_y)
    var_x_result.append(var_x_string)
```

[18]:
```
var_x_result
```

[18]:
```
[' datamining',
 ' art',
 ' language',
 ' french',
 ' artdesign',
 ' religiouseducation',
 ' mathematics',
 ' languages12201200',
 ' religiousstudies',
 ' res1010',
 ' history',
 ' languages',
 ' geography',
 ' academicwritingstudyskills',
 ' civiceducation']
```

[19]:
```
# 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
```

[19]:
```
['data mining',
 'art',
 'language',
 'french',
 'art design',
 'religious education',
 'mathematics',
 'languages 1220 1200',
 'religious studies',
 'res1010',
 'history',
 'languages',
 'geography',
 'academic writing study skills',
 'civic education']
```

Stemming

[20]:
```
# 5. Stemming

# Import NLTKs PorterStemmer: implements the Porter stemming algorithm
### from nltk.stem.porter import PorterStemmer
var_stemmer = PorterStemmer()
var_stemmer.stem("program")
var_stemmer.stem("programs")
var_stemmer.stem("programmer")
var_stemmer.stem("programmers")
var_stemmer.stem("programming")
```

[20]: 'program'

[20]: 'program'

[20]: 'programm'

[20]: 'programm'

[20]: 'program'

[21]:
```
# Check length of list
len(var_ict1110_minors_stop)
```

[21]: 15

[22]:
```
# Stem single words only [...] for illustration purposes
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]

#
var_ict1110_minors_stem
```

[22]:
```
['data mining',
 'art',
 'languag',
 'french',
 'art design',
 'religious education',
 'mathemat',
 'languages 1220 1200',
 'religious studies',
 'res1010',
 'histori',
 'languag',
 'geographi',
 'academic writing study skills',
 'civic education']
```

[23]:
```
var_ict1110_minors_stem = list(set(var_ict1110_minors_stem))
var_ict1110_minors_stem
len(var_ict1110_minors_stem)
```

[23]:
```
['data mining',
 'mathemat',
 'histori',
 'french',
 'religious education',
 'languages 1220 1200',
 'art design',
 'geographi',
 'religious studies',
```

```
'academic writing study skills',
'res1010',
'art',
'languag',
'civic education']
```

[23]:
```
14
```

Exercise 1: Preprocessing Students' Interets in 2018/19 ICT 1110 Preliminary Survey

1. Using the example dataset and queestions above, work towards the following
    1. Identify outliers
    2. Remove duplicate entries
2. Using the 2018/19 ICT 1110 Information Survey dataset
    1. Cleanup the data related to students' interests—"List one interesting fact about yourself (e.g. I cycle everyday!):"

Dataset #2: University of Zambia Institutional Repository Digital Objects

Link to dataset: http://bit.ly/2Wmhw6v

The University of Zambia (UNZA) uses an Institutional Repository (IR) to archive scholarly research output generated by faculty staff and students. The UNZA IR is hosted using a DSpace instance. The types of research output archived in the UNZA IR include journal article pre-prints and post-prints, books, book chapters, technical reports, postgraduate Electronic Theses and Dissertations (ETDs) and undergraduate student capstone project reports.

The digital objects ingested into the IR are broadly composed of bitstreams (PDF documents) and corresponding metadata (descriptive, administrative and structural information about the PDF manuscripts).

Dataset Description

- This dataset comprises of sample 2000 digital objects harvested from the UNZA IR using the OAI-PMH protocol.
- The dataset has observations presented in CSV format, using "|" as the separator.
- The descriptive metadata is encoded using Dublin Core, which has the characteristic of being repeatable and optional. Repeatable data attributes use the "=" as the separator.
- Each observation is associated with the following 12 data attributes:
    - _identifier—global unique identifier for the digital object
    - _datestamp—timestamp when digital object was ingested into the repository
    - _setSpec —container structure housing the digital object
    - title—the title of the digital object
    - creator—the author of the digital object
    - subject—the subject categories associated with the digital object
    - description—descriptive information about the digital object, usually the abstract
    - date—the date when the digital object was published
    - type—the type of the digital object, e.g. Thesis, Article, Conference Paper
    - identifier—Handle URL the uniquely identifiers the digital object
    - language—the language used to author the digital object
    - format—the document format for the digital object

[24]:
```
!head -n 2 db-unza21-csc5741-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|application
/pdf
```

[25]:
```
# Use pandas to pluck out abstracts
var_unza_etds = pd.read_csv("db-unza21-csc5741-dspace_unza_zm.csv", sep="|")
var_unza_etds.columns

# Explore abstracts
var_unza_etds["description"].head(20)
len(var_unza_etds)
```

[25]:
```
Index(['_identifier', '_datestamp', '_setSpec', 'title', 'creator', 'subject',
       'description', 'date', 'type', 'identifier', 'language', 'format'],
      dtype='object')
```

[25]:

	description
0	Morphological characteriza...
1	The purpose of the study w...
2	M.ED=The purpose of the st...
3	The purpose of the study w...
4	Past Exams for the departm...
5	Background and Objective: ...
6	Effects of Bacillus thurin...
7	Student Project Report=Far...
8	The report is as a result ...
9	The language-in-education ...
10	Third world countries have...
11	Acceptability of Antiretro...
12	past exams for the school ...
13	Master of Science degree i...
14	Masters in Clinical Pharma...

Continued on next page

	description
15	Zambia similar to other su...
16	Cassava is an important cr...
17	NaN
18	This study investigates th...
19	This study investigated th...

[25]:
```
2000
```

Missing Values

[26]:
```
#1.Missing Values
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
type(var_unza_etds_dict)
```

[26]:
```
Index(['description'], dtype='object')

/home/lightonphiri/.local/lib/python3.6/site-packages/pandas/core/series.py:4536:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  downcast=downcast,
```

[26]:
```
pandas.core.frame.DataFrame
```

[27]:
```
# Extract relevant columns
var_unza_etds_dict = var_unza_etds_dict.to_dict()

type(var_unza_etds_dict)
```

[27]:
```
dict
```

[28]:
```
var_unza_etds_dict["description"][1]
```

[28]:
```
'The purpose of the study was to evaluate the use of Instruction Based Formative
Assessment in  Colleges of Education in Zambia. The objectives of the study were to
establish the use of Instruction Based Formative Assessment during lectures, to␣
 ↪determine
the predominant Instruction Based Formative Assessment strategies being used and to
examine factors affecting the use of Instruction Based Formative Assessment in Colleges
of Education in Zambia.\\r\\nThe target population was Lecturers and Coordinators of
Continuous Professional Development (CPD) and Open Distance Learning (CODEL) in colleges
of Education in Zambia. A total of 120 respondents participated in the study. There were
100 lecturers and 20 coordinators (10 CPD, 10 CODEL). Quantitative survey research␣
 ↪design
```

```
was used to capture national wide data covering 80% of the provinces of Zambia. Data was
collected using  structured questionnaires. Data was analysed using Statistical Package
for Social Science (SPSS) software that generated frequencies and percentages which were
used in describing distributions of single and summated variables.\\r\\nThe study
established that: (i) Instruction Based Formative Assessment was used by both
coordinators and lecturers during lectures in Colleges of Education in Zambia. However
the frequency of using this type of assessment, varied among coordinators and lecturers.
(ii) The predominant  Formative Assessment technique used in Colleges of Education was
the technique that involved providing feedback that moves learners forward in their
learning while the predominant Formative Assessment activity was that which involves
getting students to peer assess their work. (iii) The predominant factors that affected
the use of Instruction Based Formative Assessment in Colleges of Education in Zambia␣
 ↪were
time limitation and large class size. In view of the findings of the study, the
following recommendations were made:  (i) Administrators of colleges of education
should ensure that lecturers in Colleges are given in-service training in student␣
 ↪centred
instructional and assessment strategies, which include instruction based formative
assessment. (ii) Administrators of colleges of education should establish CPD policy of
training and orientation of new lecturers in Instructional and assessment strategies.
(iii) Ministry of Education, Science, Vocational Training and Early Education in
collaboration with Colleges of Education administrators, should ensure that, learner
centred Instructional and assessment strategies, with emphasis on an orientation towards
formative assessment, is included as a specific component of the teaching methods
training curriculum and should be included in the school experience appraisal monitoring
tool.'
```

Case Folding

[29]:
```
# 2. Case Folding
# Good idea to implement a function here [...]

def fxn_etd_case_folding(var_input):
    return var_input.lower()


# Testing function
var_unza_etds_dict["description"][0]
fxn_etd_case_folding(var_unza_etds_dict["description"][0])

# Apply function to dictionary items
#
var_etds_dict_case = {}
for var_etd in var_unza_etds_dict["description"]:
    var_etds_dict_case[var_etd] =␣
 ↪fxn_etd_case_folding(var_unza_etds_dict["description"][var_etd])

len(var_etds_dict_case)

#
# Compare results before and after case folding
```

```
var_unza_etds_dict["description"][0]
var_etds_dict_case[0]
```

[29]: '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'

[29]: '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'

[29]: 2000

[29]: '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'

[29]: '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'

Punctuation

[30]:
```
# 3. Punctuation

# import re library for making the most out of regular expressions and string for␣
 ↪punctuations
###import re
###import string

# Check list of punctuation marks
string.punctuation

# Experiment with removing punctuations
var_example_text = " I got 25% in that useless test we wrote in 2010.! August2010 to␣
 ↪be exact"
var_example_text = re.sub("[%s]" % re.escape(string.punctuation), "", var_example_text)
var_example_text

# Experiment with removing numbers
re.sub('\w*\d\w*', '', var_example_text)

# Function for removing stopwords from string of text
def fxn_etd_punctuation(var_input_text):
    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


# Test function
var_unza_etds_dict["description"][0]
len(var_unza_etds_dict["description"][0])

fxn_etd_punctuation(fxn_etd_case_folding(var_unza_etds_dict["description"][0]))
len(fxn_etd_punctuation(fxn_etd_case_folding(var_unza_etds_dict["description"][0])))
```

[30]: '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'

[30]: ' I got 25 in that useless test we wrote in 2010 August2010 to be exact'

[30]: ' I got in that useless test we wrote in   to be exact'

[30]: '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'

[30]: 910

[30]: 'morphological characterization was done on three sunflower varieties 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 flowering and maturity seed colour presence of seed stripes colour of seed stripes position of seed stripes shape of seed weight of seeds kernel and oil percentages significant differences were noted in leaf size plant height days to flowering and maturity record had the highest oil percentage of milika and 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 agroclimatic regions'

[30]: 853

Stopwords

[31]:
```
# 4. Stopwords

# import stopwokds from nltk library
###from nltk.corpus import stopwords

# Function for removing stopwords from string of text
def fxn_etd_stopwords(var_input_text):
    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


# Test function
var_unza_etds_dict["description"][0]
len(var_unza_etds_dict["description"][0])

fxn_etd_stopwords(
    fxn_etd_punctuation(
        fxn_etd_case_folding(var_unza_etds_dict["description"][0])))
```

```
len(fxn_etd_stopwords(
    fxn_etd_punctuation(
        fxn_etd_case_folding(var_unza_etds_dict["description"][0]))))
```

[31]: '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'

[31]: 910

[31]: 'morphological characterization done three sunflower varieties milika record order see morphological differences possible use marker assisted selection parameters looked leaf size leaf shape colour leaves number leaves per plant hairiness top stem days flowering maturity seed colour presence seed stripes colour seed stripes position seed stripes shape seed weight seeds kernel oil percentages significant differences noted leaf size plant height days flowering maturity record highest oil percentage milika parameters significant differences established variation among characteristics important allows development varieties adapted specific environments agroclimatic regions'

[31]: 676

Stemming

[32]:
```
# 5. Stemming

# Import NLTKs PorterStemmer: implements the Porter stemming algorithm
###from nltk.stem.porter import PorterStemmer

var_stemmer = PorterStemmer()
var_stemmer.stem("country")
var_stemmer.stem("countries")

# Function for removing stopwords from string of text
# Remember: input will be chunck of text
def fxn_etd_stem(var_input_text):
    var_output_text = " ".join([
        var_stemmer.stem(var_etd_word) for var_etd_word in var_input_text.split()
    ])
    return var_output_text


# Test function
```

```
var_unza_etds_dict["description"][0]
len(var_unza_etds_dict["description"][0])

fxn_etd_stem(
    fxn_etd_stopwords(
        fxn_etd_punctuation(
            fxn_etd_case_folding(var_unza_etds_dict["description"][0]))))
len(fxn_etd_stem(
    fxn_etd_stopwords(
        fxn_etd_punctuation(
            fxn_etd_case_folding(var_unza_etds_dict["description"][0])))))
```

[32]: 'countri'

[32]: 'countri'

[32]: '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'

[32]: 910

[32]: 'morpholog character done three sunflow varieti milika record order see morpholog differ possibl use marker assist select paramet look leaf size leaf shape colour leav number leav per plant hairi top stem day flower matur seed colour presenc seed stripe colour seed stripe posit seed stripe shape seed weight seed kernel oil percentag signific␣ ↪differ
note leaf size plant height day flower matur record highest oil percentag milika paramet signific differ establish variat among characterist import allow develop varieti adapt specif environ agroclimat region'

[32]: 558

Exercise 2: Preprocessing The University of Zambia Institutional Repository Objects

Using the UNZA IR Digital Objects dataset and questions above, work towards the following 1. Apply all data pre-processing tasks to the title field of all the digital objects 2. Apply all data pre-processing tasks to the subject field of all the digital objects