

Analiza istniejących rozwiązań opisujących grupowanie kwalifikacji

Michał Marcińczuk¹, Grzegorz Wojdyga², Łukasz Kobyliński²

¹Politechnika Wrocławska

²IPI PAN

Table of contents

1. Opis zadania
2. Zbiór ZRK
3. Grupowanie kompetencji z bazy ZRK
4. Odnoszenie ZRK do baz zewnętrznych
5. Wnioski

Opis zadania

1. Jak wykorzystać zasoby i narzędzia (w tym szczególnie te opracowane w ramach sieci CLARIN) do automatycznego porównywania i identyfikacji części wspólnych kwalifikacji w ZRK?
2. Jaki zakres narzędzi i zasobów jest potrzebny, by możliwe było wykorzystania algorytmów NLP do automatycznego odnoszenia kwalifikacji do wybranych klasyfikacji, w tym ESCO
3. Metody i narzędzia do automatycznego klastrowania kwalifikacji w oparciu o ich charakterystyki, bez odniesienia do zewnętrznych kryteriów
4. Możliwość dostosowania narzędzi służących do wykrywania plagiatu, do automatycznego porównywania i identyfikacji części wspólnych kwalifikacji w ZRK

Zbiór ZRK

**Fundusze Europejskie**
Wiedza. Edukacja. Rozwój.

**Rzeczpospolita Polska**

IBE  **INSTYTUT BADAŃ EDUKACYJNYCH**

Unia Europejska
Europejski Fundusz Społeczny 

Zintegrowany Rejestr Kwalifikacji

[Strona główna ZRK](#) [Szukam kwalifikacji](#) [Szukam podmiotu](#) [Warto wiedzieć](#) [ZRK](#) [Przejdź na stronę ZSK](#) [Seminaria i warsztaty](#) [Zaloguj się](#)



Kwalifikacje

Przeglądanie/wyszukiwanie kwalifikacji



Podmioty

Przeglądanie/wyszukiwanie instytucji systemu kwalifikacji



Dokumenty

Najważniejsze informacje dla użytkownika rejestru

ZRK w liczbach

**9997**
Kwalifikacje

**24**
Dziedziny

**1797**
Podmioty

**632**
Użytkownicy zarejestrowani

WAŻNE:  Przygotowanie opisów syntetycznych charakterystyk kwalifikacji pełnych właściwych dla szkolnictwa wyższego

- **wersja** — maj 2019,
- **liczba kompetencji** — 9682,
- **kluczowe pola:**
 - id,
 - nazwa,
 - charakterystyka.

Strona główna / Kwalifikacje

Kwalifikacje

Eksport

Wyświetlone 21-30 z 9997.

Wyszukiwanie zaawansowane

Lp.	Nazwa kwalifikacji ⓘ	Kategoria kwalifikacji	Rodzaj kwalifikacji	Poziom PRK/ERK	
21	Aktor scen muzycznych (343601)	ze szkolnictwa zawodowego	pełna	5	👁
22	Aktywne prowadzenie sprzedaży skierowanej do klientów biznesowych – przedstawiciel handlowy	rynkowe	częstkowa	4	👁
23	Animator kultury (343901)	ze szkolnictwa zawodowego	pełna	6	👁
24	Asystent Fryzjera (932919) od 1 września 2017r.	ze szkolnictwa zawodowego	pełna	2	👁
25	Asystent kierownika produkcji filmowej / telewizyjnej (343902)	ze szkolnictwa zawodowego	pełna	4	👁
26	Asystent kierownika produkcji filmowej (343902) od 1 września 2019r.	ze szkolnictwa zawodowego	pełna	4	👁
27	Asystent osoby niepełnosprawnej (341201)	ze szkolnictwa zawodowego	pełna	4	👁
28	Asystent osoby niepełnosprawnej (341201) od 1 września 2019r.	ze szkolnictwa zawodowego	pełna	4	👁
29	Asystentka stomatologiczna (325101)	ze szkolnictwa zawodowego	pełna	4	👁
30	Asystentka stomatologiczna (325101) od 1 września 2019r.	ze szkolnictwa zawodowego	pełna	4	👁

« 1 2 3 4 5 6 7 8 9 10 »

Figure 1: Przykładowe kwalifikacje z bazy ZRK

<https://rejestr.kwalifikacje.gov.pl/frontend/index.php?r=kwalifikacja%2Fview&id=12545>

[Strona główna](#) / [Kwalifikacje](#) / Technik hutnik (311704) od 1 września 2017r.

Kwalifikacja - podgląd

Informacje ogólne

Efekty uczenia się

Podmioty

Pozostałe informacje

Nazwa kwalifikacji

Technik hutnik (311704) od 1 września 2017r.

Skrót nazwy

Rodzaj kwalifikacji

kwalifikacja pełna

Poziom PRK/ERK

4

Krótką charakterystyką kwalifikacji, obejmującą informacje o działaniach lub zadaniach, które potrafi wykonywać osoba posiadająca tę kwalifikację

Absolwent szkoły kształcącej w zawodzie Technik hutnik powinien być przygotowany do wykonywania następujących zadań zawodowych:

- 1) obsługiwanie maszyn i urządzeń w procesach technologicznych;
- 2) konserwacji i przeglądów bieżących maszyn i urządzeń hutniczych;
- 3) prowadzenia kontroli wyrobu;
- 4) organizowania i prowadzenia procesów hutniczych;
- 5) wykonywania rozruchu urządzeń i systemów mechatronicznych;
- 6) organizowania i prowadzenie konserwacji oraz przeglądów bieżących maszyn i urządzeń hutniczych;
- 7) nadzorowania procesów hutniczych zgodnie z systemem zarządzania jakością.

Figure 2: Przykładowa kwalifikacje z bazy ZRK

Grupowanie kompetencji z bazy ZRK

- podobieństwo kompetencji przez analizę podobieństwa charakterystyk — pole tekstowe, opisowe,
- metody porównywania tekstów (charakterystyk kompetencji):
 - model wektorowy z ważeniem tf-idf,
 - modele językowe,
 - pokrywanie się tekstów.

I. Wstępna analiza z wykorzystaniem systemu WebSty

- **WebSty** (ws.clarin-pl.eu/websty.shtml) — webowy system do analizy podobieństwa treści,
- predefiniowane metody porównywania:
 - Analiza autorstwa,
 - Analiza stylu gramatycznego,
 - **Grupy podobieństwa treści**,
 - Klasyczna analiza autorstwa
- konfiguracja dla *Grupy podobieństwa treści*:
 - tokenizacja i analiza morfologiczna przy użyciu WCRFT2,
 - cechy: 1000 najczęstszych lematów rzeczowników,
 - lematy występujące tylko w jednym dokumencie były pomijane.
 - ważenie cech przy użyciu tf-idf.
- **Efekt:** brak wyników grupowania ze względu na zbyt dużą liczbę podobnych tekstów

Analiza zbioru

- krótkie (do 12 słów) — prawie 97% całego zbioru,
- długie (powyżej 12 słów) — ok 3% całego zbioru (ok 320 tekstów).

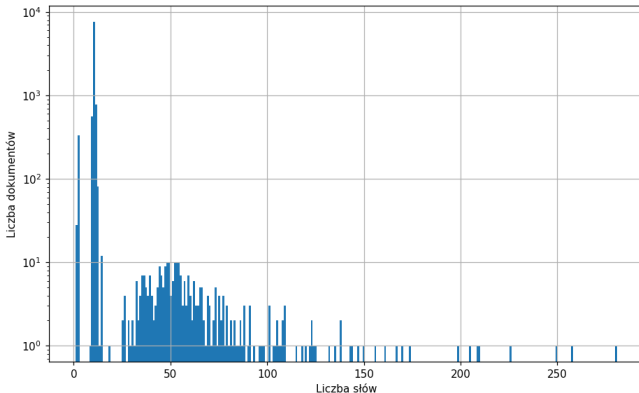


Figure 3: Rozkład długości charakterystyk kompetencji

Analiza zbioru

- krótkie (do 12 słów) — prawie 97% całego zbioru,
- długie (powyżej 12 słów) — ok 3% całego zbioru (ok 320 tekstów).

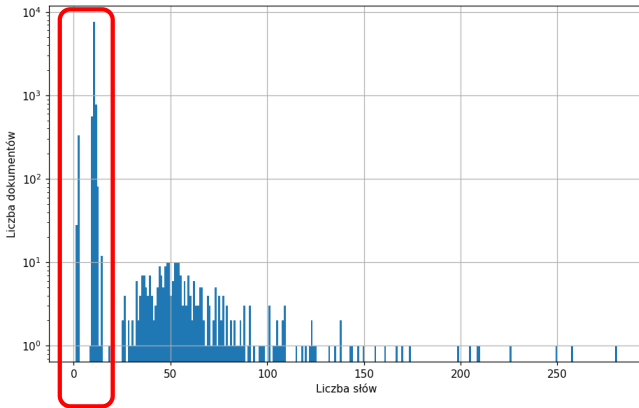


Figure 3: Rozkład długości charakterystyk kompetencji

Analiza zbioru

- krótkie (do 12 słów) — prawie 97% całego zbioru,
- długie (powyżej 12 słów) — ok 3% całego zbioru (ok 320 tekstów).

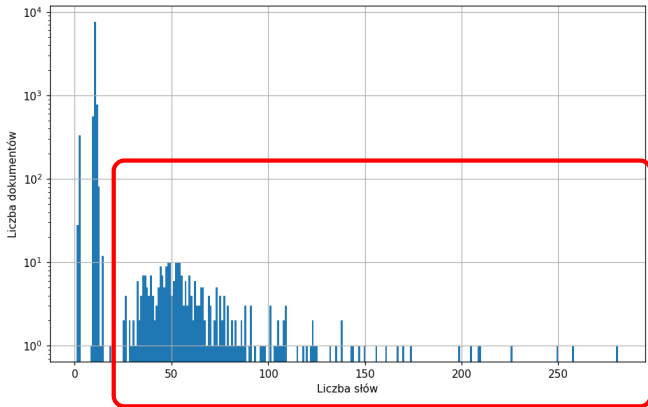


Figure 3: Rozkład długości charakterystyk kompetencji

Podobieństwo krótkich tekstów

Grupa 1 (85)

- tytuł zawodowy/naukowy: doktor nauk **biologicznych** w zakresie **biochemia**
- tytuł zawodowy/naukowy: doktor nauk **biologicznych** w zakresie **biotechnologia**
- tytuł zawodowy/naukowy: doktor nauk **chemicznych** w zakresie **biochemia**
- tytuł zawodowy/naukowy: doktor nauk **rolniczych** w zakresie **biotechnologia**
- tytuł zawodowy/naukowy: doktor nauk **rolniczych** w zakresie **inżynieria rolnicza**,
- (...)

Grupa 2 (9)

- ogólnoakademicki; tytuł zawodowy/naukowy: magister lub inny równorzędny;
- profil kształcenia: ogólnoakademicki; tytuł zawodowy/naukowy: inżynier lub inny równorzędny;
- profil kształcenia: praktyczny; tytuł zawodowy/naukowy: inżynier lub inny równorzędny;
- (...)

Grupa 3

- nie dotyczy
- nan

Podobieństwo długich tekstów (WebSty)

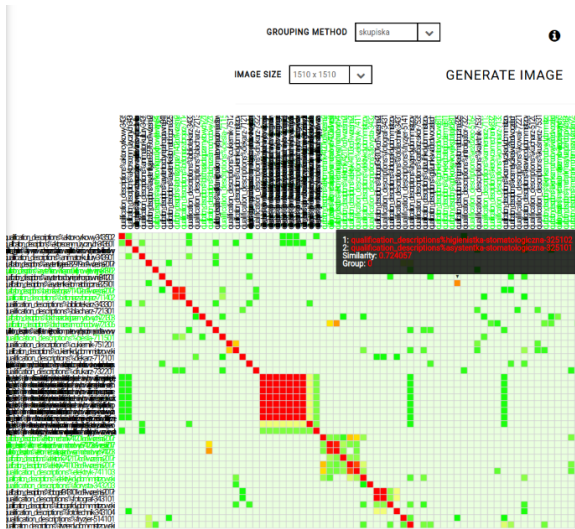


Figure 4: Mapa ciepła przedstawiająca podobieństwo tekstów

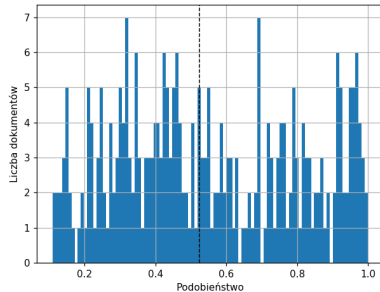


Figure 5: Rozkład podobieństwa dla najbardziej podobnych propozycji

- wyniki dla 286 z 320 tekstów,

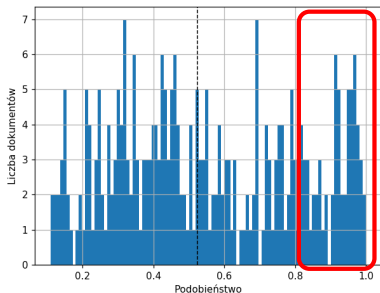


Figure 5: Rozkład podobieństwa dla najbardziej podobnych propozycji

- wyniki dla 286 z 320 tekstów,
- 0,8–1,0 (87 par) — poprawnie wskazane jako podobne,

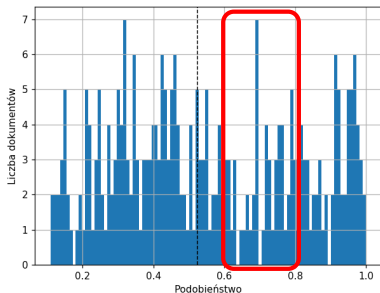


Figure 5: Rozkład podobieństwa dla najbardziej podobnych propozycji

- wyniki dla 286 z 320 tekstów,
- 0,8–1,0 (87 par) — poprawnie wskazane jako podobne,
- 0,6–0,8 (121 par) — niższa precyzja, **kryterium podobieństwa stało się podobieństwo struktury opisów**

Podobieństwo 0,65 dla pary:

[12543]: Operator maszyn i urządzeń hutniczych (812121) od 1 września 2017r.

[1]Absolwent szkoły kształcącej w zawodzie [2]Operator [3]maszyn [4]i urządzeń hutniczych [5]powinien być przygotowany do wykonywania następujących zadań zawodowych:
1) [7]obsługiwanie maszyn i urządzeń [9]w [10]procesach technologicznych[11];
[12]2[13]) [15]konserwacji [16]i przeglądów bieżących [17]maszyn i urządzeń [18]hutniczych;
3) prowadzenia kontroli wyrobu[19].

[12368] Mechanik – operator maszyn do produkcji drzewnej (817212)

[1]Absolwent szkoły kształcącej w zawodzie [2]mechanik-operator [3]maszyn [4]do produkcji drzewnej [5]powinien być przygotowany do wykonywania następujących zadań zawodowych:
1) [6]montowania maszyn i urządzeń do produkcji drzewnej;
2) kontrolowania i nadzorowania pracy maszyn i urządzeń do produkcji drzewnej;
3) [7]obsługiwanie maszyn i urządzeń[8], stosowanych [9]w [10]procesie produkcji materiałów drzewnych[11];
[12]4[13]) [14]prowadzenia bieżącej kontroli jakości surowców i produktów;
5) wykonywania przeglądów i [15]konserwacji [17]maszyn i urządzeń[19].

- **odcięcie listy cech do 1000 pozycji** — odcięcie po frekwencji powoduje, że terminy dziedzinowe są niewystarczająco częste, aby znalazły się na liście cech,
- **brak uwzględnienia przymiotników** — przymiotniki są dobrym wyznacznikiem dziedziny, np. *urządzenie **hutnicze***, a *urządzenie **rolnicze***,
- **brak uwzględnienia terminów wielowyrazowych** — wyrażenia identyfikujące określone zawody znajdowały się na liście terminów wskazanych przez TermoPL, np. *pojazd samochodowy, wyrób włókienniczy, obsługa ciągnika rolniczego*.

Rozszerzona analiza długich tekstów (TfidfVectorizer)

- lokalny potok przetwarzania:
 - tager morfologiczny — WCRFT2 poprzez API NlpRest2 (CLARIN-PL),
 - TfidfVectorizer z pakietu Gensim,
- konfiguracja:
 - n-gramy lematów długości od 1 do 3 tokenów,
 - ignorowanie wartości pos: interp, num, conj i prep,
 - ograniczenie do 4 tys. najczęstszych cech,
 - wykorzystanie stoplisty,
 - ważenie cech przy pomocy tf-idf.

TermoPL

<http://zil.ipipan.waw.pl/TermoPL>

<https://ws.clarin-pl.eu/termopl.shtml>

#	Ranking	Pojęcie	Pojęcie zlematyzowane	C-value	Długość	Freq_s	Freq_in	Kontekst
5	5	wykonywać	[wykonywanie]	65	1	657	363	74
13	13	zadanie	[zadanie]	27	1	289	284	12
12	12	zawód	[zawód]	27	1	284	108	9
1	1	zadanie zawodowy	[zadanie zawodowe]	2.6e+2	2	265	3	1
26	26	absolwent	[absolwent]	17	1	262	262	3
20	20	szkola	[szkola]	20	1	261	261	4
14	14	urządzenie	[urządzenie]	26	1	259	197	60
2	2	absolwent szkoła	[absolwent szkoły]	2.6e+2	2	256	0	0
18	18	praca	[praca]	21	1	218	198	56
34	30	przewodzić	[przewodzenie]	15	1	158	142	41
35	31	wyrób	[wyrób]	14	1	146	143	57
36	32	osoba	[osoba]	14	1	143	55	25
39	34	maszyna	[maszyna]	13	1	133	104	24
46	40	organizować	[organizowanie]	11	1	116	37	12
54	44	technika	[technika]	9.8	1	100	99	62

Porównanie wyników

- **wyższa liczba wyników** — wzrost z 286 do 309,
- **wyniki dla pozycji, dla których wcześniej było brak wyników:**
Monter sieci i instalacji sanitarnych (712618) od 1 września 2017r.
 - (0.90) *Monter sieci, instalacji i urządzeń sanitarnych (712616)*
 - (0.78) *Technik inżynierii sanitarnej (311218) od 1 września 2017r.*
 - (0.71) *Technik urządzeń sanitarnych (311209)*
- **lepsze propozycje**, np. dla *Technik włókiennik (311932)* wcześniej było (0.77) *Piekarz (751204)*, a obecnie:
 - (0.88) *Technik włókiennik (311932) od 1 września 2017r.*
 - (0.61) *Operator maszyn w przemyśle włókienniczym (815204) od 1 września 2017r.*
 - (0.60) *Operator maszyn w przemyśle włókienniczym (815204)*
- **brak wyników** dla tekstów, dla których wcześniejsze wyniki były niepoprawne, np. dla *Kominiarz (713303)* było *Technik urządzeń dźwigowych (311940)*.

- zauważalny wpływ przymiotników oraz fraz wielowyrazowych, np. dla *Operator maszyn i urządzeń hutniczych (812121) od 1 września 2017r.* najbardziej istotnymi cechami były frazy:
 - 0.306 urządzenie hutniczy
 - 0.306 maszyna urządzenie hutniczy
 - 0.260 hutniczy
 - 0.181 urządzenie proces technologiczny
 - 0.181 urządzenie hutniczy prowadzenie
- dalsze pomysły:
 - lepsze frazowanie, np. z użyciem gramatyki NKJP i narzędzia Spejd,
 - uzupełnienie stoplisty o pełne frazy, a nie tylko pojedyncze słowa.

II. Podobieństwo z użyciem modeli językowych

- **przeznaczone dla tekstów z rozłącznymi zbiorami terminów istotnych** — głównie krótkie opisy, np. *Technik budowy jednostek pływających* i *Technik budownictwa okrętowego*
- **transformacja dokumentu do postaci wektora liczb rzeczywistych** z wykorzystaniem modeli językowych,
- **podobieństwo jako wartość cosinus kąta między wektorami.**

Dostępne modele dla j. polskiego

- **word2vec**
 - dsmodels.nlp.ipipan.waw.pl — kolekcja ponad 100 modeli word2vec wytrenowana na korpusie NKJP i/lub polskiej Wikipedii,
 - clarin-pl.eu/dspace/handle/11321/327 — model wyuczony na korpusie KGR10
- **fastText**
 - fasttext.cc/docs/en/crawl-vectors.html — modele dla 157 języków uczonych na Common Crawl,
 - fasttext.cc/docs/en/pretrained-vectors.html — modele dla 294 języków uczone na wersjach językowych Wikipedii
 - clarin-pl.eu/dspace/handle/11321/600 — model uczony na KGR10.
- **ELMo**
 - clarin-pl.eu/dspace/handle/11321/690 — model uczony na KGR10; **dostępny od lipca 2019**,
- **BERT**
 - BERT-Base, Multilingual Cased — wielojęzyczny modeli obsługujący 104 języki, w tym j. polski

- **rozmiar korpusów**

Common Crawl (j. polski)	–	21,0 mld.
KGR10	–	4,0 mld.
NKJP	–	1,8 mld.
Wikipedia (j. polski)	–	0,4 mld.

- **charakterystyka:**

- **Common Crawl** — teksty ze stron internetowych, automatycznie czyszczone, dużo zanieczyszczeń i brak właściwej tokenizacji,
- **KGR10** — książki, artykuły prasowe, artykuły i posty z blogów z internetu,
- **NKJP** — książki, artykuły prasowe.

Zbiory terminów:

1. kosmetyczny masaż medyczny profilaktyczny sportowy
2. blacharstwo kontrola naprawa samochód serwis
3. ciało pielęgnacja spa zabieg

Porównanie:

Para	word2vec <i>odległość</i>	fastText <i>podobieństwo</i>
1 i 2	39,46	0,43
1 i 3	37,07	0,59
2 i 3	37,33	0,46

Tytuł zawodowy/naukowy: Doktor nauk społecznych w zakresie nauki o bezpieczeństwie

- tf-idf
 - 0.56 *Tytuł zawodowy/naukowy: Doktor nauk społecznych w zakresie nauki o polityce*
 - 0.54 *Tytuł zawodowy/naukowy: Doktor nauk społecznych w zakresie nauki o obronności*
 - 0.54 *Tytuł zawodowy/naukowy: Doktor nauk społecznych w zakresie nauki o mediach*
- fastText
 - 0.53 *Tytuł zawodowy/naukowy: Doktor nauk społecznych w zakresie nauki o obronności*
 - 0.45 *Tytuł zawodowy/naukowy: Doktor nauk o zdrowiu w zakresie nauk o zdroviu*
 - 0.45 *Tytuł zawodowy/naukowy: Doktor nauk ekonomicznych w zakresie finanse*

- **elementy strukturalne zawyżają podobieństwo opisów** — kluczowe jest odfiltrowanie terminów strukturalnych przed porównaniem,
- **podobieństwo między terminami z dziedziny jest niższe niż między terminami pokrewnymi**, np.:

	1	2	3	4	5	6
1. dekarz	—	0.72	0.69	0.42	0.41	0.47
2. cieśla	—	—	0.60	0.32	0.28	0.34
3. mechanik	—	—	—	0.25	0.34	0.24
4. dach	—	—	—	—	0.49	0.53
5. montaż rynien	—	—	—	—	—	0.62
6. kładzenie dachówki	—	—	—	—	—	—

- **brak wektorów dla fraz wielowyrazowych** — średnia wektorów dla poszczególnych słów.

III. Pokrywanie się opisów

- sekwencja lematów,
- wykorzystanie pakietu textdiff do identyfikacji części wspólnych,
- miarą pokrycia była liczba tokenów wspólnych w stosunku do liczby wszystkich tokenów w dokumencie,
- miara niesymetryczna.

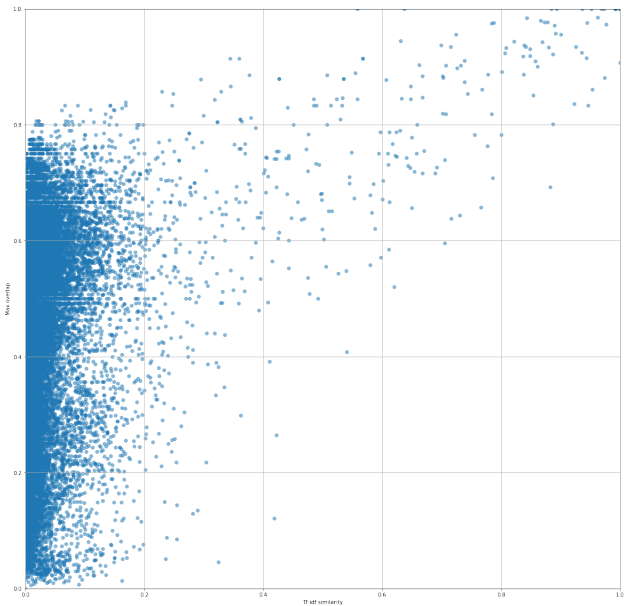
[12543]: Operator maszyn i urządzeń hutniczych (812121) od 1 września 2017r.

[1]Absolwent szkoły kształcącej w zawodzie [2]Operator [3]maszyn [4]i urządzeń hutniczych [5]powinien być przygotowany do wykonywania następujących zadań zawodowych:
1) [7]obsługiwanie maszyn i urządzeń [9]w [10]procesach technologicznych[11];
[12]2[13]) [15]konserwacji [16]i przeglądów bieżących [17]maszyn i urządzeń [18]hutniczych;
3) prowadzenia kontroli wyrobu[19].

[12368] Mechanik – operator maszyn do produkcji drzewnej (817212)

[1]Absolwent szkoły kształcącej w zawodzie [2]mechanik-operator [3]maszyn [4]do produkcji drzewnej [5]powinien być przygotowany do wykonywania następujących zadań zawodowych:
1) [6]montowania maszyn i urządzeń do produkcji drzewnej;
2) kontrolowania i nadzorowania pracy maszyn i urządzeń do produkcji drzewnej;
3) [7]obsługiwanie maszyn i urządzeń[8], stosowanych [9]w [10]procesie produkcji materiałów drzewnych[11];
[12]4[13]) [14]prowadzenia bieżącej kontroli jakości surowców i produktów;
5) wykonywania przeglądów i [15]konserwacji [17]maszyn i urządzeń[19].

Wizualizacja podobieństwa *tf-idf* vs *pokrywanie*



- tf-idf: 0,14
- A w B: 0,63
- B w A: 0,83

item_12356

```
[1]Absolwent szkoły kształcącej _ zawodzie [2]stolarz [3]powinien być przygotowany _ wykonywania następujących z  
adań zawodowych_  
— wykonywania wyrobów _ [4]drewna _ tworzyw drzewnych[5]  
— wykonywania [6]prac związanych [7] _ [8]obsługą _ konserwacją maszyn _ urządzeń stosowanych _ stolarstwie[9]  
— wykonywania napraw _ [10]renowacji _ [11]konserwacji wyrobów [12]stołarskich[13]_[14]
```

item_11971

```
[1]Absolwent szkoły kształcącej _ zawodzie [2]blacharz [3]powinien być przygotowany _ wykonywania następujących  
zadań zawodowych_  
— wykonywania wyrobów _ [4]blachy[5]  
— wykonywania [6]pokryć [7] _ [8]blachy[9]  
— wykonywania naprawy _ [10][11]konserwacji wyrobów [12][13] _ [14]pokryć _ blachy
```

Odnoszenie ZRK do baz zewnętrznych

Baza ESCO (European Skills, Competences, Qualifications and Occupations) dzieli się na 3 części:

- 942 zawodów
- 13485 umiejętności
- 8155 kwalifikacji wymaganych do tych zawodów

Q Search

OCCUPATIONS

- 0 Siły zbrojne
- 1 Kierownicy
- 2 Specjaliści
- 3 Technicy i inny średni personel
- 4 Pracownicy biurowi
- 5 Pracownicy usług i sprzedawcy
- 6 Wykwalifikowani rolnicy, robotnicy leś...
- 7 Robotnicy przemysłowi i rzemieślnicy
- 8 Operatorzy i monterzy maszyn i urz...
- 9 Pracownicy przy pracach prostych

Polish (pl)

inżynier języka naturalnego

Discuss this topic in the Online Forum

Description

Language engineers work within the field of computing science, and more specifically in the field of natural language processing. They aim to close the gap in translation between accurate human translations to machine-operated translators. They parse texts, compare and map translations, and improve the linguistics of translations through programming and code.

Regulatory aspect

W bazie zawodów regulowanych prowadzonej przez Komisję można sprawdzić, czy i w jaki sposób dany zawód jest regulowany w państwach członkowskich UE, państwach EOG lub Szwajcarii. Baza zawodów regulowanych: http://ec.europa.eu/growth/single-market/services/free-movement-professionals/qualifications-recognition_pl

Hierarchy

- 2 Specjaliści
 - 21 Specjaliści nauk fizycznych, matematycznych i technicznych
 - 215 Inżynierowie elektrotechnologii
 - 2152 Inżynierowie elektroniki
 - inżynier elektronik
 - inżynier języka naturalnego

Essential skills and competences

[interpretować wymogi techniczne](#)
[korzystać z oprogramowania do rysunków technicznych](#)
[oceniać technologie tłumaczeniowe](#)
[określać wymagania techniczne](#)
[prowadzić badania naukowe](#)

Figure 6: Przykładowe kwalifikacje z katalogu IBE

Baza w języku angielskim, zawierająca ok. 1100 zawodów. Przy testach były pod uwagę jedynie dane z bazy ESCO, jednak po przetłumaczeniu, można stosować wszystkie wypróbowane przez nas metody również na bazie O*NET.

W naszych testach sprawdzaliśmy czy można przy pomocy najbardziej popularnych metryk wyznaczyć analogiczne kwalifikacje w innej bazie danych (ESCO).

Przy porównaniu braliśmy pod uwagę nazwy kwalifikacji oraz ich charakterystyki.

Przy testach chcieliśmy porównać działanie 4 metryk:

- Odległość Jaccarda
- TF-IDF
- *Embeddings* cosine distance
- Word mover's distance

Word Mover's distance



Figure 7: Przykład WMD

Testy na bazie ESCO

ZRK	ESCO
Opiekun osoby starszej (341202)	opiekun osób starszych w placówce całodobowej
Technik inżynierii środowiska i melioracji (311208)	inżynier inżynierii środowiska
Technik mechanik (311504)	mechanik pojazdów samochodowych
Operator obrabiarek skrawających (722307)	operator maszyn do obróbki skrawaniem
Monter budownictwa wodnego (711701)	inżynier budownictwa wodnego
Murarz-tylnik (711204)	murarz
Stolarz meblowy - dyplom mistrzowski Stolarz meblowy - świadectwo czeladnicze	cieśla

Dane wymagały preprocessingu, który polegał na:

- zamianie wszystkich dużych liter na małe
- usunięciu numerów i znaków interpunkcyjnych, a także wszystkich wyrazów, które nie są rzeczownikami, przymiotnikami oraz czasownikami
- zamianie wszystkich pozostałych wyrazów na ich formę bazową.

Odległość Jaccarda - nazwy kwalifikacji

ESCO/ZRK	opiekun osoba starsza	technik mechanik	murarz	monter budownictwo wodny	operator obrabiarka	stolarz meblowy
opiekun osoba starsza placówka całodobowa	0.6	0	0	0	0	0
mechanik pojazd samochodowy	0	0.25	0	0	0	0
murarz	0	0	1	0	0	0
inżynier wodny budownictwo	0	0	0	0.5	0	0
operator maszyna obróbka skrawanie	0	0	0	0	0.2	0
cieśla	0	0	0	0	0	0

Odległość Jaccarda - charakterystyki kwalifikacji

ESCO/ZRK	opiekun osoba starsza	technik mechanik	murarz	monter budownictwo wodny	operator obrabiarka	stolarz meblowy
opiekun osoba starsza placówka całodobowa	0.14	0.05	0.01	0.01	0.01	0.06
mechanik pojazd samochodowy	0.05	0.11	0.08	0.04	0.11	0.07
murarz	0.02	0.05	0.1	0.02	0.06	0.08
inżynier wodny budownictwo	0.05	0.07	0.03	0.07	0.06	0.06
operator maszyna obróbka skrawanie	0.04	0.07	0.01	0.01	0.08	0.02
cieśla	0.01	0.06	0.12	0.05	0.07	0.1

Embeddings cosine similarity - nazwy kwalifikacji

ESCO/ZRK	opiekun osoba starsza	technik mechanik	murarz	monter budownictwo wodny	operator obrabiarka	stolarz meblowy
opiekun osoba starsza placówka całodobowa	0.83	0.22	0.11	0.06	0.17	0.15
mechanik pojazd samochodowy	0.11	0.65	0.3	0.42	0.49	0.35
murarz	0.14	0.52	1	0.41	0.29	0.53
inżynier wodny budownictwo	-0.02	0.46	0.31	0.87	0.33	0.24
operator maszyna obróbka skrawanie	-0.02	0.5	0.3	0.39	0.82	0.37
cieśla	0.15	0.42	0.61	0.35	0.2	0.46

Embeddings cosine similarity - charakterystyki kwalifikacji

ESCO/ZRK	opiekun osoba starsza	technik mechanik	murarz	monter budownictwo wodny	operator obrabiarka	stolarz meblowy
opiekun osoba starsza placówka całodobowa	0.7	0,27	0.19	0.32	0.29	0.43
mechanik pojazd samochodowy	0.27	0.71	0.49	0.52	0.52	0.56
murarz	0.31	0.51	0.7	0.53	0.52	0.57
inżynier wodny budownictwo	0.45	0.64	0.56	0.75	0.61	0.52
operator maszyna obróbka skrawanie	0.35	0.73	0.52	0.54	0.66	0.61
cieśla	0.28	0.58	0.72	0.56	0.51	0.65

Word mover's distance - nazwy kwalifikacji

Im mniejsza miara tym większe podobieństwo.

ESCO/ZRK	opiekun osoba starsza	technik mechanik	murarz	monter budownictwo wodny	operator obrabiarka	stolarz meblowy
opiekun osoba starsza placówka całodobowa	0.52	1.31	1.37	1.35	1.33	1.32
mechanik pojazd samochodowy	1.3	0.83	1.23	1.13	1.16	1.17
murarz	1.35	1.03	0	1.19	1.23	1.14
inżynier wodny budownictwo	1.37	1.15	1.25	0.34	1.26	1.28
operator maszyna obróbka skrawanie	1.4	1.14	1.26	1.25	0.76	1.26
cieśla	1.34	1.11	0.89	1.23	1.29	1.17

Word mover's distance - charakterystyki kwalifikacji

Im mniejsza miara tym większe podobieństwo.

ESCO/ZRK	opiekun osoba starsza	technik mechanik	murarz	monter budownictwo wodny	operator obrabiaarka	stolarz meblowy
opiekun osoba starsza placówka całodobowa	0.94	1.12	1.2	1.2	1.15	1.13
mechanik pojazd samochodowy	1.08	1	1.04	1.13	1	1.03
murarz	1.15	1.09	1.02	1.17	1.08	1.07
inżynier wodny budownictwo	1.06	1.03	1.1	1.01	1.04	1.09
operator maszyna obróbka skrawanie	1.16	1	1.13	1.15	0.99	1.14
cieśla	1.21	1.09	1	1.12	1.07	1

Testy pokazały, że przy porównaniach dłuższych tekstów (charakterystyk czy zestawów uczenia się) metody te potrafią wskazywać na błędną kwalifikację, jednak zawsze docelowa kwalifikacja jest punktowana dość wysoko.

Docelowo najlepiej, gdyby program proponował kilka najbliższych wyników, z których użytkownik mógłby wybrać jego zdaniem najwłaściwszą.

Wnioski

- Czynnikiem ułatwiającym porównywanie kwalifikacji byłoby rozbitcie bazy ZRK na kilka, nawzajem się do siebie odwołujących wzorem ESCO.
- Narzędzia i zasoby CLARIN-PL, które miały zastosowanie: WebSty, NlpRest2, WCRFT2, TermoPL, modele językowe (KGR10, NKJP).
- Dużym wyzwaniem jest rozdzielenie terminów strukturalnych od dziedzinowych, tj. charakterystycznych dla poszczególnych zawodów.