CSC 5741: Lecture #04—Data Processing and Transformation

Lighton Phiri
<lighton.phiri@unza.zm>

April 16 2019

Contents

Introduction

During these "hands-on" activities, we look at practical examples of how to clean data and transform it into a form a computer—READ: algorithms—will be able to understand.

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.

[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

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

Data Preprocessing

Example 1: 2018/19 ICT 1110 Information Survey

Dataset

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

```
2019/04/05 9:01:53 AM GMT+2|Participant37|1ffea8832ef6ac0d73a09a441562393a|8 miles /
chibombo/ Central province |History |It's seemed like the best option |Computers interest
me|Yes|No||No Experience|No|I am ambidextrous
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.
```

[3]:
```
# 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)

var_ict1110_survey.columns

# Inspect some of the records
var_ict1110_survey.tail(3).T
```

[3]:
```
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')
```

[3]:
```
Index(['Timestamp', 'StudentName', 'StudentID', 'HomeTown', 'MinorProgramme',
       'MinorProgrammeMotivation', 'MajorProgrammeMotivation',
       'DidComputerStudies', 'HasComputerTraining', 'ComputerTrainingType',
       'ExperienceWithComputers', 'HasComputerAccess', 'AboutMe'],
      dtype='object')
```

[3]:

	36	37
Timestamp	2019/04/05 9:01:53 AM GMT+2	2019/04/08 4:53:14 AM GMT+2
StudentName	Participant37	Participant38
StudentID	1ffea8832ef6ac0d73a09a441562393a	5f59b44f0c6a470fe410c0df68985274
HomeTown	8 miles / chibombo/ Central province	Chamba valley, Lusaka
MinorProgramme	History	Mathematics
MinorProgrammeMotivation	It's seemed like the best option	I felt maths would combine well with ICTs.
MajorProgrammeMotivation	Computers interest me	Though i didn't choose to do ict in the first

	36	37
DidComputerStudies	Yes	No
HasComputerTraining	No	No
ComputerTrainingType	NaN	NaN
ExperienceWithComputers	No Experience	1 to 2 years
HasComputerAccess	No	Yes
AboutMe	I am ambidextrous	Am a guitarist

[4]:
```
# 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()))
```

[4]:

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

[4]:
```
39
```

[4]:
```
['Religious studies ',
 'LANGUAGES',
 'Data Mining',
 'History',
 'GEOGRAPHY',
 'Languages 1220 and 1200',
 'French',
 'Languages ',
 'Civic education ',
 'RES1010',
```

```
 'art and design',
 'MATHEMATICS',
 'french',
 'Academic writing and study skills',
 'History ',
 'Art',
 'Mathematics',
 'Religious studies',
 'Language',
 'civic education',
 'Religious Studies',
 'Religious Education',
 'RELIGIOUS STUDIES',
 'Mathematics ',
 'Geography']
```

Case Folding

[5]:
```
# 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
```

[5]:
```
25
```

[5]:
```
['religious studies ',
 'languages',
 'data mining',
 'history',
 'geography',
 'languages 1220 and 1200',
 'french',
 'languages ',
 'civic education ',
 'res1010',
 'art and design',
 'mathematics',
 'french',
 'academic writing and study skills',
 'history ',
 'art',
 'mathematics',
 'religious studies',
 'language',
 'civic education',
 'religious studies',
 'religious education',
```

```
 'religious studies',
 'mathematics ',
 'geography']
```

[6]:
```
[var_minor.lower() for var_minor in var_ict1110_minors]
```

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

Deduplication

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

[7]:
```
20
```

[7]:
```
['languages',
 'history',
 'geography',
 'art',
 'religious studies ',
 'civic education',
 'res1010',
 'art and design',
 'history ',
```

```
 'french',
 'religious studies',
 'language',
 'religious education',
 'data mining',
 'languages 1220 and 1200',
 'languages ',
 'mathematics ',
 'civic education ',
 'academic writing and study skills',
 'mathematics']
```

Punctuation

[8]:
```
# 3. Punctuation
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
```

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

[8]:
```
20
```

[8]:
```
15
```

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

Stopwords

[9]:
```
# 4. Stopwords

# import stopwokds from nltk library
# from nltk.corpus import stopwords
stopwords.words('english')[0:20] # Lozi, IciBemba, IciTonga???
```

[9]:
```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his']
```

[10]:
```
# 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
```

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

Stemming

[11]:
```
# 5. Stemming

# Import NLTKs PorterStemmer: implements the Porter stemming algorithm
### from nltk.stem.porter import PorterStemmer
var_stemmer = PorterStemmer()
var_stemmer.stem("languages")
var_stemmer.stem("language")
```

[11]:
```
'languag'
```

[11]:
```
'languag'
```

[12]:
```
# Check length of list
len(var_ict1110_minors_stop)
```

[12]:
```
15
```

[13]:
```
# 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
```

[13]:
```
['languag',
 'histori',
 'art',
 'geographi',
 'religious education',
 'civic education',
 'res1010',
 'art design',
```

```
 'religious studies',
 'french',
 'data mining',
 'languages 1220 1200',
 'languag',
 'academic writing study skills',
 'mathemat']
```

[14]:
```
var_ict1110_minors_stem = list(set(var_ict1110_minors_stem))
var_ict1110_minors_stem
len(var_ict1110_minors_stem)
```

[14]:
```
['academic writing study skills',
 'languages 1220 1200',
 'art',
 'mathemat',
 'religious education',
 'civic education',
 'res1010',
 'religious studies',
 'french',
 'data mining',
 'geographi',
 'histori',
 'art design',
 'languag']
```

[14]:
```
14
```

Exercise 1: Preprocessing Students' Interets in 2018/19 ICT 1110 Information 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 (dataset available on http://lis.unza.zm/~lightonphiri/teaching/unza/2019/csc5741/resources/db-unza19-ict1110_2018_19-preliminary_survey.csv):
    1. Cleanup the data related to students' interests—"List one interesting fact about yourself (e.g. I cycle everyday!):"

Data Transformation

Example 2: University of Zambia ETD Abstracts

Dataset

[15]:
```
!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
```

[16]:
```
# Use pandas to pluck out abstracts
var_unza_etds = pd.read_csv("db-unza19-dspace_unza_zm.csv", sep="|")
var_unza_etds.columns

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

[16]:
```
Index(['_identifier', '_datestamp', '_setSpec', 'title', 'creator', 'subject',
       'description', 'date', 'type', 'identifier', 'language', 'format'],
      dtype='object')
```

[16]:

	description
0	Morphological characterization was done on thr...
1	The purpose of the study was to evaluate the u...
2	M.ED=The purpose of the study was to assess th...
3	The purpose of the study was to investigate th...
4	Past Exams for the department of Library and i...
5	Background and Objective: There is paucity of ...
6	Effects of Bacillus thuringiensis var. israele...
7	Student Project Report=Farm credit can stimula...
8	The report is as a result of the study on HIV/...

Continued on next page

	description
9	The language-in-education policy in Zambia is ...
10	Third world countries have always sought to be...
11	Acceptability of Antiretrovirals (ARVs) has be...
12	past exams for the school of Humanities and so...
13	Master of Science degree in Pathology (Haemato...
14	Masters in Clinical Pharmacy=Poor sleep plays ...
15	Zambia similar to other sub Saharan countries ...
16	Cassava is an important crop in many parts of ...
17	NaN
18	This study investigates the factors that affec...
19	This study investigated the role of student re...

[16]:
```
1699
```

Missing Values

[17]:
```
#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)
```

[17]:
```
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)
```

[17]:
```
pandas.core.frame.DataFrame
```

[18]:
```
# Extract relevant columns
var_unza_etds_dict = var_unza_etds_dict.to_dict()

type(var_unza_etds_dict)
```

[18]:
```
dict
```

[19]:
```
var_unza_etds_dict["description"][1]
```

[19]:
```
'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

[20]:
```
# 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]
```

[20]: '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'

[20]: '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'

[20]: 1699

[20]: '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'

[20]:
```
'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

[21]:
```
# 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])))
```

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

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

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

[21]: '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'

[21]: 910

[21]: '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'

[21]: 853

Stopwords

[22]:
```
# 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]))))
```

[22]:
```
'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'
```

[22]:
```
910
```

[22]:
```
'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'
```

[22]:
```
676
```

Stemming

[23]:
```
# 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])))))
```

[23]: 'countri'

[23]: 'countri'

[23]: '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'

[23]: 910

[23]: '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'

[23]: 558

Exercise 2: Preprocessing The University of Zambia ETD Abstracts

1. Using the example dataset and queestions above, work towards the following
    1. Apply all data preprocessing tasks to the entire dataset of ETDs
2. Note on Mini Projects: this could just as well be applied to (i) YouTube video titles, descriptions and comments (ii) News posts and comments (iii) OAI-PMH extracted ETD titles, subjects and abstracts (iv) Google Scholar extracted publications

Bag-of-Words Model

[24]:
```
# 6. Bag-of-Words

# Recap!
# We are working with a corpus of 1,699 ETD abstracts
var_unza_etds["description"].head(5)
len(var_unza_etds)
```

[24]:

	description
0	Morphological characterization was done on thr...
1	The purpose of the study was to evaluate the u...
2	M.ED=The purpose of the study was to assess th...
3	The purpose of the study was to investigate th...
4	Past Exams for the department of Library and i...

[24]:
```
1699
```

[25]:
```
# Explore original corpus entries in order to compare with preprocessed corpus
#
var_unza_etds_dict["description"][11]
var_etds_cleaned = {}
```

[25]:
```
'Acceptability of Antiretrovirals (ARVs) has been found to be associated with several
factors. In this study we investigated the level of willingness among adults living in
Chawama and factors likely to be associated with willingness to taking ARVs\\r\\nThis␣
 ↪was
a cross sectional study. Only eligible adults 18 years and above were recruited by a
simple random sampling. A structured questionnaire was used to collect data socio-
demographic and other factors likely to influence willingness The Chi square test was
used to determine association between variables of interest and multivariate analysis␣
 ↪was
performed to determine predictors of willingness\\r\\nOverall (n=409), 52.8% females and
46.9% males participated in the study. The non response rate was less than 1%. Overall
(n=409), 52.8% females and 46.9% males participated in this study. The non response rate
was less than 1%. A high level of willingness was observed with more than 50% of
participants willing to take ARVs if they were found legible for ART. The mean age of
participants was 31 years (SDś11.60). Some of the key factors that were found
significantly associated with willingness were, the aspect of being male or female [OR:
2.27 (95%CI, 1.10 - 4.70)] with females being more likely to be willing than males, the
perceived effectiveness of ARVs [OR: 3.50(1.71 - 7.82))], the need for consent to begin
```

```
ARV treatment [OR: 1.30(95% CI, 1.40-2.72)] with females being more likely to needing
consent than men, and fear of discrimination [OR: 2.47(95% CI,1.22  5.00)]\\r\\nA high
willingness to take Antiretroviral drugs among community members was observed but there
is need to increase intervention programs that promote acceptability and uptake of ARVs.
Furthermore stigmatizing attitudes, gender and socio-cultural influences towards people
taking ARVs still persist and interventions to reduce these influences are needed.'
```

[26]:
```
# Apply preprocessing functions to corpus
for var_etd_item in var_unza_etds_dict["description"]:
    var_etds_cleaned[var_etd_item] = fxn_etd_stem(
        fxn_etd_stopwords(
            fxn_etd_punctuation(
                ␣
 ↪fxn_etd_case_folding(var_unza_etds_dict["description"][var_etd_item]))))


var_etds_cleaned[5]
```

[26]:
```
'background object pauciti data outcom combin vp insert myelomeningocoel repair whether
reduc morbid mortal studi design address research questionrnmethod prospect descript
intervent studi use patient recruit januari octob give total inform sociodemograph␣
 ↪referr
statu preoper postop outcom document analysedresult male constitut femal case youngest
age present week oldest week major refer clinic hospit outsid lusaka hail poor
socioeconom background malform occur lumbar sacral region patient present normal mild
form neurolog impair ultrasound examin show mild find moder form hydrocephalu patient
shunt surgic closur sac postop complic seen patient oedamat infect wound one patient csf
leakag later die mening averag hospit stay patient referr hospit centr neurosurg unit
public institut children born defect outsid lusaka continu seen latealthough sampl size
small could show even come late combin surgic approach still recommend patient patient
oper recov well postop period hospit stay'
```

[27]:
```
# Import Counter from collections library
###from collections import Counter

# Create list that will hold tokenized abstracts
var_etd_corpus = []
for var_etd_item in var_etds_cleaned:
    var_etd_corpus += var_etds_cleaned[var_etd_item].split()

# Explore corpus
len(var_etd_corpus)

# Explore most frequent words in corpus
var_counter = Counter(var_etd_corpus)
var_counter.most_common(20)

var_corpus_dictionary = var_counter.most_common(20)


# Compare cleaned and unclean corpara
```

```
var_etd_corpus_dirty = []
for var_etd_item in var_unza_etds_dict["description"]:
    var_etd_corpus_dirty += var_unza_etds_dict["description"][var_etd_item].split()

var_etd_corpus_clean = []
for var_etd_item in var_etds_cleaned:
    var_etd_corpus_clean += var_etds_cleaned[var_etd_item].split()

print("Cleaned Corpus: ", len(var_etd_corpus_clean))
print("Dirty Corpus: ", len(var_etd_corpus_dirty))
```

[27]:
```
291025
```

[27]:
```
[('studi', 4922),
 ('use', 3373),
 ('school', 1901),
 ('zambia', 1860),
 ('educ', 1552),
 ('teacher', 1461),
 ('also', 1459),
 ('data', 1410),
 ('research', 1131),
 ('develop', 1017),
 ('health', 1010),
 ('effect', 1000),
 ('sampl', 993),
 ('find', 991),
 ('commun', 975),
 ('level', 954),
 ('women', 938),
 ('factor', 904),
 ('inform', 890),
 ('collect', 884)]

Cleaned Corpus:  291025
Dirty Corpus:  510207
```

Document Term Frequency

[28]:
```
# 7. Term Frequency

def fxn_transform(var_input_dictionary):
    var_transformed_dataset = []
    for var_etd_entry in var_etds_cleaned:
        #print (var_etd_entry)
        var_dataset = []
        var_etd_tokens = var_etds_cleaned[var_etd_entry].split()
        for var_dictionary_entry in var_input_dictionary:
            #print (var_dictionary_entry)
```

```
            var_dataset.append(var_etd_tokens.count(var_dictionary_entry[0]))
        var_transformed_dataset.append(var_dataset)
    return var_transformed_dataset

var_X = fxn_transform(var_corpus_dictionary)
```

[29]:
```
# Inspect the first couple of documents
print (var_X[0:20])
```

```
[[0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [5, 12, 1, 5, 11, 0, 0, 3,
1, 1, 0, 0, 0, 1, 0, 0, 0, 2, 0, 1], [3, 7, 6, 1, 1, 17, 0, 2, 1, 0, 0, 0, 3, 2, 0, 2, 0,
0, 0, 2], [4, 1, 11, 0, 2, 10, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0], [1, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0], [2, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1,
0, 0, 0, 0, 1, 0], [17, 4, 0, 3, 0, 0, 0, 0, 0, 0, 3, 1, 0, 0, 0, 0, 0, 0, 0, 1], [4, 1,
0, 1, 2, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 3, 1, 2, 0, 0], [3, 6, 0, 1, 2, 0, 1, 0, 0, 0, 2,
8, 0, 2, 8, 0, 0, 0, 3, 0], [3, 2, 4, 1, 0, 3, 1, 2, 2, 0, 0, 0, 3, 2, 0, 0, 0, 0, 0, 1],
[1, 0, 0, 0, 0, 0, 3, 0, 0, 4, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0], [4, 2, 0, 0, 0, 0, 0, 1, 0,
0, 0, 1, 1, 0, 1, 2, 0, 4, 0, 1], [0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0], [3, 0, 0, 2, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 3, 0, 0, 0, 0], [2, 2, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0], [2, 2, 0, 3, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 1,
0, 0, 1, 0, 0], [2, 4, 1, 1, 0, 0, 2, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 4, 0, 1], [0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [2, 2, 0, 5, 0, 0, 1, 1, 0, 0, 0, 1,
0, 1, 0, 0, 0, 3, 2, 0], [2, 4, 2, 1, 5, 1, 3, 3, 1, 0, 0, 1, 1, 2, 0, 0, 0, 0, 2, 1]]
```

[30]:
```
#var_unza_etds_dict["description"][0]
var_unza_etds_dict["description"][0]
```

[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'
```

[31]:
```
var_etds_cleaned[0]
var_etds_cleaned[0].split().count("use")
type(var_etds_cleaned)
```

[31]:
```
'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'
```

[31]:
```
1
```

[31]:
```
dict
```

Plot twist!

We can do what was previously manually done using scikit-learn's CountVectorizer, where every row will represent a different document and every column will represent a different word.

CountVectorizer can also be used to remove stop words.

[32]:
```
###from sklearn.feature_extraction.text import CountVectorizer

# Exclude the use of fxn_etd_stopwords() since CountVectorizer handles this
var_etds_for_vectoriser = {}
for var_etd_item in var_unza_etds_dict["description"]:
    var_etds_for_vectoriser[var_etd_item] = fxn_etd_stem(
        fxn_etd_punctuation(
            fxn_etd_case_folding(var_unza_etds_dict["description"][var_etd_item])))


var_etds_dataframe = pd.DataFrame(list(var_etds_for_vectoriser.items()),␣
 ↪columns=['identifier', 'abstract'])
var_etd_vectoriser = CountVectorizer(stop_words='english')
var_etd_vectoriser_data = var_etd_vectoriser.
 ↪fit_transform(var_etds_dataframe["abstract"])
var_etd_vectoriser_data_tf = pd.DataFrame(var_etd_vectoriser_data.toarray(),␣
 ↪columns=var_etd_vectoriser.get_feature_names())

var_etd_vectoriser_data[0]
print (var_etd_vectoriser_data[0])

var_etd_vectoriser_data_tf.columns
len(var_etd_vectoriser_data_tf.columns)
```

[32]:
```
<1x20564 sparse matrix of type '<class 'numpy.int64'>'
        with 52 stored elements in Compressed Sparse Row format>

  (0, 14624)    1
  (0, 445)      1
  (0, 5601)     1
  (0, 16846)    1
  (0, 216)      1
  (0, 4563)     1
  (0, 547)      1
  (0, 1772)     1
  (0, 8096)     1
  (0, 2780)     1
  (0, 19541)    1
  (0, 5716)     1
  (0, 7570)     1
```

```
  (0, 7466)     1
  (0, 11823)    1
  (0, 16428)    2
  (0, 12937)    2
  (0, 12225)    2
  (0, 9222)     1
  (0, 19936)    1
  (0, 13428)    1
  (0, 17215)    3
  (0, 13661)    1
  (0, 16034)    6
  (0, 10425)    2
  :     :
  (0, 17088)    1
  (0, 7301)     1
  (0, 13220)    2
  (0, 11904)    1
  (0, 9554)     2
  (0, 3303)     3
  (0, 16323)    2
  (0, 16556)    2
  (0, 9520)     3
  (0, 9919)     1
  (0, 12665)    2
  (0, 16058)    1
  (0, 1389)     1
  (0, 10327)    1
  (0, 19407)    1
  (0, 13444)    1
  (0, 4682)     3
  (0, 12357)    1
  (0, 14513)    2
  (0, 10777)    2
  (0, 19545)    2
  (0, 17475)    1
  (0, 19778)    1
  (0, 2778)     1
  (0, 11031)    2
```

[32]:
```
Index(['aa', 'aasgf', 'aat', 'aatrna', 'ab', 'abandon', 'abat', 'abatingrnth',
       'abattoir', 'abbott',
       ...
       'zwpc', 'ţgml', 'lattic', 'lactamas', 'cyhalothrin', 'g', 'gml',
       'm', 'm', 'š'],
      dtype='object', length=20564)
```

[32]:
```
20564
```

[33]:
```
###var_etds_for_vectoriser.items()
###var_etds_dataframe["abstract"]
#var_etd_vectoriser.get_feature_names()
len(var_etd_vectoriser_data_tf.columns)
```

```
var_etd_vectoriser_data_tf.columns[20400:]
```

[33]:
```
20564
```

[33]:
```
Index(['zab', 'zair', 'zambeef', 'zambesiaca', 'zambezi', 'zambeziriv',
       'zambi', 'zambia', 'zambiaan', 'zambiaand',
       ...
       'zwpc', 'ţgml', 'lattic', 'lactamas', 'cyhalothrin', 'g', 'gml',
       'm', 'm', 'š'],
      dtype='object', length=164)
```

TF-IDF

[34]:
```
# 8. TF-IDF

###from sklearn.feature_extraction.text import TfidfVectorizer

# Exclude the use of fxn_etd_stopwords() since CountVectorizer handles this
var_etds_for_vectoriser = {}
for var_etd_item in var_unza_etds_dict["description"]:
    var_etds_for_vectoriser[var_etd_item] = fxn_etd_stem(
        fxn_etd_punctuation(
            fxn_etd_case_folding(var_unza_etds_dict["description"][var_etd_item])))


var_etds_dataframe = pd.DataFrame(list(var_etds_for_vectoriser.items()),␣
 ↪columns=['identifier', 'abstract'])
# Notice the difference with CountVectorizer
var_etd_vectoriser = TfidfVectorizer(stop_words='english', use_idf=True)

var_etd_vectoriser_data = var_etd_vectoriser.
 ↪fit_transform(var_etds_dataframe["abstract"])
var_etd_vectoriser_data_tfidf = pd.DataFrame(var_etd_vectoriser_data.toarray(),␣
 ↪columns=var_etd_vectoriser.get_feature_names())

var_etd_vectoriser_data[0]
print (var_etd_vectoriser_data[0])

var_etd_vectoriser_data_tfidf.columns
len(var_etd_vectoriser_data_tfidf.columns)
```

[34]:
```
<1x20564 sparse matrix of type '<class 'numpy.float64'>'
        with 52 stored elements in Compressed Sparse Row format>

  (0, 11031)    0.16800519068679132
  (0, 2778)     0.07763305187986462
  (0, 19778)    0.02131231479231516
  (0, 17475)    0.10441332680548807
  (0, 19545)    0.13979325387021127
```

```
  (0, 10777)    0.24334335348135858
  (0, 14513)    0.10376211847172608
  (0, 12357)    0.04090011627076904
  (0, 4682)     0.11835916883439296
  (0, 13444)    0.056586232426748984
  (0, 19407)    0.023398678667304818
  (0, 10327)    0.09804378334195703
  (0, 1389)     0.06452800966949129
  (0, 16058)    0.03793677706243364
  (0, 12665)    0.13979325387021127
  (0, 9919)     0.054931100159132155
  (0, 9520)     0.25420016650080113
  (0, 16556)    0.10686770179969557
  (0, 16323)    0.1743099537405937
  (0, 3303)     0.28916595322302063
  (0, 9554)     0.15167150738173418
  (0, 11904)    0.041319053820484654
  (0, 13220)    0.1334884908164089
  (0, 7301)     0.12167167674067929
  (0, 17088)    0.08715497687029684
  :     :
  (0, 10425)    0.17099968920536007
  (0, 16034)    0.46298450941321273
  (0, 13661)    0.06537736392697652
  (0, 17215)    0.33234861080705735
  (0, 13428)    0.041413688004852726
  (0, 19936)    0.06850552150598152
  (0, 9222)     0.09989406379735895
  (0, 12225)    0.19608756668391406
  (0, 12937)    0.12204519389294512
  (0, 16428)    0.08282737600970545
  (0, 11823)    0.06069530954393918
  (0, 7466)     0.07966286831339223
  (0, 7570)     0.06452800966949129
  (0, 5716)     0.0378608869288612
  (0, 19541)    0.07276075767561793
  (0, 2780)     0.06258612996273759
  (0, 8096)     0.04155671685721763
  (0, 1772)     0.04275039672617095
  (0, 547)      0.0635270834715746
  (0, 4563)     0.03521913520959087
  (0, 216)      0.0781164508540386
  (0, 16846)    0.0476885972840157
  (0, 5601)     0.05683758542260356
  (0, 445)      0.12167167674067929
  (0, 14624)    0.06258612996273759
```

[34]:
```
Index(['aa', 'aasgf', 'aat', 'aatrna', 'ab', 'abandon', 'abat', 'abatingrnth',
       'abattoir', 'abbott',
       ...
       'zwpc', 'ţgml', 'lattic', 'lactamas', 'cyhalothrin', 'g', 'gml',
       'm', 'm', 'š'],
```

```
       dtype='object', length=20564)
```

[34]:
```
20564
```