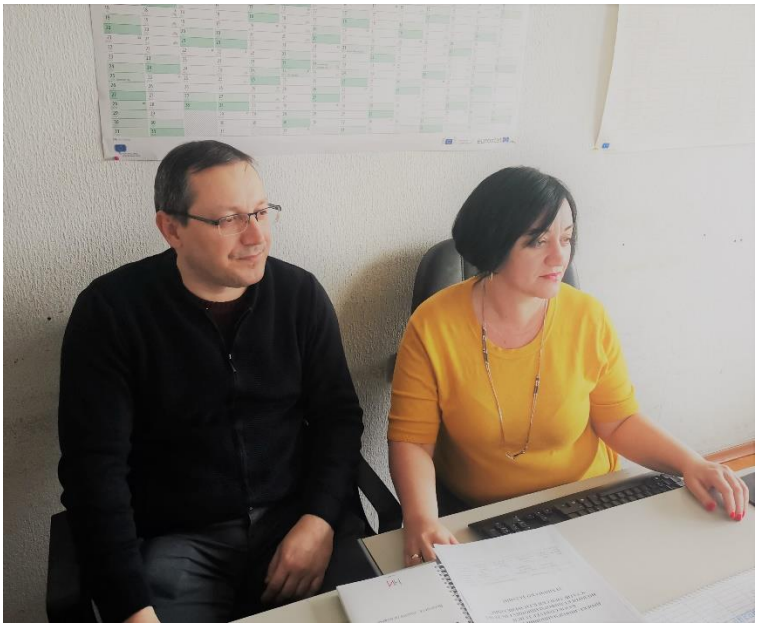
ПРОИЗВОДСТВО НА ЕКСПЕРИМЕНТАЛНА СТАТИСТИКА ЗА ХАРАКТЕРИСТИКИ НА ПРЕДПРИЯТИЯТА С ДАННИ ОТ ИНТЕРНЕТ

Костадин Георгиев[*], Галя Статева*[**]



I. Въведение

С всяка изминала година онлайн бизнесът става все по-важен, а с наличието на глобалната пандемия COVID е по-важен от всякога. Работният пакет С (WPC) за онлайн базирани характеристики на предприятието (OBEC[1]) в рамките на европейския проект ESSnet on Big data II[2] е свързан с разбирането на онлайн икономическата и бизнес активността на предприятията от гледна точка на националната статистика.

Производството на официална статистика за бизнес характеристиките обикновено е резултат от провеждането на класическо статистическо изследване и/или административни данни. Статистическото изследване има някои недостатъци като увеличаване на тежестта за респондентите или реализиране на високи разходи за националните статистически организации или други национални органи на статистиката. От своя страна, административните данни може да не включват всички променливи, необходими за производството на всеки статистически продукт, а наличните променливи понякога имат значително забавяне във времето. Независимо от това Статистическият

[*] Главен експерт в отдел „Информационни системи и приложен софтуер“, НСИ; e-mail: kgeorgiev@nsi.bg
[**] Държавен експерт в дирекция „Обща методология, координация и анализ на статистическите изследвания“, НСИ; e-mail: gstateva@nsi.bg.

[1] **O**nline **B**ased **E**nterprise **C**haracteristics (OBEC)
[2] https://ec.europa.eu/eurostat/cros/content/essnet-big-data-1_en

бизнес регистър (СБР) обикновено се използва като рамка за формиране на извадки, анализ на неотговорилите единици и калибриране, както и за статистически оценки. Извлечените от интернет мрежата данни от уебсайтовете на предприятията имат потенциала да „смекчат“ недостатъците в двата вече добре познати източника на данни: статистическото изследване и административните регистри. Процесът по извличане на данни от корпоративните уебсайтове включва незначителна тежест за предприятието (респондента) и използването на актуални „извлечени“ онлайн данни може да доведе до получаване на навременни характеристики на предприятието. Обаче включването на „извлечени“ данни от интернет за конкретен и понякога вече съществуващ статистически продукт определено не е лесна задача. Например може да не е възможно да се свърже еднозначно дадена уебстраница с предприятие, тъй като едно предприятие може да използва много уебстраници или една уебстраница може да се използва от много предприятия.

Ключов резултат от работата по WPC е подобряването на качеството на статистическия бизнес регистър по отношение на характеристики за онлайн присъствието на национално регистрираните фирми, като наличието на уебсайтове, електронна търговия или акаунти в социални медии. Тази проста иновация е от значение за всички членове на Европейската статистическа система (ЕСС), тъй като всяка национална статистическа институция поддържа СБР и едновременно с това е мощно средство, тъй като СБР вече са свързани с набори от данни, които са в основата на икономиката и бизнеса на всяка страна. Това означава, че интеграцията на данните от интернет мрежата в СБР е автоматична, незабавна и безпроблемна.

Основната цел на WPC да използва техники за web-scraping, извличане на знания от текст (text mining) и статистически изводи за събиране и обработка на корпоративна информация с цел подобряване или актуализиране на съществуващата информация, като присъствие в интернет мрежата, вид на икономическата дейност, информация за адресна информация, структура на собствеността и др. в националните СБР, беше постигната успешно.

В рамките на WPC методологията от предишния проект ESSnet on Big data I беше обобщена и разширена с цел използване във всяка държава от ЕСС, като се вземе предвид разнообразието, необходимо за поддържане на различните случаи на използване (use-cases) в статистическата практика. Тъй като уебскрапингът е сравнително нов метод за извличане на данни за статистическите организации, който изисква необходимото внимание по отношение на защита на данните, беше разработен и публикуван ЕСС макет на политиката за уебскрапване[3], който съдържа основни правни и етични съображения, както и стабилен набор от принципи и практики, които официалните статистически организации могат да следват.

[3] ESS Web scraping policy template, https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WPC_Deliverable_C1_ESS_webscraping_policy_template_2019_07_15.pdf

Националният статистически институт имаше честта и привилегията да бъде водеща институция и активен партньор по изпълнението на дейностите в рамките на работен пакет С.

Настоящата статия има за цел да запознае читателя с постигнатите резултати относно възможностите и предизвикателствата за производство на експериментална статистика за онлайн характеристики на предприятията на европейско и национално ниво.

II. Статистически контекст на данните за онлайн базирани характеристики на предприятията

Данните, „извлечени“ от интернет (web scraped) могат да бъдат допълнителен източник за производство на статистически продукти, като за тази цел е необходимо да бъдат предварително дефинирани основните използвани концепции и дефиниции. Акцентът е върху статистическите продукти, които са пряк резултат от жизнения цикъл на обработката на големи данни за получаване на онлайн характеристики на предприятието (дефинирани чрез т.нар. „случаи на използване“).

Основните елементи, които формират статистическите продукти, са: статистическа единица, целева съвкупност и наблюдавани променливи (на входно ниво), периодичността и статистическите показатели (на изходно ниво). Стандартизираните основни концепции са допълнителните елементи, гарантиращи, че разпространяваната експериментална статистика е хармонизирана и сравнима между страните от ЕСС.

Първоначално бяха дефинирани **пет „случаи на използване“** (use case):

1) Списък на URLs адреси (Uniform Resource Locator);
2) Променливи в изследването „Използване на ИКТ в предприятията“;
3) Данни за откриване на възникващи корпоративни класификации;
4) Експериментална статистика чрез невролингвистично програмиране;
5) Информация за всички предприятия, продаващи чрез платформата amazon.de.

Първият и вторият „случаи на използване“ са обусловени от контекста на изследването „Използване на ИКТ в предприятията“ и могат да бъдат внедрени в регулярното статистическо производство, докато третият, четвъртият и петият „случаи на използване“ по-скоро имат проучвателен характер, което не води непременно до производство на нова статистика. За апробиране на всеки use-case бяха разработени функционални производствени прототипи, където са описани подробно прилаганите методи, използваните софтуерни решения и работната процедура за изпълнение.

За производството на експериментална статистика за онлайн характеристиките на предприятията са изпълнени само първите два „случая на използване“, а именно: *Списък на URLs адреси* и *Променливи в изследването „Използване на ИКТ в предприятията“*.

Списъкът на URLs **(use case 1)** представлява база данни, съдържаща URL адреси за всяко предприятие в целевата съвкупност, където може да има един, много или нула

URL адреси за дадено предприятие. Списъкът с URL адреси може вече да е наличен в СБР от друг източник или доставчик на административни данни. В същото време този списък може да бъде изграден и от нулата, като се търсят предприятия чрез уебтърсачки като Bing или Google. Резултатите от търсенето могат да бъдат използвани като успешни опити за отговор на следните два въпроса:

- *Има ли предприятието уебсайт?*
- *Кой URL адрес е най-вероятно да принадлежи на предприятието?*

Обикновено уебтърсенето връща няколко резултата, водещи до различни базови URL адреси за едно предприятие. Основен URL адрес - например www.xyz.com или www.abc.com, може да бъде включен в няколко резултата от търсенето. Като правило уебтърсачката (web scraper) не е необходимо да търси в други страници на даден уеб-сайт освен на заглавната страница, тъй като информацията може да бъде намерена в основната страница или в раздел „Контакти“.

От даден списък с URL адреси уебсъдържанието може да бъде „извлечено“ с цел да се определи дали дадено URL принадлежи на конкретно предприятие. Ако извлечената информация е в съответствие със съхранената информация в СБР, тогава URL адресът е валиден; в противен случай URL адресът се заменя с URL с най-висока степен на доверителност или се премахва, ако резултатът е под някакъв предварително определен праг.

Тъй като при производството на официална статистика е ясно регламентирано, че могат да се съхраняват само релевантни данни, генерираните в процеса на скрапване данни могат да бъдат изтрити след изграждането на хранилището с URL адреси. Ако общото скрапване не е възможно поради правни проблеми, проблеми с връзката, блокиране на достъпа или др., тогава извличането на URL адреси от друг наличен набор от данни може също да бъде опция. В този случай ще са необходими прилагане на допълнителни методи и модели, както и разработване на специализиран класификатор в контекста на на алгоритмите за машинно самообучение (machine learning).

***Променливи в изследването „Използване на ИКТ в предприятията“ (*use case 2)** се фокусира върху извличането на информация от URL адрес на предприятие, съответстваща на променливите от традиционното статистическо изследване „Използване на ИКТ в предприятията“. Докато списъкът с URL адреси може да бъде създаден за всички предприятия от генералната съвкупност, то изследването „Използване на ИКТ в предприятията“ съдържа само предприятия с 10 или повече заети в дадени категории на NACE Rev. 2[4]. Извлечените онлайн базирани характеристики на предприятието трябва да отговарят на следните въпроси от изследването „Използване на ИКТ в предприятията“:

- *Уебсайтът на предприятието съдържа ли нещо от изброените функционалности: онлайн поръчки, резервации или електронен магазин (например уебсайтът има ли пазарска количка)?*

[4] Statistical classification of economic activities in the European Community (NACE).

- *Използва ли предприятието социални медии?*
- *Използва ли предприятието Twitter за конкретна цел:*

- *създаване на имидж или реклама на пазарни продукти на предприятието;*
- *набиране на персонал;*
- *други - всеки туит, който не се вписва в другите две категории.*

- *Има ли предприятието специфични характеристики на уебсайта като:*

- *описание на стоки или услуги, ценоразписи;*
- *възможност за посетителите да персонализират или проектират онлайн стоки или услуги;*
- *проследяване или състояние на направени поръчки;*
- *персонализирано съдържание в уебсайта за редовни/повтарящи се посетители;*
- *връзки или препратки към профилите на предприятието в социалните медии;*
- *обяви за свободни работни места или онлайн кандидатстване за работа.*

- *Работи ли предприятието върху предстоящи/нови явления, по-специално свързани с изкуствения интелект (AI) и машинното самообучение (ML).*

Последният въпрос не е част от въпросника на ИКТ изследването, но е включен като допълнителна характеристика, за да се идентифицират иновативните предприятия, които използват AI и ML в текущата си дейност.

1. Основни понятия

Основното значение при използването на скрапнати от уебмрежата данни като допълнителен източник за статистическа продукция се крие в процеса на извличане на онлайн базирани характеристики на предприятието (ОВЕС). Както подсказва името, ОВЕС може да се разглежда като всяка характеристика на предприятието, която е извлечена от корпоративните уебстраници с помощта на методите за скрапване. Например URL адресът на предприятието, ако съществува.

Извлечената информация може да бъде директно достъпна чрез друг доставчик на данни или да бъде събрана от самата статистическа организация. Процесът на събиране на данни изисква прилагане на софтуер за скрапване, който директно извлича съдържанието на URL или изобразява уебстраница чрез браузър и тогава извлича изобразеното съдържание. В зависимост от вида на ОВЕС различни части на уеб-страницата могат да бъдат подходящи за директно скрапване на интересуващото ни съдържание (ако е идентифицируемо), което би предотвратило натрупване на ненужно големи масиви от данни. Когато скрапнатите данни са вече събрани, извличането на ОВЕС включва или използване на детерминистичен подход, или прилагане на статистическо или машинно самообучение. Както детерминистичният, така и самообучителният подход трябва да бъдат щателно тествани и да са достигнали определено ниво на качество.

Като обобщение, процедурата за извличане на онлайн характеристики на предприятията може да бъде разделена на две основни части:

- Процедура по скрапване: извличане на информация за предприятия чрез скрапване на интернет мрежата - например URL адрес на предприятието (Uniform Resource Locator);

- Процес на извличане: извличане на ОВЕС с помощта на скрапваните данни от стъпка 1.

Извлечените онлайн характеристики на предприятията, ако процесът е бил успешен, след това могат да бъдат свързани със съществуващи единици (предприятия) в СБР и да се използват в статистическия бизнес процес или директно, или като допълнителен източник на данни.

2. Вид единица

Статистическият бизнес регистър е единственият източник за генериране на статистически единици за производство на официална бизнес статистика. В резултат на това се постига съгласувана икономическа статистика, съпоставима между секторите, страните, географските области и във времето. Информация от мрежата може да бъде получена за два вида статистически единици от СБР: за предприятия и групи предприятия.

Предприятието е най-малката комбинация от юридически единици, която е организационна единица, произвеждаща стоки или услуги, с известна степен на автономност при вземането на решения, особено за разпределение на текущите си ресурси. Всяко предприятие може да се състои от една или повече правни единици или да съдържа само част от правна единица. През последното десетилетие правната организация на предприятията се усложни, тъй като предприятията все повече следват тенденцията да диверсифицират своите дейности на множество правни единици.

Във фазата на идентификация на единиците основно се използва данъчният/ДДС код за потвърждаване на кореспондиращата връзка уебсайтове-предприятия. Трябва да се обърне внимание на случаите, когато в едно и също предприятие има повече от една юридическа единица, тъй като всяка от тях ще използва свой собствен данъчен код.

Група предприятия е набор от правни единици, правно и/или финансово обвързани, но с единно управление и контрол, т.е. управление, което определя общата корпоративна политика. Всяко предприятие в групата може да има свой собствен уебсайт, но е по-вероятно за групата предприятия да има единен уебсайт за цялата група, където цялата необходима информация да се намира по-лесно (заедно с връзките към уебсайтове на компаниите в групата).

Въпреки че основното предназначение на ОВЕС е да замести част от въпросите на изследването „Използване на ИКТ в предприятията“, единицата, за която се отнася ОВЕС, невинаги може да бъде самото предприятие. В такива случаи уебсайтът на предприятието може да се разглежда като респондент, за да се изведе статистическа оценка, отнасяща се до URL адресите на предприятията в определена целева съвкупност. Като се има предвид традиционното ИКТ изследване, някои променливи на наблюдение се отнасят не само до уебсайта на самото предприятие, но и до уебсайта на компанията майка или холдинг. По този начин един URL адрес може да доведе до ОВЕС за

множество предприятия. Накратко казано, единиците за извличане на онлайн характеристики са **предприятия и/или уебстраници.** За целите на коректните сравнения предприятието като единица за извличане на ОВЕС е в съответствие с Регламент 7 (ЕИО) № 696/93 на Съвета от 15 март 1993 г. относно статистическите единици за наблюдение и анализ на производствената система в Общността.

Уебсайтът на предприятието като единица за извличане на ОВЕС се определя като някакво цифрово решение (в World Wide Web), което едно предприятие има или предлага на своите клиенти.

3. Съвкупност и извадка от целевата съвкупност

Целевата съвкупност за официална бизнес статистика обикновено са всички единици в СБР или подсъвкупност от тях, какъвто е случаят с изследването „Използване на ИКТ в предприятията“. По-специално, когато се изчисляват статистически показатели от ОВЕС, може да се извлече оценка въз основа на подгрупата от съвкупността, съставена от предприятия, които имат един или повече URL адреси.

Първата задача е да се отчитат правилно събраните данни от интернет за единиците на съвкупността, т.е. данните, събрани от уебсайтове, трябва да се отнасят към единицата „предприятие“. Наличието на СБР, рамката на съвкупността, съдържаща всички предприятия, включени в целевата съвкупност, предполага използването на СБР като основа за търсене на съответните уебсайтове.

В зависимост от целта на изследването е по-подходящо производството на експериментална ОВЕС статистика да се ограничи до част от съвкупността, която представлява конкретна подгрупа от бизнес статистиката, вместо да се отнася до всички единици в СБР. В зависимост от вида на статистическия показател ОВЕС могат да бъдат извлечени за цялата целева съвкупност или само за извадка от нея.

След като се определи целевата съвкупност, от нея се конструира рамката на извадката. Наблюдава се, че разликата в структурата и съдържанието на уебсайтовете зависи до голяма степен от размера на предприятията, сложността на тяхната организация и икономическите дейности, които те извършват. Следователно, за да се оценят правилно резултатите от дейността по скрапване от интернет мрежата, се препоръчва да се направи стратифицирана извадка от СБР. Извличането на ОВЕС за тази извадка от целевата съвкупност може да бъде от полза по различни причини. Разходите по отношение на изчислителното време може да са значително по-ниски и би било възможно ръчно да се провери дали алгоритъмът за скрапване е дал коректни резултати. В зависимост от статистическия показател статистическата оценка на базата на извадката вече има потенциал да отговори на задължителните изисквания за качество.

Като правило се избират активни предприятия през референтната година, които имат юридическа форма на корпорация или партньорство. За да се подобри точността, стратифицирана извадка от тези единици в СБР може да бъде формирана от референтната съвкупност по няколко възможни начина, например чрез използване на пропорционално разпределение. По този начин броят на единиците във всяка страта е пропорционален на броя на единиците в общата съвкупност. Поради тази причина е

препоръчително да се работи с извадка от предприятия, стратифицирани както по размер по отношение на заетостта и оборота, така и по икономическа активност, за да се вземат предвид разликите в структурата и съдържанието на уебсайтовете, евентуално причинени от тези фактори.

Извадките на ОВЕС включва предприятия с потенциално свързани URL адреси, които имат процент на оценяване над даден праг (оценяване, дадено от онлайн търсачката за всяко предприятие). Това може да зависи още от потенциалните възможности на основния производствен процес - например ако технически е възможно да се скрапват много голям брой уебстраници по време на периода на наблюдение. Статистическият показател може също да бъде оценен с помощта на ОВЕС от извадката на целевата съвкупност и съответно да се калибрират теглата на извадката.

Подходът, основан на регистър към големите данни, категорично означава, че цялата статистическа, административна и уебинформация ще бъде „каталогизирана“ в СБР за многократно използване, като по този начин се гарантира последователност и ще даде конкретна подкрепа на статистическото производство въз основа на СБР.

В този смисъл има „двупосочен“ информационен поток към и от регистъра. От една страна, регистърът придобива нова информация по-рано от интернет мрежата и я каталогизира по последователен начин за множество цели, увеличавайки съдържанието и способността си да подпомага статистическото производство. От друга страна, данните от мрежата, попаднали в обхвата на регистъра, се интегрират с всички останали променливи от административни и статистически източници. По този начин неструктурираните данни от мрежата придобиват „структура“ и биха могли да се възползват от цялата налична статистическа класификация: скрапнатите данни получават име, размерност, местоположение и т.н., използвайки всички класификации в статистическия Бизнес регистър. Широкият обхват на СБР се превръща в поддържаща платформа за големи данни.

4. Периодичност

Когато ОВЕС се използват като алтернативен източник на данни за официална статистика, времевата рамка за извличане на ОВЕС трябва да бъде в съответствие с периода на наблюдение, както е дефиниран в методологичното ръководство на изследването „Използване на ИКТ в предприятията“. Периодичността обаче може да бъде увеличена, ако скрапването не натоварва много посетените URL адреси.

5. Променливи на наблюдение

За разлика от класическото изследване за използване на ИКТ в предприятията променливите не се наблюдават чрез събиране на отговори от въпросник, а чрез използване на интернет търсачки, приложно-програмен интерфейс (API) и/или софтуер за скрапване на данни, които потенциално съдържат информация за наблюдаваната променлива. От скрапнатите данни ОВЕС могат да бъдат извлечени или с помощта на алгоритми за машинно самообучение или чрез детерминистичен подход и от една или повече ОВЕС наблюдения. Например променливата „Има ли предприятието уебсайт“ е двумерна променлива със стойности 0/1 или „да“/„не“, която се извлича чрез използване

на списъкът с URL адреси, състоящ се от един, много или нула URL адреси за дадено предприятие.

Други потенциални ОВЕС производни, целеви променливи (да бъдат извлечени от сурови текстови данни, скрапнати от уебстраници) могат да бъдат следните:

- *Какво продават предприятията: основни продадени продукти/услуги;*
- *Как предприятията продават: канали за продажба (например карта за онлайн пазаруване, услуги за резервация, услуги за доставка);*
- *На кого продават предприятията - например бизнес към бизнес, бизнес към потребител;*
- *Къде продават предприятията: национални/мултинационални пазари.*

6. Статистически показатели

За производство на експерименатална статистика (за use case 1 и use case 2) бяха генерирани следните статистически показатели:

- Процент на предприятията, които имат уебсайтове;
- Процент на предприятията, занимаващи се с електронна търговия чрез своя уебсайт;
- Процент на предприятията, които присъстват в социалните медии;
- Процент на предприятията, използващи Twitter за конкретна цел;
- Процент на предприятията със специфични характеристики на уебсайта;
- Процент на предприятията, работещи по предстоящи/нови явления, по-специално AI и ML.

Всички изброени по-горе статистически показатели могат да бъдат директно оценени от съответните онлайн характеристики на предприятията.

В случая, когато ОВЕС е извлечена от извадка e_1,…,e_n от от цялата целева съвкупност и когато всяко предприятие e_i,i=1,…,n има кореспондиращо тегло в извадката w_i.

Единственият показател за ***use case 1*** е процентът на предприятията, които имат уебсайт:

$$R_{web} = \frac{\sum_{i=1}^{n} w_i \, I[u_i \neq \emptyset]}{\sum_{i=i}^{n} w_i} \qquad (1)$$

където $\sum_{i=i}^{n} w_i = N$ е броят на единиците в целевата съвкупност, $I[.]$ функция на показателя и u_i е набор от всички уебсайтове за единица i, който е празен, ако няма намерен поне един уебсайт. Този показател може лесно да бъде изчислен по NUTS[5] региони, NACE категории или групиран по броя на заетите в предприятието. Нека $A_1, \ldots, A_p$ бъдат части от цялата целева съвкупност $\mathcal{A}$ т.е.

$$\bigcup_{i=1}^{p} A_i = \mathcal{A} \text{ and } A_i \cap A_j = \emptyset \quad \forall i \neq j \qquad (2)$$

[5] Номенклатура на териториалните единици за статистически цели (NUTS).

тогава процентът на предприятията, имащи уебсайт за специфична група, A_k е дефинирана чрез:

$$R_{web,k} = \frac{\sum_{i=1}^{n} w_i \, I[u_i \neq \emptyset \wedge e_i \in A_k]}{\sum_{i=1}^{n} w_i \, I[e_i \in A_k]} \quad . \qquad (3)$$

По подобен начин могат да бъдат дефинирани показателите за **use case 2,** но целевата съвкупност се променя и вече съдържа предприятия, имащи уебсайт. Повечето от показателите в use case 2 могат да бъдат дефинирани чрез

$$R_{ICT,q} = \frac{\sum_{i=1}^{n} w_i \, I[u_i \neq \emptyset \wedge V_i^q = 1]}{\sum_{i=1}^{n} w_i \, I[u_i \neq \emptyset]} \quad . \qquad (4)$$

с V_i^q равно на 1, където ОВЕС за дадено предприятие e_i отговаря на въпрос q. q може да бъде всеки от следните въпроси:

- *Уебсайтът на предприятието съдържа ли нещо от изброените: онлайн поръчки, резервации или електронен магазин (например уебсайтът има ли пазарска количка)?*
- *Има ли уебсайтът на предприятието връзки или препратки към профилите на предприятието в социалните медии?*
- *Има ли предприятието специфични функции на уебсайта?*
- *Работи ли предприятието върху предстоящи/нови явления - например: AI и ML?*

За показателя *Използва ли предприятието Twitter за конкретна цел* е необходимо целевата съвкупност да бъде адаптирана, така че да съдържа само предприятия, за които онлайн характеристиката за използване на социални медии е равно на 1.

Може да се случи така, че за определени предприятия e_ (i (1)), ..., e_ (i (k)) не се скрапват достатъчно данни, така че съответните ОВЕС да не отговарят на нито един от горните въпроси. За да се елиминират техническите причини, се препоръчва използването на софтуер за скрапване, който може да се справи с вградения JavaScript, и уебстраниците да се „идвличата“ (скрапват) по няколко пъти и по различно време. Дори и да се елеминират техническите проблеми необходимата информация все още може да не е скрапната, тъй като структурата на уебстраниците не е стандартизирана и по този начин процедурата по извличане на ОВЕС се проваля.

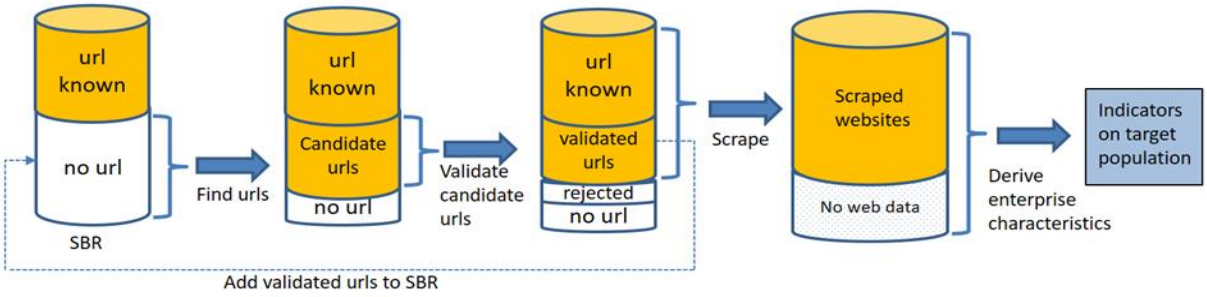
III. Жизнен цикъл за обработка на големи данни: онлайн характеристики на предприятията

Процесът на уебскрапване за откриване на характеристики на предприятието се вписва естествено в по-общия процес на жизнения цикъл на обработка на големи данни. Този обобщен процес осигурява отправна точка за всяка държава - членка на ЕСС, която би искала да извлече онлайн базирани характеристики на предприятията на национално

ниво. За да улесни приложимостта, в която и да е държава от ЕСС, екипът на WPC схематично представи процеса за определяне на онлайн характеристики на предприятията по възможно най-стандартизиран начин (вж. фиг. 1), но при следните две допускания:

- Всяка национална статистическа служба има статистически бизнес регистър (СБР), който съдържа поне имената и идентификационните номера на фирмите. Прието е, че процентът на предприятията с известен URL адрес може да варира между 0 и 100%.
- Всяка държава има някаква национална или международна търсачка (и), която може да се използва във фазата на извличане на URL адреси.

Фиг. 1. Обща схема за процеса на скрапване на онлайн характеристики на предприятията



Основните **фази на процеса** на определяне на онлайн характеристики на предприятието са, както следва:

1. Извличане на URL адреси за фирми без налични URL адреси. Резултатите от търсенията са **кандидат-URL адреси** за включване в списъка с URL. На схемата ясно се вижда, че дори след няколко търсения, че все още има предприятия, за които не може да бъде намерен URL адрес. Възможните причини за това могат да бъдат: че предприятието наистина няма уебсайт или че прилаганата стратегия за търсене не е достатъчно ефективна и не може да го намери.

2. **Валидиране или отхвърляне** на кандидат-URL адреси чрез сравняване на познатите за предприятието данни в СБР с данни от резултатите от търсенето. Извличането и валидирането/отхвърлянето на URL адреси може да бъде итеративно.

3. Във фазата на **скрапване** предприятията, за които URL адресът вече е известен в СБР и потвърдените, намерени URL адреси, се използват за скрапване на уеб-съдържание и съхраняване на изтритите уебсайтове.

4. **Извличане на онлайн характеристики на предприятието** за целевата съвкупност: това може да бъде целият СБР или част от него, например съвкупността на изследването „Използване на ИКТ предприятията“. На този етап е важно да се отбележи, че предприятията, за които няма уебданни, също трябва да бъдат отчетени и оценени чрез калибриране, използвайки общия брой единици в СБР.

5. **Актуализиране на СБР** с валидираните URL адреси, получени от последователни итерации и други статистически процеси в допълнение към основния цикъл. Това е показано на фиг. 1 като стрелка от валидираните URL адреси към СБР. Тук би могъл да се включи и допълнителен индикатор дали даден URL адрес в СБР произхожда от административен източник, или от процеса на извличане от уебмрежата.

Най-важните решения за различните фази на процеса могат да бъдат изразени и чрез фазите на GSBPM[6] модела по отношение на жизнения цикъл на големите данни: Събиране, Обработване, Анализ и Разпространение.

Процесът на **събиране** на ОВЕС данни (*фаза 4. Събиране, GSBPM*) е съставен от четири подпроцеса. Първо се изисква идентифициране на списък на компаниите, за които ще се събират данни (целева съвкупност), с основни атрибути като например името на предприятието. След това се конструира списък с потенциални адреси на уебсайтове, като се използват отговорите от търсачката на уебсайтове за всяко предприятие в СБР или други административни регистри. В подпроцес 3 се извършва частично „обхождане“ на потенциални адреси на уебсайтове (начални страници) чрез прилагане на механизъм за класиране на вероятности кой уебсайт е възможно най-добрият избор (като по този начин се търсят идентификационни данни на уебсайта). На последната стъпка се избира „първият най-добър“ уебсайт за всяко предприятие, след което може да се прави разширено събиране/скрапване на данни от намерените уебсайтове за получаване на специфични онлайн характеристики на предприятия.

В процеса на **обработка** *(извличане, почистване, интегриране, агрегиране и представяне, фаза 5. Обработване, GSBPM)* събраните уебданни първо се почистват технически (премахване на html тагове) и съдържателно (премахване на „стоп“ думи и т.н.). Вземат се решения за това кои части от уебстраниците се запазват като входни данни и кои части повече не са ни необходими. Текстовите части, които могат да се използват като бизнес идентификатори, представляват особен интерес. Необходими са и експертни решения какви методи да са прилагат с цел трансформиране на суровите текстове в структурирано съдържание. Тези методи могат да варират от базовия подход „кошница от думи“ (bag-of-words) до методи, запазващи контекстуална информация, като doc2vec, word2vec, sentence2vec или предварително обучени мрежи за текстови анализ. Процесът на преобразуване на текст в променливи се извършва чрез прилагане на детерминистичен подход и методи за машинно самообучение за извличане на характеристики.

Анализирането на уебданни *(фаза 6. Анализ, GSBPM)* в комбинация с вече съществуващите данни за предприятията в СБР е от решаващо значение в много от етапите на процеса: валидирането/отхвърлянето на кандидат-URL адреси се нуждае от внимателен анализ. Уебсайт на предприятие и предприятие са две различни концепции, и поради тази причина се изисква предварително двете понятия да се свържат на концептуално ниво. Един уебсайт може да съдържа информация за множество

[6] Generic Statistical Buisiness Process Model (GSBPM),
https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model

предприятия или обратно, едно предприятие може да притежава или публикува на множество уебсайтове.

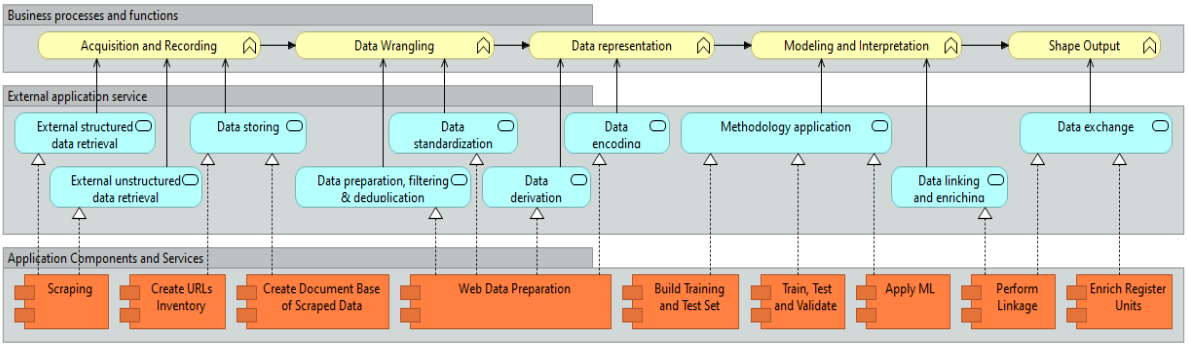
Изчисляването на ОВЕС за определена съвкупност въз основа на скрапнати уеб-данни се нуждае от моделиране и интерпретация. Изискват се решения относно дефинициите на характеристиките на предприятията - например на микро ниво - дефиниране на критерии кога дадено предприятие се счита, че извършва електронна търговия чрез своя уебсайт или е представено в някоя от социалните медии. Освен това в тази фаза, трябва да се има предвид, че почти винаги има подмножество на СБР, за което не може да бъде намерен URL адрес. Коригирането на това на макрониво може да се извърши чрез свързване на измерваната съвкупност към целевата съвкупност.

Разпространението на ОВЕС не се различава съществено от традиционните процеси на разпространение на статистика с изключение на това, че техниките за разпространение на уебданни са сравнително нови и би било добре да се добави разширено обяснение на използваните методи, което да улесни потребителите на експерименатлната статистика за ОВЕС.

IV. Референтна ИТ архитектура за ОВЕС данни

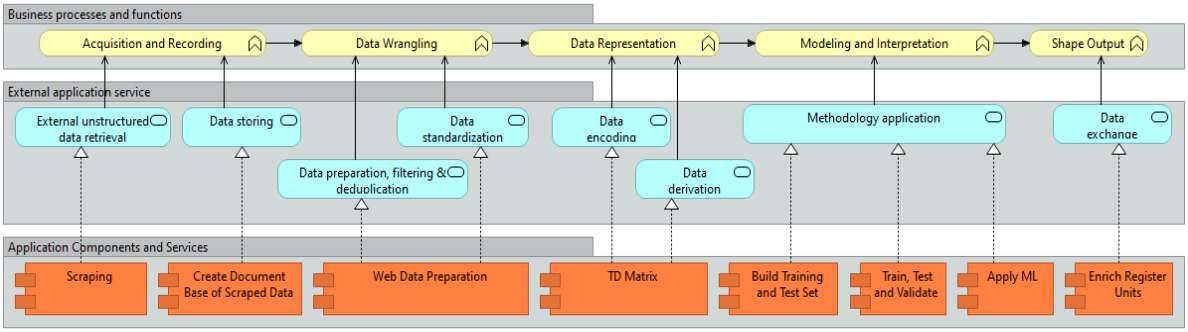
Референтната архитектура за ОВЕС данни се състои от описание на приложната и информационната архитектура на процеса на производство на експериментална статистика за случаите на използване 1 и 2. Описанието се основава на BREAL[7] архитектурата, резултат от работата по работен пакет F, проект ESSnet on BD II (вж. фиг. 2 и 3).

Фиг. 2. Общо графично представяне на приложната архитектура на Списък на URLs (use case 1)



[7] Big Data REference Architecture and Layers (BREAL), https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WPF_Deliverable_F1_BREAL_Big_Data_REference_Architecture_and_Layers_v.03012020.pdf

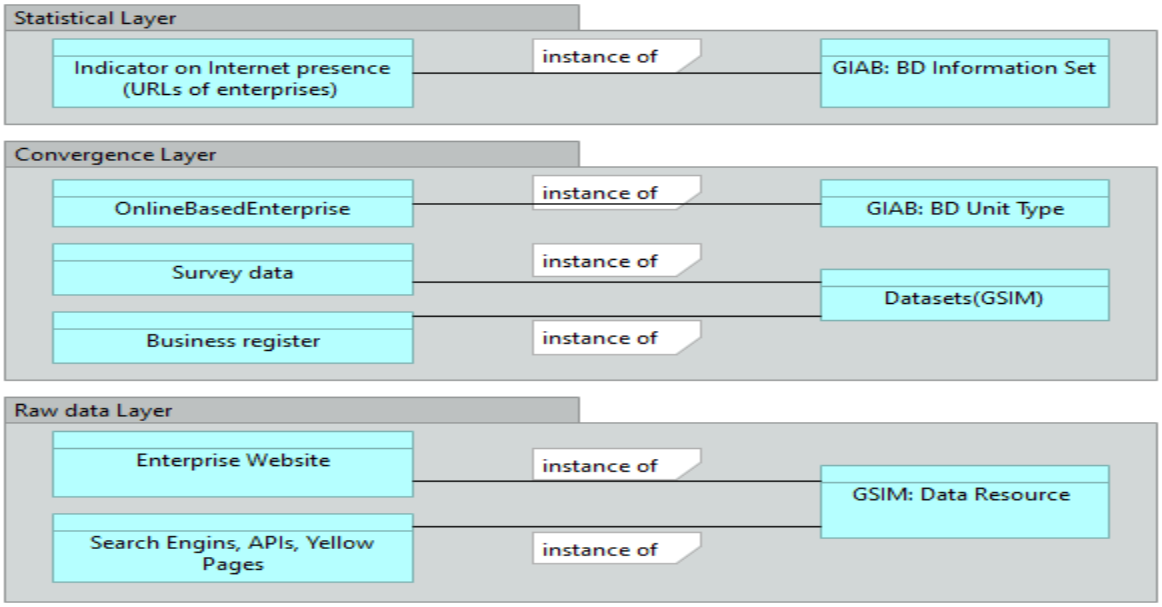
Фиг. 3. Общо графично представяне на приложната архитектура на Променливи в изследването „Използване на ИКТ в предприятията“ (use case 2)



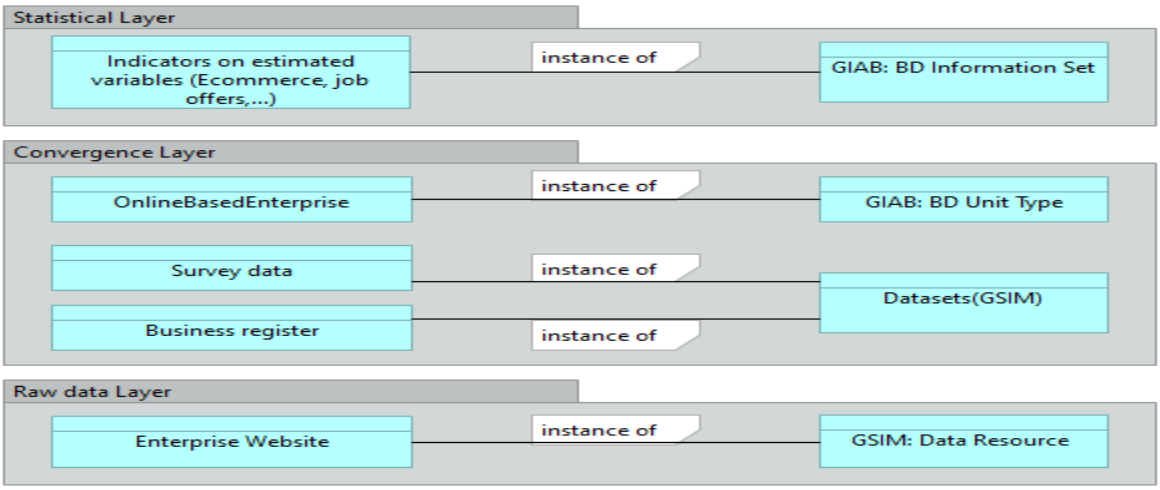
Бизнес процесите и функциите описват OBEC бизнес функциите, производни от модела BREAL. Всяка бизнес функция задейства следващата, като всяка бизнес функция се обслужва от една или повече услуги. Всяка услуга за приложения се реализира от един или повече компоненти на приложението. Компонентите на приложението могат да бъдат програми, модули, скриптове, класове или функции. Те могат да бъдат самостоятелни или част от система и да бъдат написани на различни програмни езици.

Описанието на общата информационна архитектура за OBEC за случаите на използване 1 и 2 (фиг. 4 и 5, респективно) също се основава на BREAL.

Фиг. 4. Информационна архитектура на Списък на URLs (use case 1)



Фиг. 5. Информационна архитектура на Променливи в изследването „Използване на ИКТ в предприятията“ (use case 2)



Нивото за „суровите“ данни описва първоначалните източници на данни на ОВЕС по отношение на BREAL. То обхваща процеса на придобиване и запис на ОВЕС. Междинното архитектурно ниво описва информационните обекти, получени от изходни данни чрез процесите за пренасочване и представяне на данни на ОВЕС. Статистическото ниво описва информационни обекти, получени от информационните източници на междинното ниво чрез процесите за моделиране, интерпретиране и извеждане на изходи за ОВЕС.

Почти всички приложни услуги и за двата случая на използване на ОВЕС са проектирани да бъдат достатъчно гъвкави, за да могат да бъдат оперативно съвместими, репликирани или споделяни между националните статистически организации в рамките на ЕСС.

По време на проекта всички ИТ услуги и данни се управляваха на местно ниво, но в бъдеще всички услуги (с изключение на услугите за „Свързване и обогатяване на данни“ и „Обмен на данни“, използващи локални данни) могат да бъдат споделени сред националните статистически организации в ЕСС.

В рамките на проектните дейности са внедрени някои обобщени софтуерни решения, които са публично достъпни чрез специализирания WPC GitHub на следния адрес: https://github.com/EnterpriseCharacteristicsESSnetBigData.

V. Производство на експериментална статистика за ОВЕС в Националния статистически институт

Експертният екип от НСИ, работещ по дейностите на работен пакет С, произведоха и публикуваха експериментална статистика за ОВЕС за 2019 и 2020 година. Всички получени резултати в табличен вид са достъпни на CROS портала на Евростат (рубрика „Експериментална статистика“): https://ec.europa.eu/eurostat/cros/content/WPC_Experimental_statistics_en, заедно с кратки методологични бележки и справочни метаданни (отчети за качество във формат ESQRS

и метаданни за потребителите във формат ESMS). Методологичните бележки описват процесите, използваните методи и софтуерни решения за производството на експериментални данни за ОВЕС.

Изчислените експериментални показатели са получени на базата на описаните методи, понятия и допускания в първата част на настоящата статия.

В изложението, което следва, са илюстрирани основните процесни стъпки, необходими за изпълнение на use-case 1 и use-case 2 с цел достигане до експериментални резултати за онлайн характеристики на предприятията.

1. Актуализиране на URLs адреси на предприятия (use-case 1)

За извличането на онлайн характеристики на предприятията е необходимо първоначално да се намерят техните URL адреси и да се конструира списък. Целта на това упражнение се състои от следните основни стъпки:

Подготовка на начални данни и софтуер

Процесът на актуализиране на интернет адресите на предприятията започва с дефиниране на съвкупността им. Съвкупността се състои от предприятията с 10 и повече заети, които към пролетта на 2020 г. са 28 251 на брой. За тях е взета информация от статистическия Бизнес регистър[8] в НСИ със следните полета: ЕИК[9], наименование, интернет адрес, електронна поща, пощенски адрес, стационарен телефон, мобилен телефон, населено място, код по NUTS 3[10] и код по икономическа дейност[11].

След определяне на съвкупността се подготвят операционната среда и софтуерът за обработка. Използван е обикновен компютър с Windows операционна система и език за програмиране Python[12]. Избраният софтуер за обработка е URLs Finder[13], част от Starter Kit[14], разработен на Python. Използвана е версия 1.0 на Starter Kit (специализирано средство за статистически експерти и програмисти), като софтуерът е доработен, а доработките по-късно са включени във версия Starter Kit 2.0.

За да работи URLs Finder се нуждае от слените модули и компоненти на Python:

- pandas[15] - бърз, мощен, гъвкав и лесен за използване инструмент с отворен код за анализ и манипулация на данни;

[8] https://www.nsi.bg/node/13207/
[9] Единен идентификационен код от Регистър Булстат (https://www.registryagency.bg/bg/registri/registar-bulstat/)
[10] Номенклатура на териториалните единици за статистика - малки региони (https://ec.europa.eu/eurostat/web/nuts/background)
[11] NACE Rev. 2 (https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF, https://www.nsi.bg/sites/default/files/files/pages/uplf/Methodology_KID.pdf)
[12] https://www.python.org/
[13] https://github.com/EnterpriseCharacteristicsESSnetBigData/StarterKit/tree/master/URLsFinder
[14] https://github.com/EnterpriseCharacteristicsESSnetBigData/StarterKit
[15] https://www.python.org/

- glob[16] - разширение за файлова пътека в стил Unix[17];
- BeautifulSoup[18] - библиотека, която улеснява изрязването на информация от уебстраници;
- requests[19] - позволява да се изпращат HTTP[20]/1.1 заявки изключително лесно;
- re[21] - предоставя операции за регулярни изрази върху текст;
- numpy[22] - основният пакет за научни изчисления с Python;
- time[23] - предоставя различни функции, свързани с времето;
- unquote[24] - заменя %xx кодирания с еквивалента им от един символ;
- urlparse[25] - разделя URL[26] на шест компонента в съответствие на общата структура на URL със следните имена scheme://netloc/path;parameters?query#fragment;
- tqdm[27] - предоставя лента за напредъка;
- datetime[28] - предоставя класове за манипулиране на дати и часове;
- logging[29] - дефинира функции и класове, които прилагат гъвкава система за регистриране на събития за приложения и библиотеки;
- sklearn[30] - инструмент за Машинно самообучение[31] и анализ на Python.

Променливите с входна информация и данни за работата на URLs Finder са следните:

- version - идентификация на файловете с извлечена информация по дата, идентификационен номер или друго;
- title - име на проекта, използвано и за имена на файлове;
- startpath - директория, в която се намира csv[32] файл с информация за предприятия от съвкупността;
- scrapepath - директория, в която се запазват csv файлове с извлечена информация за предприятия;
- logpath - директория, в която се записва информация за регистриране на събития от работата на софтуера;

[16] https://docs.python.org/3/library/glob.html
[17] https://bg.wikipedia.org/wiki/Unix
[18] https://pypi.org/project/beautifulsoup4/
[19] https://pypi.org/project/requests/
[20] https://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol
[21] https://docs.python.org/3/library/re.html
[22] https://numpy.org/
[23] https://docs.python.org/3/library/time.html
[24] https://docs.python.org/3/library/urllib.parse.html
[25] https://docs.python.org/3/library/urllib.parse.html
[26] https://en.wikipedia.org/wiki/URL
[27] https://pypi.org/project/tqdm/
[28] https://docs.python.org/3/library/datetime.html
[29] https://docs.python.org/3/library/logging.html
[30] https://scikit-learn.org/stable/
[31] https://en.wikipedia.org/wiki/Machine_learning
[32] https://en.wikipedia.org/wiki/Comma-separated_values

- blacklistpath - директория, в която се намира csv файл с черен списък с URL адреси;
- startfile - име на csv файл с информация за предприятия от съвкупността;
- scrapefile - име на csv файл с извлечена информация от уебсайтове;
- sapifile - име на csv файл с извлечена информация от търсачката Duck Duck Go[33];
- toscrapefile - име на csv файл с URL адреси от уебсайтове, от които ще бъде извличана информация;
- logfile - име на регистрационен файл за събития;
- blacklistfile - име на csv файл с черни списъци с URL адреси;
- csv_delimiter - разделител на csv файла, например: „;“;
- csv_encoding - кодиране на csv файла, например: „utf-8“;
- headers - информация за HTTP заявка.

За изпълнение на софтуер URLs Finder е използван Jupyter Notebook[34]. В Jupyter Notebook последователно се извикват команди за зареждане на модулите на софтуер и за изпълнение на техните методи. Първо се посочва пътят до директорията, в която се намира софтуерът. Дават се стойности на променливите. Зарежда се модулът за регистриране на събития за приложения и библиотеки и се изпълнява неговият метод за започване на работа за регистриране.

Намиране на кандидат-интернет адреси на предприятия

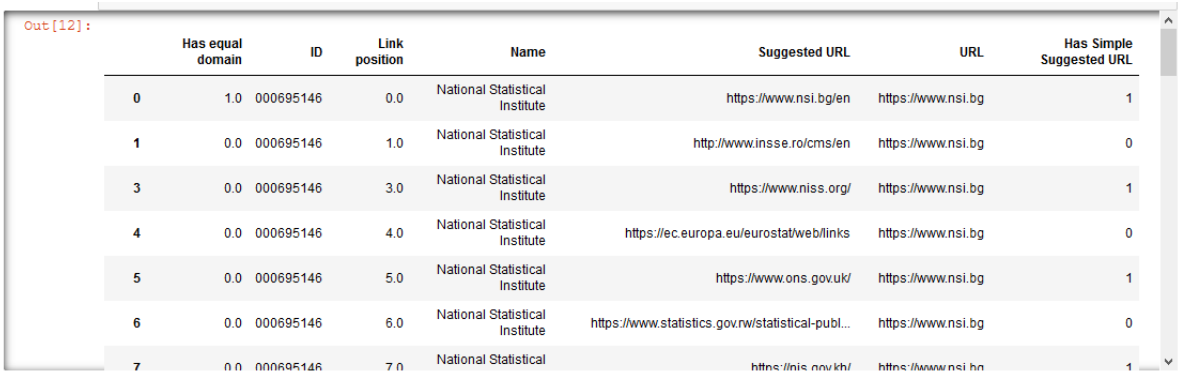
Намирането на кандидат-интернет адреси на предприятия започна със зареждане на модула на софтуера за извличане на данни от интернет и с инициализиране на променливите. Продължи със зареждането на информацията за предприятията от съвкупността взета, от статистическия Бизнес регистър и с информацията за нежеланите интернет адреси от черния списък като жълти страници, новинарски сайтове, интернет директории и други.

Изпращане на заявки с информация за предприятията към търсеща машина

Методът querySearchEngine на URLs Finder е използван за изграждане на списък с кандидат-интернет адреси на предприятията. Методът изпраща заявки с имената на предприятията към търсещата машина. Търсещата машина, използвана от софтуера, е Duck Duck Go. Търсачката връща списък с до 10 предполагаеми интернет адреса за всяко предприятие (фиг. 6).

[33] https://duckduckgo.com
[34] https://jupyter.org/ - в URLs Finder е включен Jupyter Notebook файл с тестови данни, команди и резултати. В тази статия няма да бъдат посочвани командите. Ще бъдат демонстрирани само резултати.

Фиг. 6. Предложения за интернет адреси на НСИ от уебтърсачката

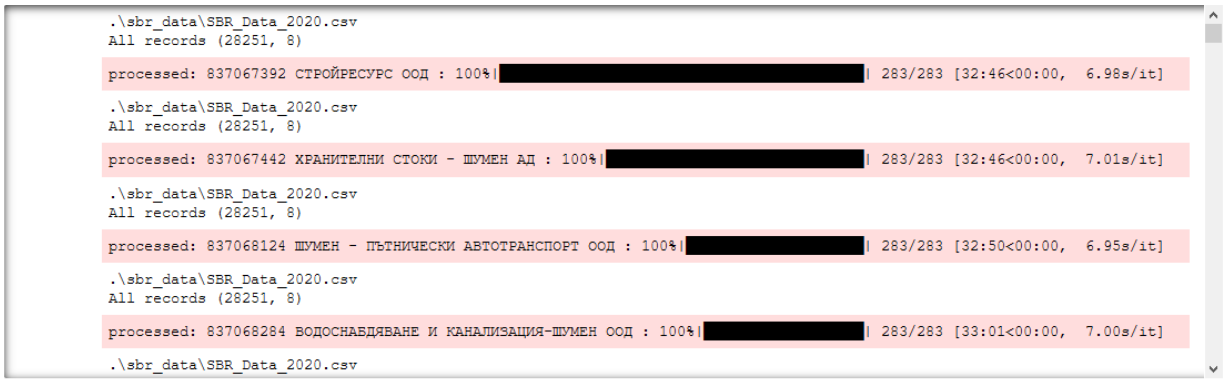


Out[12]:

	Has equal domain	ID	Link position	Name	Suggested URL	URL	Has Simple Suggested URL
0	1.0	000695146	0.0	National Statistical Institute	https://www.nsi.bg/en	https://www.nsi.bg	1
1	0.0	000695146	1.0	National Statistical Institute	http://www.insse.ro/cms/en	https://www.nsi.bg	0
3	0.0	000695146	3.0	National Statistical Institute	https://www.niss.org/	https://www.nsi.bg	1
4	0.0	000695146	4.0	National Statistical Institute	https://ec.europa.eu/eurostat/web/links	https://www.nsi.bg	0
5	0.0	000695146	5.0	National Statistical Institute	https://www.ons.gov.uk/	https://www.nsi.bg	1
6	0.0	000695146	6.0	National Statistical Institute	https://www.statistics.gov.rw/statistical-publ...	https://www.nsi.bg	0
7	0.0	000695146	7.0	National Statistical	https://nis.gov.kh/	https://www.nsi.bg	1

За да се избегне злоупотреба с ресурсите на търсещата машина и евентуална забрана за използване, заявките към търсачката по име за всяко предприятие се изпълняват през 6 секунди. Това увеличава времето за работа на две денонощия, повишава риска от прекъсване поради грешка в софтуера, загуба на захранване, загуба на интернет свързаност и други. Използваният софтуер не поддържа функция за продължаване от мястото на прекъсване. Поради това съвкупността от предприятия беше разделена на 100 множества без повторение, всяко с 282 или 283 предприятия. Така при прекъсване може да се продължи от множеството, където е възникнало прекъсване, без да се налага повторно търсене за всички вече намерени предложения. Търсенето върху всяко множество продължава средно по 33 минути (фиг. 7). В резултат на търсенето за съвкупността от 28 251 предприятия бяха предложени около 250 хил. потенциални интернет адреса.

Фиг. 7. Успешно приключване на търсене на кандидат-URLs адреси на предприятия с търсеща машина за 4 от 100-те подмножества от изследваната съвкупност от предприятия



```
.\sbr_data\SBR_Data_2020.csv
All records (28251, 8)
processed: 837067392 СТРОЙРЕСУРС ООД : 100%|[redacted]| 283/283 [32:46<00:00,  6.98s/it]
.\sbr_data\SBR_Data_2020.csv
All records (28251, 8)
processed: 837067442 ХРАНИТЕЛНИ СТОКИ - ШУМЕН АД : 100%|[redacted]| 283/283 [32:46<00:00,  7.01s/it]
.\sbr_data\SBR_Data_2020.csv
All records (28251, 8)
processed: 837068124 ШУМЕН - ПЪТНИЧЕСКИ АВТОТРАНСПОРТ ООД : 100%|[redacted]| 283/283 [32:50<00:00,  6.95s/it]
.\sbr_data\SBR_Data_2020.csv
All records (28251, 8)
processed: 837068284 ВОДОСНАБДЯВАНЕ И КАНАЛИЗАЦИЯ-ШУМЕН ООД : 100%|[redacted]| 283/283 [33:01<00:00,  7.00s/it]
.\sbr_data\SBR_Data_2020.csv
```

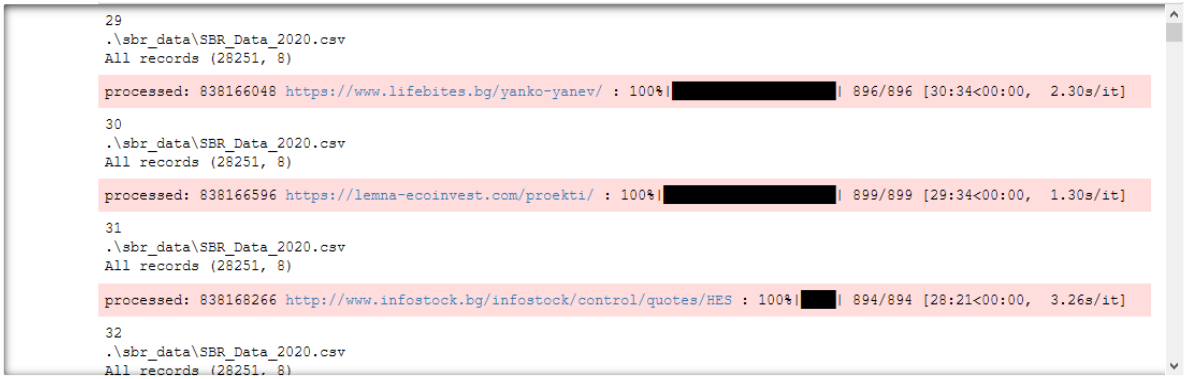
Всички неуспешни търсения се натрупват в обект с табличен формат. Използваната версия на софтуера не записва таблицата във файл и тази информация е загубена. В новата версия на софтуера този недостатък е отстранен. Грешките се

класифицират като проблеми с връзката, HTTP[35] грешка, изчерпване на времето, твърде много пренасочвания, грешка в заявката и обща грешка.

Намиране на страници за извличане на информация за предприятията

За всеки предложен потенциален интернет адрес на предприятие са извлечени първите 10 и последните 10 интернет връзки от страницата на интернет адреса. Извлечените връзки се филтрират и се запазват само тези с еднакъв домейн, като например потенциалният интернет адрес, а дублираните адреси се премахват. Отново процесът се извършва върху 100-те множества на съвкупността от предприятията с цел възобновяване след прекъсване без загуба на информация и повторно извличане. Всеки предложен адрес се проверява дали присъства в черния списък с адреси. Ако присъства, се изключва от предложените възможни адреси за предприятието. Така количеството на потенциалните интернет адреси на предприятията е сведено до около 90 хиляди. Търсенето върху всяко множество продължава средно по 30 минути (фиг. 8). В резултат на търсенето за съвкупността от 28 251 предприятия бяха предложени около 700 хил. страници за извличане на информация за потенциални интернет адреси.

Фиг. 8. Успешно приключване на търсене на страници на кандидат-интернет адреси на предприятия на 3 от 100-те подмножества от изследваната съвкупност от предприятия



Отново информацията за възникналите грешки не е достъпна, като проблемът е отстранен в новата версия на софтуера.

Извличане на данни за предприятията от страници на кандидат-интернет адреси на предприятия

Извличане на информация за предприятията от предложените интернет адреси и техните страници се проведе върху всяко от 100-те множества без повторение на цялата съвкупност от предприятия. Продължи 2 седмици, като всяко множество се обработва за около 3 часа средно (фиг. 9).

[35] https://bg.wikipedia.org/wiki/HTTP

Фиг. 9. Успешно приключване на извличане на информация за предприятията от предложените интернет адреси и техните страници

След приключване на работата на този етап, софтуерът връща данни за открита информация за предприятията на всяка страница от списъците с кандидат-URLs адреси (фиг. 10). Структурата на таблицата е следната:

- ID - ЕИК на предприятието;
- Name - име на предприятието;
- URL - интернет адрес на предприятието, който знаем от статистическия Бизнес регистър;
- Suggested URL - кандидат- (потенциалният, предложеният от търсачката) интернет адрес на предприятието;
- Link position - пореден номер на предложения от търсачката интернет адрес за предприятието;
- Has equal domain - 1, ако домейнът на известния и предложения адрес са равни, иначе 0;
- Has Simple Suggested URL – 1, ако предложеният интернет адрес съдържа само протокол, домейн и език „en“, иначе 0;
- URL to scrape - интернет страницата, от която са получени данните;
- Status code - код за статус от HTTP заявката към URL to scrape;
- Has ID - 1, ако ЕИК на предприятието е открит в текста на интернет страницата, от която са получени данните, иначе 0;
- Has Name - 1, ако името на предприятието е открито в текста на страницата, иначе 0;
- Has Phone - 1, ако стационарният телефон на предприятието е открито в текста на страницата, иначе 0;
- Has GSM - 1, ако мобилният телефон на предприятието е открито в текста на страницата, иначе 0;
- Has Address - 1, ако адресът на предприятието е открито в текста на страницата, иначе 0;
- Has Populated place - 1, ако населеното място на предприятието е открито в текста на страницата, иначе 0;

- Has Email - 1, ако електронната поща на предприятието е открито в текста на страницата, иначе 0;
- Has equal Email and URL Domains - 1, ако домейните на електронната поща на предприятието и предложения интернет адрес са равни, иначе 0.

Фиг. 10. Данни за информация за предприятията за всяка страница за кандидат-интернет адресите на предприятията

...ail	Has GSM	Has ID	Has Name	Has Phone	Has Populated place	Has Simple Suggested URL	Has equal Email and URL Domains	Has equal domain	ID	Link position	Name	Status code	Suggested URL	URL	URL to scrape
0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	127041015	7.0	ИНТЕРНЕТ СЪРВИС ООД	200.0	https://optinet.bg/	NaN	https://optinet.bg/
0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	127041015	7.0	ИНТЕРНЕТ СЪРВИС ООД	200.0	https://optinet.bg/	NaN	https://optinet.bg/интернет/
0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	127041015	7.0	ИНТЕРНЕТ СЪРВИС ООД	200.0	https://optinet.bg/	NaN	https://optinet.bg/видеонаблюдение/
0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	127041015	7.0	ИНТЕРНЕТ СЪРВИС ООД	200.0	https://optinet.bg/	NaN	https://optinet.bg/контакти/
0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	127041015	7.0	ИНТЕРНЕТ СЪРВИС ООД	200.0	https://optinet.bg/	NaN	https://optinet.bg/общи-условия/
0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	127041015	7.0	ИНТЕРНЕТ СЪРВИС ООД	200.0	https://optinet.bg/	NaN	https://optinet.bg/политика-за-поверителност
0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	127041015	7.0	ИНТЕРНЕТ СЪРВИС ООД	200.0	https://optinet.bg/	NaN	https://optinet.bg/политика-за-поверителност/
0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	127041015	9.0	ИНТЕРНЕТ СЪРВИС ООД	200.0	https://inetdec.nra.bg/	NaN	https://inetdec.nra.bg/

Отново информацията за възникналите грешки не е достъпна, като проблемът е отстранен в новата версия на софтуера.

Извличането на информацията се извършва чрез сравняване на низове с малки букви, като при съвпадение на низовете се присвоява стойност 1, а при разлика се присвоява стойност 0. В бъдеще може да се приложи алгоритъм[36] за Fuzzy logic[37] за откриване на алтернативно изписване на низове за телефони, GSM, адреси и други. Получените резултати от този алгоритъм могат да се използват за по-прецизно оценяване на наличието на търсената информация. Също така може да се пренапише софтуерът, за да открива повече от един низ за дадено поле - например телефон, GSM и други.

Следващата стъпка обединява множествата без повторение от предприятия в един масив. Софтуерът добавя две нови полета към таблицата:

- Has URL - 1, ако знаем интернет адреса на предприятието от статистическия Бизнес регистър, иначе 0;
- sum - сумата по редове на колоните Has equal Email and URL Domains, Has Email, Has Name, Has Phone, Has Address, Has ID and Has Populated place, групирани по полета ID, Name, Suggested URL, Link position and Status code и агрегирани в колона sum.

[36] https://pypi.org/project/fuzzywuzzy/
[37] https://en.wikipedia.org/wiki/Fuzzy_logic

След това софтуерът извършва премахване на дублираните записи по всички полета, като запазва само първия намерен. Резултатът е таблица с 62 160 записа в 18 колони (фиг. 11).

Фиг. 11. Таблица с данни за открита информация за предприятията по страници на кандидат-интернет адреси

	Suggested URL	Link position	Status code	Has Address	Has Email	Has GSM	Has ID	Has Name	Has Phone	Has Populated place	Has Simple Suggested URL	Has equal Email and URL Domains	Has equal domain	sum	URL	Has URL
0	https://free-images.com/display/atanas_teshovs...	8.0	200.0	0	0	0.0	0	0	1	0	0	0	0	20.0	http://osnatpk.com/gd_teshovski.php	1
1	https://www.wikiwand.com/bg/Потребителска_кооп...	1.0	200.0	0	0	0.0	0	0	1	0	0	0	0	9.0	NaN	0
2	https://www.wikizero.com/bg/Потребителска_кооп...	4.0	200.0	0	0	0.0	0	0	1	0	0	0	0	2.0	NaN	0
3	https://www.multitran.com/m.exe?a=3&l1=15&l2=3...	6.0	200.0	0	0	0.0	0	0	1	0	0	0	0	20.0	NaN	0

<u>Подготовка на модел за Логистична регресия за намиране на интернет адреси на предприятията</u>

Получените данни от предишния етап се използват за намиране на интернет адреси на предприятията с използване на Логистична регресия. За целта е приложен методът prepareLR[38] на софтуера URLs Finder, като данните са разделени на 70% обучително множество и 30% тестово множество. Методът изчислява променлива Score, като:

*Score колона = sum колона - sum колона * Link position колона/100*

Методът избира само записите с най-висок Score от всички уникални ЕИК на предприятията и ги зарежда в модела на логистичната регресия. Методът използва полетата Has Simple Suggested URL, Has Address, Has Email, Has ID, Has Name, Has Phone, Has Populated place и Has equal Email and URL Domains за независимата променлива и Has equal domain за зависимата променлива на модела на логистичната регресия. Методът връща следните обекти в резултат на прилагането му, които се използват за предсказване на интернет адресите на предприятията от наличните данни и за оценяване на качеството на полученото предсказване:

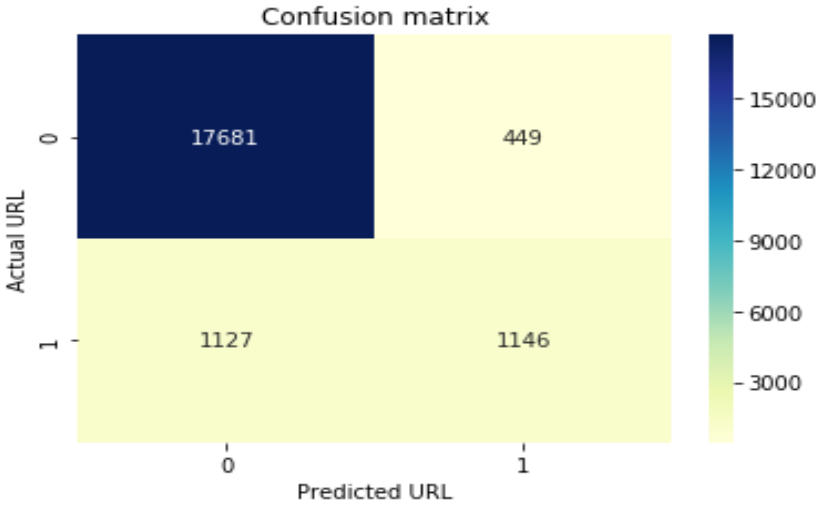
- обучително множество на независимата променлива;
- тестово множество на независимата променлива;
- обучително множество на зависимата променлива;
- тестово множество на зависимата променлива;
- предсказани резултати от тестово множество на зависимата променлива;
- табличен обект с независимата променлива;

[38] Методът е пренаписан в новата версия на софтуера, като са променени начинът на формиране на Score променливата, независимата и зависимата променлива.

- обект с моделът на логистичната регресия.

От тестовото множество на зависимата променлива и предсказаните резултати от тестово множество на зависимата променлива е получена следната Матрица на неточностите[39] (фиг. 12):

Фиг. 12. Матрица на неточностите за use-case 2



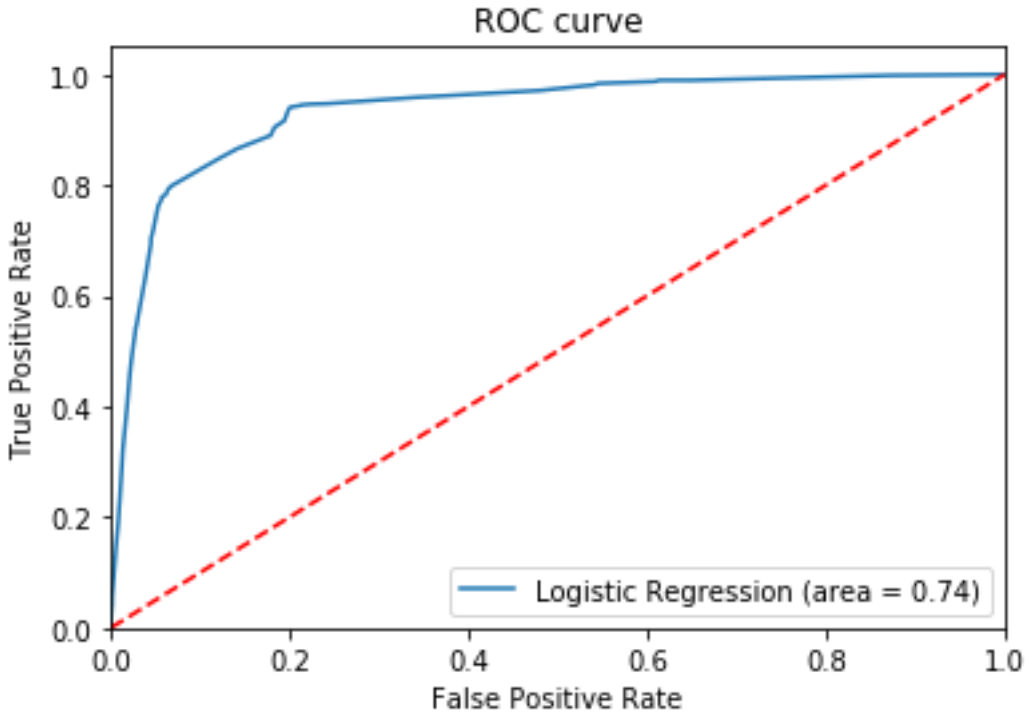
От матрицата може да се заключи, че:

- точност - моделът е коректен в 92% от случаите;
- прецизност - когато моделът предсказва интернет адреси на предприятия, то моделът е коректен в 72% от случаите;
- чувствителност - когато знаем интернет адреса на предприятието, то моделът коректно предсказва интернет адреса на предприятието само в 50% от случаите;
- специфичност - когато не знаем интернет адреса на предприятието, то моделът коректно не предсказва интернет адреса на предприятието в 98% от случаите и предсказва некоректно само в 2% от случаите;
- F1 оценка - със стойност от 0.59 (скала от 0 до 1) показва значителен дисбаланс между прецизността и чувствителността на модел и има какво да се желае за подобряването му въпреки високата точност, която отчита;
- MCC[40] - със стойност от 0.56 (скала от -1 до 1) показва, че моделът е по-близо до перфектното предсказване (1) отколкото до произволното (0), но все още е далече от идеалните резултати.

[39] Confusion matrix - https://en.wikipedia.org/wiki/Confusion_matrix
[40] Matthews Correlation Coefficient (MCC) - https://en.wikipedia.org/wiki/Matthews_correlation_coefficient

От ROC-кривата[41] (фиг. 13) се вижда, че площта под синята линия е значителна и тя се приближава към стойност 1. Това показва, че моделът има потенциал за предсказване на наличието на URLs адреси на предприятията.

Фиг. 13. ROC-крива



<u>Прилагане на избрания модел върху данните</u>

С метода prepareP на URLsFinder моделът на логистичната регресия се прилага върху таблицата с данни за открита информация за предприятията по страници на кандидат URLs адресите. Резултатът е таблица със следната структура (фиг. 14):

- ID - ЕИК на предприятието;
- Name - име на предприятието;
- URL - интернет на адрес на предприятието, който знаем от статистическия Бизнес регистър;
- Suggested URL - кандидат- (потенциалният, предложеният от търсачката) интернет адрес на предприятието;
- Link position - пореден номер на предложения от търсачката интернет адрес за предприятието;
- Score - променливата score от модела на логистичната регресия;
- predict - 1, ако логистичната регресия предсказва, че кандидат-интернет адрес на предприятието е търсеният интернет адрес на предприятието, иначе 0;
- 0 - вероятността предсказанието на логистичната регресия да е грешно;

[41] Receiver Operating Characteristic curve - https://en.wikipedia.org/wiki/Receiver_operating_characteristic

- 1 - вероятността предсказанието на логистичната регресия да е вярно.

Фиг. 14. Таблица с предсказания за интернет адреси на предприятия

Out[21]:

	ID	Name	Link position	Score	Suggested URL	URL	predict	0	1
0	000000858	ТПКИ АТАНАС ТЕШОВСКИ	8.0	18.40	https://free-images.com/display/atanas_teshovs...	http://osnatpk.com/gd_teshovski.php	0	0.994551	0.005449
1	000003402	ПОТРЕБИТЕЛСКА КООПЕРАЦИЯ МАКЕДОНИЯ	1.0	8.91	https://www.wikiwand.com/bg/Потребителска_кооп...	NaN	0	0.994551	0.005449
2	000003402	ПОТРЕБИТЕЛСКА КООПЕРАЦИЯ МАКЕДОНИЯ	4.0	1.92	https://www.wikizero.com/bg/Потребителска_кооп...	NaN	0	0.994551	0.005449
3	000003402	ПОТРЕБИТЕЛСКА КООПЕРАЦИЯ МАКЕДОНИЯ	6.0	18.80	https://www.multitran.com/m.exe?a=3&l1=15&l2=3...	NaN	0	0.994551	0.005449
4	000025060	ОБЛАСТЕН КООПЕРАТИВЕН СЪЮЗ - БЛАГОЕВГРАД	3.0	3.88	https://www.alo.bg/obiavi/zapoznanstva-eskort/...	NaN	0	0.994551	0.005449

62152	838178157	ИНФРАСТРОЙ - ИНЖЕНЕРИНГ ООД	9.0	0.00	http://www.eurobuildingengineering.com/bg/proj...	http://www.ise-yambol.com/	0	0.993792	0.006208
62153	838180083	ПРОМИШЛЕНА ЕНЕРГЕТИКА АД	0.0	18.00	https://www.prom-energo.bg/	https://www.prom-energo.bg/bg/	1	0.085908	0.914092
62154	838180083	ПРОМИШЛЕНА ЕНЕРГЕТИКА АД	8.0	3.68	https://www.euba.bg/index.php?option=com_conte...	https://www.prom-energo.bg/bg/	0	0.948508	0.051492
62155	838185899	СТАНЕВ КОНСЕРВ ЕООД	3.0	0.00	https://dawhois.com/site/stanevkonserv.com.html	http://stanevkonserv.com/	0	0.993792	0.006208
62156	838188870	НЕВА - МЕТАЛ ООД	5.0	4.75	https://enametal.com/	NaN	0	0.975192	0.024808
62157	838190394	МУЛТИФАРМ - 95 ООД	0.0	18.00	http://multipharm.eu/page/3/za-nas.html	http://multipharm.eu/	1	0.407451	0.592549
62158	838190394	МУЛТИФАРМ - 95 ООД	6.0	0.00	https://www.puls.bg/reference/pharmacy/	http://multipharm.eu/	0	0.993792	0.006208
62159	838190394	МУЛТИФАРМ - 95 ООД	7.0	0.93	http://www.zdrave.bg/?c=h&a=t&d=������	http://multipharm.eu/	0	0.989774	0.010226

62160 rows × 9 columns

От таблицата са вземат само тези редове, за които полето predict има стойност 1 и полетата URL и Suggested URL се различават, като например ред 163 и 184 (фиг. 15). За тези филтрирани редове предложените интернет адреси се сравняват с известните от СБР регистър и се подготвя таблица с актуализирани адреси на предприятията. Ръчно са прегледани 280 нови предложения за URLs адреси на предприятията, от които 230 са потвърдени; 267 адреси са ревизирани, от които 75 адреси са по-добри от вече известните адреси. Така от известни 12 058 URLs адреси на предприятия бяха актуализирани 75 и бяха добавени нови 230, с което общият брой на адресите станаха 12 288.

Фиг. 15. Таблица с предсказания за интернет адреси на предприятия, за които полето predict има стойност 1

Out[23]:

	ID	Name	Link position	Score	Suggested URL	URL	predict	0	1
21	000065763	СМА МИНЕРАЛ БУРГАС ВАР ЕООД	0.0	29.00	http://smamineralbg.com/	http://smamineralbg.com/	1	0.150551	0.849449
44	000110952	ПОДЕМ ЕООД	2.0	4.90	http://podem-eood.com/	http://podem-eood.com/	1	0.272833	0.727167
46	000111036	ПРОИЗВОДИТЕЛНА КООПЕРАЦИЯ НА ИНВАЛИДИТЕ ЦАРЕВЕЦ	0.0	4.00	http://pki-tzarevetz.com/	http://pki-tzarevetz.com/	1	0.247597	0.752403
82	000275929	АВТОМОТОР КОРПОРАЦИЯ АД	0.0	4.00	http://leasing.citroen.bg/	http://www.citroen.bg/	1	0.449579	0.550421
163	000620115	ИХБ ЕЛЕКТРИК АД	0.0	39.00	http://www.ihbelectric.com/	http://www.ihbelectric.com/bg/	1	0.054059	0.945941
184	000627259	НАЦИОНАЛНА ПОТРЕБИТЕЛНА КООПЕРАЦИЯ НА СЛЕПИТЕ ...	0.0	42.00	http://npksb.com/	NaN	1	0.352453	0.647547

Резултати

На базата на събраните, обработени и анализирани данни за URLs адреси на предприятията се установи, че за 2020 г. 43.5% от предприятията от съвкупността на изследването „Използване на ИКТ в предприятията“ имат интернет адрес. За 37.8% от предприятията, имащи между 10 и 49 заети, са намерени интернет адреси, за предприятията, имащи между 50 и 249, този процент е 67.8, а големите предприятия с повече от 250 заети имат и най-голямо отношение на интернет адресите с 85.5%. Над половината предприятия в област София (столица) имат интернет адреси, докато в областите Благоевград и Видин предприятията, имащи интернет адреси, е под една четвърт (фиг. 16). Близо две трети от предприятията в сектори на икономическа дейност „Създаване и разпространение на информация и творчески продукти; Далекосъобщения“ и „Производство и разпространение на електрическа и топлинна енергия и на газообразни горива“ имат интернет адрес. Едва една четвърт от предприятията в сектор „Хотелиерство и ресторантьорство“ имат интернет адрес, което е възможно да се дължи на използване на споделени платформи за резервация или използване на интернет адреси с имената на обектите, а не с имената на предприятията (фиг. 16).

За пълнота на анализа е направено сравнение между официалните данни от изследването „Използване на ИКТ в предприятията“ и експерименталните данни от този процес (вж. https://ec.europa.eu/eurostat/cros/sites/crosportal/files/WPC_Experimental_statistics_BG_2020_Results.pdf).

Фиг. 16. Интернет адреси на предприятия с 10 и повече заети по области

Интернет адреси на предприятия с 10 и повече заети по области

Код по NUTS	Област	Предприятия брой	Интернет адреси брой	Отношение на Интернет адресите към Предприятията (%)
BG311	Видин	177	43	24.3
BG312	Монтана	336	90	26.8
BG313	Враца	379	145	38.3
BG314	Плевен	666	204	30.6
BG315	Ловеч	392	163	41.6
BG321	Велико Търново	744	283	38
BG322	Габрово	471	231	49
BG323	Русе	848	400	47.2
BG324	Разград	270	77	28.5
BG325	Силистра	243	64	26.3
BG331	Варна	2208	1001	45.3
BG332	Добрич	468	151	32.3
BG333	Шумен	476	179	37.6
BG334	Търговище	298	84	28.2
BG341	Бургас	1678	617	36.8
BG342	Сливен	454	166	36.6
BG343	Ямбол	330	125	37.9
BG344	Стара Загора	1057	470	44.5
BG411	София (столица)	8760	4932	56.3
BG412	София	682	254	37.2
BG413	Благоевград	1400	337	24.1
BG414	Перник	361	99	27.4
BG415	Кюстендил	373	106	28.4
BG421	Пловдив	2908	1347	46.3
BG422	Хасково	699	231	33
BG423	Пазарджик	792	287	36.2
BG424	Смолян	409	104	25.4
BG425	Кърджали	372	98	26.3

Фиг. 17. Интернет адреси на предприятия с 10 и повече заети по сектори на икономическа дейност

Интернет адреси на предприятия с 10 и повече заети по сектори на икономическа дейност

Сектор на икономическа дейност	Предприятия брой	Интернет адреси брой	Отношение на Интернет адресите към Предприятията (%)
Административни и спомагателни дейности (N)	1256	528	42
Доставяне на води; Канализационни услуги, управление на отпадъци и възстановяване (E)	253	138	54.5
Други дейности (S)	15	9	60
Операции с недвижими имоти (L)	542	239	44.1
Преработваща промишленост (C)	7466	3842	51.5
Производство и разпространение на електрическа и топлинна енергия и на газообразни горива (D)	133	86	64.7
Професионални дейности и научни изследвания (M)	1278	739	57.8
Строителство (F)	3180	1260	39.6
Създаване и разпространение на информация и творчески продукти; Далекосъобщения (J)	1235	813	65.8
Транспорт, складиране и пощи (H)	2107	687	32.6
Търговия; Ремонт на автомобили и мотоциклети (G)	7761	3164	40.8
Хотелиерство и ресторантьорство (I)	3025	783	25.9

2. Извличане на онлайн характеристики на предприятията от интернет (use-case 2)

Подготовка на начални данни и софтуер

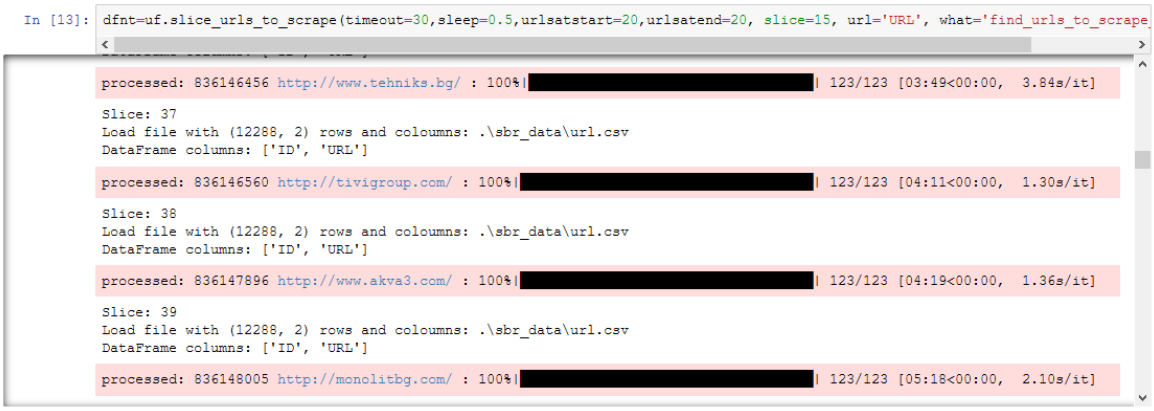
Процесът на намиране на онлайн характеристиките на предприятията започва с определяне на съвкупността им. Съвкупността от СБР е същата използвана за актуализирането на интернет адресите на предприятията, а именно предприятия с 10 и повече заети, които към пролетта на 2020 г. са 28 251 на брой. Информацията е същата, като са актуализирани URLs адресите на предприятията съгласно резултатите от предишния процес (use-case 1).

Използваните операционна среда, софтуер, модули и компоненти са същите като избраните за предишния процес. Входните променливи за софтуера са същите. Софтуерът се изпълнява като в предишния процес. Пропуска се само стъпката със събиране на предложения за интернет адреси на предприятията от търсачка, тъй като уебсайтовете на предприятията, от които ще извличаме информация са вече известни.

Намиране на страници за извличане на информация за предприятията

За всяко предприятие от съвкупността, на което знаем интернет адреса (общо 12 288 на брой), са извлечени първите 20 и последните 20 интернет връзки от страницата на URL адреса. Извлечените връзки се филтрират и се запазват само тези с идентичен домейн и интернет адрес, а дублираните URLs адреси се премахват. Отново стъпката се извършва върху разделената на 100 множества без повторение съвкупност на предприятията с цел възобновяване след прекъсване без загуба на информация и повторно извличане. Търсенето върху всяко множество продължава средно по четири минути и половина (фиг. 18). В резултат на търсенето за съвкупността от 12 288 адреса на предприятия са предложени около 220 хил. страници за извличане на информация за онлайн характеристики на предприятията (ОВЕС).

Фиг. 18. Успешно приключване на търсене на страници за характеристики на предприятия на 4 от 100-те подмножества от изследваната съвкупност от предприятия



За разлика от предишния процес за use-case 1 в този процес грешките се съхраняват във файл (фиг. 19), като общият им брой е 551, или 4.5% не отговорили. С подобряване на софтуера (направено в новата версия) могат да бъдат намалени грешки от типа Request exception. Грешки от вида Connection problem и Timeout occurred могат да бъдат намалени, като стъпката се изпълни повторно за тези URLs адреси на предприятията.

Фиг. 19. Грешки при намиране на страници за извличане на информация за характеристики на предприятията от интернет адресите им

	ID	URL	Error
0	200225006	pizzasiciliana.bg	Request exception
1	200608912	https://www.plama.bg/	Connection problems
2	201088101	https://fashionsupreme.co.uk/contact/	Timeout occurred
3	202634919	www.chemcos.eu	Request exception
4	202854154	www.globewilliams.com	Request exception
5	203089329	www.megastroi.eu	Request exception
6	203352806	https://markanpro.bg/	Connection problems

Извличане на данни за характеристики на предприятията от уебсайтовете им

Преди да започне извличане на данни за характеристиките на предприятията от уебсайтовете им, се подготвят различни тематични списъци с ключови думи, които служат за търсене на различни он-лайн характеристики (Фигура 20), такива като:

- извършване на онлайн търговия през сайта на предприятието (88 ключови думи);
- наличие на обяви за работа на сайта на предприятието (37 ключови думи);
- профили на предприятието в социални мрежи (14 ключови думи);
- информация за контакт;
- политика за бисквитки;
- политика за лични данни;
- условия за използване;
- сертификация на предприятието по различни стандарти;
- профили на потребители на сайта.

От изброените характеристики, само за първите три бяха изчислени експериментални статистически данни.

Фиг. 20. Ключови думи за характеристики на предприятията от уебсайтовете им

```
uf.get_obecwords()
```

```
Load file with (11, 1) rows and coloumns: .\sbr_data\OBEC_words_contact.txt
Load file with (13, 1) rows and coloumns: .\sbr_data\OBEC_words_cookie.txt
Load file with (88, 1) rows and coloumns: .\sbr_data\OBEC_words_ecommerce.txt
Load file with (12, 1) rows and coloumns: .\sbr_data\OBEC_words_gdpr.txt
Load file with (12, 1) rows and coloumns: .\sbr_data\OBEC_words_iso.txt
Load file with (37, 1) rows and coloumns: .\sbr_data\OBEC_words_job.txt
Load file with (14, 1) rows and coloumns: .\sbr_data\OBEC_words_socialmedia.txt
Load file with (10, 1) rows and coloumns: .\sbr_data\OBEC_words_tou.txt
Load file with (6, 1) rows and coloumns: .\sbr_data\OBEC_words_user.txt

OBEC term matrix words:
['address', 'contact', 'mail', 'phone', 'адрес', 'връзка', 'поща', 'мейл', 'контакт', 'свържеш', 'телефон', 'cookie', 'биск
витк', 'браузър', 'изтрива', 'индивидуалн', 'настройк', 'поверителност', 'политика', 'сайт', 'сигурност', 'съхранява', 'тре
ти', 'функционалност', 'american', 'bag', 'card', 'cart', 'commerce', 'dhl', 'easypay', 'ecommerce', 'e-commerce', 'econt',
'epay', 'eshop', 'e-shop', 'estore', 'e-store', 'euro', 'express', 'mastercard', 'online', 'order', 'password', 'paypal', '
postepay', 'rapido', 'shop', 'speedy', 'store', 'value', 'valuta', 'visa', 'акаунт', 'безплатно', 'валута', 'вземи', 'влиза
не', 'вноски', 'връщане', 'вход', 'ддс', 'добави', 'достав', 'достъп', 'еконт', 'желания', 'записи', 'изплащане', 'идентифи
кация', 'количка', 'комисионна', 'кредит', 'куриер', 'лв', 'лева', 'лист', 'любими', 'магазин', 'марк', 'наличност', 'намал
ени', 'оферт', 'парола', 'плащане', 'пратка', 'пополярни', 'поръчай', 'поръчка', 'потвържд', 'проверка', 'превод', 'предлож
ени', 'прода', 'продукт', 'промоци', 'проследяване', 'път', 'разход', 'рекламаци', 'слагам', 'спииди', 'спийди', 'списък',
'стоки', 'търго', 'търсен', 'цена', 'цени', 'ценова', '679', 'cvc', 'данни', 'егн', 'забравен', 'закон', 'защита', 'лични',
'потребител', 'право', 'регламент', '9001', '13485', '14001', '15050', '18001', '22000', '25424', '27001', '50001', 'iso',
'бдс', 'сертификат', 'candidate', 'career', 'experience', 'job', 'looking', 'position', 'vacanc', 'work', 'възможност', 'въ
знаграждение', 'екип', 'завършил', 'кадри', 'кандидат', 'кариер', 'конкурс', 'места', 'място', 'местожителство', 'наш', 'об
яв', 'опит', 'отворени', 'позици', 'предложе', 'при нас', 'работа', 'работни', 'свободн', 'стаж', 'стани', 'търси', 'умения
', 'част', 'човек', 'човешки', 'обучени', 'facebook', 'flickr', 'google', 'instagram', 'linkedin', 'myspace', 'picasa', 'pi
nterest', 'slideshare', 'tumbler', 'twitter', 'xing', 'yammer', 'youtube', 'гаранци', 'договор', 'използване', 'клиент', 'о
бщи', 'ползване', 'правила', 'продавач', 'условия', 'влез', 'забравена', 'регистр']
```

Извличане на информация за характеристики на предприятията от уебсайтовете и техните страници върху избраните ключови думи е проведено върху всяко от 100-те множества без повторение на съвкупността от предприятия с URLs адреси. Продължителността на този процес е 4 денонощия, като всяко множество се обработва за около 50 минути средно (фиг. 21).

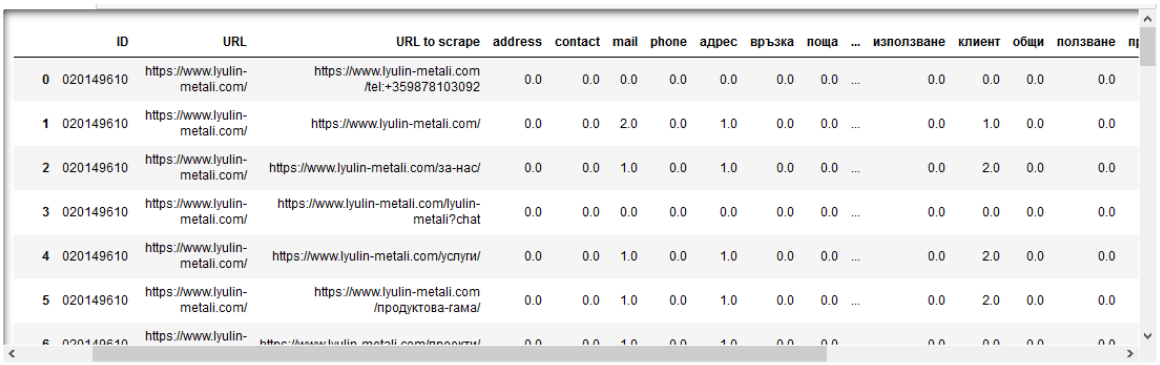
Фиг. 21. Успешно приключване на извличане на информация за предприятията от предложените интернет адреси и техните страници

```
dfnes=uf.slice_urls_to_scrape(timeout=10, sleep=0.5, slice=90, what='word_count')
```

```
Slice: 90
DataFrame columns: ['ID', 'URL', 'URL to scrape']
processed: 833160422 https://www.swe-flex.com/сервиз/клиентска-служба/ : 100%|[redacted]| 2240/2240 [55:57<00:00,  1.14s/it]
Slice: 91
DataFrame columns: ['ID', 'URL', 'URL to scrape']
processed: 833161111 https://www.dagaplus.com/za-nas/ : 100%|[redacted]| 2091/2091 [48:41<00:00,  1.36s/it]
Slice: 92
DataFrame columns: ['ID', 'URL', 'URL to scrape']
processed: 123635807 https://www.informator.bg/cgi-bin/index.pl?_state=AjaxMapAddress&q=гр. Стара Загора,  ул. Димчо Стаев
№30Б :  34%|▌| 709/2108 [18:04<29:57,  1.29s/it]
processed: 123635807 https://www.informator.bg/cgi-bin/index.pl?_state=AjaxMapAddress&q=t]
processed: 131262158 https://plesio.bg/search.html?codes=1352407,2035359,2049546,2159724,3023974,3023990,2342901,2343010,23
43029,2411172,2901986,2901994,2977613&sort=priceasc
processed: 131262158 https://plesio.bg/search.html?codes=1352407,2035359,2049546,2159724,3023974,3023990,2342901,2343010,23
43029,2411172,2901986,2901994,2977613&sort=priceasc
processed: 833163129 https://www.raya-press.com/inovacii-konkurentnosposobnost/ : 100%|[redacted]| 2108/2108 [52:58<00:00,  1.42s/i
t]
```

След приключване на работата на този етап софтуерът връща данни за броя на откритите ключови думи на всяка от избраните страници от уебсайтовете на предприятията (фиг. 22).

Фиг. 22. Брой на откритите ключови думи на всяка от избраните страници от уебсайтовете на предприятията



	ID	URL	URL to scrape	address	contact	mail	phone	адрес	връзка	поща	...	използване	клиент	общи	ползване	пр
0	020149610	https://www.lyulin-metali.com/	https://www.lyulin-metali.com/tel:+359878103092	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
1	020149610	https://www.lyulin-metali.com/	https://www.lyulin-metali.com/	0.0	0.0	2.0	0.0	1.0	0.0	0.0	...	0.0	1.0	0.0	0.0	
2	020149610	https://www.lyulin-metali.com/	https://www.lyulin-metali.com/за-нас/	0.0	0.0	1.0	0.0	1.0	0.0	0.0	...	0.0	2.0	0.0	0.0	
3	020149610	https://www.lyulin-metali.com/	https://www.lyulin-metali.com/lyulin-metali?chat	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	
4	020149610	https://www.lyulin-metali.com/	https://www.lyulin-metali.com/услуги/	0.0	0.0	1.0	0.0	1.0	0.0	0.0	...	0.0	2.0	0.0	0.0	
5	020149610	https://www.lyulin-metali.com/	https://www.lyulin-metali.com/продуктова-гама/	0.0	0.0	1.0	0.0	1.0	0.0	0.0	...	0.0	2.0	0.0	0.0	
6	020149610	https://www.lyulin-	https://www.lyulin-metali.com/проекти/	0.0	0.0	1.0	0.0	1.0	0.0	0.0		0.0	0.0	0.0	0.0	

Установените грешки при работата на софтуера са 921, или по-малко от 0.5% от всички проверени страници (фиг. 23).

Фиг. 23. Грешки при извличане на ключови думи от страниците на уебсайтовете на предприятията

	ID	URL	URL to scrape	Error
0	108692434	http://www.techno-lux.com/	http://www.techno-lux.com/	Timeout occurred
1	115545438	https://tvsatcom.bg	https://tvsatcom.bg/category/shows/kinomaraton/	Connection problems
2	131335001	https://efellows.bg/	https://efellows.bg/about-us	Timeout occurred
3	201302854	https://www.oplus.bg	https://www.oplus.bg/catalogue/hartieni_izdeli...	Timeout occurred
4	201302854	https://www.oplus.bg	https://www.oplus.bg/catalogue/hartieni_izdeli...	Timeout occurred
5	201302854	https://www.oplus.bg	https://www.oplus.bg/catalogue/hartieni_izdeli...	Timeout occurred
6	201302854	https://www.oplus.bg	https://www.oplus.bg/catalogue/hartieni_izdeli...	Timeout occurred
7	201302854	https://www.oplus.bg	https://www.oplus.bg/catalogue/hartieni_izdeli...	Timeout occurred
8	201302854	https://www.oplus.bg	https://www.oplus.bg/catalogue/hartieni_izdeli...	Timeout occurred
9	201302854	https://www.oplus.bg	https://www.oplus.bg/catalogue/hartieni_izdeli...	Timeout occurred
10	201302854	https://www.oplus.bg	https://www.oplus.bg/catalogue/hartieni_izdeli...	Timeout occurred
11	201302854	https://www.oplus.bg	https://www.oplus.bg/catalogue/hartieni_izdeli...	Timeout occurred

Подготовка на модел за Логистична регресия за намиране на ОВЕС

Получените данни от предишния етап са използвани за намиране на онлайн характеристики на предприятията с използване на Логистична регресия. За целта с помощта на методи от софтуера е приготвен файл, съдържащ следните полета (фиг. 24):

- ID - ЕИК на предприятието;
- URL - интернет адреса;

- полета с тематични ключови думи за съответната търсена онлайн характеристика, съдържащи единица, ако думата е намерена в наблюдаваните страници от уебсайта на предприятието, иначе 0;
- OBEC - известната страница от сайта на предприятието, на която се среща търсената онлайн характеристика;
- Known OBEC - поле с 1, ако знаем търсената характеристика за даденото предприятие, иначе 0;
- Link position - необходимо поле за предишния процес за намиране на интернет адрес;
- Sum - сума от полетата с ключови думи за характеристиката;
- Score - необходимо поле за предишния процес за намиране на интернет адрес.

Фиг. 24. Файл за ML с Логистична регресия за намиране на OBEC (в случая за наличие на обяви за работа на сайта на предприятието)

```
df[0].info(verbose=False)
df[0].sample(5)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10533 entries, 0 to 10532
Columns: 44 entries, ID to Score
dtypes: float64(39), int64(2), object(3)
memory usage: 3.5+ MB
```

	ID	URL	candidate	career	experience	job	looking	position	vacanc	work	...	умения	част	човек	човешки	обучени	
9657	816089656	https://www.holcim.bg/bg	0.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	...	1.0	1.0	1.0	1.0	1.0	h /
10122	831430207	https://amperel.net/bg/	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	0.0	1.0	0.0	
4738	130589420	https://antares-bg.net/	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	1.0	1.0	1.0	
9748	822106665	https://www.vikpz.com/	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	h
6306	175264381	http://www.call-ex.com/	0.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	...	1.0	1.0	0.0	1.0	1.0	

5 rows × 44 columns

idate	career	experience	job	looking	position	vacanc	work	...	умения	част	човек	човешки	обучени	OBEC	Known OBEC	Link position	sum	Score
0.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	...	1.0	1.0	1.0	1.0	1.0	https://www.holcim.bg/bg/kariera-v-kholsim/svo...	1	1	26.0	25.74
0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	0.0	1.0	0.0	NaN	0	2	14.0	13.86
0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	1.0	1.0	1.0	NaN	0	2	22.0	21.78
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	https://www.vikpz.com/page/15/karieri.html	1	1	7.0	6.93
0.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	...	1.0	1.0	0.0	1.0	1.0	http://www.call-ex.com/bg/careers/	1	1	23.0	22.77

Полученият файл се използва за намиране на онлайн характеристика на предприятието. С помощта на метод от софтуера данните са разделени на 70% обучително множество и 30% тестово множество за машинно самообучение с Логистична регресия. За избраното произволно състояние[42] 3333 са получени 244

[42] random_state - https://scikit-learn.org/stable/glossary.html#term-random_state

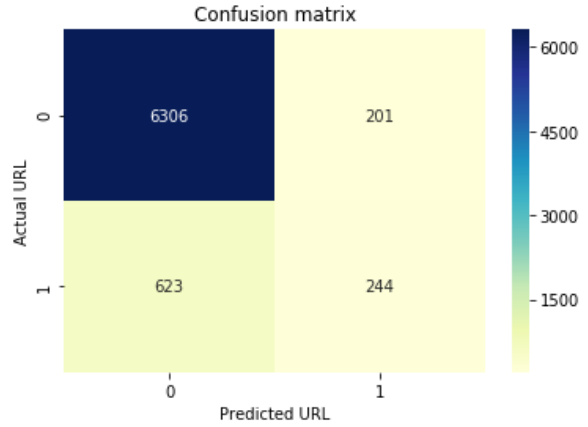
истински положителни, 6 306 истински отрицателни, 201 фалшиво положителни и 623 фалшиво отрицателни резултата за онлайн характеристиката за наличие на обяви за работа на сайта на предприятието (фиг. 25).

Фиг. 25. Матрица на неточностите за use-case 2

```
lr = ml.logistic_regression_fit(test_size=0.7, random_state=3333,
```

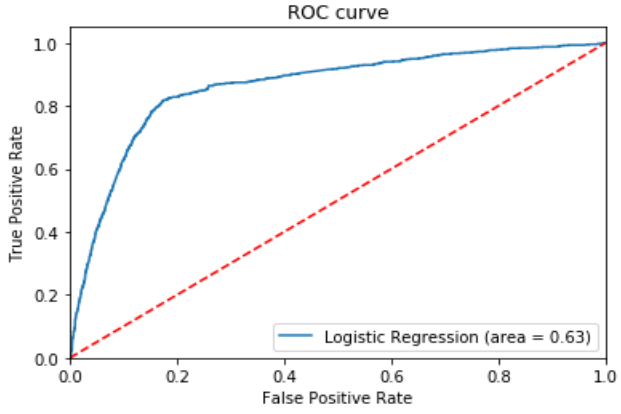
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10533 entries, 0 to 10532
Columns: 44 entries, ID to Score
dtypes: float64(39), int64(2), object(3)
memory usage: 3.6+ MB
None
```

```
ml.prepareCM(lr[3],lr[4])
```



От матрицата може да се направи заключението, че моделът по-скоро предсказва произволно отколкото точно, което може да се дължи на неподходящи ключови думи за онлайн характеристиката на предприятието или на голям шум в използването на тези думи по интернет страниците. Същият извод се налага и от ROC-кривата (фиг. 26), където се вижда, че площта над синята линия е значителна и синята линия е далече от стойност 1.

Фиг. 26. ROC-крива за намиране на ОВЕС (в случая за наличие на обяви за работа на сайта на предприятието)



Въпреки не толкова добрите резултати на модела той е приложен за произволни състояния от 1 до 100 на Логистичната регресия за намиране на онлайн характеристики на предприятията за наличие на електронен магазин и обяви за работа на сайтовете на предприятията. Резултатите от 100-те итерации са интегрирани и изчистени от повторения по ЕИК на предприятието за всяка от двете наблюдавани характеристики (фиг. 27).

Фиг. 27. Резултат от 100 итерации на Логистичния модел за наличие на обяви за работа на сайтовете на предприятията

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1956 entries, 368 to 1052772
Columns: 10 entries, ID to 1
dtypes: float64(4), int64(3), object(3)
memory usage: 168.1+ KB
```

Out[19]:

	ID	URL	OBEC	Known OBEC	Link position	sum	Score	predict	0	1
368	000070536	http://www.galateabg.net/	http://galateabg.net/main_bg/10870_karieri.html	1	1	13.0	12.87	1	0.163721	0.836279
980	000179553	http://gt-vratza.com/	http://gt-vratza.com/tyrgove-konkursi/	1	1	10.0	9.90	0	0.861697	0.138303
1297	000220021	http://monek-bg.com/	NaN	0	2	13.0	12.87	1	0.477315	0.522685
2743	000443113	https://www.addisan.bg/	NaN	0	2	13.0	12.87	1	0.382640	0.617360
3195	000551929	https://www.silistra.avtogara.eu/bg/	https://www.silistra.avtogara.eu/bg/	1	1	3.0	2.97	0	0.928014	0.071986
3347	000600649	http://mihalkovo.com/bg/	NaN	0	2	23.0	22.77	1	0.361865	0.638135

След ръчно експертно валидиране е установено, че от предложените 1 917 URLs адреса за електронна търговия на предприятията 1 329 са верни, а от предложените 1 956 интернет адреса за страници с обяви за работа на сайтовете на предприятията само 1 652 са верни.

3. Присъствие/профили на предприятията в социалните мрежи

За намиране на профилите на предприятията в социалните мрежи е използван SocialMediaProfiles[43] софтуер на Python, част от средството Starter Kit. Софтуерът проверява сайтове за наличие на връзки към профили в следните социални мрежи:

- Facebook;
- Twitter;
- Youtube;
- LinkedIn;
- Instagram;
- Xing;
- Pinterest.

След подаване на 12 288-те URLs адреса на предприятия, които са известни от use-case 1, софтуерът намира общо 4 398 профила в социални мрежи на предприятия (фиг. 28).

[43] https://github.com/EnterpriseCharacteristicsESSnetBigData/StarterKit/tree/master/SocialMediaProfiles

Фиг. 28. Намиране на профили в социалните мрежи на предприятия със софтуер SocialMediaProfiles

Резултати

На базата на получените данни от уебсайтовете на предприятията с 10 и повече заети е установено, че:

- 4.7% от предприятията имат собствен електронен магазин;
- 10.8% от предприятията с уебсайт извършват електронна търговия през уебсайта си;
- 5.8% от предприятията имат собствена страница с обяви за работа в интернет;
- 13.4% от предприятията с уебсайт имат страница с обяви за работа на него;
- Най-много предприятия имат електронни магазини в София (столица) - 6.9%, а най-малко - в област Видин - 0.6% (фиг. 29);
- От предприятията с уебсайтове най-много електронни магазини имат в област Хасково - 15.2%, а най-малко - отново в област Видин - 2.3% (фиг. 29);
- За предприятията от сектор „Производство и разпространение на електрическа и топлинна енергия и на газообразни горива“ (D) не са открити електронни магазини, докато предприятията от сектор „Други дейности“ (S) е най-вероятно да имат такъв (фиг. 30);
- Най-много предприятия имат страници с обяви за работа в област София (столица) - 11%, а най-малко - в област Видин - 0.3% (фиг. 31);
- От предприятията с уебсайтове най-много страници с обяви за работа имат в област София (столица) - 19.5%, а най-малко - отново в област Кърджали - 1% (фиг. 32);
- Предприятията от сектор „Производство и разпространение на електрическа и топлинна енергия и на газообразни горива“ (D) имат най-много страници с обяви за работа на техните сайтове - 27.9%, докато предприятията от сектор „Хотелиерство и ресторантьорство“ (I) имат най-малко - 7.4% (фиг. 33);

- 19.9% от предприятията с уебсайт имат профили в социални мрежи, като най-популярна сред тях е Facebook - с 18.5% (фиг. 34);
- От предприятията с уебсайтове най-много профили в социални мрежи имат в област Силистра - 29.7%, а най-малко - в област Габрово - 13% (фиг. 34);
- Предприятията от сектор „Хотелиерство и ресторантьорство“ (I) имат най-много профили в социални мрежи на техните сайтове - 29.1%, докато предприятията от сектор „Доставяне на води; Канализационни услуги, управление на отпадъци и възстановяване“ (E) имат най-малко - 12.3% (фиг. 35).

Фиг. 29. Електронни магазини на предприятията с 10 и повече заети по области

Електронни магазини на предприятията с 10 и повече заети по области

Код по NUTS	Област	Предприятия брой	Интернет адреси брой	Електронни магазини брой	Отношение на Електронните магазини към Предприятията (%)	Отношение на Електронните магазини към Интернет адресите (%)
BG311	Видин	177	43	1	0.6	2.3
BG312	Монтана	336	90	6	1.8	6.7
BG313	Враца	379	145	11	2.9	7.6
BG314	Плевен	666	204	24	3.6	11.8
BG315	Ловеч	392	163	10	2.6	6.1
BG321	Велико Търново	744	283	26	3.5	9.2
BG322	Габрово	471	231	21	4.5	9.1
BG323	Русе	848	400	34	4	8.5
BG324	Разград	270	77	7	2.6	9.1
BG325	Силистра	243	64	3	1.2	4.7
BG331	Варна	2208	1001	113	5.1	11.3
BG332	Добрич	468	151	17	3.6	11.3
BG333	Шумен	476	179	11	2.3	6.1
BG334	Търговище	298	84	9	3	10.7
BG341	Бургас	1678	617	63	3.8	10.2
BG342	Сливен	454	166	13	2.9	7.8
BG343	Ямбол	330	125	14	4.2	11.2
BG344	Стара Загора	1057	470	54	5.1	11.5
BG411	София (столица)	8760	4932	606	6.9	12.3
BG412	София	682	254	12	1.8	4.7
BG413	Благоевград	1400	337	30	2.1	8.9
BG414	Перник	361	99	8	2.2	8.1
BG415	Кюстендил	373	106	9	2.4	8.5
BG421	Пловдив	2908	1347	158	5.4	11.7
BG422	Хасково	699	231	35	5	15.2
BG423	Пазарджик	792	287	22	2.8	7.7
BG424	Смолян	409	104	9	2.2	8.7
BG425	Кърджали	372	98	3	0.8	3.1

Фиг. 30. Електронни магазини на предприятията с 10 и повече заети по сектори на икономическата дейност

Електронни магазини на предприятията с 10 и повече заети по сектори на икономическа дейност

Сектор на икономическа дейност	Предприятия брой	Интернет адреси брой	Електронни магазини брой	Отношение на Електронните магазини към Предприятията (%)	Отношение на Електронните магазини към Интернет адресите (%)
Административни и спомагателни дейности (N)	1256	528	20	1.6	3.8
Доставяне на води; Канализационни услуги, управление на отпадъци и възстановяване (E)	253	138	1	0.4	0.7
Други дейности (S)	15	9	4	26.7	44.4
Операции с недвижими имоти (L)	542	239	7	1.3	2.9
Преработваща промишленост (C)	7466	3842	337	4.5	8.8
Производство и разпространение на електрическа и топлинна енергия и на газообразни горива (D)	133	86	0	0	0
Професионални дейности и научни изследвания (M)	1278	739	6	0.5	0.8
Строителство (F)	3180	1260	19	0.6	1.5
Създаване и разпространение на информация и творчески продукти; Далекосъобщения (J)	1235	813	44	3.6	5.4
Транспорт, складиране и пощи (H)	2107	687	15	0.7	2.2
Търговия; Ремонт на автомобили и мотоциклети (G)	7761	3164	818	10.5	25.9
Хотелиерство и ресторантьорство (I)	3025	783	58	1.9	7.4

Фиг. 31. Страници с обяви за работа на предприятията с 10 и повече заети по области

Страници с обяви за работа на предприятията с 10 и повече заети по области

Код по NUTS	Област	Предприятия брой	Интернет адреси брой	Страници с обяви за работа брой	Отношение на Страници с обяви за работа към Предприятията (%)	Отношение на Страници с обяви за работа към Интернет адресите (%)
BG311	Видин	177	43	1	0.6	2.3
BG312	Монтана	336	90	4	1.2	4.4
BG313	Враца	379	145	18	4.7	12.4
BG314	Плевен	666	204	15	2.3	7.4
BG315	Ловеч	392	163	9	2.3	5.5
BG321	Велико Търново	744	283	27	3.6	9.5
BG322	Габрово	471	231	15	3.2	6.5
BG323	Русе	848	400	53	6.2	13.2
BG324	Разград	270	77	7	2.6	9.1
BG325	Силистра	243	64	6	2.5	9.4
BG331	Варна	2208	1001	123	5.6	12.3
BG332	Добрич	468	151	13	2.8	8.6
BG333	Шумен	476	179	16	3.4	8.9
BG334	Търговище	298	84	5	1.7	6
BG341	Бургас	1678	617	49	2.9	7.9
BG342	Сливен	454	166	10	2.2	6
BG343	Ямбол	330	125	13	3.9	10.4
BG344	Стара Загора	1057	470	41	3.9	8.7
BG411	София (столица)	8760	4932	961	11	19.5
BG412	София	682	254	27	4	10.6
BG413	Благоевград	1400	337	23	1.6	6.8
BG414	Перник	361	99	8	2.2	8.1
BG415	Кюстендил	373	106	4	1.1	3.8
BG421	Пловдив	2908	1347	154	5.3	11.4
BG422	Хасково	699	231	14	2	6.1
BG423	Пазарджик	792	287	29	3.7	10.1
BG424	Смолян	409	104	6	1.5	5.8
BG425	Кърджали	372	98	1	0.3	1

Фиг. 32. Страници с обяви за работа на предприятията с 10 и повече заети по сектори на икономическата дейност

Страници с обяви за работа на предприятията с 10 и повече заети по сектори на икономическа дейност

Сектор на икономическа дейност	Предприятия брой	Интернет адреси брой	Страници с обяви за работа брой	Отношение на Страници с обяви за работа към Предприятията (%)	Отношение на Страници с обяви за работа към Интернет адресите (%)
Административни и спомагателни дейности (N)	1256	528	105	8.4	19.9
Доставяне на води; Канализационни услуги, управление на отпадъци и възстановяване (E)	253	138	23	9.1	16.7
Други дейности (S)	15	9	1	6.7	11.1
Операции с недвижими имоти (L)	542	239	26	4.8	10.9
Преработваща промишленост (C)	7466	3842	327	4.4	8.5
Производство и разпространение на електрическа и топлинна енергия и на газообразни горива (D)	133	86	24	18	27.9
Професионални дейности и научни изследвания (M)	1278	739	191	14.9	25.8
Строителство (F)	3180	1260	113	3.6	9
Създаване и разпространение на информация и творчески продукти; Далекосъобщения (J)	1235	813	197	16	24.2
Транспорт, складиране и пощи (H)	2107	687	127	6	18.5
Търговия; Ремонт на автомобили и мотоциклети (G)	7761	3164	460	5.9	14.5
Хотелиерство и ресторантьорство (I)	3025	783	58	1.9	7.4

Фиг. 33. Присъствие в социалните мрежи на предприятия с 10 и повече заети

Присъствие в социални мрежи на предприятия с 10 и повече заети

Социални мрежи Присъствие	брой	Отношение към всички предприятия (%)	Отношение към предприятията с интернет адрес (%)
Предприятия	2450	8.7	19.9
Facebook	2269	8	18.5
Twitter	558	2	4.5
Youtube	592	2.1	4.8
LinkedIn	551	2	4.5
Instagram	330	1.2	2.7
Xing	9	0	0.1
Pinterest	89	0.3	0.7

Фиг. 34. Присъствие в социалните мрежи на предприятията с 10 и повече заети по области

Присъствие в социални мрежи на предприятията с 10 и повече заети по области

Код по NUTS	Област	Предприятия брой	Интернет адреси брой	Присъствие в социални мрежи брой	Отношение на Присъствие в социални мрежи към Предприятията (%)	Отношение на Присъствие в социални мрежи към Интернет адресите (%)
BG311	Видин	177	43	6	3.4	14
BG312	Монтана	336	90	17	5.1	18.9
BG313	Враца	379	145	27	7.1	18.6
BG314	Плевен	666	204	33	5	16.2
BG315	Ловеч	392	163	30	7.7	18.4
BG321	Велико Търново	744	283	52	7	18.4
BG322	Габрово	471	231	30	6.4	13
BG323	Русе	848	400	74	8.7	18.5
BG324	Разград	270	77	15	5.6	19.5
BG325	Силистра	243	64	19	7.8	29.7
BG331	Варна	2208	1001	216	9.8	21.6
BG332	Добрич	468	151	35	7.5	23.2
BG333	Шумен	476	179	34	7.1	19
BG334	Търговище	298	84	17	5.7	20.2
BG341	Бургас	1678	617	120	7.2	19.4
BG342	Сливен	454	166	29	6.4	17.5
BG343	Ямбол	330	125	26	7.9	20.8
BG344	Стара Загора	1057	470	73	6.9	15.5
BG411	София (столица)	8760	4932	1007	11.5	20.4
BG412	София	682	254	46	6.7	18.1
BG413	Благоевград	1400	337	62	4.4	18.4
BG414	Перник	361	99	15	4.2	15.2
BG415	Кюстендил	373	106	25	6.7	23.6
BG421	Пловдив	2908	1347	295	10.1	21.9
BG422	Хасково	699	231	57	8.2	24.7
BG423	Пазарджик	792	287	48	6.1	16.7
BG424	Смолян	409	104	24	5.9	23.1
BG425	Кърджали	372	98	18	4.8	18.4

Фиг. 35. Присъствие в социалните мрежи на предприятията с 10 и повече заети по сектори на икономическата дейност

Присъствие в социални мрежи на предприятията с 10 и повече заети по сектори на икономическа дейност

Сектор на икономическа дейност	Предприятия брой	Интернет адреси брой	Присъствие в социални мрежи брой	Отношение на Присъствие в социални мрежи към Предприятията (%)	Отношение на Присъствие в социални мрежи към Интернет адресите (%)
Административни и спомагателни дейности (N)	1256	528	115	9.2	21.8
Доставяне на води; Канализационни услуги, управление на отпадъци и възстановяване (E)	253	138	17	6.7	12.3
Други дейности (S)	15	9	2	13.3	22.2
Операции с недвижими имоти (L)	542	239	40	7.4	16.7
Преработваща промишленост (C)	7466	3842	734	9.8	19.1
Производство и разпространение на електрическа и топлинна енергия и на газообразни горива (D)	133	86	12	9	14
Професионални дейности и научни изследвания (M)	1278	739	147	11.5	19.9
Строителство (F)	3180	1260	208	6.5	16.5
Създаване и разпространение на информация и творчески продукти; Далекосъобщения (J)	1235	813	157	12.7	19.3
Транспорт, складиране и пощи (H)	2107	687	134	6.4	19.5
Търговия; Ремонт на автомобили и мотоциклети (G)	7761	3164	656	8.5	20.7
Хотелиерство и ресторантьорство (I)	3025	783	228	7.5	29.1

VI. Заключение

През последните десетилетия комбинираното използване на данни от различни източници за статистически цели се превърна в консолидирана практика. Наред с използването на административни данни за създаване на статистически регистри експериментирането по отношение на използването на източници на големи данни продължава, за да се актуализира, разшири и валидира наличната информация в статистическия Бизнес регистър. Най-използваният и най-достъпният източник за намиране на големи данни е интернет мрежата. Голямото количество налична информация предоставя нови възможности, но също така и нови предизвикателства пред статистическите експерти по интеграция на данни, предвид структурните разлики между административните и уебданните.

В административните източници идентификацията на единиците е сигурна. В допълнение, интегрирането на данните за единицата е лесно, тъй като те използват общ идентификационен код (данъчен код, ДДС номер и др.). Чрез консолидиран процес на интеграция се присвоява идентификационен код за правните единици и евентуално идентификационен код за предприятия.

За разлика от административните данни големите данни крият сериозни рискове. Някои от тях са очевидни като трудността да се управляват бързо нарастващи обеми данни, водещи до голямо потребление на изчислителни ресурси и ресурси за съхранение. Освен това има технически ограничения за решаване като например дългосрочното време, необходимо на скрапера да обходи цялото съдържание, и ограниченията, свързани с информационната сигурност на уебсайтовете, които пречат на автоматичния достъп. Не са за подценяване и статистическите проблеми като например: трудността да се удостовери качеството на уебинформацията и надеждността на данните, как да се съчетава уебинформацията за дадено предприятие с тази в СБР и как да се гарантира сигурността на класифицирането в съвкупността, към която извлечените от уебмрежата данни принадлежат.

Въпреки посочените рискове използването на уебданни има и очевидни ползи като например обогатяване на статистическото производство с нова информация, подобряване на навременността на статистическите продукти и увеличаване на приложимостта на бизнес статистиката за сметка на по-ниски разходи в сравнение с увеличаването на съществуващите данни. Силен аргумент в тази посока е също, че интернет мрежата е независим източник на данни, докато всички останали източници - административни и статистически, могат да се считат за свързани по някакъв начин и да си влияят взаимно. Цялостната картина, предоставена от уебданните, е със сигурност по-реална, а именно: как едно предприятие вижда себе си и как иска да се представи пред потребителите си.

Сравнително новата идея да се използват големи данни като допълнителен източник за СБР чрез използване на методите за уебскрапване и технологии за извличане на текст с цел интегриране на „структурираните“ бизнес данни с „неструктурираните“ уебданни е добре приета от статистическата общност в ЕСС. Техниките за големи данни променят начина, по който се събират, обработват, анализират и интегрират данни.

Добавената стойност в това отношение се крие точно в информацията, която е скрита в данните и в тяхното проактивно използване, т.е. четенето и използването на данните като отправна точка за създаване на стратегия. Следователно интегрирането на инструментите за анализ на големи данни в контекста на традиционния статистически производствен процес е трудна задача. Това всъщност означава да се комбинират управлявани от данни процеси - въз основа на входни данни, които не идват от статистически източници, нееднородни са, неструктурирани и нестабилни във времето - с процеси, базирани на ориентиран към изхода подход, тъй като в контекста на официалната статистика производственият процес е изграден с оглед получаване на статистически изходи.

Интернет като източник на данни представлява нови възможности и предизвикателства за официалната статистика, която трябва да включи всички иновационни, потенциални източници на данни, колкото е възможно повече в концептуалния дизайн на своите изследвания. Все повече национални статистически организации експериментират с използването на алтернативни източници на данни, за да произвеждат една и съща или нова статистическа информация по-ефективно и с по-високо ниво на качество в една наситена среда с много източници на данни.

Има няколко важни задачи, които предстои да бъдат решени от националните статистически служби по отношение на ефективното използване на източници на големи данни като цяло. Основната задача обаче е как на практика да се премине от експериментиране към реално производство на статистика от големи данни. Тази стъпка включва различни аспекти, вариращи от спазване на поверителността на личните данни до необходимостта от изграждане на цялостно нова инфраструктура (методологична, технологична, организационна), както и придобиване на нови умения от експертния състав.

На базата на получените резултати може да се твърди, че извлечените от интернет мрежата данни за онлайн характеристиките на предприятията от техните уебсайтове притежават необходимия потенциал да бъдат допълнителен източник за производство на официална статистика. Това е и основната цел, която се постига чрез изпълнението на първите два случая на използване (описани в настоящата статия), а именно: да се вземе решение за използване на новата информация за производство на още по-подробна статистика за използването на ИКТ в предприятията.