Individual report: Statistics Poland

Statistics on companies undertaking activities in the field of corporate social responsibility (CSR) using web scraping and machine learning

Organisation: Statistics Poland

Authors: Bartosz Grancow,
Emilia Murawska,
Klaudia Peszat

Date: 15.11.2022

Version: 2.0

Introduction

The traditional data sources do not provide information on many new and fast changing socio-economic phenomena, therefore national statistical offices more often reach for new data sources, such as web data.

Statistics Poland within the ML2022 web scraping theme group made an attempt to explore web data sources and new machine learning techniques to recognize the possibilities to augment official statistics on enterprises. The experimental research was focused on Corporate Social Responsibility activities.

Corporate Social Responsibility (CSR) is defined as „a company's sense of responsibility towards the community and environment (both ecological and social) in which it operates.

Companies express this citizenship:

(1) through their waste and pollution reduction processes,
(2) by contributing educational and social programs, and
(3) by earning adequate returns on the employed resources" (*the Business Dictionary*).

The report presents the results of the research which aimed at the classification of companies into two groups: those which carry out activities in a CSR field (and inform about it) and those which more likely do not have any CSR strategies or at least do not communicate their activities in this area. The type of such activity was not the subject of the project.

Data

Input data

The data source used to build the population of companies was the Business Register – Database of Statistical Units (BJS), including name of enterprise, address, e-mail address, URL address (if available). Due to the fact that CSR-related activities are most often carried out by large and medium-sized companies (small companies often do not have websites), we focused our attention on this group of enterprises (employing more than 10 employees).

The first step was the extraction from the Business Register of the population on enterprises with URL addresses. It is worth mentioning that, the URLs can be also obtained from email address. The methodology for obtaining URLs among others from email addresses has been described in detail in the report delivered within the ESSnet Trusted Smart Statistics – Web Intelligence Network project (Kuhnemann et al. 2022). Taking into account the exploratory purpose of our research and the fact that the database of large and medium companies covered over 20 thous. enterprises with URL addresses, we did not decide to obtain additional URLs via email addresses. The total number of URLs in our database consisted of 6 thous. URLs for large enterprises and nearly 14 thous. for medium ones from all over Poland.

Next, we checked validity of the URLs using web scraping tools. There were 11,380 valid URLs (with response code 200). Inactive URLs were removed from the database.

Search engine results

We have decided that in a scope of our interests are only websites held by the companies to provide information on their business. The enterprise should have control over the content of such websites. External web pages providing basic information on the companies, such as yellow pages, government domains etc. were excluded.

Information on CSR activities can be found in various places on the website. Some enterprises post them on the home page, others have a dedicated tab or subpage. The descriptions of the CSR activities conducted by companies are also published in various ways. Some enterprises publish them as a plain text on the website, others attach pdf reports.

Taking into consideration the enormous variety of websites, we decided to use an internet search engine to find texts from the official company's web sites referring to their activities in the CSR field. Only search results (snippets) were used as data source and no additional scraping was done. This approach has been implemented first by CBS (Delden et al. 2019). This enabled to save time and resources needed to web scrape the entire content of webpages.

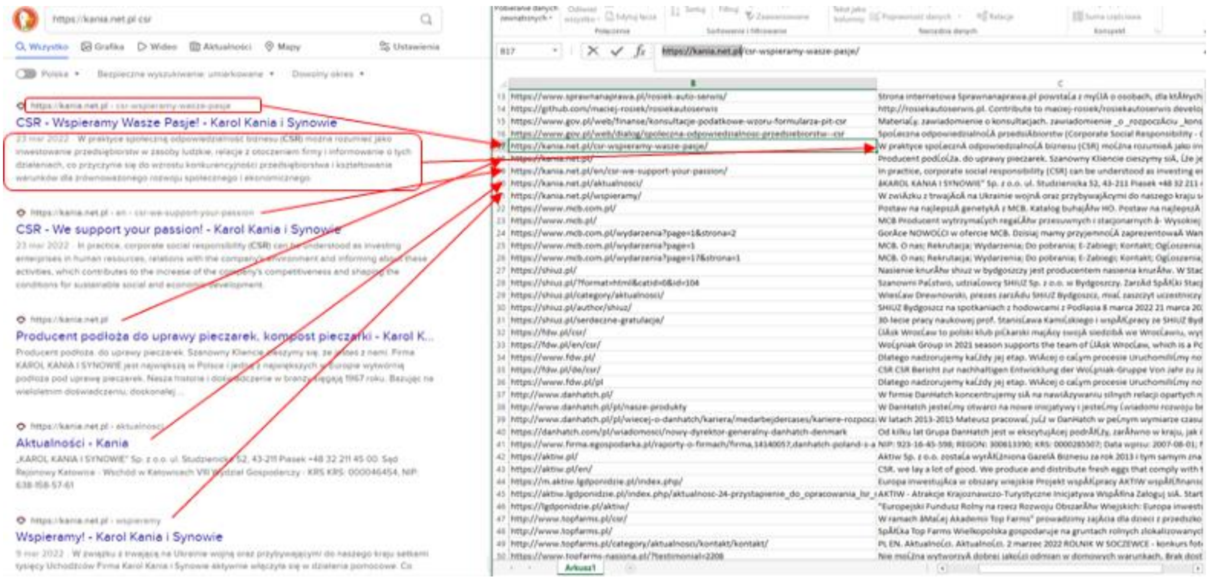
The choice of the search engine and the search term have significant impact on the results. However, due to the exploratory purpose of the research and payment requirements for using popular search engines, such as Google or Bing for mass queries, we decided to use the less known search engine – Duck Duck Go[1]. Its advantage is not collecting data about users, and thus not profiling the results. Each user can get the same search results, which made it easier to work on the project.

We also tested various search options by entering company's URL and keyword or name of a company and keyword. The first variant brought more accurate results. When entering a company's name, contact details were usually in the first places of the search.

To prepare a catalogue of keywords that appeared most often in the descriptions of CSR activities we also searched manually on companies websites. The directory consisted of words and phrases such as: 'csr', 'społeczna odpowiedzialność', 'odpowiedzialność biznesu', 'społeczna', 'zaangażowanie', 'środowisko', 'ekologiczne', 'zrównoważony rozwój', 'csr raport', 'charytatywny'. We tested all above mentioned keywords, however the search term 'URL address' + keyword: 'csr' gave the best results.

We also noticed that if a company conducts CSR activities, information about it appears in the first five search results (most often in the first three descriptions, but there are some cases when it appears in the fourth or fifth ones). For this reason, we limited the search to the first five results.

Below is an example of the search results – URL address and keyword entered manually in the search engine Duck Duck Go and the results of descriptions retrieved and saved in an Excel file.



https://kania.net.pl › csr-wspieramy-wasze-pasje
CSR - Wspieramy Wasze Pasje! - Karol Kania i Synowie
23 mar 2022 · W praktyce społeczną odpowiedzialność biznesu (CSR) można rozumieć jako inwestowanie przedsiębiorstw w zasoby ludzkie, relacje z otoczeniem firmy i informowanie o tych działaniach, co przyczynia się do wzrostu konkurencyjności przedsiębiorstwa i kształtowania warunków dla zrównoważonego rozwoju społecznego i ekonomicznego.

https://kania.net.pl › en › csr-we-support-your-passion
CSR - We support your passion! - Karol Kania i Synowie
23 mar 2022 · In practice, corporate social responsibility (CSR) can be understood as investing enterprises in human resources, relations with the company's environment and informing about these activities, which contributes to the increase of the company's competitiveness and shaping the conditions for sustainable social and economic development.

https://kania.net.pl
Producent podłoża do uprawy pieczarek, kompost pieczarki - Karol K...
Producent podłoża. do uprawy pieczarek. Szanowny Kliencie cieszymy się, że jesteś z nami. Firma KAROL KANIA I SYNOWIE jest największą w Polsce i jedną z największych w Europie wytwórnią podłoża pod uprawę pieczarek. Nasza historia i doświadczenie w branży sięgają 1967 roku. Bazując na wieloletnim doświadczeniu, doskonalej ...

https://kania.net.pl › aktualnosci
Aktualności - Kania
„KAROL KANIA I SYNOWIE" Sp. z o.o. ul. Studzienicka 52, 43-211 Piasek +48 32 211 45 00. Sąd Rejonowy Katowice - Wschód w Katowicach VIII Wydział Gospodarczy · KRS KRS: 000046454, NIP: 638-158-57-61

https://kania.net.pl › wspieramy
Wspieramy! - Karol Kania i Synowie
9 mar 2022 · W związku z trwającą na Ukrainie wojną oraz przybywającymi do naszego kraju setkami tysięcy Uchodźców Firma Karol Kania i Synowie aktywnie włączyła się w działania pomocowe. Co

https://kania.net.pl/csr-wspieramy-wasze-pasje/

	B	C
13	https://www.sprawnanaprawa.pl/rosiek-auto-serwis/	Strona internetowa Sprawnanaprawa.pl powstaĹa z myĹlĂ o osobach, dla ktĂłrych
14	https://github.com/maciej-rosiek/rosiekautoserwis	http://rosiekautoserwis.pl. Contribute to maciej-rosiek/rosiekautoserwis develo
15	https://www.gov.pl/web/finanse/konsultacje-podatkowe-wzoru-formularza-pit-csr	MateriaĹy: zawiadomienie o konsultacjach. zawiadomienie _o _rozpoczÄciu _kons
16	https://www.gov.pl/web/dialog/spoleczna-odpowiedzialnosc-przedsiebiorstw---csr	SpoĹeczna odpowiedzialnoĹÄ przedsiÄbiorstw (Corporate Social Responsibility - (
17	https://kania.net.pl/csr-wspieramy-wasze-pasje/	W praktyce spoĹecznÄ odpowiedzialnoĹÄ biznesu (CSR) moĹźna rozumieÄ jako in
18	https://kania.net.pl/	Producent podĹoĹźa. do uprawy pieczarek. Szanowny Kliencie cieszymy siÄ, Ĺźe je
19	https://kania.net.pl/en/csr-we-support-your-passion/	In practice, corporate social responsibility (CSR) can be understood as investing e
20	https://kania.net.pl/aktualnosci/	âKAROL KANIA I SYNOWIE" Sp. z o.o. ul. Studzienicka 52, 43-211 Piasek +48 32 211
21	https://kania.net.pl/wspieramy/	W zwiÄzku z trwajÄcÄ na Ukrainie wojnÄ oraz przybywajÄcymi do naszego kraju s
22	https://www.mcb.com.pl/	Postaw na najlepszÄ genetykÄ z MCB. Katalog buhajĂłw HO. Postaw na najlepszÄ
23	https://www.mcb.pl/	MCB Producent wytrzymaĹych regaĹĂłw przesuwnych i stacjonarnych â Wysokiej
24	https://www.mcb.com.pl/wydarzenia?page=1&strona=2	GorÄce NOWOĹCI w ofercie MCB. Dzisiaj mamy przyjemnoĹÄ zaprezentowaÄ Wan
25	https://www.mcb.com.pl/wydarzenia?page=1	MCB. O nas; Rekrutacja; Wydarzenia; Do pobrania; E-Zabiegi; Kontakt; OgĹoszenia;
26	https://www.mcb.com.pl/wydarzenia?page=17&strona=1	MCB. O nas; Rekrutacja; Wydarzenia; Do pobrania; E-Zabiegi; Kontakt; OgĹoszenia;
27	https://shiuz.pl/	Nasienie knurĂłw shiuz w bydgoszczy jest producentem nasienia knurĂłw. W Stac
28	https://shiuz.pl/?format=html&catid=0&id=104	Szanowni PaĹstwo, udziaĹowcy SHIUZ Sp. z o.o. w Bydgoszczy. ZarzÄd SpĂłĹki Stac
29	https://shiuz.pl/category/aktualnosci/	WiesĹaw Drewnowski, prezes zarzÄdu SHIUZ Bydgoszcz, miaĹ zaszczyt uczestniczy
30	https://shiuz.pl/author/shiuz/	SHIUZ Bydgoszcz na spotkaniach z hodowcami z Podlasia 8 marca 2022 21 marca 20
31	https://shiuz.pl/serdeczne-gratulacje/	30-lecie pracy naukowej prof. StanisĹawa KamiĹskiego i wspĂłĹpracy ze SHIUZ Byd
32	https://fdw.pl/csr/	ĹlÄsk WrocĹaw to polski klub piĹkarski majÄcy swojÄ siedzibÄ we WrocĹawiu, wy
33	https://fdw.pl/en/csr/	WoĹşniak Group in 2021 season supports the team of ĹlÄsk WrocĹaw, which is a Po
34	https://www.fdw.pl/	Dlatego nadzorujemy kaĹźdy jej etap. WiÄcej o caĹym procesie UruchomiliĹmy no
35	https://fdw.pl/de/csr/	CSR CSR Bericht zur nachhaltigen Entwicklung der WoĹşniak-Gruppe Von Jahr zu J
36	https://www.fdw.pl/pl	Dlatego nadzorujemy kaĹźdy jej etap. WiÄcej o caĹym procesie UruchomiliĹmy no
37	http://www.danhatch.pl/	W firmie DanHatch koncentrujemy siÄ na nawiÄzywaniu silnych relacji opartych n
38	http://www.danhatch.pl/pl/nasze-produkty	W DanHatch jesteĹmy otwarci na nowe inicjatywy i jesteĹmy Ĺwiadomi rozwoju be
39	http://www.danhatch.pl/pl/wiecej-o-danhatch/kariera/medarbejdercases/kariere-rozpocz	W latach 2013-2015 Mateusz pracowaĹ juĹź w DanHatch w peĹnym wymiarze czasu
40	https://danhatch.com/pl/wiadomosci/nowy-dyrektor-generalny-danhatch-denmark	Od kilku lat Grupa DanHatch jest w ekscytujÄcej podrĂłĹźy, zarĂłwno w kraju, jak i
41	https://www.firma.egospodarka.pl/raporty-o-firmach/firma,14140057,danhatch-poland-s-a	NIP: 923-16-45-598; REGON: 300613390; KRS: 0000285507; Data wpisu: 2007-08-01;
42	https://aktiw.pl/	Aktiw Sp. z o.o. zostaĹa wyrĂłĹźniona GazelÄ Biznesu za rok 2013 i tym samym zna
43	https://aktiw.pl/en/	CSR. we lay a lot of good. We produce and distribute fresh eggs that comply with t
44	https://m.aktiw.lgdponidzie.pl/index.php/	Europa inwestujÄca w obszary wiejskie Projekt wspĂłĹpracy AKTIW wspĂłĹfinans
45	https://aktiw.lgdponidzie.pl/index.php/aktualnosc-24-przystapienie_do_opracowania_lsr_	AKTIW - Atrakcje Krajoznawczo-Turystyczne Inicjatywa WspĂłlna Zaloguj siÄ. Start
46	https://lgdponidzie.pl/aktiw/	"Europejski Fundusz Rolny na rzecz Rozwoju ObszarĂłw Wiejskich: Europa inwest
47	http://www.topfarms.pl/csr/	W ramach âMaĹej Akademii Top Farms" prowadzimy zajÄcia dla dzieci z przedszko
48	http://www.topfarms.pl/	SpĂłĹka Top Farms Wielkopolska gospodaruje na gruntach rolnych zlokalizowanyc
49	http://www.topfarms.pl/category/aktualnosci/kontakt/kontakt/	PL EN. AktualnoĹci. AktualnoĹci. 2 marzec 2022 ROLNIK W SOCZEWCE - konkurs fot
50	https://www.topfarms-nasiona.pl/?testimonial=2208	Nie moĹźna wytworzyÄ dobrej jakoĹci odmian w domowych warunkach. Brak dost

Figure 1. Search engine results

Data preparation

Data stored in an Excel file had to be further processed. At this step, results which did not come from company's' webpages were removed from the database. It was: yellow pages, phonebook websites (e.g. panoramafirm.pl, oferteo.pl, fimy-pl.com), job offer pages (e.g. gowork.pl, pulshr.pl), social media (e.g. Facebook, Linkedin), social campaign pages (e.g. kampaniespoleczne.pl), non-governmental organizations pages (addresses with the org.pl domain), as well as government and educational institution pages (addresses with the gov.pl and edu.pl domain).

In the next step, the pre-processing techniques, such as tokenization, reduction of words to their basic form by lemmatization and stop words removal were applied.

Machine learning solution

Model Top2Vec

Once the data was prepared, we attempted to detect proximity of texts from our database with the CSR topic. For this purpose, we applied a new unsupervised machine learning algorithm – the Top2Vec. The Top2Vec algorithm is used to automatically detect topics present in texts and documents. This algorithm was chosen as it enables working on short text (such as snippets), does not require removing stop words and pre-processing (e.g. stemming/lemmatization)[2].

Despite the fact that the Top2Vec algorithm does not require previous text processing, some of the pre-processing steps were done, mostly to compare the model results applied on the raw data and pre-processed ones.

The default embedding model applied in the Top2Vec algorithm which enables creation jointly embedded document and word vectors is Doc2vec. This is not a pre-trained model, it learns from scratch and thus it can be used for different languages. The algorithm provides also a possibility to use pre-trained embedding models, such as: universal-sentence-encoder, which currently supports 16 languages, including Polish. For the experimental research purposes different embedding models were tested, however universal-sentence-encoder-multilingual-large model proved to give the best results.

Software used

The implementation of the Top2Vec model was carried out in a Python programming language. The installation of the library seemed not to be very challenging, however several adjustments of the IT environment were necessary. First of all, Microsoft Visual C++ 14.0 or greater was required, and secondly – installation of hdbscan. The latter problem can be easily solved in Anaconda environment by running the code: conda install -c conda-forge hdbscan.

Results

Raw data

The Top2Vec algorithm was applied on the dataset consisting of 13.4 thous. raw texts. The model detected 73 topics, including one very closely related to the CSR concept (Topic 0). It consists of 477 semantically similar texts, which also makes it the largest cluster. The wordcloud for the topic 0 is presented below.

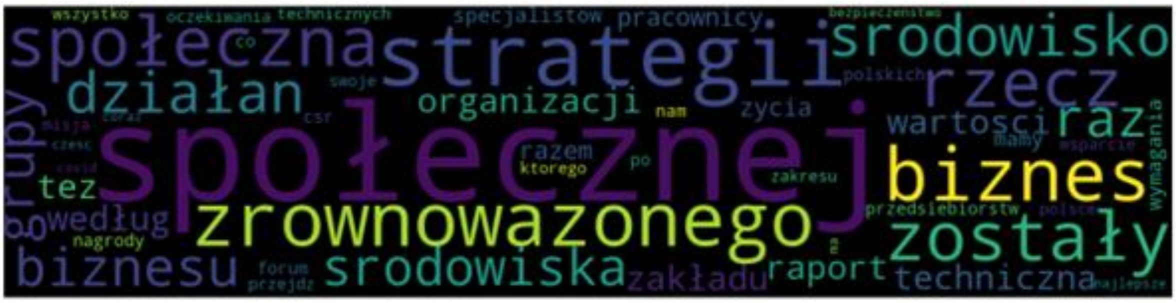
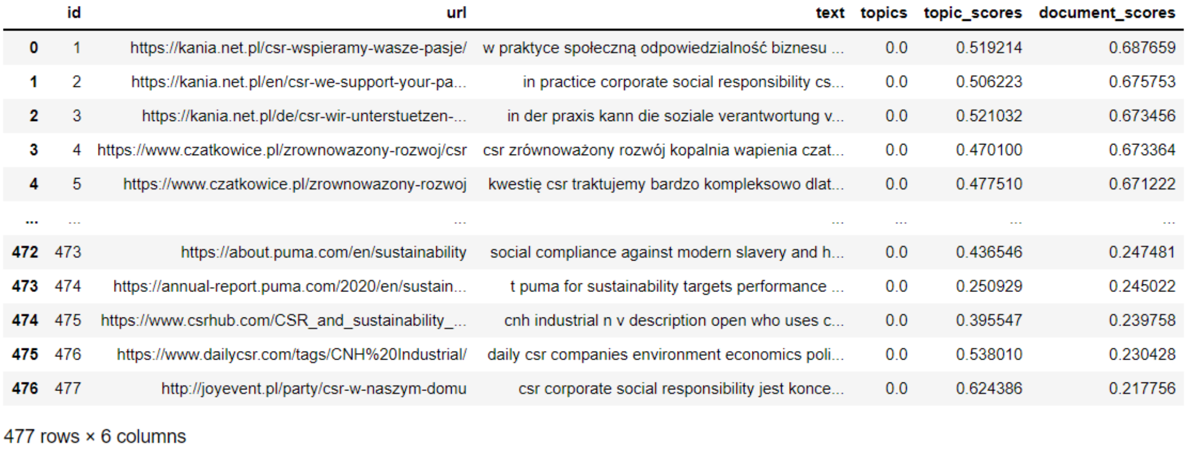


Figure 2. The wordcloud for the topic 0 (in Polish)

In the next step, we calculated the document scores, which indicate a semantic similarity of documents to the assigned topic. The table below presents the document scores for the texts clustered in the topic 0 (in descending order).



	id	url	text	topics	topic_scores	document_scores
0	1	https://kania.net.pl/csr-wspieramy-wasze-pasje/	w praktyce społeczną odpowiedzialność biznesu ...	0.0	0.519214	0.687659
1	2	https://kania.net.pl/en/csr-we-support-your-pa...	in practice corporate social responsibility cs...	0.0	0.506223	0.675753
2	3	https://kania.net.pl/de/csr-wir-unterstuetzen-...	in der praxis kann die soziale verantwortung v...	0.0	0.521032	0.673456
3	4	https://www.czatkowice.pl/zrownowazony-rozwoj/csr	csr zrównoważony rozwój kopalnia wapienia czat...	0.0	0.470100	0.673364
4	5	https://www.czatkowice.pl/zrownowazony-rozwoj	kwestię csr traktujemy bardzo kompleksowo dlat...	0.0	0.477510	0.671222
...	...	...	...	...	...	...
472	473	https://about.puma.com/en/sustainability	social compliance against modern slavery and h...	0.0	0.436546	0.247481
473	474	https://annual-report.puma.com/2020/en/sustain...	t puma for sustainability targets performance ...	0.0	0.250929	0.245022
474	475	https://www.csrhub.com/CSR_and_sustainability_...	cnh industrial n v description open who uses c...	0.0	0.395547	0.239758
475	476	https://www.dailycsr.com/tags/CNH%20Industrial/	daily csr companies environment economics poli...	0.0	0.538010	0.230428
476	477	http://joyevent.pl/party/csr-w-naszym-domu	csr corporate social responsibility jest konce...	0.0	0.624386	0.217756

477 rows × 6 columns

Figure 3. The results for the topic 0

The results can also be easily assessed in a scatter plot where the x axis is the document id and y axis is the document score.

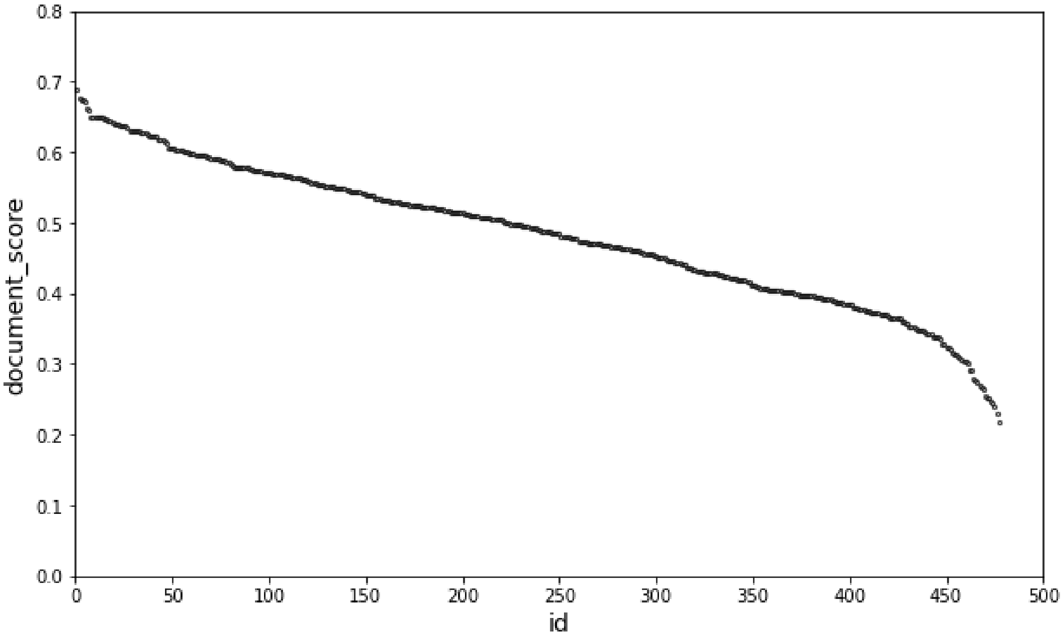


Figure 4. The scatter plot for the topic 0

The table and scatter plot contain the number of texts from companies websites, thus to calculate the number of companies which more likely conduct CSR activities we had to deduplicate information.

The number of unique companies which are assigned to the topic 0, closely related to the CSR concept is 280 (2,5% of all companies).

Pre-processed data

The same process was carried out on the pre-processed data. In this case, the model detected 93 topics, from which topic 2 seemed to be the most related to the CSR concept. It was the third topic taking into account its size (379 documents).

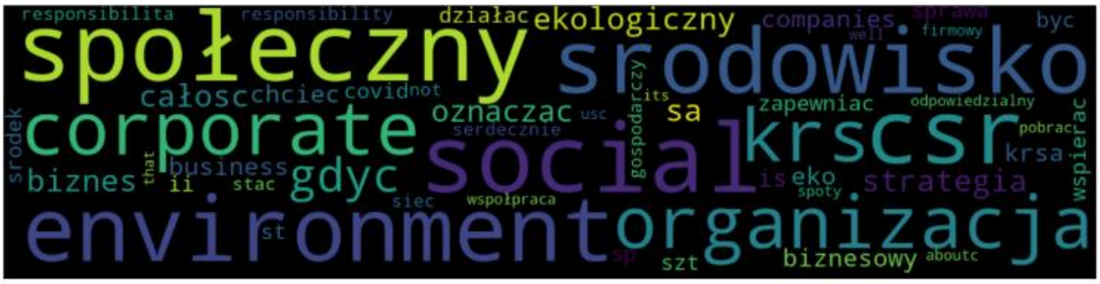


Figure 5. The wordcloud for the topic 2 (in Polish)

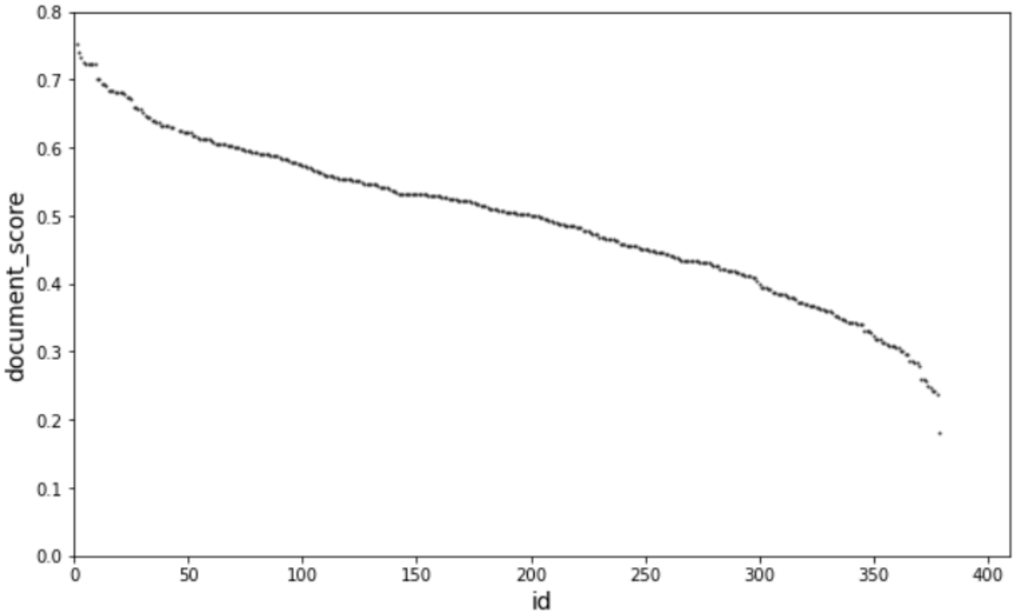


Figure 6. The scatter plot for the topic 2

The number of unique companies which are assigned to the topic 2, closely related to the CSR concept is 246 (2,2% of all companies).

The above means, that the results of the model trained on the raw data and on the pre-processed dataset are very similar. The random assessment of the assignments of texts to the CSR-related topics confirmed these outcomes, however further research as well as more advanced validation methods should be used when considering production of official statistics.

Reproducibility of the results

Basically, the Top2Vec model is stochastic, thus to achieve the reproducibility of the results for the same dataset it is necessary to set a random seed in the UMAP algorithm. Unfortunately, the Top2Vec library does not provide the possibility to set this parameter via API and it can be only done in the source code[3].

References

Delden, Arnout van; Windmeijer, Dick; ten Bosch, Olav (2019): *Finding enterprise websites*. Bilbao (European Establishment Statistics Workshop).

Kuhneman, Heidi; van Delden, Arnout; Summa, Donato; Gussenbauer, Johannes; Ils, Alexandra, Loytynoja, Katja (2022): *URL finding methodology*. Joint report for Work Package 2 (Online Based Enterprise Characteristics) and Work Package 3, Use Case 5 (Business register quality enhancement) V.5.0. ESSnet Trusted Smart Statistics – Web Intelligence Network. https://ec.europa.eu/eurostat/cros/content/url-finding-methodology_en