

# My Projects and Experience

## Introduction

The ONS-UNECE Machine Learning Group is a platform for knowledge exchange, research collaboration and capacity building in the use of machine learning in official statistics. It brings together members from national and international statistical organisations around the world to explore the value-add of machine learning for official statistics as well as how best to integrate it into existing production systems.

I currently work in Turkish Statistical Institute (TurkStat) and I have a strong interest in how machine learning can help improve statistical output and be integrated into production systems. For this reason, I joined Machine Learning Group to contribute to the group's activities. ML Group is a dynamic community and a great place to connect with and learn from other colleagues working on similar ML challenges.

I signed up for the web scraping data theme in this year's activity programme of ML Group 2022. I chose this theme because I have an interest in web scraping and I have projects in mind that can be accomplished through web scraping. In our web scraping theme group, there were two other statistical offices as implementers: Statistics Poland and Statistics Flanders. Web Scraping Data theme group coordinator was Dr Michael Reusens from Statistics Flanders. The implementing members of the web scraping data group meets up monthly to discuss web scraping of statistical unit information. Besides informative presentations associated with web scraping, meeting agenda includes presentation of each implementing member on progress, methods, tooling, roadblocks, etc. and discussion on differences and similarities in approach.

## My Projects

I ran three projects in the web scraping theme group as an implementer:

1. **Scrape an ICT variable:** Obtaining an ICT variable, namely social media presence by web scraping and comparing it to the TurkStat ICT Usage in Enterprises survey results.
2. **Gain insights from an open-ended question:** Extracting insights from responses to an open-ended question in biotechnology survey without examining individual responses one by one.
3. **Create a framework for government R&D survey:** Through web scraping, identifying public organizations that use R&D-related words or phrases on their web pages at least once and creating a government R&D survey framework with them.

### Project 1 – Scrape an ICT variable ([Link](#))

In the ICT Usage in Enterprises Survey of TurkStat, the variable "whether the enterprise website has links to their social media profiles" has been obtained via web scraping this time and compared both results.

### Review existing work

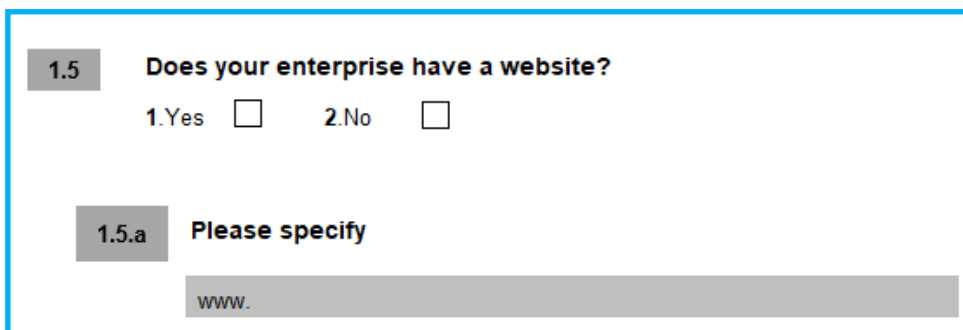
What I did during the preparation was that I first reviewed ESSNET Bigdata II WPC deliverables ([link](#)). In the link it can be found Workpackage C (WPC) of ESSnet Big Data deliverables which focus on enterprise characteristics. Its aim is to use web-scraping, text mining and inference techniques for collecting and processing enterprise information, in order to improve or update information held by the national business registers. The implementation involves massive scraping of company websites, collecting, processing, analysing unstructured data and dissemination of national-level experimental statistics. The enterprise data collected by WPC combined with existing data from multiple other sources, such as ICT usage surveys.

And then I visited Jacek Maślankowski (Ph.D., Assistant Professor, University of Gdańsk, Poland) WP2-Social-Media-Presence GitHub repository and examined the code he wrote and adapted the relevant code for TurkStat case. The application is used to scrap all the links related to social media from websites ([link](#)).

### Work steps followed

After the preparation, I followed the below work steps:

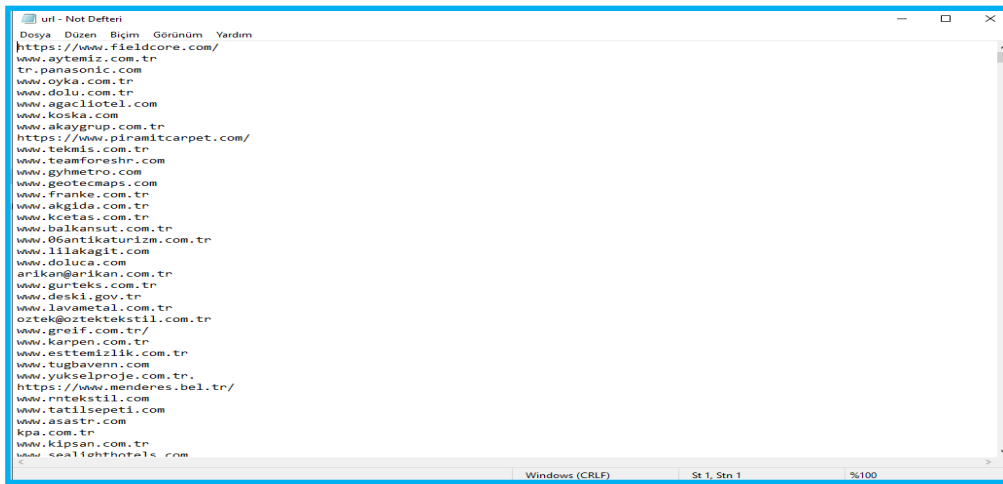
- First, I got a list of enterprise web addresses to scrape: In the 2021 TurkStat ICT usage survey, there was a question asking enterprises to specify what their web addresses are. The responses to this question created the web addresses framework (URL list). I did not have a problem getting the list of enterprises' web addresses because we already have this information since we asked this question to enterprises in the 2021 ICT usage survey. As you can see from the below screenshot we asked enterprises if they have a website or not and for those who have a website we wanted them to indicate it.



The image shows a survey interface with two questions. The first question, labeled '1.5', asks 'Does your enterprise have a website?' and provides two radio button options: '1.Yes' and '2.No'. The second question, labeled '1.5.a', asks 'Please specify' and features a text input field with the placeholder text 'www.'.

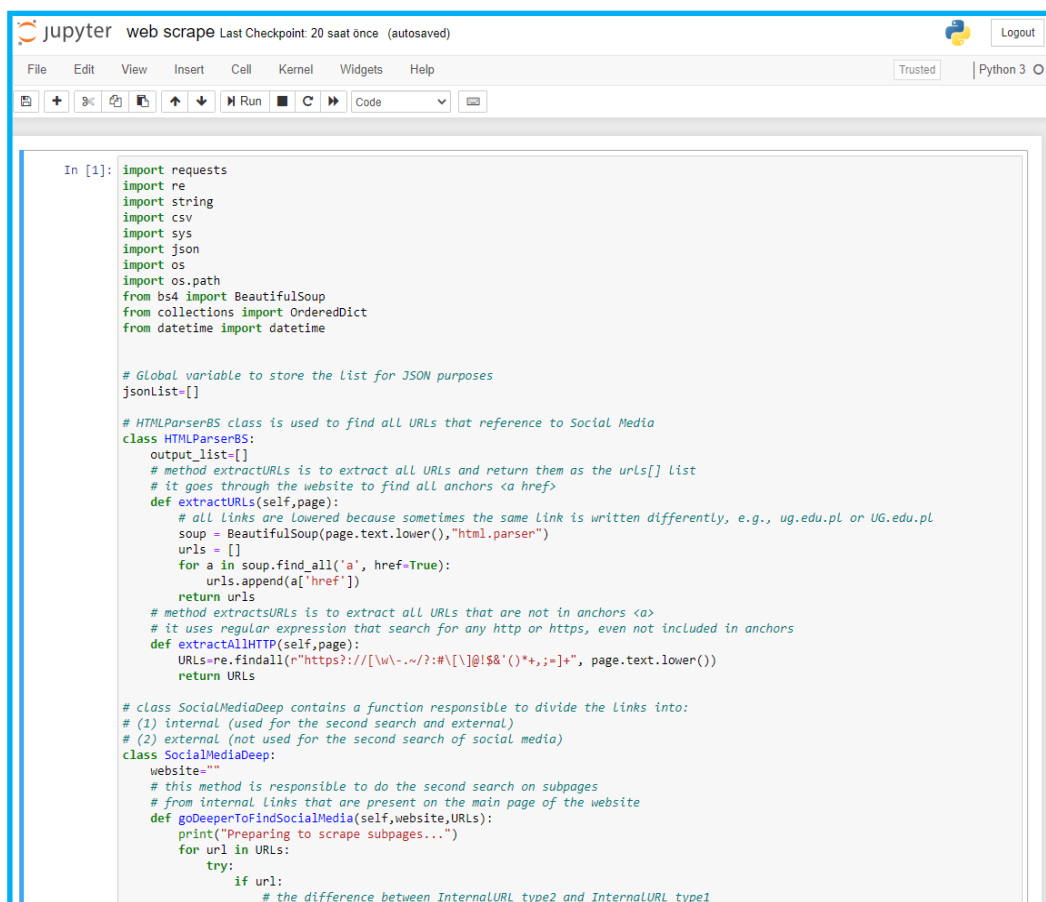
**Figure 1. TurkStat ICT usage survey question asking enterprises for their websites**

- Next, I created a text file of these web addresses (as shown below) that would be the input for the code I was going to run. A text file was created from the list of web addresses of enterprises with 250+ employees. And this file became the input file of the code. ICT Usage Survey in TurkStat is a complete census for enterprises with 250+ employees. Therefore, I limited this study to only enterprises having 250 or more employees in Türkiye. The python code was designed to take a text file with web addresses written one after the other as input. Therefore, I also created my file in this format. It is also possible to just copy/paste the ready text file.



**Figure 2. “url.txt” input file consisting of URLs**

- Then, I used Jupyter notebook in the Anaconda distribution to execute the code which is an open-source IDE (Integrated Development Environment). Very basically the purpose of the code is to find social media links on enterprise websites. Input is a list of URLs and output is a “csv” file containing domain names and found social media links. I executed the ready-made python code ([link](#)) with very minor changes using Jupyter Notebook. The application scrapped all the links related to social media from websites.



**Figure 3. Running code for web scrapping in Jupyter Notebook**

- And finally, I compared the results generated by the code with the results from the survey for the social media presence variable. As the identifier, the web address field in the output file could be used for matching and linking scraped data and survey data. I had to make some adjustments in the web addresses to bring them to the same format to achieve the maximum match rate. After combining the two datasets, all that remains was to make a comparison which was made with the scraped data and the enterprise responds to the question shown below figure in the 2021 ICT survey questionnaire. The results obtained with web-scraping were compared with those obtained from the 2021 survey.

A7. Does the website have any of the following? <i>-Optional</i>		Yes	No
	a) Description of goods or services, price information	<input type="checkbox"/>	<input type="checkbox"/>
*6	b) Online ordering or reservation or booking, e.g. shopping cart	<input type="checkbox"/>	<input type="checkbox"/>
	c) Possibility for visitors to customise or design online goods or services	<input type="checkbox"/>	<input type="checkbox"/>
	d) Tracking or status of orders placed	<input type="checkbox"/>	<input type="checkbox"/>
	e) Personalised content on the website for regular/recurrent visitors	<input type="checkbox"/>	<input type="checkbox"/>
	f) Links or references to the enterprise's social media profiles	<input type="checkbox"/>	<input type="checkbox"/>

**Figure 4. “social media presence” question in the TurkStat ICT usage survey**

The code took a total of 4 hours and 44 minutes to run and produced 1,566 pages of output. The execution of the code was interrupted only once. Then I just rerun the code for the remaining URLs and combined the two outputs when it has finished.

#### Results

As a result, web scraping was successfully performed for 1,087 out of a total of 3,640 enterprises. This corresponds to a rate of 30%. The comparison summary results are below:

**Table 1. Summary of the comparison between web scraping and the survey results for the social media presence variable**

		Web Scraping		Total
		Yes	No	
Survey	Yes	392 (36%)	325 (30%)	717
	No	89 (8%)	281 (26%)	370
Total		606	481	1,087

As can be seen from the table above, the percentage of enterprises stating in the survey that they have a link to their social media platforms and having at least one social media link in the scraped data is 36%. The percentage of enterprises stating in the survey that they don't have any link to their social media platforms and having no social media link in the scraped data is 26%. The percentage of enterprises stating in the survey that they have a link to their social media platforms but having no social media link in the scraped data is 30%. And lastly, the percentage of enterprises stating in the survey that they don't have any link to their social media platforms but having at least one social media link in the scraped data is 8%.

Eventually, 62% accuracy was achieved while the inconsistency was 38% between two different data acquisition methods.

## Project 2 – Gain insights from an open-ended question ([Link](#))

### Introduction

With the Turkish Statistical Institute (TurkStat) Biotechnology Statistics Survey, it is aimed to create statistical data in the field of biotechnology. All enterprises engaged in biotechnology activities have been covered in the survey and the data is collected directly from enterprises via web survey.

In the 2020 Biotechnology Statistics Survey, unlike previous years, an open-ended question was added to the end of the questionnaire in addition to the routinely asked questions. The purpose of adding this question was not to miss anything about this very technical domain. Because open-ended questions allow the respondent units to freely express the points that the survey designer missed on the subject. The question was: “Briefly inform us about the biotechnology activities carried out by your enterprise and the techniques and applications it uses”. All enterprises carrying out biotechnology activities were required to fill in this open-ended question at the end of the survey questionnaire. Eventually; all 499 enterprises carrying out biotechnology activities in Türkiye filled the free text box provided to them with information about their activities.

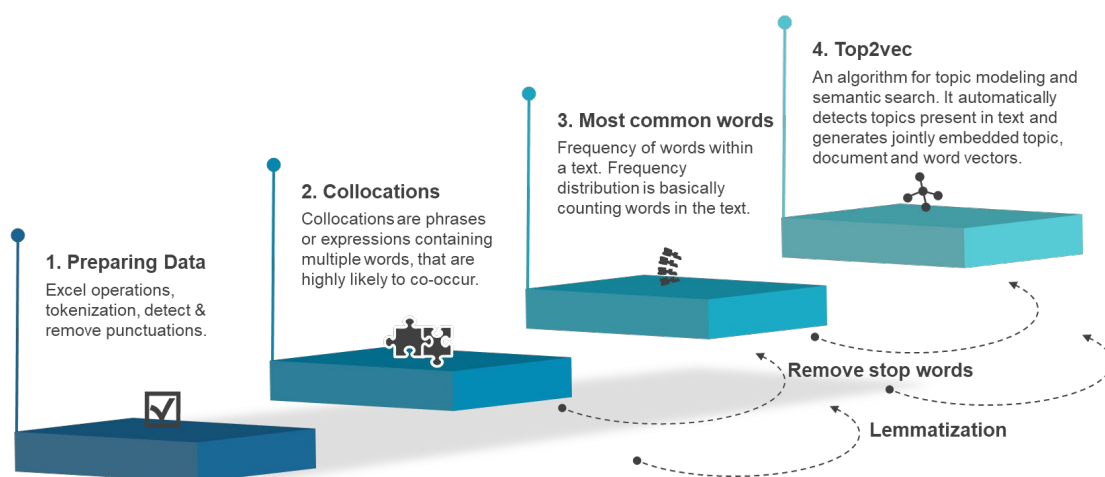
This paper presents an experimental study to analyse these free texts and try to understand in a few words what biotechnology enterprises in Türkiye are mentioning about most in these texts in 2020. Of course, these free texts would not be reviewed and analysed one by one, which was not possible anyway. For this reason, a different analysis method based on Natural Language Processing (NLP) and deep learning techniques has been developed using Python to gain insights from these free texts. Natural Language Processing—or NLP for short—in a wide sense to cover any kind of computer manipulation of natural language [1]. The reason for choosing Python programming language is that it has an excellent functionality for processing linguistic data.

### Methods

In order to extract insights from the responses provided by enterprises to the open-ended biotechnology survey question as free text, the following processes were carried out in Python programming language:

1. Preparing data by pre-processing,
2. Finding collocations,
3. Finding most common words and
4. Detecting topics present in text with Top2vec algorithm

The second, third and fourth steps mentioned above were repeated once more after both removing stop words and lemmatization.



**Figure 5. Process steps**

### *Preparing data by pre-processing*

Before insights could be gained from the free texts, first of all, some operations needed to be done on the raw data. These operations covered deleting nonsense records (the records with only "." or "x"), standardizing some words (such as removing the hyphen between words) and converting all letters to lowercase. After these operations performed on the raw data, the data was ready for pre-processing. Pre-processing stage included tokenization and detecting and removing punctuations.

The analyses carried out to extract meaning from the free texts in this study were also repeated by removing stop words and performing lemmatization. Stop words are basically a set of commonly used words in any language and in NLP and text mining applications, they are used to eliminate unimportant words. As for the lemmatization, it is the method to take any kind of word to that base root form with the context. It groups together the different inflected forms of a word so they can be analysed as a single item [2].

### *Finding collocations*

After the pre-processing was completed, analyses were started to extract meaningful information from the word list obtained. For this purpose, collocations in the word list were searched first. Collocations are expressions of multiple words which commonly co-occur [3]. Natural Language Toolkit (NLTK) library in python has been used to find collocations. NLTK is a leading platform for building Python programs to work with human language data [4]. NLTK contains *collocations* module having tools to identify collocations within corpora.

### *Finding most common words*

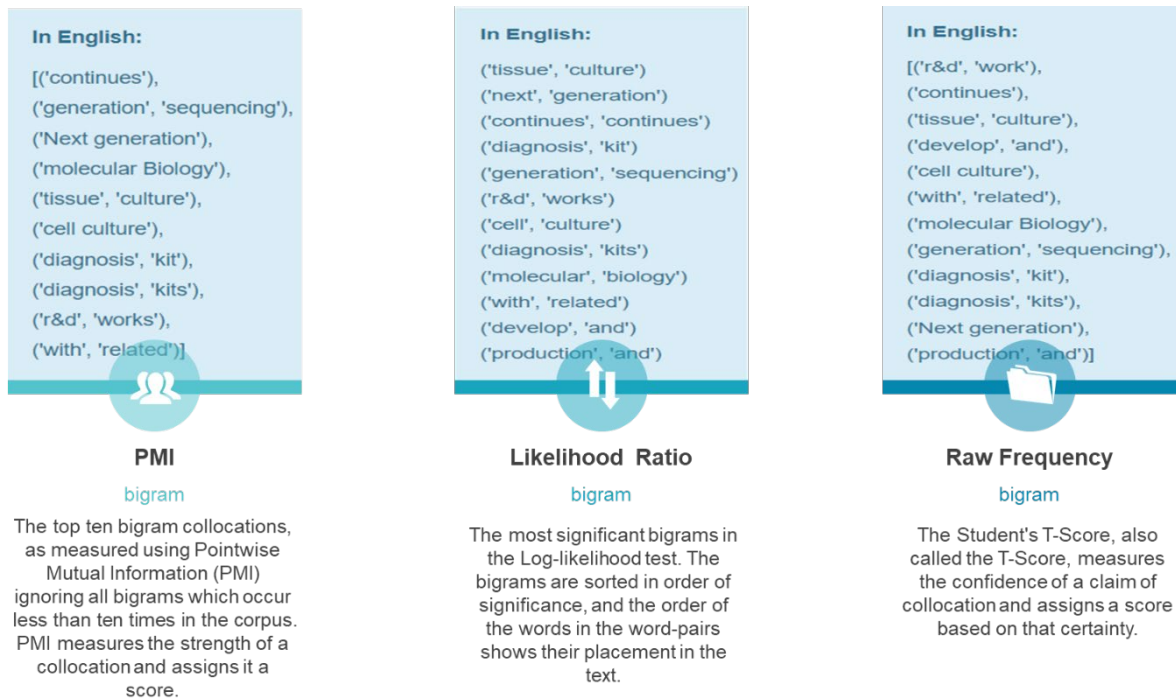
One way to find out what is most frequently mentioned in free texts is to look at the most frequently used words in these texts. For this purpose, number of occurrences of each individual of the word/word group were calculated through the *FreqDist* module in NLTK.

### *Detecting topics present in text with Top2vec algorithm*

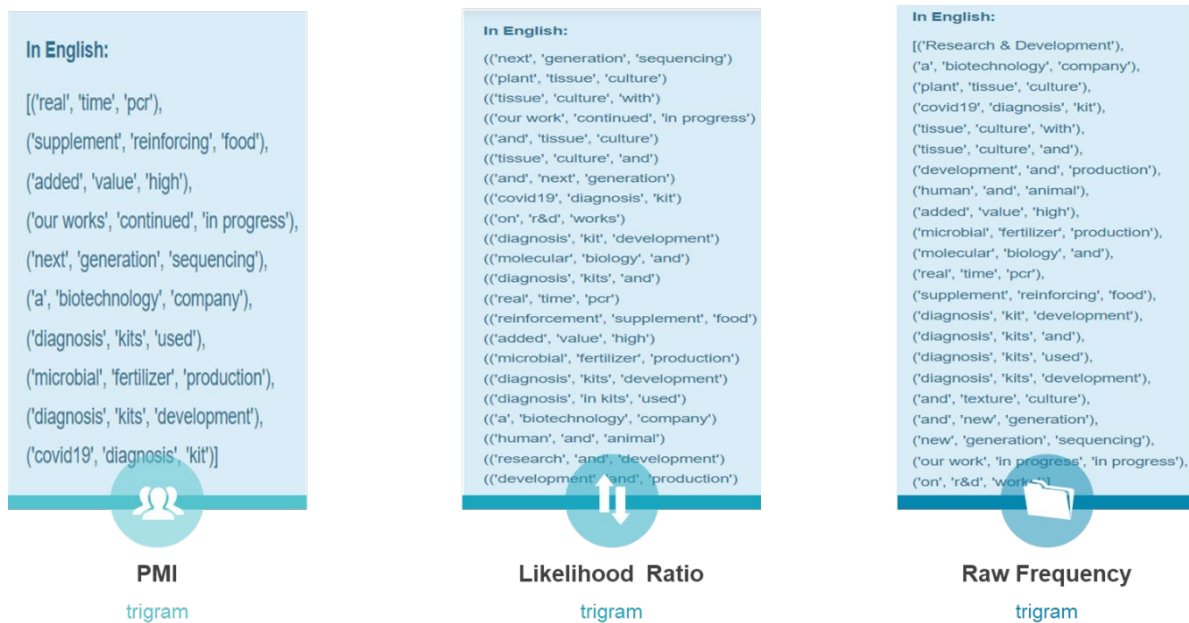
In this study, it has also been tried to find how many different topics can be produced and what these topics can be by combining similar words in free texts using the *Top2vec* algorithm. Top2Vec is an algorithm for topic modelling and semantic search. It automatically detects topics present in text and generates jointly embedded topic, document and word vectors [5].

## Results

First; the bigram and trigram collocations were obtained before removing stop words, using three different measure of association, namely Pointwise Mutual Information (PMI), Likelihood Ratio and Raw Frequency. The results are shared with the figures below.



**Figure 6. Bigram collocations**

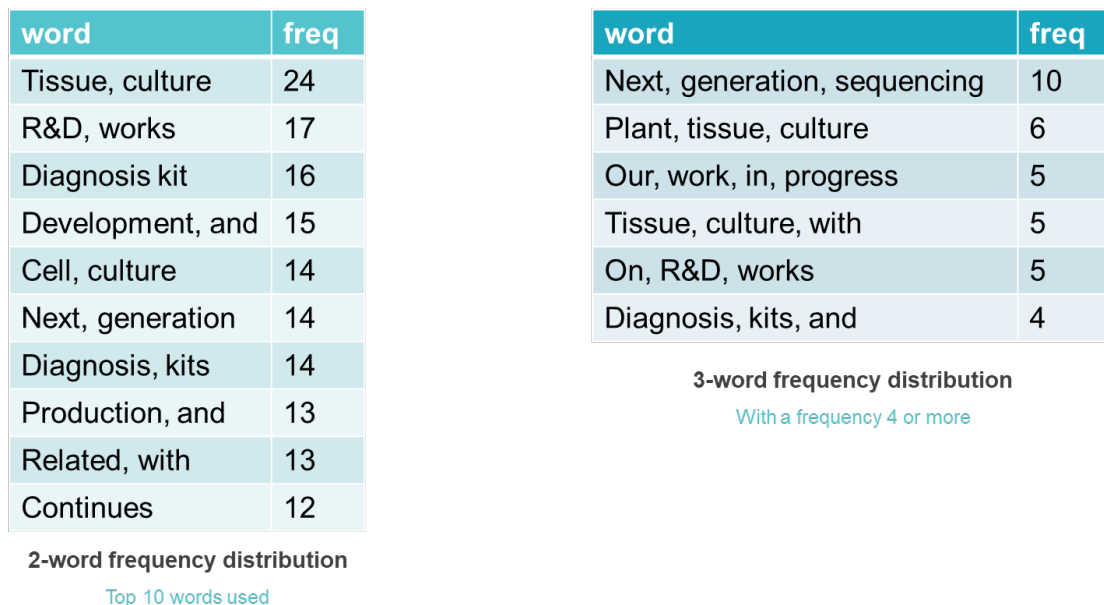


**Figure 7. Trigram collocations**

Then, the most used words in the texts entered into the open-ended question, were found as single, double and triple before and after removing stop words. The most important finding here was that after removing stop words, the three-word phrase "plant tissue culture" was repeated six times. The

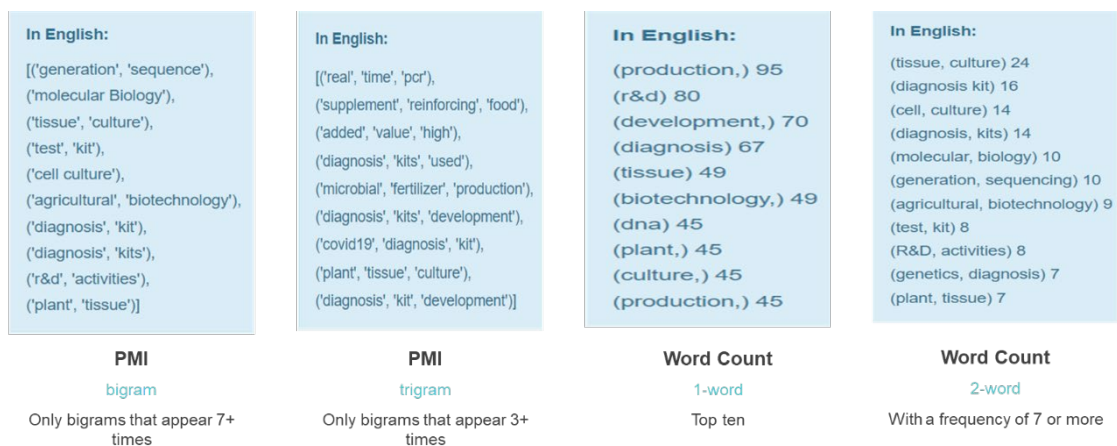


results of word counts are shared below figure. Note that, as a result of one-word count, since stop words have dominated the frequency table they are not included in the below figure.



**Figure 8. Word counts**

Collocations and word counts were obtained again after stop word elimination:



Note: the most frequent 3-word is (plant, tissue, culture) 6.

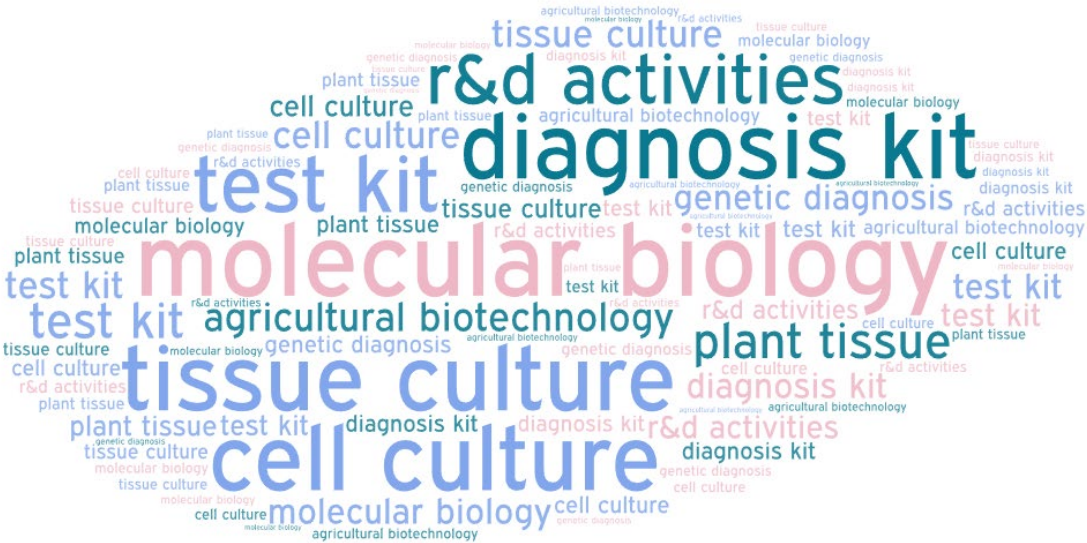
**Figure 9. Collocations and word counts after eliminating stop words**

Finally, the top2vec algorithm was used to find similar words used in free texts and to derive topic titles from these word groups. Top2vec derived four topics and also visualized the similar words it grouped.



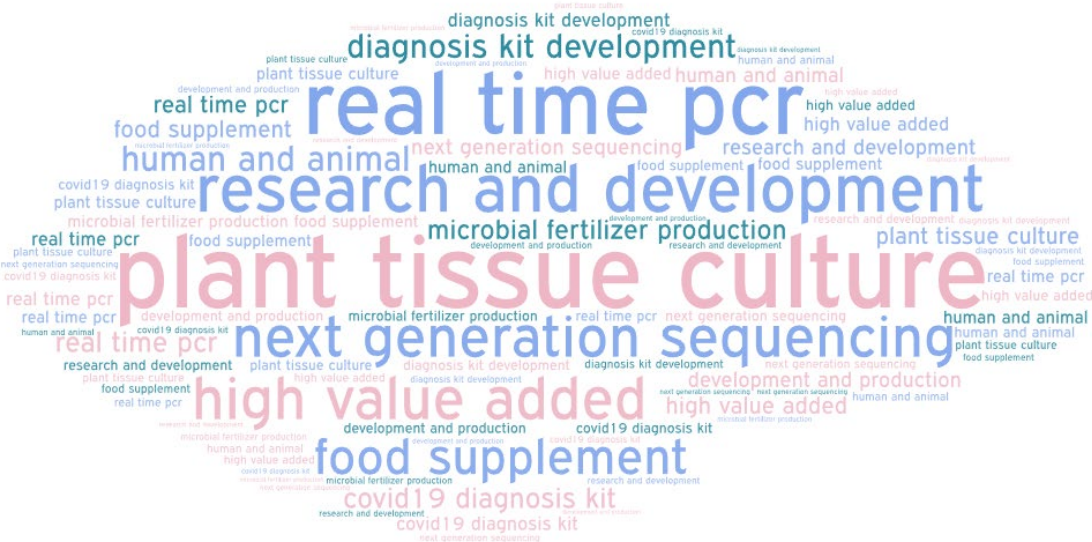


When we dig a little deeper, biotechnology enterprises mention these two words the most in order of emphasis): molecular biology, tissue culture, diagnosis kit, cell culture, R&D activities, test kit, agricultural biotechnology, plant tissue and genetic diagnosis.



**Figure 12. most emphasized two words in open-ended texts**

NLP code outputs give a deeper insight into the biotechnology activities of the enterprises. As for three-word expressions, the most emphasized expressions in open-ended texts are: plant tissue culture, research and development, next generation sequencing, real time PCR, high value added, food supplement, diagnosis kit development, microbial fertilizer production, covid19 diagnosis kit, human and animal and development and production.



**Figure 13. most emphasized three words in open-ended texts**

Since the reference period of the Biotechnology Statistics survey is 2020, when the pandemic is more intense, the effect of covid-19 on the results can be easily seen (such as real time PCR, diagnosis kit development, covid19 diagnosis kit etc.).

As a result of this work, it was concluded that NLP and deep learning techniques can be used to reduce human effort in extracting meaning from responses to an open-ended free text. They can automate and speed up an otherwise laborious or infeasible task

REFERENCES

- [1] S. Bird, E. Klein and E. Loper, Natural Language Processing with Python, (2009), ix.
- [2] <https://jaimin-ml2001.medium.com/stemming-lemmatization-stopwords-and-n-grams-in-nlp-96f8e8b6aa6f>, accessed 23 September 2022
- [3] <https://www.nltk.org/howto/collocations.html>, accessed 23 September 2022
- [4] <https://www.nltk.org/>, accessed 23 September 2022
- [5] D. Angelov, Top2Vec: Distributed Representations of Topics, (2020), arXiv.org, accessed 23 September 2022

Project 3 – Create a framework for government R&D survey ([Link](#))

Introduction

I tried to find out if Research and Development (R&D) was mentioned on the web pages of the government units via web scraping. So that I can add them to the Government R&D survey framework. Main purpose was to create an evidence-based Government R&D Survey framework. This framework is currently created by including government units that have the potential to carry out R&D activities in the framework.

It is relatively easy to set up the survey framework for financial companies because the list of business enterprises receiving R&D support is available in the administrative registers. However, we do not know which institutions and/or organizations carry out R&D activities on the government R&D side. For this reason, I am trying to find out those who use the term “R&D” on the internet pages of public institutions and how many times this R&D expression is mentioned. So that; if the Research and Development phrase is mentioned on their web pages, I can include them in the government R&D framework. Thus, both the response burden and the cost are reduced.

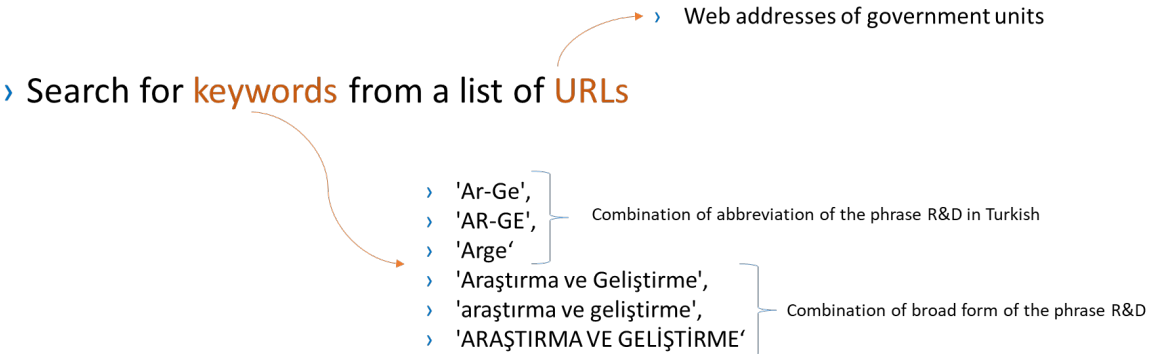


Figure 14. Project code summary

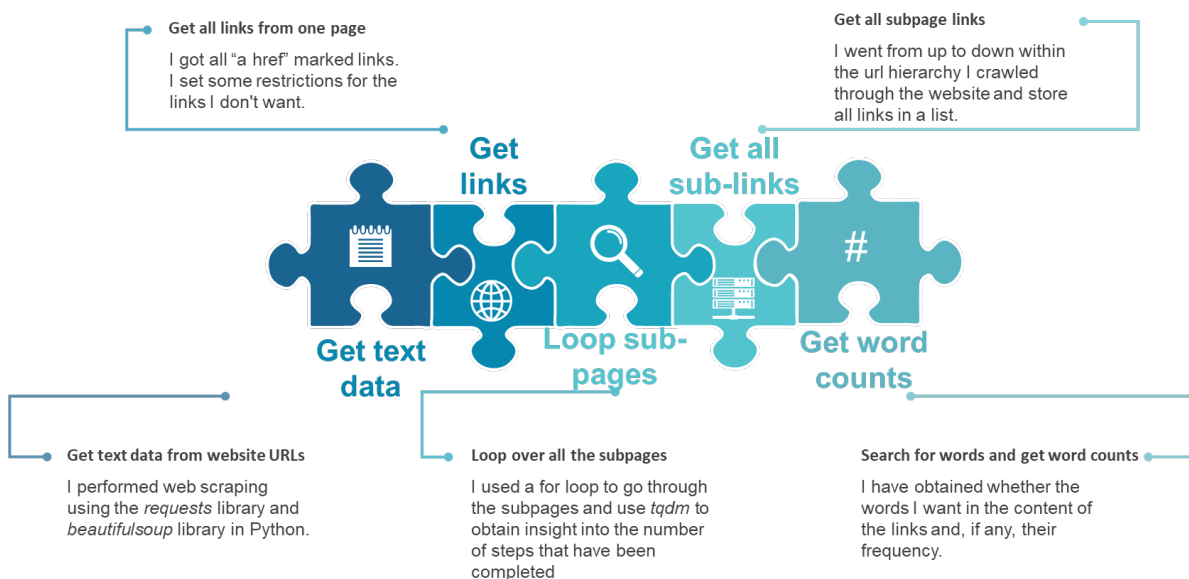
## Work steps followed

For this purpose, I developed a project with web scraping in python that searches for keywords from a list of URLs. The project consists of two main parts: searching the links and scraping the keywords in the found links. In the first part, I took advantage of [the article](#) by Kelvin Kramp (accessed 18 October 2022) and by modifying it I performed the following steps in order:

- Import necessary modules/libraries
- Write a function for getting the text data from a website URL
- Write a function for getting all links from one page and store them in a list
- Write a function that loops over all the subpages
- Import the URL list and create the loop to get all “a href” marked links.

In the second part, the code executed the following steps in order:

- Import previously produced json data containing all the links found
- Get links from imported json data
- Import the keyword list and scrape the list of words from URLs
- Print the url's scraped and cumulative number of found words
- Show and export only the links containing keywords with their frequency



**Figure 15. Process steps**

The code takes a list of URLs as input and performs web scraping on the web pages and their subpages in that list for a list of words/phrases, which is also a separate input. This is actually a generic application: It reports how many of the word/word groups you entered as a list were detected within the URLs you entered as a list, at the URL level. By default, the program uses two iterations to detect subpages. This iteration number can be changed as some URLs might require more iterations than others.

## Conclusion

In conclusion, with this project, it is possible to learn which government units (ministries, general directorates, hospitals, municipalities, etc.) mention R&D on their websites and how often. Thus, we have an evidence-based knowledge of whether to include them in the R&D survey framework or not.



At the same time, since this is a generic application, it can be used easily when certain words or word groups are searched on specified web pages.

## Overall Conclusion

The ever-growing supply of data and demand for timely and high-quality statistics are challenging governments to transform the way they produce official statistics. Traditional methods and resource alone are not enough to exploit the vast potential of this data. Significant investment in both technology and capability will be needed to harness the opportunities from administrative data and other sources of big data. National Statistical Offices around the world have stepped up their response in recent years, with more organisations establishing their own specialist data science functions to conduct exploratory research and build capability.

As more statistical organisations develop their own data science strategy and functions, it is important for them to learn from each other's experiences and identify ways to drive forward development.

Joining the ML group 2022 and being an implementer in the web scraping theme group provided a great opportunity for me and representatives from different organisations and job functions to reflect on the progress that has been made in building data science capability in official statistics in the past years, discuss common challenges, and learn about different approaches to tackling them.

In the past twelve months it has been run a varied programme of research, knowledge exchange and capacity building activities in the use of ML for official statistics. I found this group a valuable and enjoyable platform for connecting with colleagues working on ML across the world. And the skills and knowledge I built here have direct benefits for my work and organisation.