# Wydobywanie słów kluczowych z niewielkich dokumentów i frag- mentów dokumentów

Klaudia Herkt, Radosław Lemiec

25 stycznia 2018

## 1 Wprowadzenie

Celem projektu było stworzenie oraz wyuczenie modelu do wydobywania słów kluczowych z niedużych dokumentów i fragmentów dokumentu dla języka polskiego. Do predykcji słów kluczowych zastosowano metodę Conditional Random Fields. Model został wyuoczony na zbiorze przygotowanym przez Grupę Technlogii Językowych G4.19 Politechniki Wrocławskiej, zawierającym 1288 dokumentów o różnej tematyce. Do lematyzacji i analizy morfo-syntaktycznej użyty został Wrocławski CRF Tagger (WCRFT). Program napisany został w języku Python, a model CRF zimprtowany z biblioteki python-crfsuite `https://python-crfsuite.readthedocs.io/en/latest/` .

## 2 Opis metody Condition Random Fields

CRF to klasa popularnych metod modelowania statystycznego powszechnie stosowana w problemach klasyfikacji, rozpoznawania wzorców czy prognozowania strukturalnego. Służą do modelowania sekwencji uwzględniając kontekst informacji. Są rodzajem dyskryminującego, nieskierowanego modelu graficznego używanego do kodowania znanych apriori relacji między obserwacjami i konstruowania ich spójnych interpretacji. W naszej pracy narzędzie to zostało użyte do tagowania słów występujących w tekstach do formatu IOB oraz IOBS.

## 3 Zbiór danych

Korpus Języka Polskiego Politechniki Wrocławskiej (KPWr, wersja 1.1) jest zbiorem składającym się z 1288 dokumentów tekstowych, należących do 12 kategorii tematycznych: blogi, dłuższe artykuły prasowe, dialog, krótsze artykuły prasowe, popularno naukowe i podręczniki, proza dawna, proza współczesna, stenogramy, techniczne, urzędowe, ustawy oraz wikipedia. W zbiorze danych zawarte są informacje o słowach kluczowych każdego z dokumentu (również takie, które nie występują w tekście). Dostępne pliki zawierają tokenizację, analizę

morfologiczną tekstu, anotację oraz lematy. Dane dostępne są publicznie pod adresem: `http://www.nlp.pwr.wroc.pl/narzedzia-i-zasoby/zasoby/kpwr` .
Model CRF uczony był na losowo wybranych 80% tekstów KPWr (1030 dokumentów), a testowany na pozstałych 20% (258 dokumentów) oraz 5 tekstów pochodących ze strony `https://seocopywriter.com.pl/tekst-zapleczowy` .

# 4 Przeprowadzone badania

## 4.1 Przyjęte założenia

Stworzony przez nas model CRF działa dla krótkich dokumentów tekstowych napisanych w języku polskim. Wynajduje tylko te słowa kluczowe, które występują w tekście.

## 4.2 Tagowanie słów

Dane wejściowe do CRF były krotkami *(słowo występujące w tekście, lista cech)*.
Słowa zostały otagowane za pomocą 2 typów notacji:

- **IOB (Inside-outside-beginning)**
  - **Beginning (B)** - słowo jest początkiem wielowyrazowego słowa kluczowego,
  - **Inside (I)** - słowo zawiera się w wielowyrazowym słowie kluczowy, ale nie jest jego początkiem albo słowo jest jednowyrazowym słowem kluczowym,
  - **Outside (O)** - słowo nie jest ani nie należy do wielowyrazowego słowa kluczowego.
- **IOBS** - format IOB wzbogacony o klasę **S**, określającą, iż dane słowo znajduje się w *stop-liście* języka polskiego (liście wyrazów popularnych dla języka, zazwyczaj nieniosących ze sobą ważnych informacji).

Lista słów kluczowych dla języka polskiego została pobrana z biblioteki stopwords, dostępnej pod adresem `https://pypi.org/project/stop-words/` .

## 4.3 Uczenie modelu

Meotda CRF była uczona na dwóch typach danych wejściowych: oryginalna odmieniona forma słów występujących w tekście oraz ich lematy. W dokumentach pobranych z KPWr występują wszystkie możliwe lematy podanych słów. Za pomocą taggera WCRFT dokonano ich ujednoznaczniania.
Do uczenia wykorzystano następujący zbiór cech słów występujących w tekście:

1. wartość tf-idf,
2. znormalizowana długość słowa,

3. indeks pierwszego wystąpienia słowa w tekście,
4. słowo poprzedzające pierwsze występnie podanego słowa,
5. słowo występujące 2 pozycje wcześniej niż podane,
6. następne słowo po pierwszym wystąpieniu danego słowa,
7. słowo występujące 2 pozycje później niż podane,
8. liczba, w jakiej słowo występuje (pojedyncza albo mnoga),
9. stopień słowa (równy, wyższy, najwyższy),
10. część zdania (rzeczownik, czasownik, przymiotnik, zaimek, przysłówek, przyimek),
11. część zdania słowa następnego,
12. część zdania słowa poprzedzającego,
13. krota *(obecne słowo, słowo następne)*,
14. krota *(poprzednie słowo, słowo następne)*.

Znormalizowana długość słowa była liczona jako stosunek długości tego słowa do maksymalnej długości słowa występującej w danym dokumencie.

# 5 Wyniki

Skuteczność działania metody była sprawdzana za pomocą 3 miar: Precission, Recall oraz F1Score. W poniższej tabeli przedstawione są średnie arytmetyczne uzykanych wyników dla 10 modeli CRF. Do każdego modelu dane uczące i testowe były wybierane w sposób losowy. Przedstawione wyniki doyczą danych pochodzących z KPWr.
Model oceniany był za pomocą 2 metod ewaluacji:

- **restrykcyjnej** - za dobrze sklasyfikowane słowo kluczowe było uznane tylko takie słowo, którego predykowany tag I albo B był zgodny z jego prawdziwą wartością,
- **łagodnej** - słowo uznane za słowo kluczowe, ale ze złym tagem, było oceniane zarówno pozytywnie jak i negatywnie.

Przebane zostały wpływy 3 zbiorów cech słów na jakość predykcji:

1. wartość tf-idf, znormalizowana długość słowa, indeks pierwszego wystąpienia słowa w tekście, słowo poprzedzające pierwsze występnie podanego słowa, następne słowo po pierwszym wystąpieniu danego słowa, , liczba,

w jakiej słowo występuje (pojedyncza albo mnoga), stopień słowa (równy, wyższy, najwyższy), część zdania (rzeczownik, czasownik, przymiotnik, zaimek, przysłówek, przyimek), słowo występujące 2 pozycje wcześniej niż podane oraz słowo występujące 2 pozycje później niż podane,

2. powyższe cechy wzbogacone o: część zdania słowa występującego 2 pozycję wcześniej oraz część zdania słowa występującego 2 pozycję później,

3. powyższe cechy wzbogacone o: pary *(obecne słowo, następne słowo)*, *(poprzednie słowo, następne słowo)*

## 5.1 Pierwszy zbiór cech

### 5.1.1 Format IOB

Przykładowa macierz pomyłek dla zbioru testowego:

| | **I** | **O** | **B** |
|---|---|---|---|
| **I** | 195 | 3868 | 24 |
| **O** | 66 | 64104 | 44 |
| **B** | 17 | 1370 | 120 |

Wiersze w macierzy pomyłek oznaczają klasy predykowane, kolumny zaś klasy rzeczwyiste. Słów kluczowych przypisanych do tekstów z definicji jest niewiele. Łatwo zauważyć, jak znaczną część danych uczących stanowią dane oznaczone tagiem "outside".

Ocena predykcji słów przedstawionych w formie występującej w tekście oraz jako lemat:

| **Forma słów** | **Sposób oceniania** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|---|
| **Lemat** | **Łagodny** | 53.27% | 11.23% | 18.60% |
| **Oryginalna** | **Łagodny** | 54.43% | 5.92% | 10.66% |
| **Lemat** | **Restrykcyjny** | 24.68% | 4.96% | 8.25% |
| **Oryginalna** | **Restrykcyjny** | 36.66% | 3.72% | 6.75% |

### 5.1.2 Format IOBS

Przykładowa macierz pomyłek dla zbioru testowego:

| | **I** | **O** | **B** | **S** |
|---|---|---|---|---|
| **I** | 196 | 3765 | 25 | 127 |
| **O** | 112 | 53985 | 50 | 2959 |
| **B** | 8 | 1353 | 129 | 16 |
| **S** | 2 | 1370 | 1 | 4136 |

Ocena predykcji słów przedstawionych w formie występującej w tekście oraz jako lemat:

| Forma słów | Sposób oceniania | Precision | Recall | F1 Score |
|---|---|---|---|---|
| **Lemat** | **Łagodny** | 51.40% | 13.03% | 20.77% |
| **Oryginalna** | **Łagodny** | 55.76% | 6.29% | 11.29% |
| **Lemat** | **Restrykcyjny** | 26.92% | 6.35% | 10.26% |
| **Oryginalna** | **Restrykcyjny** | 35.83% | 3.74% | 6.76% |

Wyniki ekstrakcji słów kluczowych na pomocniczym zbiorze testowym (dokumenty niepochodzące z KPWr):

- Dla słów w tekście wyrażowych za pomocą lematu:

| Lp. | Rzeczywsite słowa kluczowe | Predykowane słowa kluczowe |
|---|---|---|
| **1** | "szybki pożyczka gotówkowy", "chwilówka", "pożyczka", "pożyczka przez internet" | [] |
| **2** | "kurs pierwszy pomoc" | "ratownictwo" |
| **3** | "mechanik samochodowy", "warsztat samochodowy", "przegląd auto" | [] |
| **4** | "ubezpieczyć na życie" | [] |
| **5** | "ubezpieczenie zdrowotny", "ubezpieczenie prywatny" | "obowiązkowy ubezpieczenie", "urząd praca", "abonament" |

- Dla oryginalnych form słów występujących w tekście:

| Lp. | Rzeczywsite słowa kluczowe | Predykowane słowa kluczowe |
|---|---|---|
| **1** | "szybkie pożyczki gotówkowe", "pożyczka", "pożyczka przez internet" | [] |
| **2** | "kurs pierwszej pomocy" | [] |
| **3** | "mechanik samochodowy", "warsztat samochodowy", "przegląd auta" | [] |
| **4** | "ubezpieczenie na życie" | [] |
| **5** | "ubezpieczenia zdrowotne", "ubezpieczenia prywatne" | [] |

Uzyskane wyniki są identyczne zarówno dla notacji IOB jak i IOBS. Znaki [] oznaczają, że model nie znalazł żadnych słów kluczowych w tekście. Proces lematyzacji tekstu okazał się niezbędny, aby model był w stanie wychwycić choć kilka słów kluczowych.
Jakość predykcji modelu opartego na lematach oceniana sposobem łagodnym dla przedstawia się następująco:

- Precission: 16.67%,

- Recall: 4.17%,
- F1Score: 6.67%.

Wszystkie miary ocenione metodą restykcyjną wynoszą 0%.

Ze względu na lepsze wyniki uzyskiwane za pomocą notacji IOBS, format IOB nie będzie dalej badny. Analogicznie dalsze badania nie będą przeprowadzane dla odmienionych, niebazowych słów występujących w tekście.

## 5.2 Drugi zbiór cech

Ocena predykcji słów przedstawionych w formie występującej w tekście oraz jako lemat:

| **Sposób oceniania** | **Precision** | **Recall** | **F1 Score** |
|---|---|---|---|
| **Łagodny** | 52.31% | 12.29% | 19.90% |
| **Restrykcyjny** | 24.32% | 5.44% | 8.88% |

Z powyżej tabelki można odczytać, iż dodanie nowych cech pozytywnie wpłynęło na jakość predykcji słów kluczowych.

Wyniki ekstrakcji słów kluczowych na pomocniczym zbiorze testowym:

| **Lp.** | **Rzeczywsite słowa kluczowe** | **Predykowane słowa kluczowe** |
|---|---|---|
| **1** | "szybki pożyczka gotówkowy", "chwilówka", "pożyczka", "pożyczka przez internet" | [] |
| **2** | "kurs pierwszy pomoc" | "ratownictwo" |
| **3** | "mechanik samochodowy", "warsztat samochodowy", "przegląd auto" | [] |
| **4** | "ubezpieczyć na życie" | [] |
| **5** | "ubezpieczenie zdrowotny", "ubezpieczenie prywatny" | "obowiązkowy ubezpieczenie", "zdrowotny" |

Jakość predykcji oceniona za pomocą łagodnej ewaluacji:

- Precission: 50.00%,
- Recall: 8.33%
- F1Score: 14.29% .

### 5.3 Trzeci zbiór cech

Ocena predykcji słów przedstawionych w formie występującej w tekście oraz jako lemat:

| Sposób oceniania | Precision | Recall | F1 Score |
|---|---|---|---|
| **Łagodny** | 38.65% | 19.93% | 26.29% |
| **Restrykcyjny** | 19.21% | 11.10% | 14.05% |

Wyniki ekstrakcji słów kluczowych na pomocniczym zbiorze testowym:

| Lp. | Rzeczywsite słowa kluczowe | Predykowane słowa kluczowe |
|---|---|---|
| **1** | "szybki pożyczka gotówkowy", "chwilówka", "pożyczka", "pożyczka przez internet" | "móc" |
| **2** | "kurs pierwszy pomoc" | "ratownictwo" |
| **3** | "mechanik samochodowy", "warsztat samochodowy", "przegląd auto" | [] |
| **4** | "ubezpieczyć na życie" | "urząd pracy", "narodowy", "fundusz" |
| **5** | "ubezpieczenie zdrowotny", "ubezpieczenie prywatny" | "umowa ubezpieczeniowej" |

Wszystkie wartości miar jakości modelu wyniosły 0.

Dodanie krotek *(obecne słowo, następne słowo)* oraz *(poprzednie słowo, następne słowo)* pozwoliło na zwiększenie wartości miar Recall oraz F1 Score, natomiast negatywnie wpłynęło na iare Precission. Model nauczył się wynajdywać więcej słów kluczowych.

## 6 Analiza wyników

Jakość otrzymywanych wyników jest bardzo niska, można jednak zauważyć, iż dodanie tagu "stop word" oraz lematyzacja tekstu spowodowały poprawę jakości predykcji. Pozytywnie na wyniki wpływa również dodawanie większej liczby cech słów. Nadal jednak jakość predykcji pozostawia wiele miejsca na poprawę. Omówione i zinterpretowane zostaną teraz poszczególne, zastosowane miary dla scenraiusza wykorzsytującego format IOBS oraz łagodny sposób oceniania:

- Precision 53.27% - ponad połowa wyznaczanych słów kluczowych jest poprawna, jest to niezły wynik, świadczący o tym, że model ma stosunkowo duża pewność kiedy wskazuje dane słowo jako jedno ze słów kluczowych,
- Recall 18,60% - mniej niż 1 na 5 słów kluczowych jest znajdowane przez model. Jest to kiepski wynik, mogacy wskazywać na niewyuczonie przez model większości zależności identyfikujących słowa kluczowe. Może mieć na to wpływ zestaw cech wykorzystywanych przy klasyfikacji.

Warto zwrócić uwagę na klasyfikację słów do pośredniego formatów IOB i IOBS. Ze względu na zdecydowaną dominację klasy O (słowo nie jest słowem kluczowym ani jego częścią) model mógł mieć problemy z poprawnym wyuczeniem się rozpoznawania rzadziej występujących klas. Średnie wartości metryk wszystich klas występujących w zbiorze sięgają prawie 100%, co wydaje się zaskakująco dobrym rezultatem, ale nie przekłada się na poprawne wyznaczanie słów kluczowych.

# 7 Potencjał modelu

Uczenie modelu na ciągu treningowym na CPU (Intel i7-4702MQ 2.2GHz) zajmuje ok. 3 minuty, co jest dużą zaletą wykorzystanego modelu. Jednakże model ten nie wydaje się być odpowiedni do rozwiązania problemu ekstracji słów kluczowych dla krótkich tekstów w języku polskim. Nie jest on w stanie dobrze wyuczyć się predykcji klas, które występują rzadko w zbiorze uczącym (w rozważanym rozwiązaniu są to klasy Inside i Beginning). Ze względu na naturę słów kluczowych (występują rzadko w tekście) stwarza to duże problemy w uzyskaniu zadowlających wyników.

# 8 Wnioski

Model CRF nie radzi sobie dobrze z problem wyszukiwania słów kluczowych w tekstach. Problem na pewno stano język polski, który jest językiem skomplikowanym i trudnym. Istniejące rozwiązanie można sprobować polepszyć poprzez zapewnienie większej liczby przykładów uczących, a w szczególności zwiększenie całkowitej liczby słów kluczowych. Warto byłoby dodać do zbioru danych teksty bardzo krótkie oraz zbadać wpływ innych cech słów na jakość predykcji.