Druga wersja klasyfikatora tematycznego tekstów WiKNN
Zadanie A23
Punkt kontrolny M16

Piotr Pęzik, Maciej Buczek

17 kwietnia 2015, 13:40

Spis treści

1 Zawartość

Niniejszy dokument opisuje wstępną implementację klasyfikatora tematycznego WiKNN.

2 Opis i zastosowania klasyfikatora

Zarówno dla języka polskiego jak również angielskiego istnieją duże, referencyjne korpusy, takie jak Narodowy Korpus Języka Polskiego (NKJP) [Przepiórkowski et al., 2012], czy też Brytyjski Korpus Narodowy [BNC, 2001] oraz narzędzia do ich przeszukiwania i analizy językoznawczej (np. nkjp.uni.lodz.pl). Chociaż korpusy te posiadają wiele poziomów anotacji bibliograficznej i lingwistycznej, to brakuje w nich ogólnej anotacji tematycznej. Wraz z pojawieniem się dużych, otwartych cyfrowych encyklopedii, a w szczególności Wikipedii oraz jej pochodnych, takich jak DBpedia możliwe stało się opracowanie narzędzi do automatycznej kategoryzacji tematycznej tekstów w oparciu o wspólną wielojęzyczną taksonomię. Mimo swej nieformalnej struktury i społecznościowej natury, Wikipedia stanowi zdecydowanie największy i najszybciej aktualizowany, wielojęzyczny zasób encyklopedyczny, który można wykorzystać do klasyfikacji tematycznej tekstów. Zarówno dla języka angielskiego (np. [Milne and Witten, 2008]), jak też polskiego, (np. [Ciesielski et al., 2012]) opracowano metody kategoryzacji tekstów na podstawie Wikipedii, jednak nadal brakuje, zwłaszcza dla języka polskiego, wydajnego systemu indeksowania tekstów wszystkimi kategoriami tematycznymi Wikipedii. System taki umożliwiłby analizę i przeszukiwanie korpusów referencyjnych z wykorzystaniem anotacji metajęzykowych kategorii encyklopedycznych.

W ramach zadania A23 projektu CLARIN-PL opracowywany jest klasyfikator tematyczny **WiKNN** (Wikipedia K-Nearest Neighbors), który ma spełniać wymagania automatycznego przypisywania deskryptorów tematycznych do tekstów polskich i angielskich. Należy przy tym zaznaczyć, że zgodnie z założeniami klasyfikator ma działać z dużą wydajnością dla bardzo dużych taksonomii tj. o wielkości rzędu kilkudziesięciu tysięcy taksonów (np. kategorie polskiej i angielskiej Wikipedii) i dla dużych kolekcji tekstów, takich jak korpusy referencyjne. Klasyfikator wykorzystuje mechanizmy wyszukiwania pełnotekstowego udostępniane przez platormę Solr, bazującą na bibliotece Lucene. Mechanizm jego działania oparty jest na indeksie artykułów Wikipedii i występujących w niej kategoriach. Możliwe jest również wykorzystanie innych źródeł danych, pod warunkiem, że istnieje dla nich odpowiednia ilość opisanych kategoriami lub innymi deskryptorami tematycznymi tekstów (danych uczących).

3 Mechanizm działania

3.1 Budowa indeksu Wikipedii

Do budowy indeksu wykorzystane zostały wbudowane w platformę Solr wysokowydajne narzędzia do importu danych w formacie XML (DataImportHandler) oraz zrzuty całości danych Wikipedii w tym formacie, dostępne pod adresami, odpowiednio dla wersji angielskiej i polskiej, http://dumps.wikimedia.org/enwiki/ lub http://dumps.wikimedia.org/enwiki/. Oprócz narzędzi dostępnych w podstawowej wersji platformy Solr, wykorzystane zostały także stworzone przez zespół PELCRA rozszerzenia funkcji tej platformy oraz biblioteka Java Wikipedia Library (http://www.ukp.tu-darmstadt.de/software/jwpl/). Pozwoliło to na zindeksowanie treści Wikipedii bez zbędnych znaczników formatowania tekstu, znajdujących się w udostępnionych przez fundację Wikimedia danych, a także pominięcie stron z metadanymi encyklopedii.
Poniżej przedstawiono przykładowe pole indeksu, odpowiadające jednemu artykułowi Wikipedii:

```
1
2  {
3         "content": "In machine learning, instance-based
                learning or memory-based learning is a family of
                learning algorithms that, instead of performing
                explicit generalization, compares new problem
                instances with instances seen in training, which
                have been stored in memory. Instance-based learning
                is a kind of lazy learning. It is called instance-
                based because it constructs hypotheses directly from
                 the training instances themselves. Stuart Russell
                and Peter Norvig (2003). Artificial Intelligence: A
                Modern Approach, second edition, p. 733. Prentice
                Hall. ISBN 0-13-080302-2 \nThis means that the
                hypothesis complexity can grow with the data:  in
                the worst case, a hypothesis is a list of n training
                 items and the computational complexity of
                classification a single new instance is O(n). One
                advantage that instance-based learning has over
                other methods of machine learning is its ability to
                adapt its model to previously unseen data...",
4         "id": "22589574",
5         "timestamp": "2013-12-04T14:44:36Z",
6         "revision": 584534802,
7         "category": [
8           "Machine_learning"
9         ],
```

```
10          "title": "Instance-based learning",
11          "_version_": 1458226160222601200
12      }
```

3.2 Metody klasyfikacji

Na bieżącym etapie prac dostępne są wstępne implementacje dwóch podstawowych metod klasyfikacji:

- podstawowa, oparta o algorytm k-najbliższych sąsiadów, oraz
- dodatkowa, oparta o profile tematyczne kategorii.

3.2.1 Metoda podstawowa - K-najbliższych sąsiadów

Po wysłaniu zapytania REST zawierającego klasyfikowaną treść oraz wymagane parametry do instancji platformy Solr z istniejącym indeksem Wikipedii, pierwszym krokiem wykonywanym w procesie klasyfikacji jest wykrycie słów kluczowych. Odbywa się to z wykorzystanie algorytmu wskazanego w parametrze kwMethod (domyślnie g2). Wykryte słowa kluczowe są następnie wykorzystywane w zapytaniu (BooleanQuery) do indeksu Solr. Jest ono budowane w oparciu o zbiór wszystkich n-elementowych (parametr grouping) kombinacji słów. Domyślnie, każde słowo musi wystąpić przynajmniej z jednym z pozostałych. Zapytanie zwraca k najtrafniejszych wyników. W następnej kolejności zliczana jest liczba wystąpień poszczególnych kategorii oraz suma wartości score dla artykułów, w których dana kategoria występuje. Z powstałej listy usuwane są kategorie. które nie wystąpiły przynajmniej tyle razy, ile określono to w parametrze minOcc (jego domyślna wartość wynosi 2). Powstała lista wykrytych kategorii jest następnie zwracana w odpowiedzi.

3.2.2 Metoda dodatkowa - profile tematyczne kategorii

W przypadku tej metody, wykonywane są zapytania REST. Pierwsze podzapytanie pozwala wykryć słowa kluczowe i wygląda podobnie, jak w poprzednim przypadku. Drugie z nich wykorzystuje wykryte słowa, porównując je z profilami kategorii utworzonymi przez obliczenie wspólnych słów kluczowych dla 100 najdłuższych tekstów dla danej kategorii. Ze względu na czasochłonność tego procesu, są one przechowywane w osobnym indeksie Solr. Drugie podzapytanie kierowane jest do tego właśnie indeksu. W tym przypadku wagi przydzielane przez platformę Solr nie mają tak dużego znaczenia, jak w przypadku poprzedniej metody - stopień dopasowania danego tekstu do danej kategorii określa liczbę słów kluczowych występujących jednocześnie w danym tekście i w profilu tej kategorii.

4 Interfejs programistyczny

4.1 Metoda K-najbliższych sąsiadów

Zapytanie REST jest wysyłane do instancji platformy Solr z następującymi parametrami:

Nazwa	Funkcja	Domyślnie
text	tekst do sklasyfikowania	-
kwMethod	metoda wybierania słów kluczowych	g2
k	maksymalna liczba kategorii opisujących dokument	100
kwDiv	dzielnik pozwalający ograniczyć liczbę wykorzystanych słów kluczowych	2
maxKws	maksymalna liczba słów kluczowych	20
lang	język tekstu wykorzystanego w zapytaniu. Planowane jest użycie automatycznego detektora - możliwość automatycznego wykrywania języka	pl
scale	zmienna określająca, czy przeprowadzone zostanie skalowanie wag wyników	false
grouping	liczba określająca liczbę słów kluczowych, które zostaną zgrupowane w pojedynczym podzapytaniu pełnotekstowym. W zapytaniu wykorzystywana są wszystkie n-elementowe kombinacje słów kluczowych	2
minOcc	metoda wybierania słów kluczowych	g2
fl	pola indeksu, które wykorzystywane są w zapytaniu	title, content
scaleKw	parametr określający, czy skalować słowa kluczowe	false
kwCutoff	wartość, poniżej której skalowane słowa kluczowe są usuwane	0
useKwCooc	parametr określający, czy używać wartości współwystępowania słów kluczowych do odsiewania zbyt często występujących słów	false
kwCutoff	wartość, poniżej której skalowane słowa kluczowe są usuwane	10

Tablica 1: Parametry zapytania REST - metoda podstawowa

4.2 Metoda profili tematycznych

Metoda wykorzystuje dwa zapytania REST ze względu na fakt, że dane w nim używane przechowywane są w dwóch osobnych indeksach . Pierwsze z nich służy do uzyskania listy słów kluczowych i jest wysyłane do indeksu zawierającego artykuły Wikipedii, używanego też w w/w metodzie. Wykorzystuje ono następujące parametry:

Nazwa	Funkcja	Domyślnie
text	tekst, dla którego szukane będą słowa kluczowe;	-
maxKws	maksymalna liczba słów kluczowych	20
kwMethod	metoda wybierania słów kluczowych	g2
minOcc	minimalna liczba wystąpień słowa w tekście, przy której zostanie obliczona jego kluczowość	2
fl	pola indeksu, które wykorzystywane są w zapytaniu	title, content
scaleKw	parametr określający, czy skalować słowa kluczowe	false
kwCutoff	wartość, poniżej której skalowane słowa kluczowe są usuwane	0
useKwCooc	parametr określający, czy używać wartości współwystępowania słów kluczowych do odsiewania zbyt często występujących słów	false
kwCutoff	wartość, poniżej której skalowane słowa kluczowe są usuwane	10

Tablica 2: Parametry zapytania REST - wykrywanie słów kluczowych

Kolejne zapytanie kierowane jest do indeksu zawierającego profile kategorii. Przyjmuje ono następujące parametry:

Nazwa	Funkcja	Domyślnie
k	maksymalna liczba kategorii do wykorzystania	50
keywords	lista słów kluczowych uzyskanych w poprzednim kroku	-
grouping	liczba określająca liczbę słów kluczowych, które zostaną zgrupowane w pojedynczym podzapytaniu pełnotekstowym. W zapytaniu wykorzystywana są wszystkie n-elementowe kombinacje słów kluczowych	1

Tablica 3: Parametry zapytania REST - metoda profili kategorii

keywords - lista słów kluczowych uzyskanych w poprzednim kroku;
grouping - liczba określająca liczbę słów kluczowych, które zostaną zgrupowane w pojedynczym podzapytaniu
pełnotekstowym. W zapytaniu wykorzystywana są wszystkie n-elementowe kombinacje słów kluczowych;

4.3 Format wyników

Niniejsza sekcja zawiera przykładowe wyniki pracy klasyfikatora w postaci otrzymywanej przez bezpośrednie zapytanie do serwera Solr oraz opis

zawartych w nim danych. Należy zauważyć, że rzeczywiste wyniki zawierają więcej informacji - tu przedstawione zostały wyłącznie przykłady. Do przykładowej klasyfikacji wykorzystano następujący tekst:

> “Last year, a few of Tesla’s Model S vehicles caught on fire following accidents – a situation that caused a bit of embarrassing public backlash for Elon Musk and his company. Despite the fact that Musk has steadfastly defended the safety of his vehicle, the company has announced a fix for the cause of the reported fires. In a post on Medium , Musk details a new titanium shield and aluminum deflector plates that protect the underbody of the vehicle – all cars produced after March 6th will have this new safety system in place, and existing vehicles can have it added free of charge.” (The Verge)

4.3.1 Metoda podstawowa


```
{

   "keywords":[
      "kw",
      {
         "term":"musk",
         "localFreq":3,
         "refFreq":2773,
         "refCollectionSize":1572686017,
         "localDocSize":53,
         "keynessValue":56.430152893066406,
         "scaledKeynessValue":0.0
      },
      "kw",
      {
         "term":"vehicle",
         "localFreq":4,
         "refFreq":238272,
         "refCollectionSize":1572686017,
         "localDocSize":53,
         "keynessValue":42.01165771484375,
         "scaledKeynessValue":0.0
      }
   ],
   "neighbours":[
      {
         "id":1048980,
         "score":344.3584,
         "title":"Tesla Motors"
      },
      {
         "id":2366255,
         "score":255.8872,
         "title":"Tesla Model S"
      },
      {
         "id":4201254,
         "score":245.01503,
         "title":"Plug-in electric vehicle fire incidents"
      },


   ],
   "categories":{
```

```
      "2010s_automobiles":11,
      "Rear-wheel-drive_vehicles":9,
      "2000s_automobiles":8,
      "Nikola_Tesla":6

   },
   "weighedCategories":{
      "Rear-wheel-drive_vehicles":844.5729331970215,
      "2010s_automobiles":819.0741119384766,
      "Tesla_Motors_vehicles":575.3129577636719,
      "2000s_automobiles":500.64119720458984,


   },
   "scaledCategories":{
      "Rear-wheel-drive_vehicles":3.549799534568481,
      "2010s_automobiles":3.3941638745255513,
      "Tesla_Motors_vehicles":1.9063332121529248,
      "2000s_automobiles":1.4505635660448393,

   },
   "titles":{
      "Tesla Motors":344.3583984375,
      "Tesla Model S":255.88720703125,
      "Plug-in electric vehicle fire incidents":245.01502990722
         656,

   }
}
```


Przedstawiony dokument składa się z następujących sekcji: keywords - lista słów kluczowych wykrytych w tekście wraz z ich parametrami; neighbours - lista sąsiadów wykorzystanych w algorytmie K-najbliższych sąsiadów wraz z sumą trafności (score) wyników wyszukiwania wg platformy Solr; categories - najczęściej występujące kategorie; weighedCategories- najbardziej pasujące do zapytania kategorie wg. kryterium trafności wyszukiwania SOLR. Jest to najważniejszy element dokumentu, zawierający właściwe wyniki klasyfikacji; scaledCategories - te same kategorie o wynikach przeskalowanych w celu ich normalizacji; titles - pełna lista tytułów artykułów pasujących do wykrytych słów kluczowych.

4.4 Ogólnodostępna usługa REST

Demonstracyjna wersja klasyfikatora jest dostępna jako usługa REST pod adresem:

http://pelcra.clarin-pl.eu/tools/classifier/

Usługa ta przyjmuje następujące parametry:

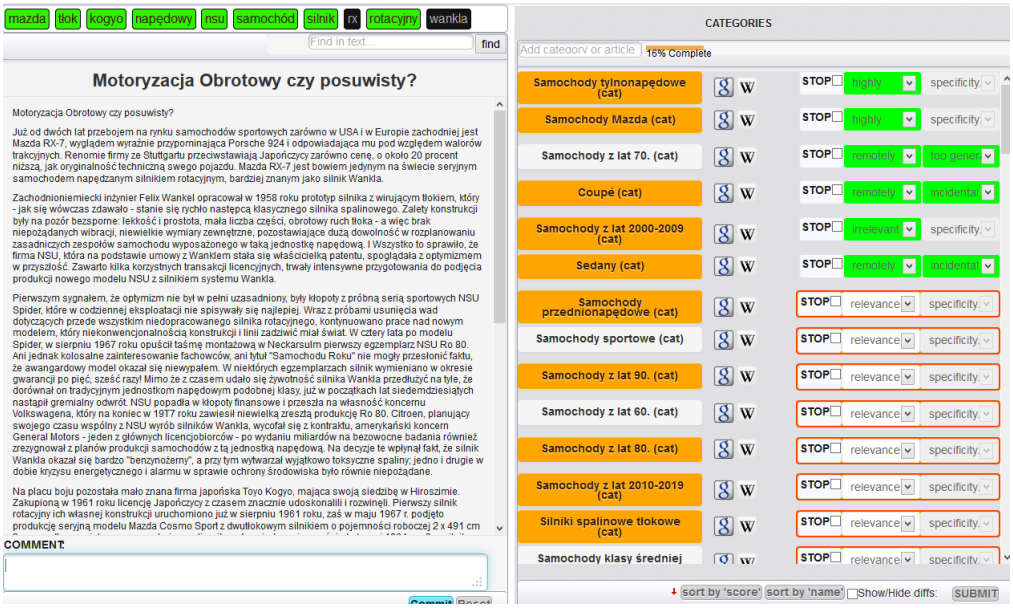
Nazwa	Funkcja	Domyślnie
text	tekst do sklasyfikowania	-
k	maksymalna liczba kategorii opisujących dokument	100
kwDiv	dzielnik pozwalający ograniczyć liczbę wykorzystanych słów kluczowych	2
maxKws	maksymalna liczba słów kluczowych	20
lang	język tekstu wykorzystanego w zapytaniu. Planowane jest użycie automatycznego detektora - możliwość automatycznego wykrywania języka	pl
scale	zmienna określająca, czy przeprowadzone zostanie skalowanie wag wyników	false
grouping	liczba słów kluczowych, które zostaną zgrupowane w pojedynczym podzapytaniu pełnotekstowym. W zapytaniu wykorzystywana są wszystkie n-elementowe kombinacje słów kluczowych	2
minOcc	metoda wybierania słów kluczowych	g2
fl	pola indeksu, które wykorzystywane są w zapytaniu	title, content
scaleKw	parametr określający, czy skalować słowa kluczowe	false
kwCutoff	wartość, poniżej której skalowane słowa kluczowe są usuwane	0
useKwCooc	parametr określający, czy używać wartości współwystępowania słów kluczowych do odsiewania zbyt często występujących słów	false
kwCutoff	wartość, poniżej której skalowane słowa kluczowe są usuwane	10

Tablica 4: Interfejs REST

5 Ewaluacja aktualnej wersji

5.1 WiKNNEval

W celu ewaluacji wyników pracy klasyfikatora stworzona została aplikacja WiKNNEval, stanowiąca graficzny interfejs umożliwiający ocenę kategorii za pomocą dowolnej przeglądarki internetowej. Ewaluacją zajmuje się zespół specjalnie dobranych i przeszkolonych w tym celu osób. Docelowo, planowane jest ocenienie ok. 1000 tekstów angielskojęzycznych i 1000 polskojęzycznych. Obecnie zostało ocenionych ok. 150 tekstów polskich i 70 angielskich.



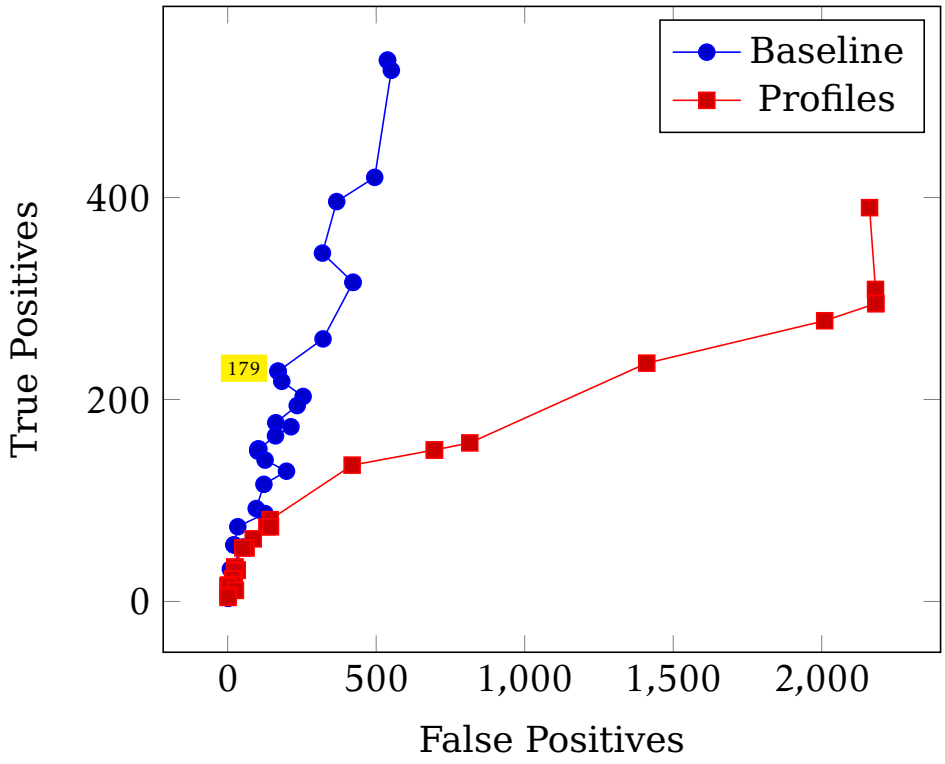
Powyższy obraz przedstawia główny ekran aplikacji WiKNNEval, służącej do specyfikacji i oceny funkcjonalności klasyfikatora sematycznego. Z lewej strony widoczny jest sklasyfikowany tekst wraz ze swoimi słowami kluczowymi. Interfejs pozwala oceniającemu na oznaczenie słów kluczowych jako nieprzydatnych, dzięki czemu możliwe będzie ich wykluczenie z kolejnych prób klasyfikacji. Oprócz tego dostępne jest okno wyszukiwania, pozwalające na przeszukiwanie wyłącznie ocenianego tekstu. Pod tekstem znajduje się pole komentarza, pozwalające oceniającemu przekazać opinie na temat wyników nie objętą przez istniejące kategorie ocen. Z prawej strony znajduje się lista kategorii wykrytych przez klasyfikator. Dane z obu metod klasyfikatora przedstawione są wspólnie, ale w przypadku , gdy zostały wykryte przez metodę profili tematycznych (wyróżnione pomarańczowym kolorem), można zobaczyć słowa kluczowe z profilu danej kategorii, zarówno te pokrywające się ze słowami kluczowymi dla tekstu, jak i te, które w nim nie występują. Istnieje również możliwość dodania kategorii przez użytkownika. W tym celu udostępniono listę podpowiedzi zawierającą istniejące kategorie. Co więcej, podpowiedzi obejmują także listę artykułów Wikipedii, na wypadek gdyby któryś z nich lepiej odzwierciedlał treść danego tekstu. Kryteria oceny odzwierciedlają problemy związane z funkcjonalnością klasyfikatora, jak i specyfiką danych Wikipedii. Pierwszy parametr do wprowadzenia to “STOP”. Zaznaczenie tego pola powoduje dodanie kategorii do listy kategorii odrzucanych, które zostaną wykluczone z finalnej wersji klasyfikatora. Ma to na celu usunięcie z wyników kategorii o niskiej zawartości informacji, jak np. popularne w Wikipedii listy dat. Właściwa ocena danej kategorii składa się z dwóch elementów. Pierwszym z nich jest trafność samej oceny. Przyjęte zostały cztery możliwe stopnie trafności - “highly relevant”, “remotely relevant”, “irrelevant” oraz “not sure”. Skala ta pozwala na precyzyjne określenie stopnia dopasowania danej kategorii do tekstu, jednocześnie zapewniając jednoznaczność, którą trudniej byłoby osiągnąć przy wykorzysta-

niu skali numerycznej. Budowa aplikacji pozwala na ocenę tego samego tekstu przez wiele osób, co pozwala na bardziej precyzyjne określenie poziomu trafności klasyfikacji, niż miałoby to miejsce w wypadku oceny tylko przez jedną osobę. Oceny danej kategorii dopełnia parametr specificity, przyjmujący trzy możliwe wartości: “OK”, “too specific” oraz “too general”. Pozwala to na określenie stopnia ziarnistości wyników. Może to być użyteczne w przypadkach, kiedy wykryte kategorie będą powiązane z tematyką klasyfikowanego tekstu silnie lub też luźno - co określa poprzedni parametr - ale jednocześnie bedą, przykładowo, dotyczyć informacji zbyt szczegółowych w stosunku do poziomu szczegółowości danego tekstu.

5.2 Aktualne wyniki ewaluacji

Test	method	kw_method	scale_kw	kw_cutoff	max_kws	grouping	minocc	k	TP	FP	Precyzja
151	baseline	g2	t	-2	10	2	2	100	536	538	0.499
147	baseline	g2	f	0	10	2	2	100	536	538	0.499
149	baseline	g2	t	-1	10	2	2	100	526	551	0.488
171	baseline	g2	t	-1	10	3	3	200	420	495	0.459
167	baseline	g2	t	-1	10	2	3	150	396	367	0.519
146	profiles	g2	t	0	50	1	2	100	390	2161	0.153
169	baseline	g2	t	-1	10	3	3	150	345	319	0.520
195	baseline	g2	t	0	10	2	2	100	316	422	0.428
145	baseline	g2	t	0	10	2	2	100	316	422	0.428
150	profiles	g2	t	-1	50	1	2	100	309	2181	0.124
152	profiles	g2	t	-2	50	1	2	100	295	2182	0.119
148	profiles	g2	f	0	50	1	2	100	295	2182	0.119
164	profiles	g2	t	-1	50	2	2	150	278	2010	0.122
173	baseline	g2	t	-1	20	4	4	200	260	321	0.448
158	profiles	g2	t	1	50	1	2	100	236	1411	0.143
179	baseline	g2	t	0	50	3	3	75	228	169	0.574
229	baseline	g2	t	0	50	3	3	75	228	170	0.573
227	baseline	g2	t	0	50	3	3	75	218	182	0.545
231	baseline	g2	f	0	50	3	3	75	203	254	0.444
245	baseline	g2	f	0	50	3	3	75	194	234	0.453
247	baseline	g2	t	0	50	3	3	75	177	162	0.522
233	baseline	g2	f	0	50	3	3	75	173	213	0.448
235	baseline	g2	t	0	50	3	3	75	164	161	0.505
172	profiles	g2	t	-1	50	3	2	200	157	815	0.162
215	baseline	g2	t	0	20	3	3	75	151	105	0.590
217	baseline	g2	t	0	20	3	3	75	151	102	0.597
170	profiles	g2	t	-1	50	3	2	150	150	696	0.177
168	profiles	g2	t	-1	50	3	2	150	150	696	0.177
219	baseline	g2	t	0	20	3	3	75	150	102	0.595
221	baseline	g2	t	0	20	3	3	75	149	102	0.594
223	baseline	g2	t	0	20	3	3	75	149	102	0.594
251	baseline	g2	t	0	50	2	3	75	140	126	0.526
174	profiles	g2	t	-2	20	3	2	200	135	419	0.244
243	baseline	g2	f	0	50	3	3	75	129	198	0.394
239	baseline	g2	t	0	50	3	3	75	116	122	0.487
225	baseline	g2	t	0	20	3	3	75	92	96	0.489
157	baseline	g2	t	1	10	2	2	100	87	125	0.410
193	baseline	g2	t	1	10	2	2	100	87	125	0.410
232	profiles	g2	f	1	50	3	2	10	81	143	0.362
202	profiles	g2	t	0	10	2	2	10	74	132	0.359
181	baseline	g2	t	0	10	3	3	75	74	34	0.685
234	profiles	g2	f	1	50	3	2	10	74	144	0.339
196	profiles	g2	t	0	10	2	2	5	62	86	0.419
191	baseline	g2	t	0	10	3	3	50	56	21	0.727
183	baseline	g2	t	0	10	3	3	50	56	21	0.727
163	baseline	g2	t	1	10	2	3	100	54	52	0.509
216	profiles	g2	t	0	50	3	2	10	53	51	0.510
194	profiles	g2	t	0	10	2	2	3	53	62	0.461
244	profiles	g2	t	0	50	3	2	5	34	24	0.586
246	profiles	g2	t	0	50	3	2	5	34	24	0.586
185	baseline	g2	t	0	10	3	3	25	32	9	0.780
186	profiles	g2	t	0	20	3	2	50	31	32	0.492
192	profiles	g2	t	0	20	3	2	50	31	32	0.492
184	profiles	g2	t	0	15	3	2	50	30	23	0.566
228	profiles	g2	t	1	50	3	2	10	22	17	0.564
240	profiles	g2	t	0	50	3	2	10	22	22	0.500
226	profiles	g2	t	1	50	3	2	10	22	17	0.564
230	profiles	g2	t	1	50	3	2	10	22	17	0.564
242	profiles	g2	t	0	50	3	2	5	17	17	0.500
182	profiles	g2	t	0	10	3	2	50	16	3	0.842
200	profiles	g2	t	0	10	3	2	10	16	3	0.842
236	profiles	g2	t	1	50	3	2	10	15	1	0.938
198	profiles	g2	t	0	10	3	2	5	15	3	0.833
252	profiles	g2	t	0	10	2	2	3	11	26	0.297
180	profiles	g2	t	0	5	3	2	50	6	2	0.750
248	profiles	g2	t	0	10	3	2	3	4	1	0.800
197	baseline	g2	t	1	10	3	2	100	3	2	0.600

Tablica 5: Wykonane testy i ich parametry



W toku dotychczasowych testów za najefektywniejsze ustawienia zostały uznane parametry podane dla testu nr 179. Zapewniają one precyzję wynoszącą 57.4%, przy jednoczesnym zapewnieniu względnie akceptowalnego poziomu pokrycia. Wykryte kategorie nie pokrywające się z występującymi w danych kontrolnych były uznawane za błędne na tym etapie ewaluacji. Kolejnym krokiem w procesie oceniania wyników będzie ocena nowo wykrytych kategorii dla już ocenionych tekstów.

6 Planowane ulepszenia

6.1 Algorytmy wykrywania słów kluczowych

W toku dotychczasowych prac nad klasyfikatorem, obok metod wykrywania słów kluczowych istniejących w podstawowej wersji klasyfikatora (g2, tf-idf) zaimplementowane zostały dwa dodatkowe algorytmy spełniające tę rolę - TextRank oraz RAKE. Utrudnieniem w przypadku pierwszego z nich jest konieczność stosowania dodatkowych narzędzi do tagowania i płytkiego parsingu tekstu, co znacznie zwiększa złożoność czasową procesu klasyfikacji. Z kolei algorytm RAKE jest skuteczny wyłącznie dla języków analitycznych. Zastosowanie go dla języka polskiego wymaga więc modyfikacji uwzględniających jego syntetyczny charakter.

6.2 Alternatywne taksonomie kategoriii

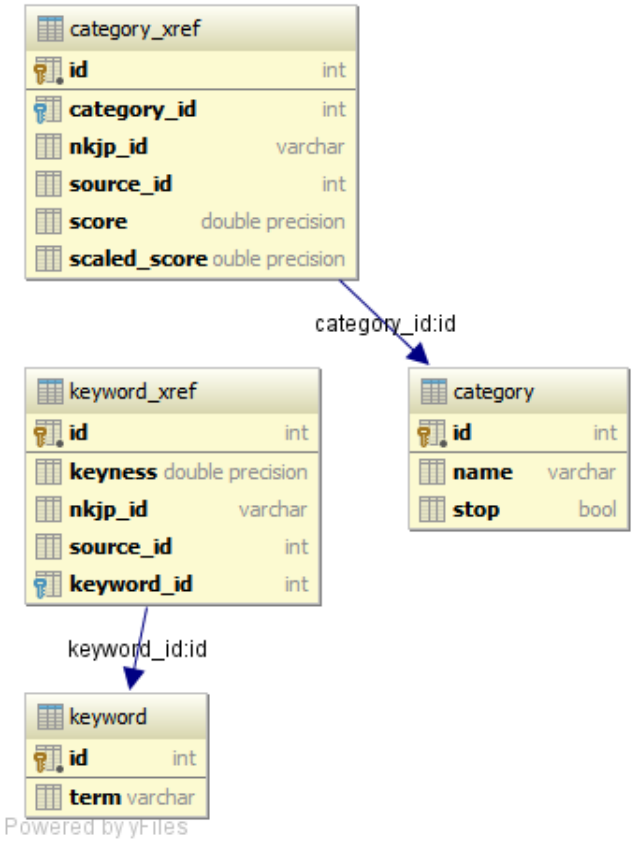
Jak wspomniano w sekcji 2, klasyfikator może wykorzystywać inne taksonomie kategorii niż pochodzące z Wikipedii. Warunkiem koniecznym do ich zastosowania jest dostępność wystarczającej ilości tekstów z przyporządkowanymi kategoriami lub słowami kluczowymi. Jednym z możliwych źródeł takich danych, opartym o model Wikipedii, jest witryna Wikinews. Próby zastosowania treści z niej pochodzących do klasyfikacji tematycznej pozwalają jednak wysunąć wniosek, że zbyt mała popularność tej witryny oraz wynikające z niej problemy, taki jak względnie niewielka ilość tekstów oraz ograniczony zakres społecznościowej kontroli nad publikowanymi artykułami ograniczają ich przydatność. Większą ilość różnorodnych, oznakowanych tematycznie tekstów pozyskano poprzez zbieranie treści z witryn prasowych oznaczonych metaznacznikiem HTML "keywords". W tym przypadku skuteczność procesu klasyfikacji jest zależna od słów kluczowych podanych w tym znaczniku. Ponieważ ich podstawową funkcją jest zapewnienie widoczności danej strony w wyszukwarkach internetowych, nie mają one charakter obiektywnych, encyklopedyczych opisów i często odzwierciedlają polityczną stronniczość danego czasopisma. Co więcej, w obu przypadkach popularność danego tematu przekłada się na częstość jego występowania w danych źródłowych w znacznie większym stopniu, niż w przypadku danych z Wikipedii. Z kolei kategorie nie będące przedmiotem zainteresowania prasy są reprezentowane tylko w niewielkim stopniu.

6.3 Drzewa decyzyjne

7 Bazy danych z wynikami

7.1 Baza nkjp_categories

Poniższy diagram przedstawia schemat bazy danych nkjp_categories, zawierającej wyniki klasyfikacji dla wybranych tekstów z Narodowego Korpusu Języka Polskiego. Tabele category i category_xref zawierają informacje o przydzielonych tekstom kategoriach. Obok samych kategorii, tabele zawierają także wagi przydzielone im podczas klasyfikacji oraz pole stop, wskazujące kategorie normalnie odrzucane w procesie klasyfikacji, ale uwzględnione w bazie w celach informacyjnych. Tabele keyword i keyword_xref zawierają słowa kluczowe przydzielone danych tekstom wraz z miarą ich kluczowości. Należy zauważyć, że korzystanie z wyników klasyfikacji zawartych w bazie wymaga dostępu do NKJP - klasyfikowane teksty nie są umieszczane w niniejszej bazie.



7.2 Baza bnc_categories

8 Bibliografia

Literatura

[BNC, 2001] BNC, C. (2001). The british national corpus, version 2 (BNC world). *Distributed by Oxford University Computing Services*.

[Ciesielski et al., 2012] Ciesielski, K., Borkowski, P., Kłopotek, M. A., Trojanowski, K., and Wysocki, K. (2012). Wikipedia-based document categorization. In *Security and Intelligent Information Systems*, pages 265–278. Springer.

[Milne and Witten, 2008] Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM.

[Przepiórkowski et al., 2012] Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B., editors (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.