Introduction

INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION

ISKO

An introduction to Machine Learning & Data Visualization

Tara McDarby

April 21st 2020

Machine Learning

Using computer programs such as python to process data

Advantage:

Can process large datasets quickly

Use

Volume/Size

- effect this has on resources
- collections growing at certain levels
- areas of dominance changing overall nature of collection
- new technical tools shaping subject headings and classification (computer generated art)

Python

(object oriented programming)

- abstraction (hiding unnecessary details from the user)
- encapsulation (combining data and methods that work on that data within one unit)
- inheritance (when an already existing class extends its features to a new class).
- polymorphism (when objects of different types can be accessed through the same interface)

https://stackify.com/oop-concept-polymorphism/

Link: https://www.python.org/

Glossary

python
jupyter
notebook
script
textual data

Libraries

A library in python is a collection of functions and methods that you can ‘import’ into your script directly. This saves you having to write the code.

Numpy - scientific computing

Pandas - data manipulation and analysis

Scikit-learn - machine learning and data mining

NLTK - Language processing

Loading a corpus

In python on your jupyter notebook...

```
from nltk.corpus import gutenberg
import matplotlib.pyplot as plt
import matpoltlib

bible = gutenberg.open(‘bible-kjv.txt’)
bible = bible.readlines ()
Bible[:5}
```

Results

```
['[The King James Bible]\n',
'\n',
'The Old Testament of the King James Bible\n',
'\n',
 'The First Book of Moses: Called Genesis\n']
```

Stopwords

```
stopwords = nltk.corpus.stopwords.words(‘english’)

words = [word.lower() for word in words if word.lower() not in
stopwords()

c = Counter (words)

c.most_common(10)
```

Results

```
[('the', 64014),
('and', 51313),
('of', 34634),
('to', 13567),
 ('that', 12784),
('in', 12503),
 ('he', 10261),
 ('shall', 9838),
('unto', 8987),
 ('for', 8810)]
```

Algorithms for text

- Bag of Words Model
- Bag of n-grams Model
- Document similarity
- Topic Models

Advanced Feature Engineering

- Word2Vec Model
- The GloVe Model
- The FastText Model

Text Classification

```
import pandas as pd
import numpy as np
import re
import nltk
import matplotlib.pyplot as plt
pd.options.display.max_colwidth = 200
%matplotlib inline
```

```
corpus = [['The sky is blue and beautiful.',
        'Love this blue and beautiful sky!',
        'The quick brown fox jumps over the lazy dog',
        'I love eggs, ham, sausage and bacon',
        'The brown fox is quick and the blue dog is lazy',
        'The sky is very blue and the sky is beautiful today',
        'The dog is lazy but the fox is quick‘
]
```

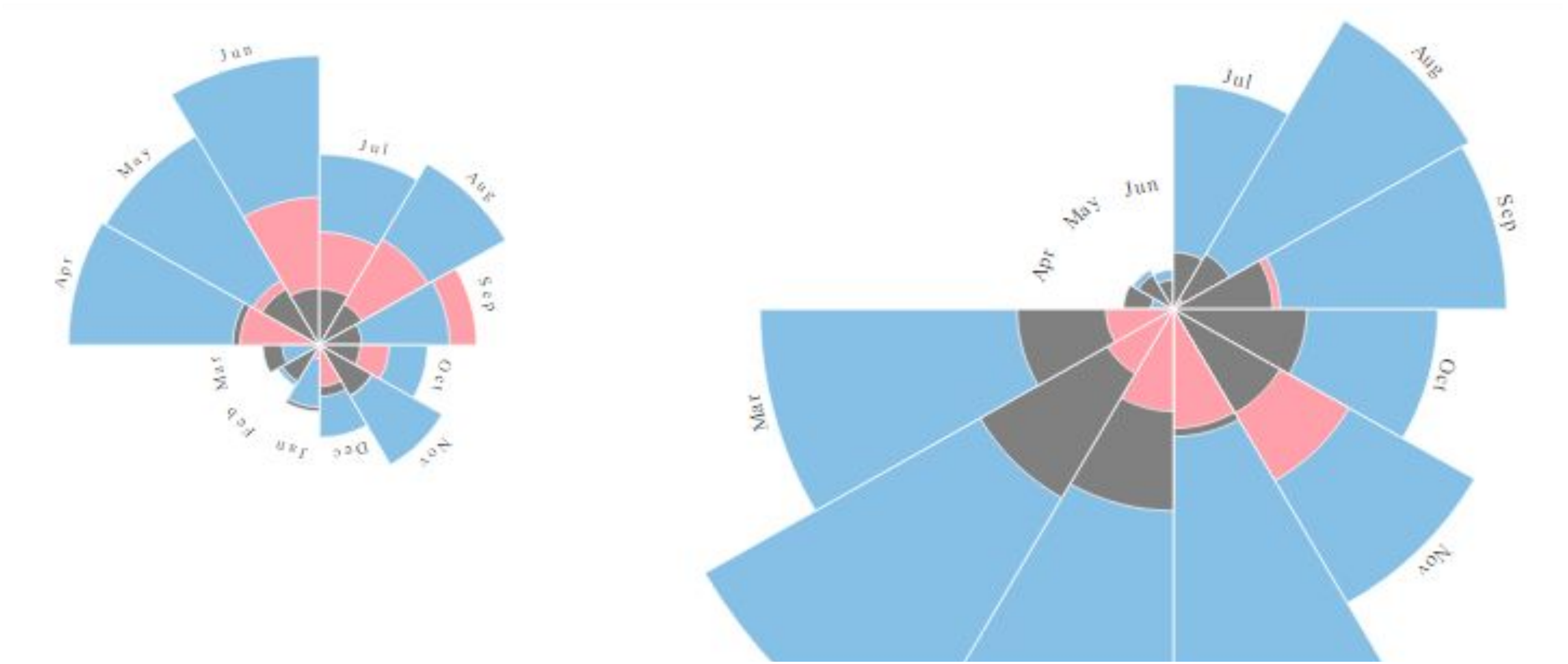
Labelling

```
labels = ['weather', 'weather', 'animals',  'food' ,
'animals',  'weather' , 'animals']

corpus = np.array(corpus)

corpus_df = pd>DataFrame({‘Document’: corpus, ‘Category’:
labels})

corpus_df = corpus_df[[‘Document’, ‘Category]]
corpus_df
```

Results

Document	Category	
0	The sky is blue and beautiful.	weather
1	Love this blue and beautiful sky!	weather
2	The quick brown fox jumps over the lazy dog	animals
3	I love eggs, ham, sausage and bacon	food
4	The brown fox is quick and the blue dog is lazy	animals
5	The sky is very blue and the sky is beautiful today	weather
6	The dog is lazy but the fox is quick	animals

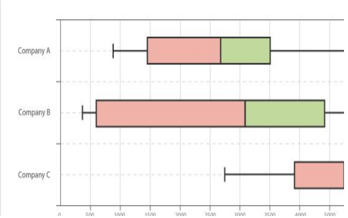
Data Visualization



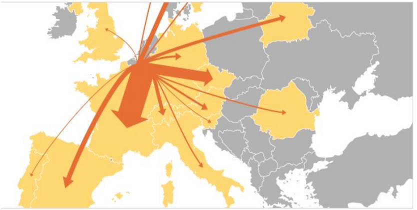
What kinds of Data Visualizations are there?

Data visualizations can be maps, plots,
diagrams and graphs. Instead of reading densely written reports, we can use visualizations to see patterns or trends in data.

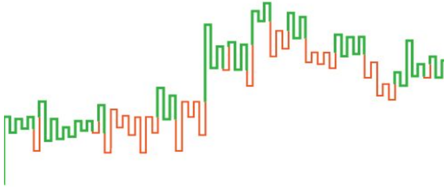
PLOT



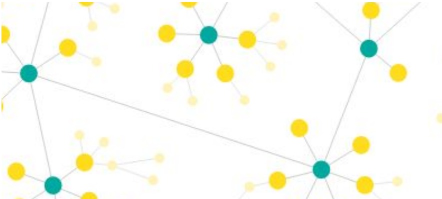
MAP



CHART



DIAGRAM



Selecting a visualization type

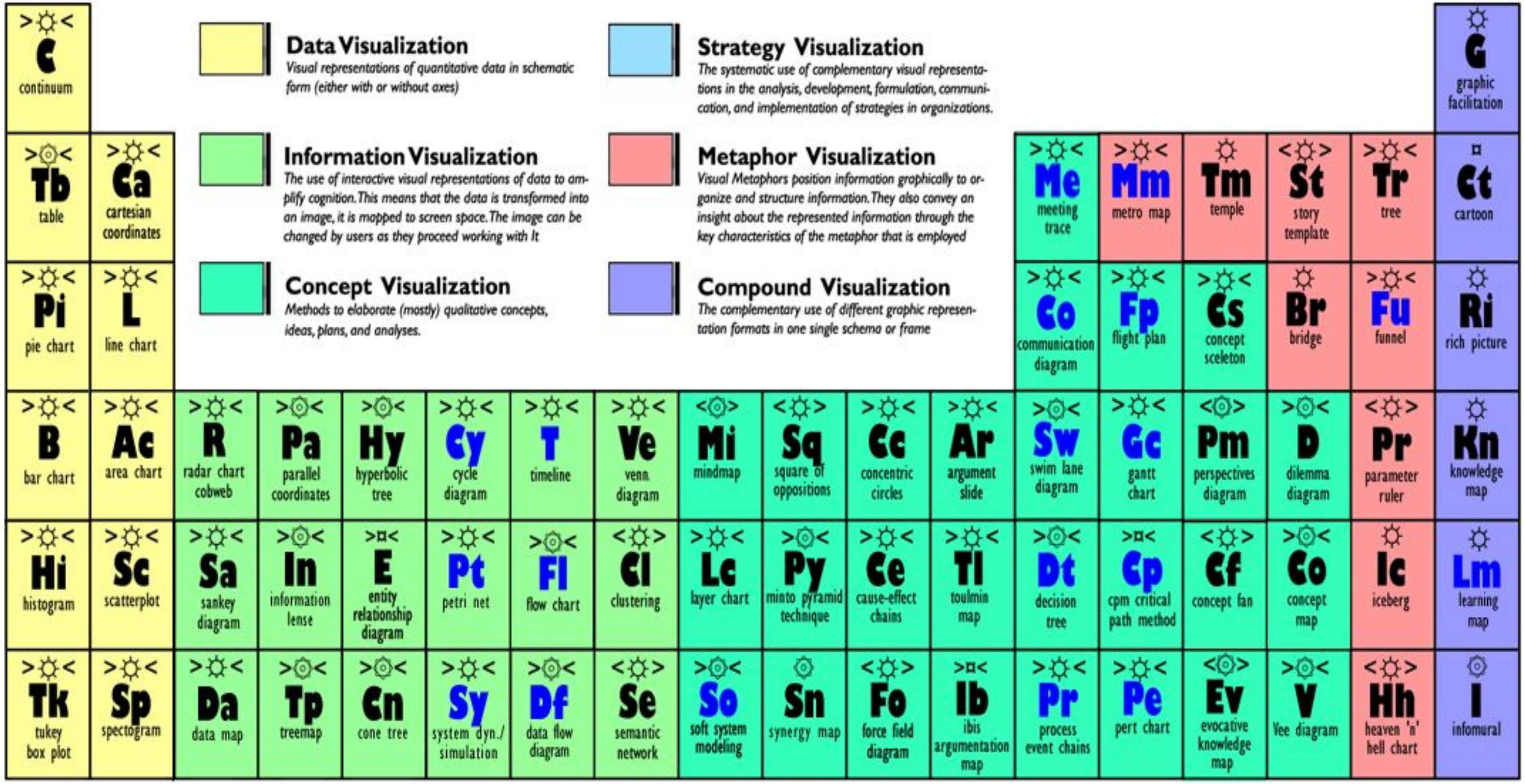
What do you want to find?

https://datavizcatalogue.com/

The data viz catalogue is a great interactive resource that can be used to discover which type of visualization suits which function best.

A PERIODIC TABLE OF VISUALIZATION METHODS

INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION



🟨 **Data Visualization**
Visual representations of quantitative data in schematic form (either with or without axes)

🟩 **Information Visualization**
The use of interactive visual representations of data to amplify cognition. This means that the data is transformed into an image, it is mapped to screen space. The image can be changed by users as they proceed working with It

🟩 **Concept Visualization**
Methods to elaborate (mostly) qualitative concepts, ideas, plans, and analyses.

🟦 **Strategy Visualization**
The systematic use of complementary visual representations in the analysis, development, formulation, communication, and implementation of strategies in organizations.

🟥 **Metaphor Visualization**
Visual Metaphors position information graphically to organize and structure information. They also convey an insight about the represented information through the key characteristics of the metaphor that is employed

🟪 **Compound Visualization**
The complementary use of different graphic representation formats in one single schema or frame

>☼< C continuum																	☼ G graphic facilitation
>⚙< Tb table	>☼< Ca cartesian coordinates											>☼< Me meeting trace	>☼< Mm metro map	☼ Tm temple	<☼> St story template	>☼< Tr tree	¤ Ct cartoon
>☼< Pi pie chart	>☼< L line chart											>☼< Co communication diagram	>☼< Fp flight plan	>☼< Cs concept sceleton	☼ Br bridge	>☼< Fu funnel	☼ Ri rich picture
>☼< B bar chart	>☼< Ac area chart	>☼< R radar chart cobweb	>⚙< Pa parallel coordinates	>⚙< Hy hyperbolic tree	>☼< Cy cycle diagram	>☼< T timeline	>☼< Ve venn diagram	<⚙> Mi mindmap	<☼> Sq square of oppositions	>☼< Cc concentric circles	>☼< Ar argument slide	>⚙< Sw swim lane diagram	>☼< Gc gantt chart	<⚙> Pm perspectives diagram	>⚙< D dilemma diagram	<☼> Pr parameter ruler	☼ Kn knowledge map
>☼< Hi histogram	>☼< Sc scatterplot	>☼< Sa sankey diagram	>⚙< In information lense	>¤< E entity relationship diagram	>☼< Pt petri net	>⚙< Fl flow chart	<☼> Cl clustering	>☼< Lc layer chart	>⚙< Py minto pyramid technique	>☼< Ce cause-effect chains	>☼< Tl toulmin map	>⚙< Dt decision tree	>¤< Cp cpm critical path method	<☼> Cf concept fan	>⚙< Co concept map	☼ Ic iceberg	☼ Lm learning map
>☼< Tk tukey box plot	>☼< Sp spectogram	>☼< Da data map	>⚙< Tp treemap	>⚙< Cn cone tree	>☼< Sy system dyn./ simulation	>⚙< Df data flow diagram	<☼> Se semantic network	>⚙< So soft system modeling	⚙ Sn synergy map	<☼> Fo force field diagram	>¤< Ib ibis argumentation map	>☼< Pr process event chains	>☼< Pe pert chart	<⚙> Ev evocative knowledge map	>⚙< V Vee diagram	<☼> Hh heaven 'n' hell chart	⚙ I infomural

https://www.visual-literacy.org/periodic_table/periodic_table.html

Functions

Comparisons Proportions Relationships

Part-to-a-whole Processes & methods

Distribution How things work Range

Patterns Locations Concepts Analysing Text

Movement or flow Data over time

Creating visualizations

select

process

mine

visualize

We select the data then process it.

We identify what we want to do with it - group the content by theme or topic, analyse the content for features of language for example. Once we know what we are looking for, we can select a classifier to classify the data accordingly.

We did this at the beginning when we grouped our sentences together into topics (food, animals, weather)

Mining

Mining methods place the data into a context that enables it to be visualized

Methods include sequences analysis, classifications, path analysis and clustering

Clustering algorithms

- **Flat clustering** (creates a set of clusters without any explicit structure that would relate clusters to each other; It's also called exclusive clustering)
- **Hierarchical clustering** (Creates a hierarchy of clusters)
- **Hard clustering** (Assigns each document/object as a member of exactly one cluster)
- **Soft clustering** (Distribute the document/object over all clusters)

https://www.codeproject.com/Articles/439890/Text-Documents-Clustering-using-K-Means-Algorithm

Algorithms

Agglomerative (Hierarchical clustering)

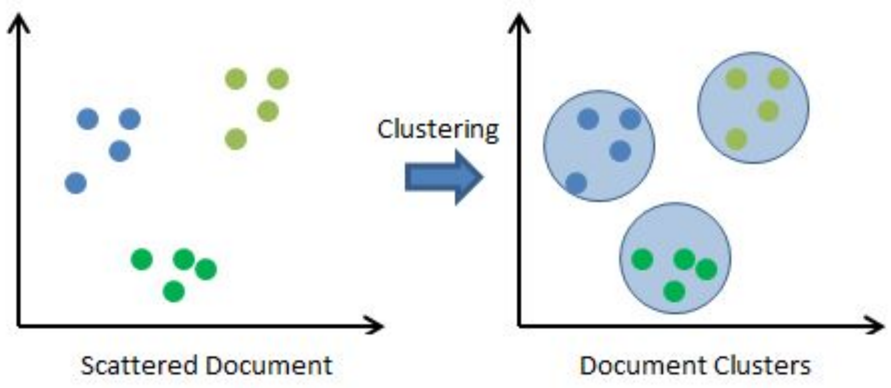
K-Means (Flat clustering, Hard clustering)

EM Algorithm (Flat clustering, Soft clustering)

https://www.codeproject.com/Articles/439890/Text-Documents-Clustering-using-K-Means-Algorithm

Clustering (unsupervised)

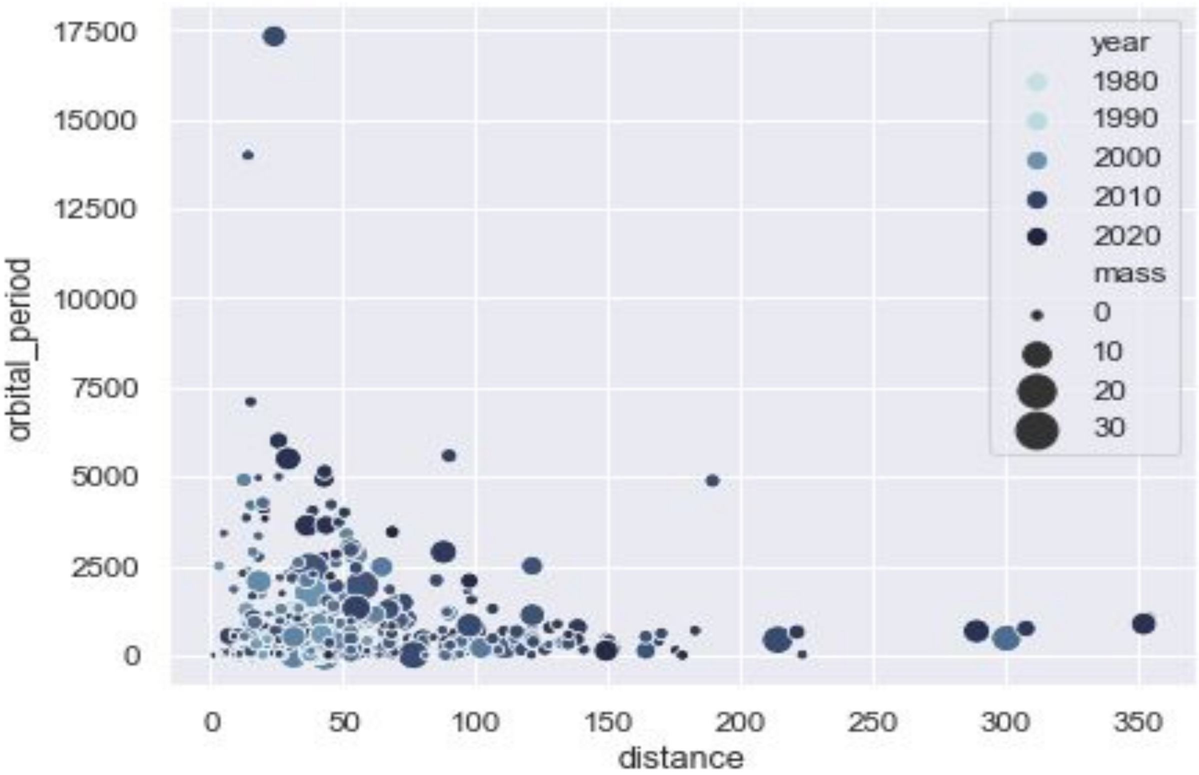
finding a *structure* in a collection of unlabeled data. The aim is to organize the data into groups based on common features or similarities.



Scatterplot

```
import seaborn as sns
sns.set()

# Load the example planets dataset
planets = sns.load_dataset("planets")

cmap = sns.cubehelix_palette(rot=-.2, as_cmap=True)
ax = sns.scatterplot(x="distance", y="orbital_period",
                     hue="year", size="mass",
                     palette=cmap, sizes=(10, 200),
                     data=planets)
```

Scatterplot



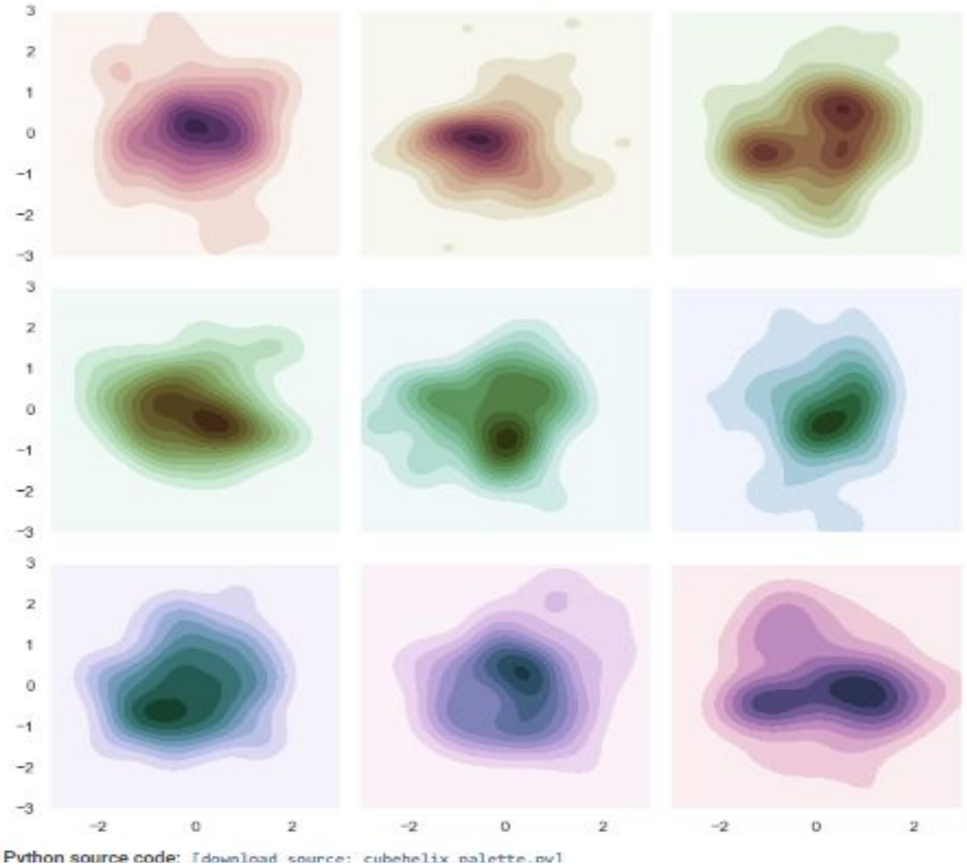
kdeplot

```
import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt


sns.set(style="dark")
rs = np.random.RandomState(50)

# Set up the matplotlib figure

f, axes = plt.subplots(3, 3, figsize=(9, 9), sharex=True, sharey=True)

# Rotate the starting point around the cubehelix hue circle
for ax, s in zip(axes.flat, np.linspace(0, 3, 10)):
```

kdeplot

```
# Create a cubehelix colormap to use with kdeplot
    cmap = sns.cubehelix_palette(start=s, light=1, as_cmap=True)

    # Generate and plot a random bivariate dataset
    x, y = rs.randn(2, 50)

  sns.kdeplot(x, y, cmap=cmap, shade=True, cut=5, ax=ax)
    ax.set(xlim=(-3, 3), ylim=(-3, 3))
f.tight_layout()
```

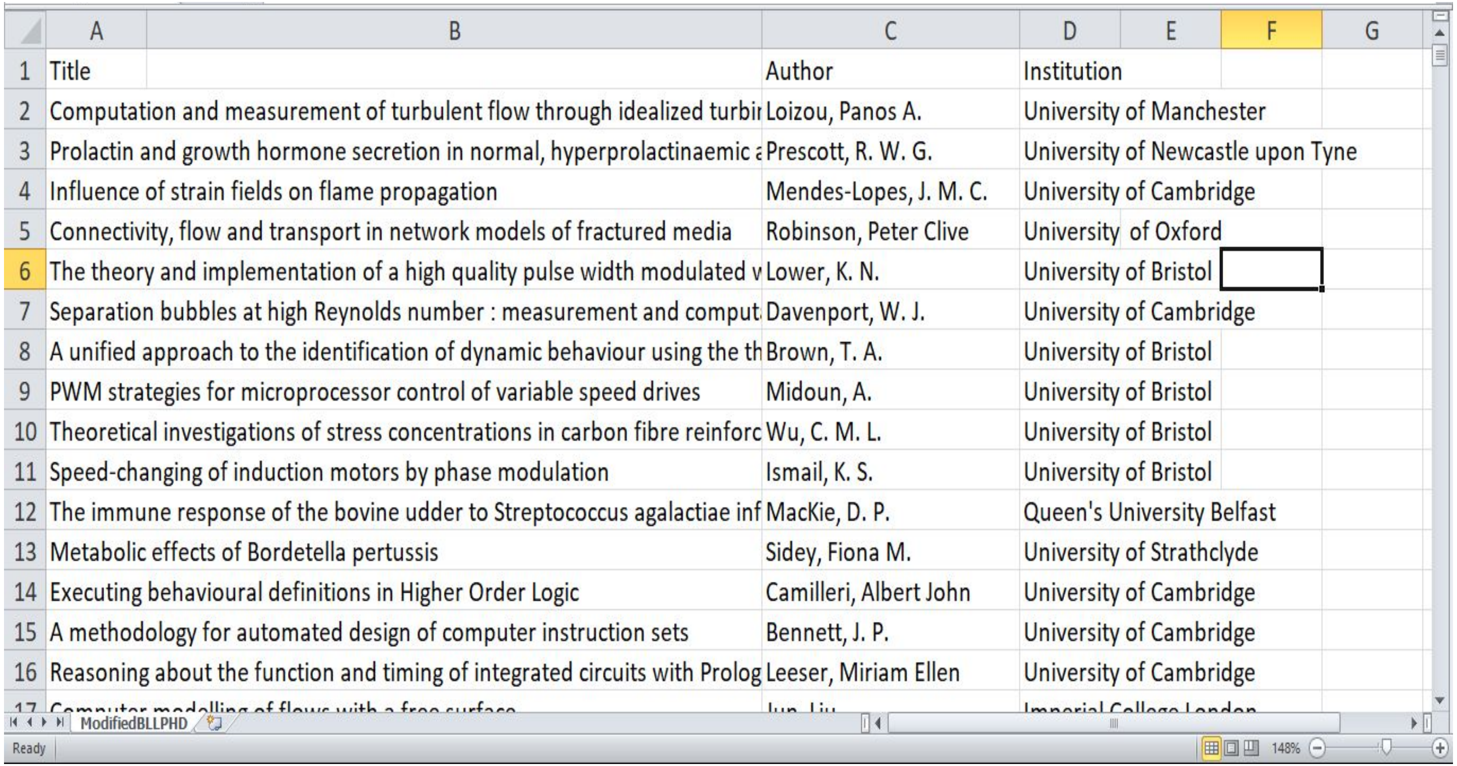
kdeplot



Python source code: [download source: cubehelix_palette.py]

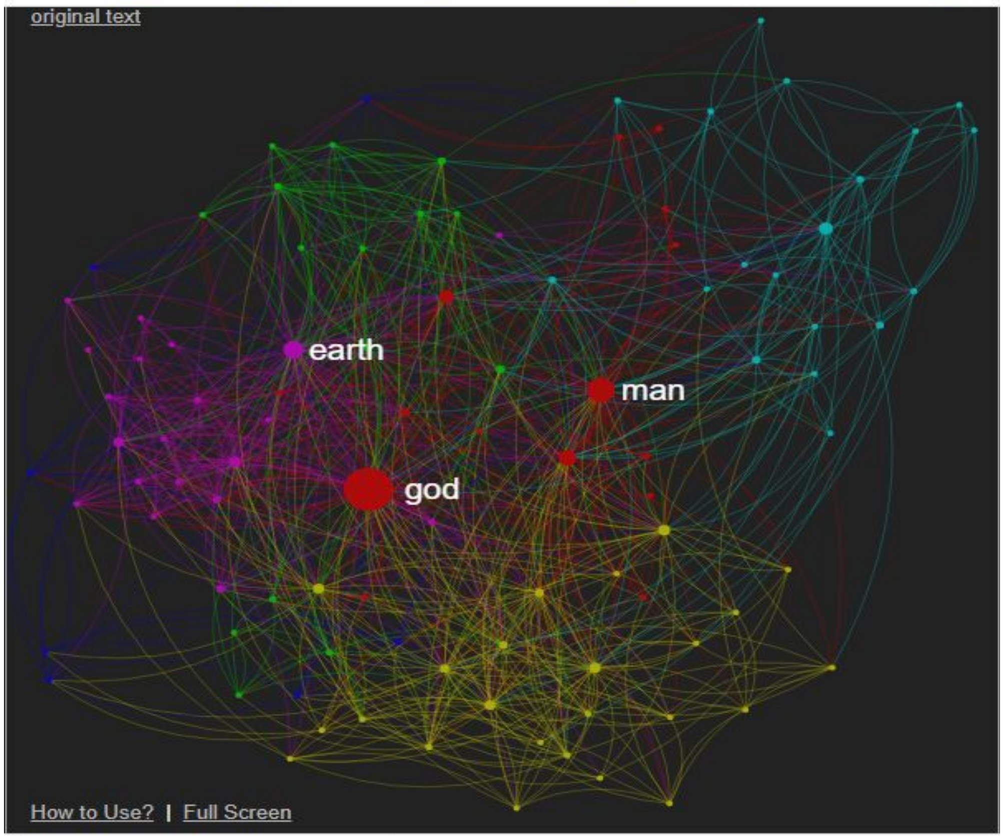
Dataset BLL Theses

https://bl.iro.bl.uk/work/86c21604-10d3-4367-a131-fb19a259ce07

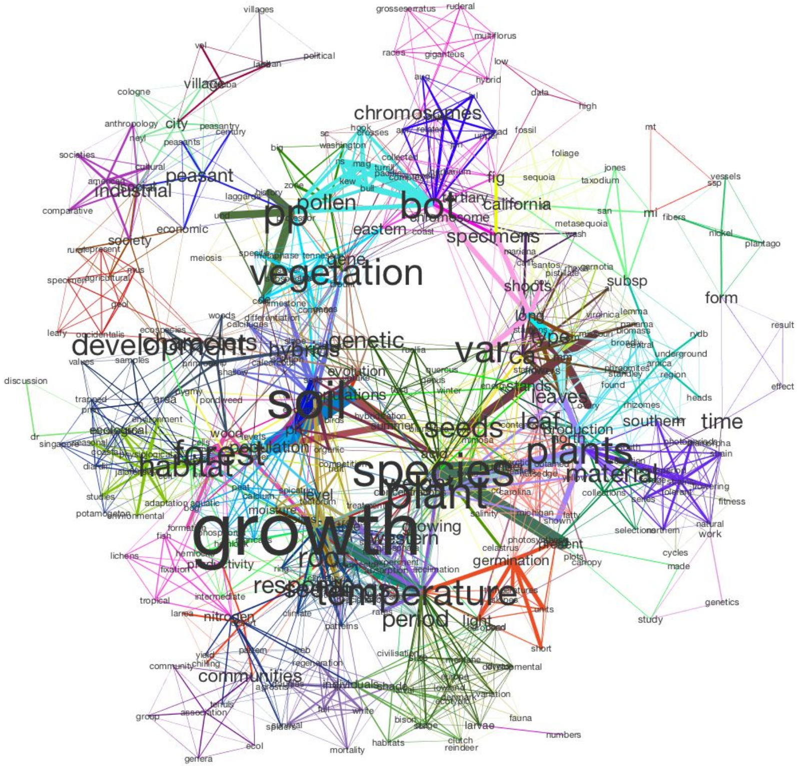


	A	B	C	D	E	F	G
1	Title		Author	Institution			
2	Computation and measurement of turbulent flow through idealized turbi		Loizou, Panos A.	University of Manchester			
3	Prolactin and growth hormone secretion in normal, hyperprolactinaemic a		Prescott, R. W. G.	University of Newcastle upon Tyne			
4	Influence of strain fields on flame propagation		Mendes-Lopes, J. M. C.	University of Cambridge			
5	Connectivity, flow and transport in network models of fractured media		Robinson, Peter Clive	University of Oxford			
6	The theory and implementation of a high quality pulse width modulated v		Lower, K. N.	University of Bristol			
7	Separation bubbles at high Reynolds number : measurement and comput		Davenport, W. J.	University of Cambridge			
8	A unified approach to the identification of dynamic behaviour using the th		Brown, T. A.	University of Bristol			
9	PWM strategies for microprocessor control of variable speed drives		Midoun, A.	University of Bristol			
10	Theoretical investigations of stress concentrations in carbon fibre reinforc		Wu, C. M. L.	University of Bristol			
11	Speed-changing of induction motors by phase modulation		Ismail, K. S.	University of Bristol			
12	The immune response of the bovine udder to Streptococcus agalactiae inf		MacKie, D. P.	Queen's University Belfast			
13	Metabolic effects of Bordetella pertussis		Sidey, Fiona M.	University of Strathclyde			
14	Executing behavioural definitions in Higher Order Logic		Camilleri, Albert John	University of Cambridge			
15	A methodology for automated design of computer instruction sets		Bennett, J. P.	University of Cambridge			
16	Reasoning about the function and timing of integrated circuits with Prolog		Leeser, Miriam Ellen	University of Cambridge			
17	Computer modelling of flows with a free surface		Jun, Liu	Imperial College London			

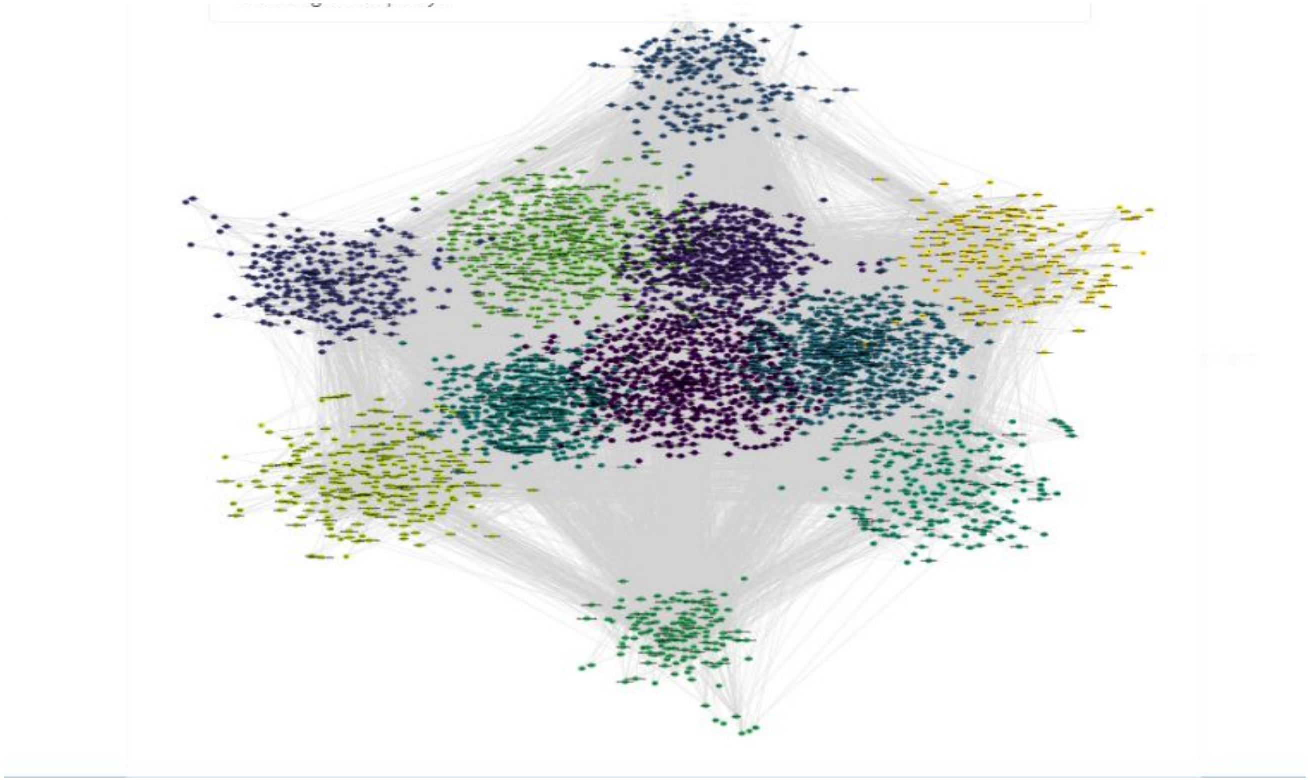
https://textexture.com



https://pythonhosted.org/tethne/tutorial.mallet.html



[https://towardsdatascience.com/getting-started-with-graph-analysis-in-pytho
etworkx-5e2d2f82f18e](https://towardsdatascience.com/getting-started-with-graph-analysis-in-pytho)



Regression

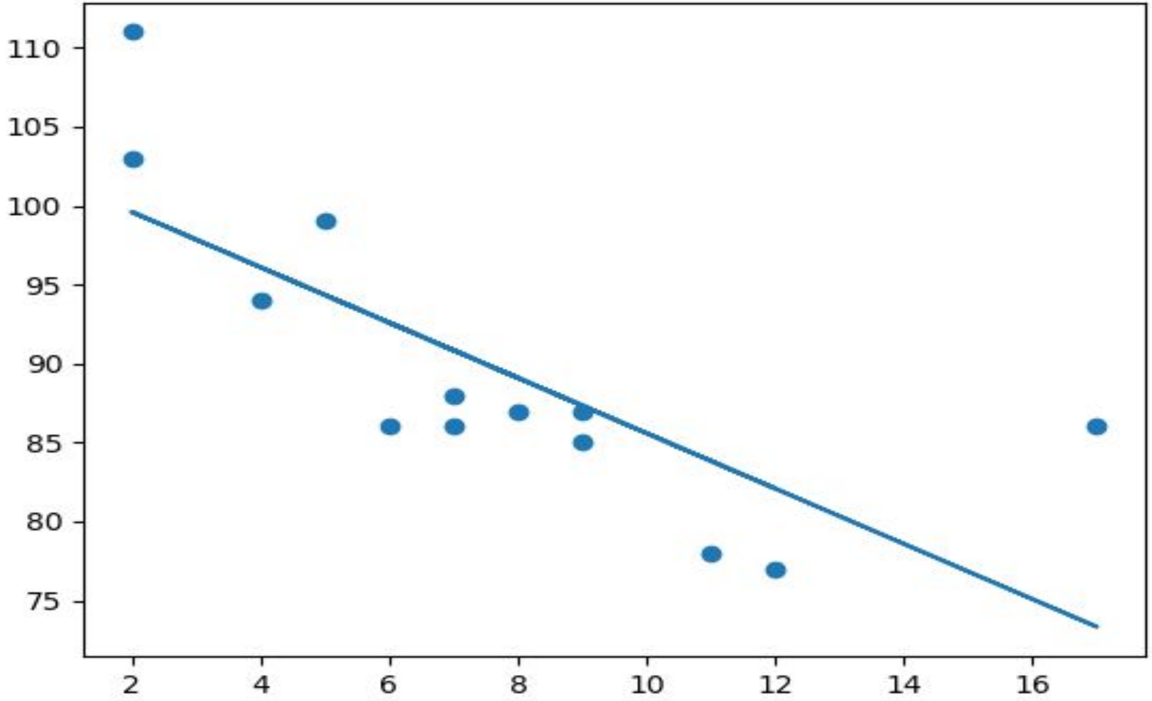
The term regression is used when you try to find the relationship between variables.

In Machine Learning, and in statistical modeling, that relationship is used to predict the outcome of future events.

Linear Regression https://www.w3schools.com/python/python_ml_polynomial_regression.asp

Linear regression uses the relationship between the data-points to draw a straight line through all them. This line can be used to predict future values.

Linear Regression



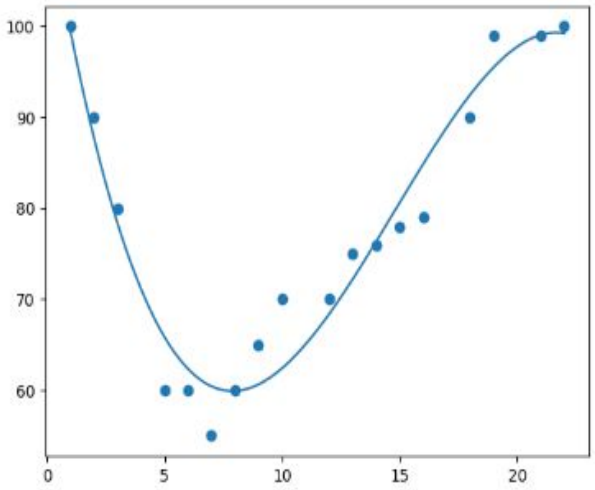
https://www.w3schools.com/python/python_ml_linear_regression.asp

Polynomial Regression

If your data points clearly will not fit a linear regression (a straight line through all data points), it might be ideal for polynomial regression.

Polynomial regression, like linear regression, uses the relationship between the variables x and y to find the best way to draw a line through the data points.

https://www.w3schools.com/python/python_ml_polynomial_regression.asp



Links & Tools

Machine Learning

- www.python.org

Visualization

- https://seaborn.pydata.org/

Excellent resources to explore

https://github.com/brianspiering/awesome-dl4nlp

https://datavizcatalogue.com/

www.tableau.com

https://densitydesign.org/

https://www.flickr.com/photos/densitydesign/sets/72157628222445801/with/6431913399/

https://www.flickr.com/photos/densitydesign/sets/72157624141332939/

https://densitydesign.org/research/minerva/

Stack Overflow

https://stackoverflow.com/questions/tagged/python

Tableau https://www.tableau.com/learn/articles/data-visualization

ttps://www.elsevier.com/connect/story/research-matters/research-data/a-5-step-guide-to-data-visualization

Links

www.iskouk.org
https://twitter.com/ISKOUK
https://www.linkedin.com/groups/2079995/