This EDA tend to prepare our dataset to be processed by our algorithms.

We want the dataset to have relevant data so we will exclude everything that doesn't contains informations about agriculture or machine learning, we will discard everything that is not obviouslt english, weirds ponctuation and everything we judge bad for our algorithm performances.

```
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", force_remount=True).
```

```
import pandas as pd
df = pd.read_json('/content/gdrive/MyDrive/Exam/AgrSmall.json')
df.head()
```

	doi	titles	abstracts	authors	keywords	sources
0	10.3390	Community Faecal Management Strategies and Per...	Most people in rural areas in South Africa (SA...	[Matthew Mamera, Johan J. van Tol, Makhosazana...	agriculture	mdpi
1	10.3390	Adoption of Sustainable Agriculture Practices ...	The aim of this study was to highlight the imp...	[Rafay Waseem, Gershom Endelani Mwalupaso, Far...	agriculture	mdpi
2	10.3390	Atlanta Residents’ Knowledge Regarding Heavy M...	Urban agriculture and gardening provide many h...	[Lauren Balotin, Samantha Distler, Antoinette ...	agriculture	mdpi
3	10.3390	Perceptions of the Challenges and Opportunitie...	Waste management has become pertinent in urban...	[Nqubeko Neville Menyuka, Melusi Sibanda, Urmi...	agriculture	mdpi
4	10.3390	An Assessment of Seaweed Extracts: Innovation ...	Plant growth regulators (PGRs) are described i...	[El Chami Daniel, Galli Fabio]	agriculture	mdpi

```
df.describe()
```

	doi	titles	abstracts	authors	keywords	sources
count	4531	4531	4531	4531	4531	4531
unique	169	4011	3879	3761	33	419
top	10.1016/	ombining computer vision and deep learning to ...	No abstract is available for this item.	[]	agronomy	arxiv
freq	1293	5	16	145	321	652

```
docs = list(df.loc[:, "abstracts"].values)
```

```
docs
```

```
['Most people in rural areas in South Africa (SA) rely on untreated drinking groundwater sources and pit latrine sanitations. A minimum basic sanitation facility should enable safe and
 'The aim of this study was to highlight the importance of socioeconomic and psychosocial factors in the adoption of sustainable agricultural practices (SAPs) in banana farm production.
 'Urban agriculture and gardening provide many health benefits, but the soil is sometimes at risk of heavy metal and metalloid (HMM) contamination. HMM, such as lead and arsenic, can re
 'Waste management has become pertinent in urban regions, along with rapid population growth. The current ways of managing waste, such as refuse collection and recycling, are failing to
 'Plant growth regulators (PGRs) are described in the literature as having a significant role in securing crop management of modern agriculture in conditions of abiotic and biotic stres
 'Dichlorvos is a toxic organophosphate insecticide that is used in agriculture and other insecticide applications. Dermal uptake is a known exposure route for dichlorvos and chemical p
 'Desert plants are able to survive under harsh environmental stresses inherent to arid and semiarid regions due to their association with bacterial endophytes. However, the identity, f
 'Species from the crested wheatgrass (Agropyron spp.) complex have been widely used for revegetation and grazing on North American rangelands for over 100 years. Focused crested wheatg
```

```
'The epidemiological dynamics followed by viruses in protected horticultural crops in the Mediterranean Arc of Spain has evolved from a majority of those transmitted by aphids to the p
'Steep slopes are the main cause of rollover incidents in agriculture. Targeted safety signs have been developed to warn machinery operators against risky slopes. However, machinery us
'Autonomous mowers are becoming increasingly common in public and private greenspaces. Autonomous mowers can provide several advantages since these machines help to save time and energ
'The use of satellites to monitor crops and support their management is gathering increasing attention. The improved temporal, spatial, and spectral resolution of the European Space Ag
'Flood recession farming is as an important supplement to rainfed agriculture in West Africa. Every year, large areas are flooded along riverbanks and temporary lakes. When water reced
'Extension services play a crucial role by improving skills and access to information that result in greater farm level innovations, especially on family farms which are the predominan
'Soil degradation is the greatest threat to agricultural production globally. The practice of applying or retaining crop residues in the field as mulch is imperative to prevent soil er
'While climate change threatens global food security, health, and nutrition outcomes, Africa is more vulnerable because its economies largely depend on rain-fed agriculture. Thus, ther
'The development of rural infrastructure plays an essential role in improving rural livelihoods and enhancing sustainable and environmentally friendly agricultural production. However,
'For decades, non-renewable resources have been the basis of worldwide economic development. The extraction rate of natural resources has increased by 113% since 1990, which has led to
'Agrivoltaic (agriculture–photovoltaic) or solar sharing has gained growing recognition as a promising means of integrating agriculture and solar-energy harvesting. Although this field
'Nitrogen use efficiency in modern agriculture is very low. It means that a lot of synthetic chemicals are wasted rather than utilized by crops. This can cause more problems where the 
'Greenhouse farming is an agricultural management system that has demonstrated its efficiency in intensifying food production. These systems constitute a feasible alternative for ensur
'Internet of Things (IoT) provides a diverse platform to automate things where smart agriculture is one of the most promising concepts in the field of Internet of Agriculture Things (I
'This paper considers the evolution of processes applied in agriculture for field operations developed from non-organized handmade activities into very specialized and organized produc
'Part-time farming has been suggested by scholars to play an important part in farmers’ decision making, but seldom empirical evidence has been done on the field of conservation agricu
'The role of agriculture in environmental degradation and climate change has been at the center of a long-lasting and controversial debate. This situation combined with the expected gr
'Many Philippine species are at risk of extinction because of habitat loss and degradation driven by agricultural land use and land-use change. The Philippines is one of the world’s pr
'The conventional tillage based rice-wheat system (RWS) in Indo-genetic plains (IGP) of South Asia is facing diverse challenges like increase in production cost and erratic climatic ev
'Climate change and food security are critical topics in sustainable agricultural development. The climate-smart agriculture initiative proposed by the Food and Agriculture Organizatio
'The negative impact of agriculture on the natural environment is not a new issue. One of the ideas to overcome this problem is the eco-efficiency concept, analyzing the agricultural o
'Environmental costs should be taken into account when measuring the achievements of China’s agricultural development, since the long-term extensive development of agriculture has caus
'The information that crops offer is turned into profitable decisions only when efficiently managed. Current advances in data management are making Smart Farming grow exponentially as 
'This paper aimed at evaluating the level of sustainability in agriculture in 28 member states of the European Union. The surveys were carried out based on a synthetic technique for or
'The aim of this paper was to analyze the main factors that affect green consumers’ choice regarding the purchase of organic agriculture products. The data collected through a survey o
'Several studies address the topic of Information and Communication Technologies (ICT) adoption in irrigated agriculture. Many of these studies testify on the growing importance of ICT
'Games are particularly relevant for field research in agriculture, where alternative experimental designs can be costly and unfeasible. Games are also popular for non-experimental pur
'Techniques such as intercropping and minimum tillage improve soil quality, including soil microbial activity, which stimulates the efficient use of soil resources by plants. However, 
'There are three main contradictions associated with urbanization: population growth and food demand, urban sprawl and production space, and production patterns and energy consumption.
'Economic, environmental and social sustainability is increasingly gaining the attention of academia and commitment in the policies of national economies. Global warming and climate ch
'Curbing emissions from agriculture, and especially from livestock production, is essential in order to fulfil the Paris Agreement. Shifting to a diet lower in meat consumption has bee
'In recent years, social and economic goals have been preferable compared to environmental issues. However, global problems with the environment, increasing pollution, and gas heating 
'The Kingdom of Tonga has one of the highest rates of diet-related non-communicable diseases (NCDs) in the world. Initiatives to promote pro-health dietary behaviour are possibly being
'The use of intensive high-yield agricultural systems has proved to be a feasible alternative to traditional systems as they able to meet the objective of guaranteeing long-term sustai
'Sustainability has been an emerging issue for years in the economy and agriculture. Making agriculture sustainable has become so essential that it has become part of the Common Agricu
'The thirty journal articles dealing with the relationship between climate change and agriculture (the latter is treated in general, i.e., as an industry) and which have gained >1000 c
'The objective of this paper is to study the impact of using micro-grid solar photovoltaic (PV) systems in rural areas in the West Bank, Palestine. These systems may have the potential
'Global warming is an unanimously accepted phenomenon by the international scientific community, being already highlighted by the analysis of observational data over long periods of ti
'Modern agriculture increasingly demands an alternative to synthetic chemicals (fertilizers and pesticides) in order to respond to the changes in international law and regulations, but
'It is crucial to actively encourage the development of agriculture green technology, which has been regarded as one of the most effective solutions to the environmental degradation ca
'This research paper focuses on providing an algorithm by which (Unmanned Aerial Vehicles) UAVs can be used to provide optimal routes for agricultural applications such as, fertilizers
'In this work, a novel, dynamic sustainability assessment tool is presented and validated in a case study. This tool combines two methods—system dynamics (SD) and temporal soil carbon 
'Luoshan Organic Agriculture Village was the first organic agriculture village in Taiwan, and it focuses on organic farming and cultivation. The village is developed through community 
'Environment, biodiversity and ecosystem services are essential to ensure food security and nutrition. Managing natural resources and mainstreaming biodiversity across agriculture sect
'Risk management in agriculture is at the heart of major reforms in many OECD countries and European agricultural policies. Price risks, which are generally not insurable per se, have 
'The objective of this work is to study the effects of traditional land uses (vineyard, cropland, and olive orchard) on soil properties, overland flow, and sediment loss in the Istria 
'The application of new technologies in precision agriculture offers the possibility to link information to very specific crop locations. The spatial representation of these agricultur
'Combining agriculture with behaviour change communication and other nutrition-sensitive interventions could improve feeding practices to reduce maternal and child undernutrition. Such
'Conservation agriculture, characterized by minimal tillage, permanent soil cover and crop diversification, has been widely adapted under rainfed conditions, but adoption under irrigat
'Organic farming systems are considered not compatible with conservation tillage mainly because of the reliance of conservative systems on herbicides. In this three-year field experime
```

Clearly some text have a very bad beggining like @@@@Highlights •

Ponctuation

```
import re
import string
print(string.punctuation)
```

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

```
def remove_punctuation(text):
    text = ''.join([char for char in text if char not in string.punctuation])
    # remove numerical values
    text = re.sub(r'\d+', '', text)
    # substitute multiple whitespace with single whitespace and remove leading and trailing whitespaces
    text = re.sub('\s+', ' ', text).strip()
    return text


df['text_clean'] = df['abstracts'].apply(lambda x: remove_punctuation(x))
df
```

	doi	titles	abstracts	authors	keywords	sources	text_clean
0	10.3390	Community Faecal Management Strategies and Per...	Most people in rural areas in South Africa (SA...	[Matthew Mamera, Johan J. van Tol, Makhosazana...	agriculture	mdpi	Most people in rural areas in South Africa SA ...
1	10.3390	Adoption of Sustainable Agriculture Practices ...	The aim of this study was to highlight the imp...	[Rafay Waseem, Gershom Endelani Mwalupaso, Far...	agriculture	mdpi	The aim of this study was to highlight the imp...
2	10.3390	Atlanta Residents' Knowledge Regarding Heavy M...	Urban agriculture and gardening provide many h...	[Lauren Balotin, Samantha Distler, Antoinette ...	agriculture	mdpi	Urban agriculture and gardening provide many h...
3	10.3390	Perceptions of the Challenges and Opportunitie...	Waste management has become pertinent in urban...	[Nqubeko Neville Menyuka, Melusi Sibanda, Urmi...	agriculture	mdpi	Waste management has become pertinent in urban...
4	10.3390	An Assessment of Seaweed Extracts: Innovation ...	Plant growth regulators (PGRs) are described i...	[El Chami Daniel, Galli Fabio]	agriculture	mdpi	Plant growth regulators PGRs are described in ...
...	...	...	...	...	...	...	...
4526	10.2478/	1. Modelling groundwater flow and nitrate tran...	The present paper discusses studies related to...	[Sieczka Anna, Bujakowski Filip, Koda Eugeniusz]	precision agriculture	Sciendo, 2018.	The present paper discusses studies related to...
4527	unknown	2. Cosechando los beneficios de la agricultura...	El objetivo del trabajo fue desarrollar una me...	[Bonilla, Camila, Terra, José A, Gutiérrez, Lu...	precision agriculture	Facultad de Agronomía - Instituto Nacional de ...	El objetivo del trabajo fue desarrollar una me...
4528	unknown	6. A Risk Analysis of Precision Agriculture Te...	Precision agriculture technology can transform...	[Yangxuan Liu, Michael R. Langemeier, Ian M. S...	precision agriculture	MDPI, Open Access Journal, 2018.	Precision agriculture technology can transform...
4529	10.1016/	7. Integrated open geospatial web service enab...	Highlights •We proposed an integrated geospati...	[Chen, Nengcheng, Zhang, Xiang, Wang, Chao]	precision agriculture	Elsevier B.V.	Highlights •We proposed an integrated geospati...
4530	10.5194/	9. UNMANNED AERIAL VEHICLE (UAV) DERIVED NORMA...	Malaysia currently is one of the biggest globa...	[S. A. Suab, M. S. Syukur, R. Avtar, A. Korom]	precision agriculture	Copernicus Publications, 2019.	Malaysia currently is one of the biggest globa...

4531 rows × 7 columns

Obvious words that should not be here

```
stop_words = ["Highlights", "Resumen", "Abstract", "Introduction", "©",'Keywords','•','INTRODUCTION','Background']
```

```
print(stop_words)
```

```
['Highlights', 'Resumen', 'Abstract', 'Introduction', '©', 'Keywords', '•', 'INTRODUCTION', 'Background']
```

```
def remove_stopwords(x):
  words = x.split(" ")
  temp = []
  for word in words:
    if word not in stop_words:
      temp.append(word)
  return " ".join(temp)
```

```
df['text_clean'] = df['text_clean'].apply(lambda x: remove_stopwords(x))
df
```

	doi	titles	abstracts	authors	keywords	sources	text_clean
0	10.3390	Community Faecal Management Strategies and Per...	Most people in rural areas in South Africa (SA...	[Matthew Mamera, Johan J. van Tol, Makhosazana...	agriculture	mdpi	Most people in rural areas in South Africa SA ...
1	10.3390	Adoption of Sustainable Agriculture Practices ...	The aim of this study was to highlight the imp...	[Rafay Waseem, Gershom Endelani Mwalupaso, Far...	agriculture	mdpi	The aim of this study was to highlight the imp...
2	10.3390	Atlanta Residents’ Knowledge Regarding Heavy M...	Urban agriculture and gardening provide many h...	[Lauren Balotin, Samantha Distler, Antoinette ...	agriculture	mdpi	Urban agriculture and gardening provide many h...
3	10.3390	Perceptions of the Challenges and Opportunitie...	Waste management has become pertinent in urban...	[Nqubeko Neville Menyuka, Melusi Sibanda, Urmi...	agriculture	mdpi	Waste management has become pertinent in urban...
4	10.3390	An Assessment of Seaweed Extracts: Innovation ...	Plant growth regulators (PGRs) are described i...	[El Chami Daniel, Galli Fabio]	agriculture	mdpi	Plant growth regulators PGRs are described in ...
...	...	...	...	...	...	...	...
4526	10.2478/	1. Modelling groundwater flow and nitrate tran...	The present paper discusses studies related to...	[Sieczka Anna, Bujakowski Filip, Koda Eugeniusz]	precision agriculture	Sciendo, 2018.	The present paper discusses studies related to...
4527	unknown	2. Cosechando los beneficios de la agricultura...	El objetivo del trabajo fue desarrollar una me...	[Bonilla, Camila, Terra, José A, Gutiérrez, Lu...	precision agriculture	Facultad de Agronomía - Instituto Nacional de ...	El objetivo del trabajo fue desarrollar una me...
4528	unknown	6. A Risk Analysis of Precision Agriculture Te...	Precision agriculture technology can transform...	[Yangxuan Liu, Michael R. Langemeier, Ian M. S...	precision agriculture	MDPI, Open Access Journal, 2018.	Precision agriculture technology can transform...
4529	10.1016/	7. Integrated open geospatial web service enab...	Highlights •We proposed an integrated geospati...	[Chen, Nengcheng, Zhang, Xiang, Wang, Chao]	precision agriculture	Elsevier B.V.	•We proposed an integrated geospatial service ...
4530	10.5194/	9. UNMANNED AERIAL VEHICLE (UAV) DERIVED NORMA...	Malaysia currently is one of the biggest globa...	[S. A. Suab, M. S. Syukur, R. Avtar, A. Korom]	precision agriculture	Copernicus Publications, 2019.	Malaysia currently is one of the biggest globa...

4531 rows × 7 columns

```
df = df[~df['text_clean'].astype(str).str.startswith('Note In lieu')]
```

Russian rows

```
# function to remove non-ASCII
def remove_non_ascii(text):
    return ''.join(i for i in text if ord(i)<128)

df['text_clean'] = df['text_clean'].apply(remove_non_ascii)
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

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

Here we take out the russian rows

```
df =df[df['text_clean'].map(len) > 50]
```

```
df = df[~df['text_clean'].astype(str).str.startswith('   ')]
```

These are deep adjustement as i read the dataset rapidly. Sentences that begins with that only indicate technical consideration like 'this paper is edited this date etccc'

```
df = df[~df['text_clean'].astype(str).str.startswith('The authors')]
```

Suppression of duplicate

```
df.drop_duplicates(subset=['text_clean'], keep='last')
```

	doi	titles	abstracts	authors	keywords	sources	text_clean
0	10.3390	Community Faecal Management Strategies and Per...	Most people in rural areas in South Africa (SA...	[Matthew Mamera, Johan J. van Tol, Makhosazana...	agriculture	mdpi	Most people in rural areas in South Africa SA ...
1	10.3390	Adoption of Sustainable Agriculture	The aim of this study was to	[Rafay Waseem, Gershom Endelani	agriculture	mdpi	The aim of this study was to
2	10.3390	Heavy M	provide many b	Antoinette	agriculture	mdpi	provide many b
		An Assessment of Seaweed Extracts:	Plant growth regulators (PGRs) are				Plant growth regulators PGRs are

```
docs = list(df.loc[:, "text_clean"].values)
```

Result

```
docs
```

```
'Agricultural activity is very important for every country that strives to create a stimulating stable abundant sustainable and equal business environment for all market participants B
'The advent of Internet of Things has propelled the agricultural domain through the integration of sensory devices capable of monitoring and wirelessly propagating information to produ
'The serine protease inhibitors SPIs are widely distributed in living organisms like bacteria fungi plants and humans The main function of SPIs as protease enzymes is to regulate the p
'In this paper we consider the effects of desertification in Mongolia where the area of degraded land has increased significantly in the recent decade Currently almost the entire terri
'Agriculture is highly dependent on climate change and Cyprus especially is experiencing its impacts on agricultural production to a greater extent mainly due to its geographical locat
'Tackling diffuse pollution from agriculture is a key challenge for governments seeking to implement the European Unions Water Framework Directive WFD In the research literature how be
'The presence of mycotoxins in cereal grain is a very important food safety issue with the occurrence of masked mycotoxins extensively investigated in recent years This study investiga
'Community supported agriculture CSA serves as a platform for local producers especially for small size farms to sell fresh local products directly to its members CSA is an important a
'Compared to rural agriculture urban agriculture UA has some distinct features eg the limited land access alternative growing media unique legal environments or the nonproductionrelate
'There are many factors involved in the release of CO emissions from the soil such as the type of soil management the soil organic matter the soil temperature and moisture conditions c
'In the area of plant protection and precision farming timely detection and classification of plant diseases and crop pests play crucial roles in the management and decisionmaking Rece
'Precisely measuring the work area of agriculture farm machinery is important for performing the authentication of machinery usage better allocation of resources measuring the effect o
'The agriculture and horticulture sector in the Netherlands is one of the most productive in the world Although the sector is one of the most advanced and intense agricultural producti
'Lowaltitude remote sensing RS using unmanned aerial vehicles UAVs is a powerful tool in precision agriculture PA In that context thermal RS has many potential uses The surface tempera
'There are numerous studies and publications about sustainable agriculture Many papers argue that sustainable agriculture is necessary and analyze how this goal could be achieved At th
'Precision agriculture is considered to be a fundamental approach in pursuing a lowinput highefficiency and sustainable kind of agriculture when performing sitespecific management prac
'Wireless sensor networks WSNs have demonstrated research and developmental interests in numerous fields like communication agriculture industry smart health monitoring and surveillanc
'In the last few decades a great deal has been written on the use of sustainable agriculture to improve the resilience of ecosystem services to climate change However no tangible and s
'Globalization and the related processes of land and capital concentration are also present in Polish agriculture As a result of the occurring changes in agriculture itself and in its
'In this paper a novel UGV unmanned ground vehicle for precision agriculture named Agriq is presented The Agriq has a multiple degrees of freedom positioning mechanism and it is equipp

'High rates of phosphorus P currently being applied to soils for the production of vegetables in the Mekong Delta Vietnam has led to concern regarding negative effects on the economy a
'The Internet of Things IoT concept has met requirements for security and reliability in domains like automotive industry food industry as well as precision agriculture Furthermore Sys
'In the context of climate change a nutritional transition and increased pressures to migrate internally and internationally this study examined the relationship between seasonal food
'The study of the carbon emission intensity of agricultural production is of great significance for the formulation of a rational agricultural carbon reduction policy This paper examin
'In this paper multispectral and multitemporal satellite data were used to assess the spatial and temporal evolution of the agriculture activities in the AlJouf region Kingdom of Saudi
'Unmanned aerial vehicle UAV platforms with sensors covering the rededge and nearinfrared NIR bands to measure vegetation indices VIs have been recently introduced in agriculture resea
'Communitysupported agriculture CSA is considered to be a new alternative mode for agricultural development which has developed rapidly in China and attracted the attention of scholars
'The main challenge faced by agriculture is to produce enough food for a continued increase in population however in the context of evergrowing competition for water and land climate c
'Proximal sensors in controlled environment agriculture CEA are used to monitor plant growth yield and water consumption with nondestructive technologies Rapid and continuous monitorin
'This study aimed to understand the perception of drought among farmers in order to support decisionmaking in the water allocation process This study was carried out in the Tabuleiro d
'The area of remote sensing techniques in agriculture has reached a significant degree of development and maturity with numerous journals conferences and organizations specialized in i
'This review article contributes new knowledge relating to the sustainability of antihail antiinsect and windbreak plastic nets in agriculture Based on the review biobased plastic nets
'Using native seed mixtures to create or recover grassland habitats in rotation to crops or in strips surrounding fields is considered a costeffective practice to enhance ecosystem res
'Almost one billion people in the world still do not have access to electricity Most of them live in rural areas of the developing world Access to electricity in the rural areas of Sub
'A growing awareness that highly intensified agricultural systems have made a substantial worldwide contribution to the worsening of the resilience capacity of natural ecosystems has o
'The technology development in wireless sensor network WSN offers a sustainable solution towards precision agriculture PA in greenhouses It helps to effectively use the agricultural re
'A field experiment consists of conservation agriculture CA and conventional tillage CT practices were set up in two areas Robit and Dangishta in subhumid Ethiopian highlands Irrigatio
'The relationships between tourism and agriculture have traditionally been studied due to the positive impacts they can potentially have on the development of rural economies This rese
'Endophytic fungi produce various mixtures of carbonbased compounds which are known as volatile organic compounds VOCs Research regarding the use of VOCs as pesticide substitutes has g
'The research was aimed at an overview and analysis of the demonstration activities in the Czech Republic dealing with the transfer of innovations for agricultural practice Several met
'To evaluate the ecological niche of photovoltaic agriculture in China an evaluation index system was constructed Based on the presentation form of interval numbers we used the interva
'Tobacco is a key cash crop for many farmers in Kenya although there is a variety of challenges associated with tobacco production This study seeks to understand alternatives to tobacc
'Excessive water consumption associated with regional agriculture and livestock development and rapid urbanization has caused significant stress to the ecological health and sustainabl
'The concept of circular economy whose model is based on three main pillars i design out waste and pollution ii keep products and materials in use and iii regenerate natural systems ha
```

```
'In sustainable agriculture seeking ecofriendly methods to promote plant growth and improve crop productivity is a priority Humic acid HA and plant growth promoting rhizobacteria PGPR
'Community Supported Agriculture CSA is a direct partnership between producers and a group of consumersmembers to share the risks and responsibilities of farming activities CSA aims at
'Currently the productivity of some European cropping systems is maintained artificially by increasing production factors like mineral fertilizers or pesticides in order to mask the lo
'AotearoaNew Zealand NZ is internationally renowned for picturesque landscapes and agricultural products Agricultural intensification has been economically beneficial to NZ but has imp
'The agricultural sector of Cyprus is seriously affected by climate change impacts In the framework of the ADAPTCLIMA project the available techniques and methods implemented worldwide
'In order to feed a growing global population projected to increase to billion by food production will need to increase from its current level The bulk of this growth will need to come
'Sustainable intensification practices SIPs involve a process to produce high yields for existing land without affecting the environment The significance and relevance of SIPs in a Pak
'Groundwater resources became a recognized enabler of important rural and socioeconomic development in Mediterranean countries However the development of this groundwater economy is cu
'Agriculture plays an important role for many countries It provides raw materials for food and provides large employment opportunities for people in the country especially for countrie
'Agriculture is an essential component of food security sustainable livelihoods and economic development in subSaharan Africa SSA Smallholder farmers however are restricted in the numb
'With climate change drought is expected to increase and its negative impacts will be particularly important in developing countries usually with rainfalldependent agriculture The Cabo
'Montado is an agrosilvopastoral system characterized by a high complexity as a result of the interactions between climate soil pasture trees and animals It is in this context that man
'It is estimated that at least one quarter of the worlds population will be affected by water shortages in the coming years and by there will be a global water deficit of if urgent act
'Although agriculture and aquaculture depend on access to increasingly scarce shared water resources to produce food for human consumption they are most often considered in isolation W
```

Exportation

```
df['abstracts']=df['text_clean']
df = df.drop(['text_clean'], axis=1)
df
```

	doi	titles	abstracts	authors	keywords	sources
0	10.3390	Community Faecal Management Strategies and Per...	Most people in rural areas in South Africa SA ...	[Matthew Mamera, Johan J. van Tol, Makhosazana...	agriculture	mdpi
1	10.3390	Adoption of Sustainable Agriculture Practices ...	The aim of this study was to highlight the imp...	[Rafay Waseem, Gershom Endelani Mwalupaso, Far...	agriculture	mdpi
2	10.3390	Atlanta Residents' Knowledge Regarding Heavy M...	Urban agriculture and gardening provide many h...	[Lauren Balotin, Samantha Distler, Antoinette ...	agriculture	mdpi
3	10.3390	Perceptions of the Challenges and Opportunitie...	Waste management has become pertinent in urban...	[Nqubeko Neville Menyuka, Melusi Sibanda, Urmi...	agriculture	mdpi
4	10.3390	An Assessment of Seaweed Extracts: Innovation ...	Plant growth regulators PGRs are described in ...	[El Chami Daniel, Galli Fabio]	agriculture	mdpi
...	...	...	...	...	...	...
4526	10.2478/	1. Modelling groundwater flow and nitrate tran...	The present paper discusses studies related to...	[Sieczka Anna, Bujakowski Filip, Koda Eugeniusz]	precision agriculture	Sciendo, 2018.
4527	unknown	2. Cosechando los beneficios de la agricultura...	El objetivo del trabajo fue desarrollar una me...	[Bonilla, Camila, Terra, José A, Gutiérrez, Lu...	precision agriculture	Facultad de Agronomía - Instituto Nacional de ...
4528	unknown	6. A Risk Analysis of Precision Agriculture Te...	Precision agriculture technology can transform...	[Yangxuan Liu, Michael R. Langemeier, Ian M. S...	precision agriculture	MDPI, Open Access Journal, 2018.
4529	10.1016/	7. Integrated open geospatial web service enab...	We proposed an integrated geospatial service e...	[Chen, Nengcheng, Zhang, Xiang, Wang, Chao]	precision agriculture	Elsevier B.V.
4530	10.5194/	9. UNMANNED AERIAL VEHICLE (UAV) DERIVED NORMA...	Malaysia currently is one of the biggest globa...	[S. A. Suab, M. S. Syukur, R. Avtar, A. Korom]	precision agriculture	Copernicus Publications, 2019.

4458 rows × 6 columns

```
df.to_json(path_or_buf="/content/gdrive/MyDrive/Exam/AgrSmall_cleaned.json")
```

```
import pandas as pd
df = pd.read_json('/content/gdrive/MyDrive/Exam/AgrSmall_cleaned.json')
df.head()
```

	doi	titles	abstracts	authors	keywords	sources
0	10.3390	Community Faecal Management Strategies and Per...	Most people in rural areas in South Africa SA ...	[Matthew Mamera, Johan J. van Tol, Makhosazana...	agriculture	mdpi
1	10.3390	Adoption of Sustainable Agriculture Practices ...	The aim of this study was to highlight the imp...	[Rafay Waseem, Gershom Endelani Mwalupaso, Far...	agriculture	mdpi
2	10.3390	Atlanta Residents’ Knowledge Regarding Heavy M...	Urban agriculture and gardening provide many h...	[Lauren Balotin, Samantha Distler, Antoinette ...	agriculture	mdpi
3	10.3390	Perceptions of the Challenges and Opportunitie...	Waste management has become pertinent in urban...	[Nqubeko Neville Menyuka, Melusi Sibanda, Urmi...	agriculture	mdpi
4	10.3390	An Assessment of Seaweed Extracts: Innovation ...	Plant growth regulators PGRs are described in ...	[El Chami Daniel, Galli Fabio]	agriculture	mdpi

```
docs = list(df.loc[:, "abstracts"].values)
```

```
docs
```

```
 'The work is devoted to the study of the mutagenic effect of potassium carbonate and laser red light on spring barley of the Bios variety The article reveals the data for three years o
 'Alloplasmic lines are a suitable model for studying molecular coevolution and interrelations between genetic systems of plant cells Whole chloroplast cp and mitochondrial mt genome se
 'Naked barley Hordeum vulgare var nudum L is a traditional culturally important climateresilient winter cereal crop of Nepal Evaluation of the naked barely genotypes for yield and dise
 'Spot blotch caused by Bipolaris sorokiniana Sacc in Sorok Shoem is an important disease of barley Hordeum vulgare L A total of barley genotypes received from Hill Crops Research Progr
 'The review presents the perspectives of using DNAmarkers in barley breeding for resistance to toxicity of aluminum boron manganese and cadmium ions Currently there have been identifie
 'Established a new method assisted by high speed centrifugalvortex HSCV to accelerate the extraction of BBGHSCV method reduced the extraction time by half and increased the yield by ab
 'Microbial community succession during highland barley wine brewingOrganic acids and volatiles varied on fermentation timeCorrelation analysis based on OPLS was conductedAcetobacter Le
 'The present study assessed the effects of glucooligosaccharides GOS derived from barley glucan on the proliferation and antimicrobial activity of probiotics All cocci examined in this

 'Understanding the genetic complexity of traits is an important objective of small grain temperate cereals yield and adaptation improvements Biparental quantitative trait loci QTL link
 'Barley is one of the most significant cereal crops of the Poaceae family and it is an important crop not only in Egypt but also all over the',
 'Most cereal breeding programmes are carried out under inversion tillage and high input agronomy resulting in the selection of high yield and quality commercial varieties This cultivar
 'Historically lucerne Medicago sativa is the most widely grown perennial pasture legume in southern Australia supplying a livestock feed source facilitating the high growth rates and',
 'Soil nitrogen N availability usually limits plant yields such that large quantities of synthetic N fertilizers are applied to ensure maximum productivity However excessive N use is a'
 'Malting is a process of forced germination conducted in order to degrade starch molecules and to obtain certain levels of amylolytic and proteolytic enzymes which are important in',
 'Barley Hordeum vulgare Epistatic QTL pairs Genomewide association study Nodal root anatomical traits Nodal root architecture Waterdeficit stress Roots perform vital roles for adaptati
 'Phenotype HDZI and HDZI promoters transcription factors transgenic barley and wheat Crepeat binding factor like protein Summary Networks of transcription factors regulate diverse phys
 'Highland barley gels were subjected to hydrothermal treatment and the rheological properties of both small and large strain amplitude oscillation shear SAOSLAOS were investigated in t
 'Barley starch was dual modified by hydroxypropylation using and propylene oxide separately based on starch weight dry basis followed by crosslinking through addition of mixture of sod
 'Improving grain yield and adaptation is achieved by synchronising crop phenology with the environment Phenology research is complex and encounters analytical challenges in characteris
 'The fungus Pyrenophora teres causal agent of net blotch of barley generates important yield losses mainly regarding grain weight and quality of malt extract for beer production In ord
 'The objective of this in vitro study was to determine the effects of different barley and oat varieties on CH production digestibility and rumen fermentation patterns in dairy cows Ou
 'The shape of an inflorescence varies among cereals ranging from a highly branched panicle in rice to a much more compact spike in barley Hordeum vulgare L and wheat Triticum aestivum
 'Kernel weight KW together with kernel number per unit area determines yield of cereal crops Here two barley recombinant inbred lines RILs populations with a shared parent were used to
 'Crop productivity is limited by several environmental constraints Among these salt stress plays a key role in limiting the growth and yield production of crop plants However the exoge
 'components in fermented barley aqueous extract were identified by UPLCHRMSPCA distinguished the samples and visualized the process with fermentation timeMetabolomics revealed the chan
 'In plants wall associated kinases WAKs form a unique subfamily of receptor likekinases RLKs In Arabidopsis thaliana WAKRLKs are known to regulate biotic stress cell expansion and meta
 'Grain kernel discoloration KD in cereal crops leads to downgrading grain quality and substantial economic losses worldwide Breeding KD tolerant varieties requires a clear understandin
 'Fusarium crown rot FCR a chronic and severe disease caused by various Fusarium species is prevalent in semiarid cropping regions worldwide One of the major QTL conferring FCR resistan
 'The flow of water through food commodity trade has been rationalized in the virtual water concept Estimates of future virtual water flows under climate land use and population changes
 'Barley bran is a good source of dietary fibers such as glucanDifferent amylolytic proteolytic and xylolytic enzymes are used in extractionExtraction affects molecular and rheological
 'In recent years the data science and remote sensing communities have started to align due to userfriendly programming tools access to highend consumer computing power and the availabi
 'Fraudulent ewallet deposit notification SMSes designed to steal money and goods from mbanking users have become pervasive in Namibia Motivated by an observed lack of mobile applicatio
 'In this work we modeled a novel approach to enhance surfaceenhanced Raman scattering SERS signals using principal component analysis PCA as a machine learning approach Zinc oxide nano
 'Modern biology frequently relies on machine learning to provide predictions and improve decision processes There have been recent calls for more scrutiny on machine learning performan
 'Chest Xray CXR images are usually used to identify the causes of patients symptoms including the classes of lung or heart disorders In visualization examination CXR imaging in anterio
 'Machine learning can precisely identify different cancer tumors at any stage by classifying cancerous and healthy samples based on their genomic profile We have developed novel method
 'Machine learning has the potential to fuel further advances in data science but it is greatly hindered by an ad hoc design process poor data hygiene and a lack of statistical rigor in
 'Uncertainty quantification is crucial to assess prediction quality of a machine learning model In the case of Extreme Learning Machines ELM most methods proposed in the literature mak
 'In recent years the data science and remote sensing communities have started to align due to userfriendly programming tools access to highend consumer computing power and the availabi
 'Fraudulent ewallet deposit notification SMSes designed to steal money and goods from mbanking users have become pervasive in Namibia Motivated by an observed lack of mobile applicatio
 'In this work we modeled a novel approach to enhance surfaceenhanced Raman scattering SERS signals using principal component analysis PCA as a machine learning approach Zinc oxide nano
 'Modern biology frequently relies on machine learning to provide predictions and improve decision processes There have been recent calls for more scrutiny on machine learning performan
 'Chest Xray CXR images are usually used to identify the causes of patients symptoms including the classes of lung or heart disorders In visualization examination CXR imaging in anterio
 'Machine learning can precisely identify different cancer tumors at any stage by classifying cancerous and healthy samples based on their genomic profile We have developed novel method
 'Machine learning has the potential to fuel further advances in data science but it is greatly hindered by an ad hoc design process poor data hygiene and a lack of statistical rigor in
 'Uncertainty quantification is crucial to assess prediction quality of a machine learning model In the case of Extreme Learning Machines ELM most methods proposed in the literature mak
 'The current practice of manually processing features for highdimensional and heterogeneous aviation data is laborintensive does not scale well to new problems and is prone to informat
 'Collaborative machine learning ML is an appealing paradigm to build highquality ML models by training on the aggregated data from many parties However these parties are only willing t
 'Machine learning plays a role in many deployed decision systems often in ways that are difficult or impossible to understand by human stakeholders Explaining in a humanunderstandable 
 'We present a brief history of the field of interpretable machine learning IML give an overview of stateoftheart interpretation methods and discuss challenges Research in IML has boome
 'As the decisions made or influenced by machine learning models increasingly impact our lives it is crucial to detect understand and mitigate unfairness But even simply determining wha
 'The unique electronic configuration endows carbon with superflexible bonding ability displaying metallic semiconducting and insulating features with unprecedented applications Inspire
 'Diagnosis of Parkinsons disease PD is commonly based on medical observations and assessment of clinical signs including the characterization of a variety of motor symptoms However tra
 'Machine learning has typically focused on developing models and algorithms that would ultimately replace humans at tasks where intelligence is required In this work rather than replac
 'Lconomtrie et lapprentissage machine semblent avoir une finalit en commun construire un modle prdictif pour une variable dintrt  laide de variables explicatives ou features Pourtant c
 'A learning machine in the form of a gating network that governs a finite number of different machine learning methods is described at the conceptual level with examples of concrete pr
 'Statistical machine learning theory often tries to give generalization guarantees of machine learning models Those models naturally underlie some fluctuation as they are based on a da
 ...]
```

✓ 13 s terminée à 12:54

Summarization_task

July 9, 2021

1 Abstractive Summarization Task

Résumé de texte (abstractive summary) : réaliser un modèle qui résume les abstracts (un seul abstract à la fois).

@Author : ERGUN Emrullah 09/07/2021 09:26

2 Loading Data

[1]:

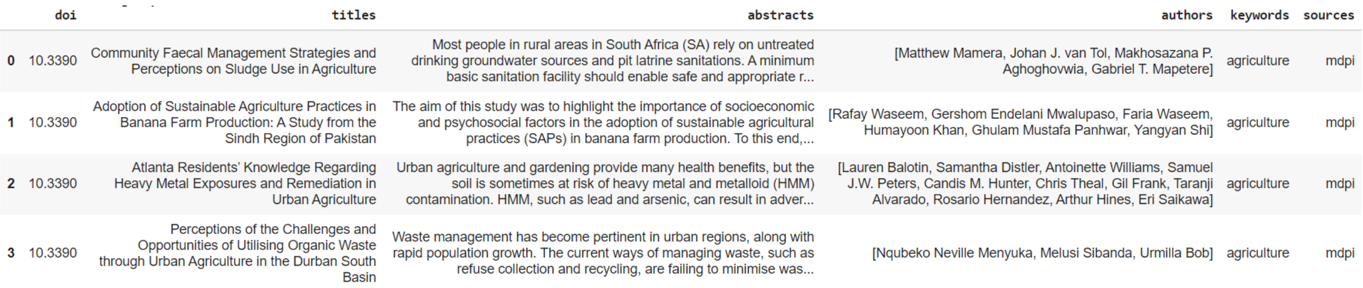
```
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Mounted at /content/gdrive
```

[4]:

```
import pandas as pd
df = pd.read_json('/content/gdrive/MyDrive/Exam/AgrSmall.json')
df.head()
```

[4]:

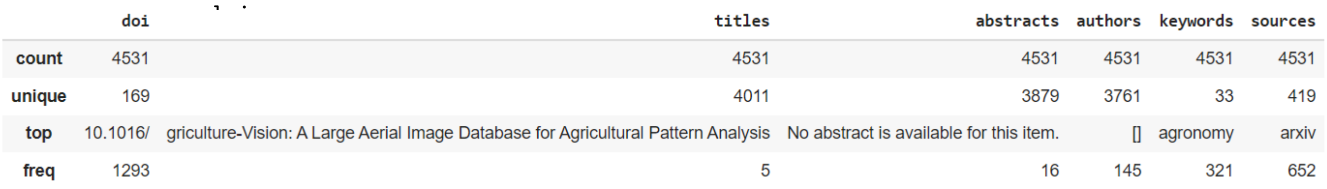


	doi	titles	abstracts	authors	keywords	sources
0	10.3390	Community Faecal Management Strategies and Perceptions on Sludge Use in Agriculture	Most people in rural areas in South Africa (SA) rely on untreated drinking groundwater sources and pit latrine sanitations. A minimum basic sanitation facility should enable safe and appropriate r...	[Matthew Mamera, Johan J. van Tol, Makhosazana P. Aghoghovwia, Gabriel T. Mapetere]	agriculture	mdpi
1	10.3390	Adoption of Sustainable Agriculture Practices in Banana Farm Production: A Study from the Sindh Region of Pakistan	The aim of this study was to highlight the importance of socioeconomic and psychosocial factors in the adoption of sustainable agricultural practices (SAPs) in banana farm production. To this end,...	[Rafay Waseem, Gershom Endelani Mwalupaso, Faria Waseem, Humayoon Khan, Ghulam Mustafa Panhwar, Yangyan Shi]	agriculture	mdpi
2	10.3390	Atlanta Residents' Knowledge Regarding Heavy Metal Exposures and Remediation in Urban Agriculture	Urban agriculture and gardening provide many health benefits, but the soil is sometimes at risk of heavy metal and metalloid (HMM) contamination. HMM, such as lead and arsenic, can result in adver...	[Lauren Balotin, Samantha Distler, Antoinette Williams, Samuel J.W. Peters, Candis M. Hunter, Chris Theal, Gil Frank, Taranji Alvarado, Rosario Hernandez, Arthur Hines, Eri Saikawa]	agriculture	mdpi
3	10.3390	Perceptions of the Challenges and Opportunities of Utilising Organic Waste through Urban Agriculture in the Durban South Basin	Waste management has become pertinent in urban regions, along with rapid population growth. The current ways of managing waste, such as refuse collection and recycling, are failing to minimise was...	[Nqubeko Neville Menyuka, Melusi Sibanda, Urmilla Bob]	agriculture	mdpi

[30]:

```
df.describe()
```

[30]:



	doi	titles	abstracts	authors	keywords	sources
count	4531	4531	4531	4531	4531	4531
unique	169	4011	3879	3761	33	419
top	10.1016/	griculture-Vision: A Large Aerial Image Database for Agricultural Pattern Analysis	No abstract is available for this item.	[]	agronomy	arxiv
freq	1293	5	16	145	321	652

```
[4 rows x 6 columns]
```

[3]:

```
```

[ ]:

```
!pip install transformers==2.8.0
!pip install torch==1.4.0
```

[6]:

```
import torch
import json
from transformers import T5Tokenizer, T5ForConditionalGeneration, T5Config
```

[7]:

```
model = T5ForConditionalGeneration.from_pretrained('t5-small')
tokenizer = T5Tokenizer.from_pretrained('t5-small')
device = torch.device('cpu')
```

```
HBox(children=(FloatProgress(value=0.0, description='Downloading', max=1197.0, style=ProgressS
```

```
HBox(children=(FloatProgress(value=0.0, description='Downloading', max=242065649.0, style=Prog
```

```
HBox(children=(FloatProgress(value=0.0, description='Downloading', max=791656.0, style=Progres
```

You need to change the value below to select the abstract you want to summarize. We can automatize this by creating a for loop and a list that will contain all summaries.

[12]:

```
text =df['abstracts'][0]
text
```

[12]:

```
'Most people in rural areas in South Africa (SA) rely on untreated drinking
groundwater sources and pit latrine sanitations. A minimum basic sanitation
facility should enable safe and appropriate removal of human waste, and although
pit latrines provide this, they are still contamination concerns. Pit latrine
sludge in SA is mostly emptied and disposed off-site as waste or buried in-situ.
Despite having knowledge of potential sludge benefits, most communities in SA
are reluctant to use it. This research captured social perceptions regarding
latrine sludge management in Monontsha village in the Free State Province of SA
through key informant interviews and questionnaires. A key informant interview
and questionnaire was done in Monontsha, SA. Eighty participants, representing
5% of all households, were selected. Water samples from four boreholes and four
rivers were analyzed for faecal coliforms and E.coli bacteria. On average, five
people in a household were sharing a pit latrine. Eighty-three percent disposed
filled pit latrines while 17% resorted to closing the filled latrines. Outbreaks
of diarrhoea (69%) and cholera (14%) were common. Sixty percent were willing to
use treated faecal sludge in agriculture. The binary logistic regression model
indicated that predictor variables significantly (p  0.05) described water
quality, faecal sludge management, sludge application in agriculture and biochar
```

```
adaption. Most drinking water sources in the study had detections  1 CFU/100
mL. It is therefore imperative to use both qualitative surveys and analytical
data. Awareness can go a long way to motivate individuals to adopt to a new
change. View Full-Text'
```

[13]:

```python
preprocess_text = text.strip().replace("\n","")

t5_prepared_Text = "summarize: " + preprocess_text
print ("original text preprocessed: \n", preprocess_text)

tokenized_text = tokenizer.encode(t5_prepared_Text, return_tensors="pt").
 ↪to(device)
```

```
original text preprocessed:
 Most people in rural areas in South Africa (SA) rely on untreated drinking
groundwater sources and pit latrine sanitations. A minimum basic sanitation
facility should enable safe and appropriate removal of human waste, and although
pit latrines provide this, they are still contamination concerns. Pit latrine
sludge in SA is mostly emptied and disposed off-site as waste or buried in-situ.
Despite having knowledge of potential sludge benefits, most communities in SA
are reluctant to use it. This research captured social perceptions regarding
latrine sludge management in Monontsha village in the Free State Province of SA
through key informant interviews and questionnaires. A key informant interview
and questionnaire was done in Monontsha, SA. Eighty participants, representing
5% of all households, were selected. Water samples from four boreholes and four
rivers were analyzed for faecal coliforms and E.coli bacteria. On average, five
people in a household were sharing a pit latrine. Eighty-three percent disposed
filled pit latrines while 17% resorted to closing the filled latrines. Outbreaks
of diarrhoea (69%) and cholera (14%) were common. Sixty percent were willing to
use treated faecal sludge in agriculture. The binary logistic regression model
indicated that predictor variables significantly (p  0.05) described water
quality, faecal sludge management, sludge application in agriculture and biochar
adaption. Most drinking water sources in the study had detections  1 CFU/100
mL. It is therefore imperative to use both qualitative surveys and analytical
data. Awareness can go a long way to motivate individuals to adopt to a new
change. View Full-Text
```

[14]:

```python
# summmarize
summary_ids = model.generate(tokenized_text,
                                    num_beams=4,
                                    no_repeat_ngram_size=2,
                                    min_length=30,
                                    max_length=100,
                                    early_stopping=True)

output = tokenizer.decode(summary_ids[0], skip_special_tokens=True)

print ("\n\nSummarized text: \n",output)
```

```
Summarized text:
 most people in rural areas in south africa (SA) rely on untreated drinking
groundwater sources and pit latrine sanitations. despite having knowledge of
potential sludge benefits, most communities in SA are reluctant to use it.
```

[38]:

```python
print("\n\nLongueur du text d'origine \n", len(text),"\n\nLongueur du text produit \n", len(output))
```

```
Longueur du text d'origine
 1675

Longueur du text produit
 226
```

We can assess the quality of the summary in a human way because we don't have target summary columns or anything that can help us with that. So, after reading the original text, we can see that the result is quite good.

[43]:

```python
%%capture
!wget -nc https://raw.githubusercontent.com/brpy/colab-pdf/master/colab_pdf.py
from colab_pdf import colab_pdf
colab_pdf('Summarization_task.ipynb',"/content/drive/MyDrive/Exam/")
```

[ ]:

```python

```

BERT-Topic-Modelling

July 9, 2021

1 Topic Modelling

Détection de thématiques (topic modeling) : extraire les thématiques pertinentes traitées par les abstracts.

[ ]:
```
!pip install bertopic[visualization] --quiet
```

[ ]:
```
!pip install numpy==1.19.5
```

2 Imports

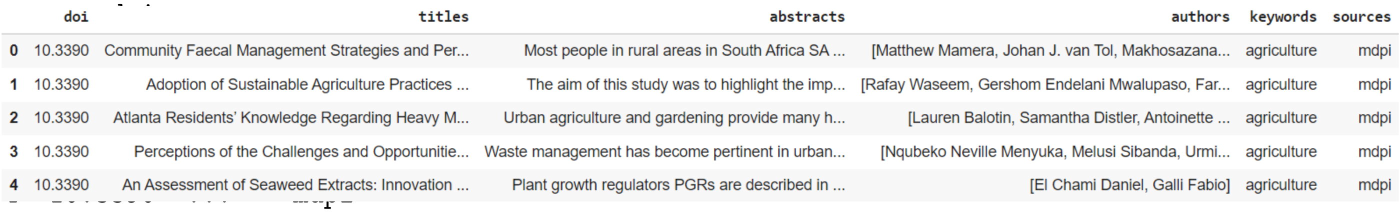
[75]:
```
import numpy as np
import pandas as pd
from copy import deepcopy
from bertopic import BERTopic
```

[76]:
```
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call
drive.mount("/content/gdrive", force_remount=True).
```

[126]:
```
import pandas as pd
df = pd.read_json('/content/gdrive/MyDrive/Exam/AgrSmall_cleaned.json')
df.head()
```

[126]:



	doi	titles	abstracts	authors	keywords	sources
0	10.3390	Community Faecal Management Strategies and Per...	Most people in rural areas in South Africa SA ...	[Matthew Mamera, Johan J. van Tol, Makhosazana...	agriculture	mdpi
1	10.3390	Adoption of Sustainable Agriculture Practices ...	The aim of this study was to highlight the imp...	[Rafay Waseem, Gershom Endelani Mwalupaso, Far...	agriculture	mdpi
2	10.3390	Atlanta Residents' Knowledge Regarding Heavy M...	Urban agriculture and gardening provide many h...	[Lauren Balotin, Samantha Distler, Antoinette ...	agriculture	mdpi
3	10.3390	Perceptions of the Challenges and Opportunitie...	Waste management has become pertinent in urban...	[Nqubeko Neville Menyuka, Melusi Sibanda, Urmi...	agriculture	mdpi
4	10.3390	An Assessment of Seaweed Extracts: Innovation ...	Plant growth regulators PGRs are described in ...	[El Chami Daniel, Galli Fabio]	agriculture	mdpi

```
[5 rows x 6 columns]
```

Fortunately this kind of methods don't need much pre-processing. That way we will just have to pass our data to model topics.

3 Load data

We need to have some idea of how many topics we need to extract from our abstracts. Fortunately, there is a column called Keywords which can be interpreted as a column of topics. We see that there are 33 topics in our abstracts.

[127]:
```
len(df['keywords'].value_counts())
```

[127]:
```
33
```

[128]:
```
docs = list(df.loc[:, "abstracts"].values)
```

4 Creating Topics

[130]:
```
model = BERTopic(language="english",nr_topics="auto")
```

[131]:
```
topics, probs = model.fit_transform(docs)
```

5 We can then extract most frequent topics:

[132]:
```
model.get_topic_freq()
```

[132]:
```
    Topic  Count
0      -1   1478
1       0    223
2       1    221
3       2    193
4       3    165
..    ...    ...
56     55     12
57     56     12
58     57     11
59     58     11
60     59     10

[61 rows x 2 columns]
```

Subject -1 means outliers. It is not relevant for our application as we are trying to group those that can be grouped and not the outliers.

6 Get Individual Topics

[133]:
```
model.get_topic(0)
```

[133]:
```
[('resistance', 0.027306180360468505),
 ('wheat', 0.022536936387512512),
 ('genes', 0.015326432149689312),
 ('gene', 0.012973224803286497),
 ('breeding', 0.010564508038693646),
 ('genetic', 0.008334712069907636),
 ('plant', 0.008172686948642576),
 ('plants', 0.0077816582570526096),
 ('cultivars', 0.007644248708109173),
 ('genotypes', 0.007554980210517708)]
```

This is the outliers topic

```
[134]: model.get_topic(2)
```

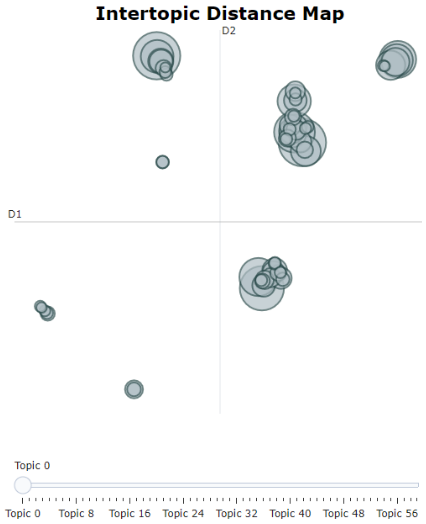
```
[134]: [('cows', 0.02477964625381168),
        ('dairy', 0.02239357360037012),
        ('milk', 0.019933434338471116),
        ('cattle', 0.012417915963207972),
        ('pig', 0.010149702434404785),
        ('animal', 0.009687357316715972),
        ('cow', 0.008901690751340622),
        ('pigs', 0.008148020096708798),
        ('farm', 0.007660743789666665),
        ('feeding', 0.007506514561680994)]
```

This one seems to be related to machine learning

```
[135]: model.get_topic(14)
```

```
[135]: [('leaf', 0.01709953169884755),
        ('plant', 0.015461347938547013),
        ('classification', 0.013272324515412761),
        ('tree', 0.012575488701553818),
        ('canopy', 0.012183447987412239),
        ('leaves', 0.01105078285855785),
        ('trees', 0.008473905512920798),
        ('dataset', 0.007336205870198295),
        ('detection', 0.007298431730194899),
        ('vegetation', 0.006829647371095687)]
```

agriculture



7 Visualize Topics

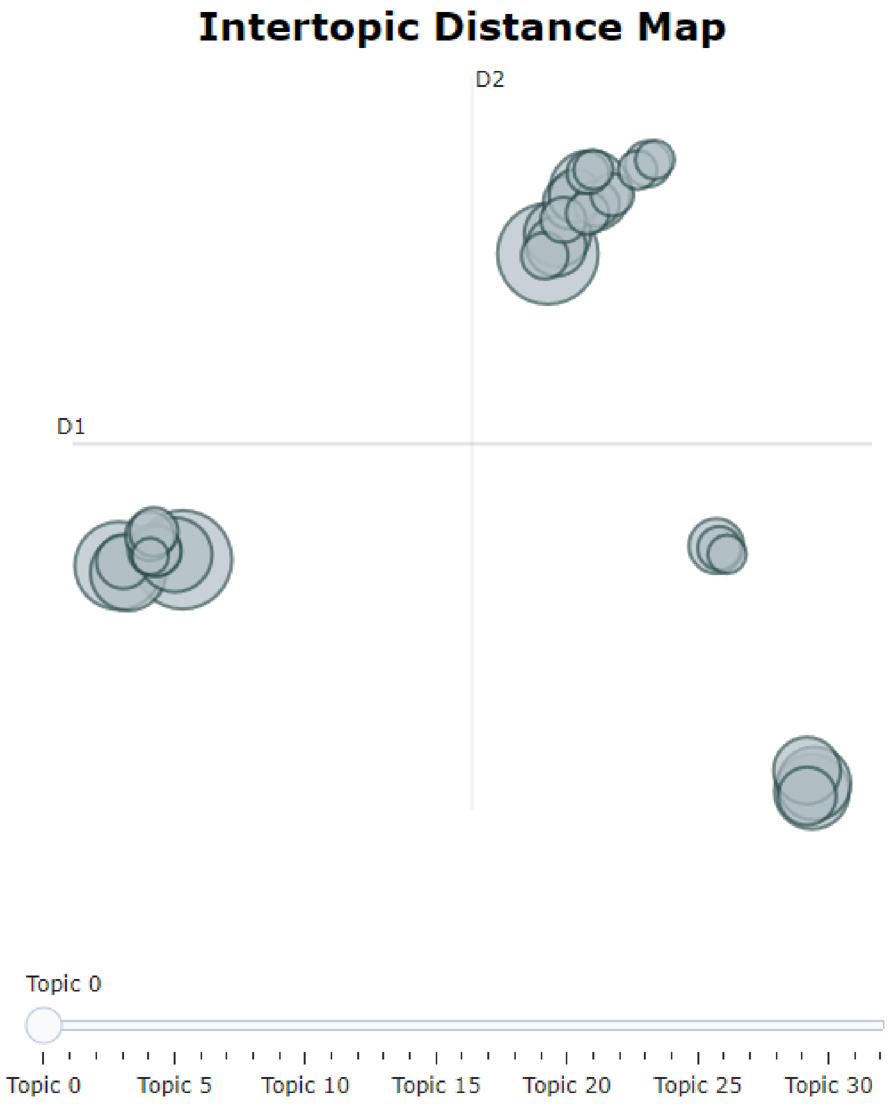
```
[136]: model.visualize_topics()
```

```
[137]: new_topics, new_probs = model.reduce_topics(docs, topics, probs, nr_topics=33)
```

```
[138]: model.get_topic_freq()
```

```
[138]:    Topic  Count
       0     -1   1535
       1      0    254
       2      1    243
       3      2    193
```

```
4     3   165
5     4   143
6     5   140
7     6   138
8     7   134
9     8   112
10    9   110
11   10    89
12   11    84
13   12    75
14   13    74
15   14    73
16   15    70
17   16    68
18   17    63
19   18    60
20   19    59
21   20    54
22   21    52
23   22    52
24   23    51
25   24    47
26   25    45
27   26    45
28   27    44
29   28    39
30   29    38
31   30    38
32   31    37
33   32    34
```



[139]:
```
model.visualize_topics()
```

Here we reduced the number to 33 to see if our first assumption is relevant.

We can see that the cluster are pretty spaced

You can see that we can still reduce the number of topics if we want a very llow topic granularity.

[143]:
```
model.find_topics("agriculture")
```

[143]:
```
([7, 5, 27, 9, 28],
 [0.8518545761899723,
  0.8431021045540084,
  0.675743308534456,
  0.6502194162663008,
  0.6107970152718672])
```

[145]:
```
model.get_topic(7)
```

[145]:

```
[('agricultural', 0.03565437019382535),
 ('technology', 0.026140116507505806),
 ('technologies', 0.025191018766316336),
 ('agriculture', 0.023997458757431l3),
 ('farmers', 0.020352165381229397),
 ('farming', 0.01690241404833638),
 ('smart', 0.012369160185100859),
 ('development', 0.010033620563684323),
 ('innovation', 0.00941452781672134),
 ('rural', 0.008290963459814 67)]
```

[144]:

```
model.find_topics("machine learning")
```

[144]:

```
([19, 28, 1, -1, 26],
 [0.6228054769533977,
  0.5893433771438448,
  0.583040589428113,
  0.5528516026756964,
  0.48521100341254775])
```

[146]:

```
model.get_topic(19)
```

[146]:

```
[('learning', 0.03819207433373763 4),
 ('data', 0.018738292850432 82),
 ('model', 0.013674783425643256),
 ('audio', 0.013612465297208966),
 ('models', 0.013596429964412085),
 ('fairness', 0.013477921456525252),
 ('as', 0.012610033948148306),
 ('ml', 0.010859556709394 49),
 ('healthcare', 0.009806047128952 94),
 ('supervised', 0.009429501947349846)]
```

The result is very interesting because the method of evaluating the proximity between a topic and a word gives probabilities that this topic can be represented by a single word.

The words that defines topics are way more logic when we create 33 topics. Lets see if we ask for 2 topics, we will obtain 2 topics that define agriculture and machine learning.

8 2 TOPICS

[117]:

```
docs = list(df.loc[:, "abstracts"].values)
```

[118]:

```
model = BERTopic(language="english",nr_topics=3)
```

[119]:

```
topics, probs = model.fit_transform(docs)
```

[120]:

```
model.get_topic_freq()
```

[120]:

```
   Topic  Count
0     -1   1540
1      0   1212
```

```
2      1   1100
3      2    606
```

```
[124]: model.find_topics("agriculture")
```

```
[124]: ([1, 0, 2, -1],
        [0.6286732631640763,
         0.41762277978959006,
         0.3311495793753889,
         0.32436338257908837])
```

```
[122]: model.get_topic(1)
```

```
[122]: [('to', 0.06128262780777263),
        ('for', 0.03765940981166114),
        ('data', 0.022153236115159 29),
        ('agriculture', 0.02195908129465209),
        ('agricultural', 0.018556788586467855),
        ('using', 0.013720183003748965),
        ('use', 0.012720749746971046),
        ('learning', 0.012604427999175504),
        ('water', 0.012456796406484245),
        ('farming', 0.012391099137114642)]
```

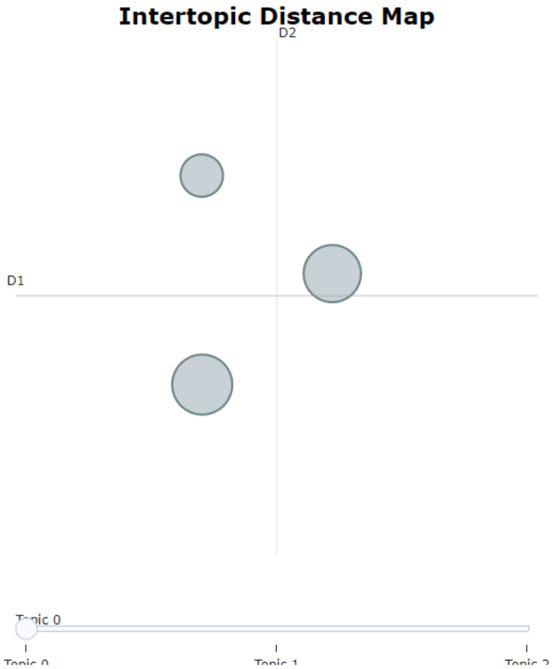
```
[125]: model.find_topics("machine learning")
```

```
[125]: ([2, 1, -1, 0],
        [0.4688787801961388,
         0.3184875039725184,
         0.29089146589564324,
         0.2871437819998832])
```

```
[123]: model.get_topic(2)
```

```
[123]: [('in', 0.04879732338921885),
        ('we', 0.034624216579477785),
        ('that', 0.029920501263070783),
        ('learning', 0.025694609610814705),
        ('neural', 0.023517840793407484),
        ('cnn', 0.021882377052382294),
        ('segmentation', 0.021457022246107817),
        ('data', 0.021047214596373723),
        ('as', 0.020567844773851995),
        ('model', 0.018847795307258816)]
```

```
[121]: model.visualize_topics()
```



After a first attempt of creating 2 topics we saw that one topic has been defined by some stopwords. I think that is normal as topwords appears everywhere and we ask the algorithm to overreduce the amount of topic so it end-up to underfit and to generalize.

```
[94]: import nltk
```

[95]:

```
from nltk.corpus import stopwords
nltk.download('stopwords')
nltk.download('punkt')
from nltk.tokenize import word_tokenize
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

[96]:

```
len(docs)
```

[96]:

```
4458
```

[97]:

```
stop_words = set(stopwords.words('english'))

def remove_stopwords(x):
  words = x.split(" ")
  temp = []
  for word in words:
    if word not in stop_words:
      temp.append(word)
  return " ".join(temp)
```

[98]:

```
df['text_clean'] = df['abstracts'].apply(lambda x: remove_stopwords(x))
df
```

[98]:

	doi	titles	abstracts	authors	keywords	sources	text_clean
0	10.3390	Community Faecal Management Strategies and Per...	Most people in rural areas in South Africa SA ...	[Matthew Mamera, Johan J. van Tol, Makhosazana...	agriculture	mdpi	Most people rural areas South Africa SA rely u...
1	10.3390	Adoption of Sustainable Agriculture Practices ...	The aim of this study was to highlight the imp...	[Rafay Waseem, Gershom Endelani Mwalupaso, Far...	agriculture	mdpi	The aim study highlight importance socioeconom...
2	10.3390	Atlanta Residents' Knowledge Regarding Heavy M...	Urban agriculture and gardening provide many h...	[Lauren Balotin, Samantha Distler, Antoinette ...	agriculture	mdpi	Urban agriculture gardening provide many healt...
3	10.3390	Perceptions of the Challenges and Opportunitie...	Waste management has become pertinent in urban...	[Nqubeko Neville Menyuka, Melusi Sibanda, Urmi...	agriculture	mdpi	Waste management become pertinent urban region...
4	10.3390	An Assessment of Seaweed Extracts: Innovation ...	Plant growth regulators PGRs are described in ...	[El Chami Daniel, Galli Fabio]	agriculture	mdpi	Plant growth regulators PGRs described literat...
...	...	...	...	...	...	...	...
4526	10.2478/	1. Modelling groundwater flow and nitrate tran...	The present paper discusses studies related to...	[Sieczka Anna, Bujakowski Filip, Koda Eugeniusz]	precision agriculture	Sciendo, 2018.	The present paper discusses studies related pr...
4527	unknown	2. Cosechando los beneficios de la agricultura...	El objetivo del trabajo fue desarrollar una me...	[Bonilla, Camila, Terra, José A, Gutiérrez, Lu...	precision agriculture	Facultad de Agronomía - Instituto Nacional de ...	El objetivo del trabajo fue desarrollar una me...
4528	unknown	6. A Risk Analysis of Precision Agriculture Te...	Precision agriculture technology can transform...	[Yangxuan Liu, Michael R. Langemeier, Ian M. S...	precision agriculture	MDPI, Open Access Journal, 2018.	Precision agriculture technology transform far...
4529	10.1016/	7. Integrated open geospatial web service enab...	We proposed an integrated geospatial service e...	[Chen, Nengcheng, Zhang, Xiang, Wang, Chao]	precision agriculture	Elsevier B.V.	We proposed integrated geospatial service enab...
4530	10.5194/	9. UNMANNED AERIAL VEHICLE (UAV) DERIVED NORMA...	Malaysia currently is one of the biggest globa...	[S. A. Suab, M. S. Syukur, R. Avtar, A. Korom]	precision agriculture	Copernicus Publications, 2019.	Malaysia currently one biggest global producer...

[99]:

```
docs = list(df.loc[:, "text_clean"].values)
```

Then we choose to get rid of them and we will see if the result is better

[112]:

```
model = BERTopic(language="english",nr_topics=3)
```

[113]:

```
topics, probs = model.fit_transform(docs)
```

[114]:

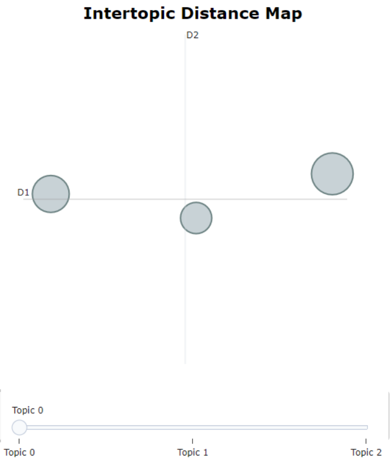
```
model.get_topic_freq()
```

[114]:

```
   Topic  Count
0     -1   3037
1      0    607
2      1    474
3      2    340
```

[115]:

```
model.visualize_topics()
```



As we can see removing stopwords is not the best idea as the algorithm yield bad results.

We still need to create an evaluation pipeline to score the topics created but I can't think of a method

9 Conclusion

As we can see in this notebook our work permits us to extract topic from abstracts. The resuults are pretty good. As the probabilities are high for the two words : ' Agriculture' and 'Machine Learning'.

[ ]:

```
!wget -nc https://raw.githubusercontent.com/brpy/colab-pdf/master/colab_pdf.py
from colab_pdf import colab_pdf
colab_pdf('BERT-Topic-Modelling.ipynb',"/content/drive/MyDrive/Exam/")
```

```
--2021-07-09 11:00:45--  https://raw.githubusercontent.com/brpy/colab-
pdf/master/colab_pdf.py
Resolving raw.githubusercontent.com (raw.githubusercontent.com)...
185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com
(raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1864 (1.8K) [text/plain]
Saving to: colab_pdf.py

colab_pdf.py        100%[===================>]   1.82K  --.-KB/s    in 0s

2021-07-09 11:00:45 (32.3 MB/s) - colab_pdf.py saved [1864/1864]

Mounted at /content/drive/

WARNING: apt does not have a stable CLI interface. Use with caution in scripts.


WARNING: apt does not have a stable CLI interface. Use with caution in scripts.

Extracting templates from packages: 100%
```

[ ]:

```

```