WIKIQUIZ

Algorytm tworzenia testu z sugerowanymi opcjami odpowiedzi na podstawie artykułów z Wikipedii

autor: Mariusz Andziak

Założenia projektu

- korzystanie wyłącznie z artykułów pisanych w języku polskim
- możliwość tworzenia quizu dla praktycznie dowolnego artykułu Wikipedii

Techniczne założenia projektu

- pominięcie rozwiązań związanych z Machine Learningiem, Deep Learningiem
- stworzenie projektu na podstawie możliwości zewnętrznych bibliotek, ale opieranie ich na własnych algorytmach napisać od zera (skupienie się na rozwoju własnych algorytmów)

Poboczny cel projektu:

- implementacja narzędzia do odmiany przez przypadki rzeczowników

Wykorzystane technologie/języki/biblioteki

Wykorzystane zasoby

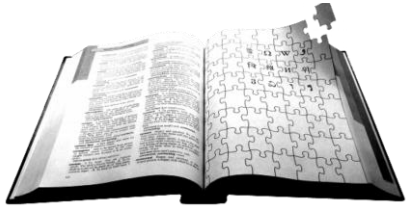
Lokalne

Polish language model for SpaCy:
http://zil.ipipan.waw.pl/SpacyPL

Sparsowane zbiory przymiotników i czasowników JSON:
https://github.com/tombusby/PolishCaseTrainer/tree/master/polish_case_trainer/word/word-data

Sieciowe

WIKIPEDIA



WIKTIONARY
the free dictionary

Idea działania algorytmu

Algorytm analizuje sparsowany artykuł z Wikipedii pod kątem wystąpienia trzech rodzajów słów lub ich grup:

- sekwencji słów bazujących na łańcuchu Marcova
- słów pojawiających się często w tekście
- podmiotów, które dało radę się zidentyfikować, tzn. nazw geograficznych, nazw organizacji, dat, nazwisk ludzi itp.

Tworzenie enteties - Named Entity Recognition

Oryginalny SpaCy bazuje na korpusie OntoNotes 5.0
https://catalog.ldc.upenn.edu/LDC2013T19

Tagger polskiego modelu został wytrenowany na 1 milionie słów z National Corpus of Polish
http://clip.ipipan.waw.pl/NationalCorpusOfPolish%7D

oraz na 0,5 milionie słów z Polish Language of the 1960s
http://clip.ipipan.waw.pl/PL196x

Rozpoznawane enteties - Named Entity Recognition

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.

Rozpoznawane enteties - Named Entity Recognition

LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including ”%“.
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	“first”, “second”, etc.
CARDINAL	Numerals that do not fall under another type.

Znajdowanie niepoprawnych odpowiedzi

1. Analiza zdania pod kątem występowania słów, które mogą zostać zidentyfikowane (enteties), jak nazwy geograficzne, osoby itp. ale także słowa częste i elementy z łańcucha Marcova.
2. Losowanie innego słowa z tego samego słownika enteties przy założeniu, że obydwa słowa zaczynają się takimi samymi literami (mała/wielka) i mają inną lemmę.

Przykłady działania

https://pl.wikipedia.org/wiki/Przemiana_adiabatyczna

```
Przemiana adiabatyczna jest przemianą, w której zmieniają się parametry stanu gazu, m.in. ciśnienie, objętość właściwa, temperatura, energia wewnętrzna, entalpia.
'Równanie _____ można dla takiego przypadku zapisać następująco:Wstawiając równania Clapeyrona i odpowiednio przekształcając można uzyskać inne postacie równania _____, wiążące ze sobą temperaturę i objętość oraz temperaturę i ciśnienie czynnika:Krzywe obrazujące procesy adiabatyczne zwane są adiabatami.'
```

```
chosen_word
```

```
'Poissona'
```

```
write_bad_words_for_single(chosen_word)
```

```
{'Poissona': ['Clapeyrona']}
```

Przykłady działania

https://pl.wikipedia.org/wiki/Hiszpania

```
Zgodnie z zapisem w konstytucji Hiszpanii, każdy obywatel tego kraju ma obowiązek znać język kastylijski.
'Język kataloński jako urzędowy występuje w Katalonii i na _____.'
```

```
write_bad_words_for_single(chosen_word)
```

```
{'Balearach': ['Atlantyku', 'Covadonga', 'El Caracol']}
```

Przykłady działania

https://pl.wikipedia.org/wiki/Linux

```
Linux jest jednym z przykładów wolnego i otwartego oprogramowania (FLOSS): jego kod źródłowy może być dowolnie wykorzystywany, modyfikowany i rozpowszechniany.
'Od kwietnia 2017 roku _____ (a tym samym Linux) oficjalnie jest najpopularniejszym systemem operacyjnym na świecie.'
```

```
write_bad_words_for_single(chosen_word)
```

```
{'Android': ['FLOSS', 'SRPM', 'Mac OS X']}
```

Przykłady działania - krótki artykuł

https://pl.wikipedia.org/wiki/Lanmodez

Lanmodez [edytuj]

Lanmodez (bret. *Lanvaodez*) – miejscowość i gmina we Francji, w regionie Bretania, w departamencie Côtes-d'Armor.

Według danych na rok 1990 gminę zamieszkiwały 392 osoby, a gęstość zaludnienia wynosiła 94 osoby/km² (wśród 1269 gmin Bretanii Lanmodez plasuje się na 891. miejscu pod względem liczby ludności, natomiast pod względem powierzchni na miejscu 1042.).

Bibliografia [edytuj | edytuj kod]

- Francuski urząd statystyczny (fr.).

```
'Według danych na rok 1990 gminę zamieszkiwały 392 osoby, a gęstość zaludnienia wynosiła 94osoby/km² (wśr
ód 1269 gmin _______________ Lanmodez plasuje się na 891. miejscu pod względem liczby ludności, natomiast
pod względem powierzchni na miejscu 1042.).'
```

```
write_bad_words_for_single(chosen_word)
```

```
{'Bretanii': ['Lanvaodez', 'Francji', 'Lanmodez']}
```

Co się dzieje jeżeli algorytm nie znajdzie odpowiedzi

https://pl.wikipedia.org/wiki/Marynarka_Wojenna

```
Poza wyjątkami wskazanymi w ratyfikowanych przez Rzeczpospolitą Polską przepi
sach prawa międzynarodowego organy obcego państwa nie mogą wykonywać w stosun
ku do okrętów aktów władczych ani ingerować w ich życie wewnętrzne.
'Wszelkie próby ingerencji stosunku do _____ powinny być zdecydowan
ie odparte, a w razie zagrożenia życia załogi, bądź siłowego naruszenia immun
itetu przysługującego okrętowi należy postępować zgodnie z zasadami użycia si
ły (ang. „Rules Of Engagement”) oraz przepisami prawa międzynarodowego.'
```

```
write_bad_words_for_single(chosen_word)
```

```
{'okrętów': ()}
```

Co się dzieje jeżeli algorytm nie znajdzie odpowiedzi

Proponuje się zastosować podobieństwo wyrazów i wybrać przynajmniej jedną odpowiedź o najwyższym stopniu podobieństwa inną niż właściwa:

```
Poza wyjątkami wskazanymi w ratyfikowanych przez Rzeczpospolitą Polską przepi
sach prawa międzynarodowego organy obcego państwa nie mogą wykonywać w stosun
ku do okrętów aktów władczych ani ingerować w ich życie wewnętrzne.

'Wszelkie próby ingerencji stosunku do _______________ powinny być zdecydowan
ie odparte, a w razie zagrożenia życia załogi, bądź siłowego naruszenia immun
itetu przysługującego okrętowi należy postępować zgodnie z zasadami użycia si
ły (ang. „Rules Of Engagement”) oraz przepisami prawa międzynarodowego.'
```

```
write_bad_words_for_single(chosen_word)
```

```
{'okrętów': ()}
```

```
[('okrętów', 1.0),
 ('niszczycieli', 0.9158970819055983),
 ('flotylli', 0.858886492561332),
 ('okrętach', 0.8527440373090631),
 ('kutrów', 0.8502905383004363),
 ('floty', 0.8275401477314428),
 ('okręty', 0.8270314671786636),
 ('fregaty', 0.8231612686765043),
 ('okrętu', 0.7951390927840019),
 ('korwety', 0.789201610969838),
 ('eskadry', 0.7787919879615797),
 ('okręt', 0.7710204085222641),
 ('niszczyciele', 0.7516646576491781),
 ('Okręty', 0.748263947753357),
 ('dywizjonu', 0.7339920300669945),
 ('marynarzy', 0.7286280528145127),
 ('uzbrojenia', 0.7201205558971374),
```

Co się dzieje jeżeli algorytm nie znajdzie odpowiedzi

Ale nie rozwiązuje to problemu polskiej deklinacji. Co by było, gdyby zaznaczony wyraz miał największy stopień podobieństwa?

```
Poza wyjątkami wskazanymi w ratyfikowanych przez Rzeczpospolitą Polską przepi
sach prawa międzynarodowego organy obcego państwa nie mogą wykonywać w stosun
ku do okrętów aktów władczych ani ingerować w ich życie wewnętrzne.

'Wszelkie próby ingerencji stosunku do _______________ powinny być zdecydowan
ie odparte, a w razie zagrożenia życia załogi, bądź siłowego naruszenia immun
itetu przysługującego okrętowi należy postępować zgodnie z zasadami użycia si
ły (ang. „Rules Of Engagement”) oraz przepisami prawa międzynarodowego.'
```

```
write_bad_words_for_single(chosen_word)
```

```
{'okrętów': ()}
```

```
[('okrętów', 1.0),
 ('niszczycieli', 0.9158970819055983),
 ('flotylli', 0.858886492561332),
 ('okrętach', 0.8527440373090631),
 ('kutrów', 0.8502905383004363),
 ('floty', 0.8275401477314428),
 ('okręty', 0.8270314671786636),
 ('fregaty', 0.8231612686765043),
 ('okrętu', 0.7951390927840019),
 ('korwety', 0.789201610969838),
 ('eskadry', 0.7787919879615797),
 ('okręt', 0.7710204085222641),
 ('niszczyciele', 0.7516646576491781),
 ('Okręty', 0.748263947753357),
 ('dywizjonu', 0.7339920300669945),
 ('marynarzy', 0.7286280528145127),
 ('uzbrojenia', 0.7201205558971374),
```

Co się dzieje jeżeli algorytm nie znajdzie odpowiedzi

https://pl.wikipedia.org/wiki/Energetyka_s%C5%82oneczna

```
Pod uwagę brana jest między innymi efuzja możliwa dzięki dużej różnicy mas atomów wodoru i tlenu, oraz użycie wirówek.
'Konieczność pracy w tak wysokiej temperaturze powoduje duże straty _____, wysokie koszty budowy urządzeń, ich
szybkie zużywanie się i małą sprawność.'
```

```
write_bad_words_for_single(chosen_word)
```

```
{'energii': ()}
```

Co się dzieje jeżeli algorytm nie znajdzie odpowiedzi

Proponuje się zatem zmodyfikowanie algorytmu i wzbogacenie go o możliwość deklinacji

Algorytm nie znalazł alternatywnych odpowiedzi

```
Pod uwagę brana jest między innymi efuzja możliwa dzięki dużej różnicy mas atomów wodoru i tlenu, oraz użycie wirówek.
'Konieczność pracy w tak wysokiej temperaturze powoduje duże straty _____, wysokie koszty budowy urządzeń, ich
szybkie zużywanie się i małą sprawność.'
```

```
write_bad_words_for_single(chosen_word)
```

```
{'energii': ()}
```

Algorytm, po uwzględnieniu podobieństwa i polskiej deklinacji znalazł sensowne odpowiedzi alternatywne

```
write_bad_words_if_none()
```

```
['węgla', 'ciepła', 'wody']
```

Jak starano się rozwiązać problem deklinacji polskiej

Dwa ujęcia:

- lokalne
- sieciowe

Jak starano się rozwiązać problem deklinacji polskiej

Ujęcie lokalne - pliki JSON

Jak starano się rozwiązać problem deklinacji polskiej

Ujęcie lokalne - pliki JSON

```
ret_declin_form('nosorożca')
```

```
'biernik'
```

```
ret_declin_form('monitorze')
```

```
'miejscownik'
```

```
ret_declin_form('konstytucją')
```

```
'narzędnik'
```

```
ret_declin_form('gęsi')
```

```
'biernik'
```

```
ret_declin_form('samochodowi')
```

```
'celownik'
```

```
ret_declin_form('ognia')
```

```
'dopełniacz'
```

Jak starano się rozwiązać problem deklinacji polskiej

Ujęcie sieciowe - Wiktionary

```
ret_declin_form_wiktionary('żółw', True)
```

```
['żółw', 'żółwia', 'żółwiowi', 'żółwia', 'żółwiem', 'żółwiu', 'żółwiu']
```

```
ret_declin_form_wiktionary('przełącznikiem', False)
```

```
'narzędnik'
```

```
change_declination('korkociągi', 'celownik')
```

```
'korkociągom'
```

Przykłady działania

https://pl.wikipedia.org/wiki/Psychoza

```
Przykładem omamu rzekomego jest „głos mówiący w głowie”, „widok ludzi na Księżycu”, „zadanie słysza
ne od szwagra z Gdyni (podczas gdy pacjent przebywa w Krakowie)” i podobne – doznania zmysłowe (wzr
ok, słuch – w odniesieniu do pozostałych zmysłów nie ma to zastosowania) dochodzące z przestrzeni,
dla której „zdrowy” nie percepowałby z powodu ograniczenia znanego mu zasięgu zmysłu.
'Przyjmuje się, że omamy rzekome należą do _____ spostrzegania, choć istnieją co do tego
wątpliwości.'
```

```
write_bad_words_for_single(chosen_word)
```

```
{'zaburzeń': ()}
```

```
bad_answers
```

```
['psychoz', 'objawów', 'psychopatologii']
```

Namiastka frontendu (bonus)

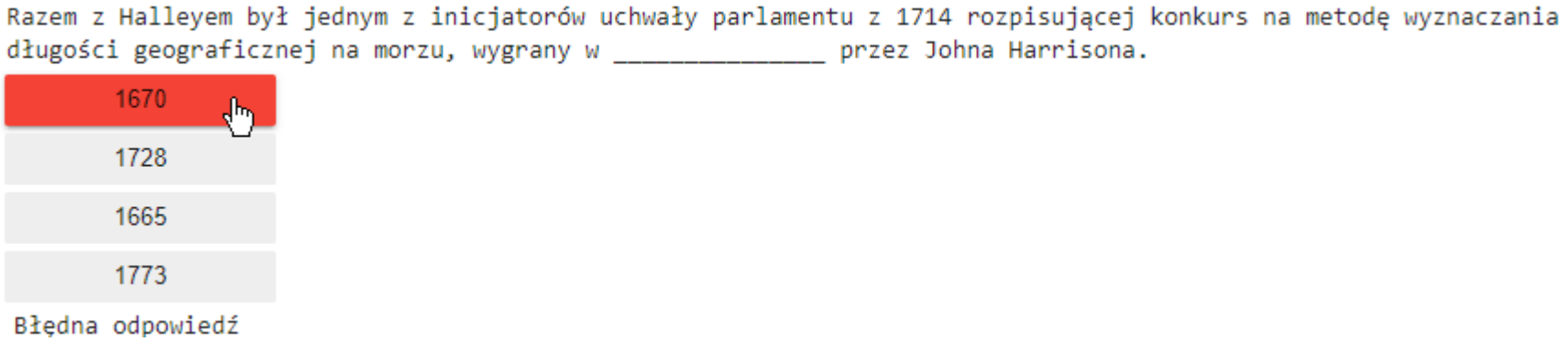
https://pl.wikipedia.org/wiki/Isaac_Newton

Razem z Halleyem był jednym z inicjatorów uchwały parlamentu z 1714 rozpisującej konkurs na metodę wyznaczania długości geograficznej na morzu, wygrany w _____ przez Johna Harrisona.

- 1670
- 1728
- 1665
- 1773

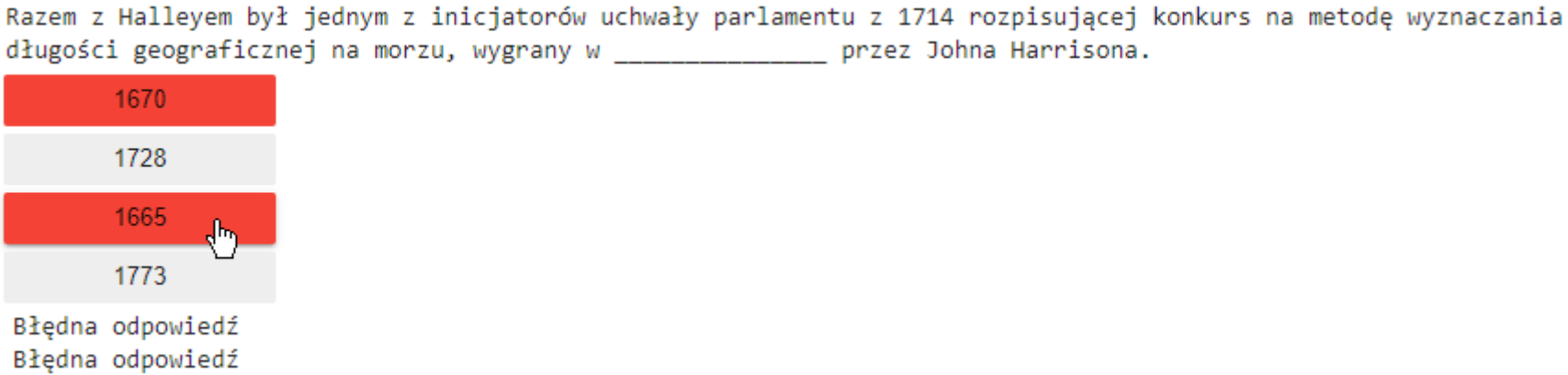
Namiastka frontendu (bonus)

https://pl.wikipedia.org/wiki/Isaac_Newton



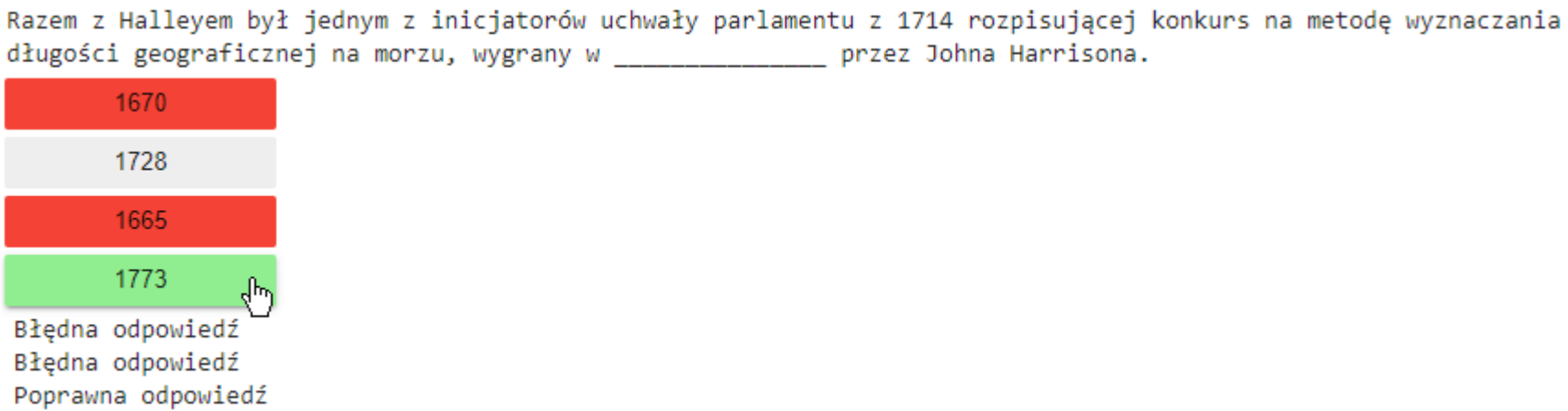
Namiastka frontendu (bonus)

https://pl.wikipedia.org/wiki/Isaac_Newton



Namiastka frontendu (bonus)

https://pl.wikipedia.org/wiki/Isaac_Newton



Wnioski

Algorytm pomimo sprawnego działania w większości artykułów Wikipedii, w przypadku niektórych nie zwraca zadowalających wyników, co może powodować występowanie błędów.

Powodem powyższego zachowania jest nie do końca właściwe odnajdywanie lemmy słowa zaimplementowane w Spacy:

```
ret_lemma('woda')
'woda'

ret_lemma('wodą')
'woda'

ret_lemma('wody')
'woda'
```

```
ret_lemma('ciecz')
'ciecz'

ret_lemma('cieczą')
'ciecza'

ret_lemma('cieczy')
'ciecza'
```

Algorytm odmiany przez przypadki bazuje na Wikisłowniku, do którego powinno być wpisane hasło ‘ciecz’, aby można było wygenerować jego prawidłową odmianę i zastosować pozostałe algorytmy np. do odnajdywania słów podobnych i zamiany ich na odpowiednią formę deklinacyjną. Niestety zwracane jest ‘ciecza’, co nie jest zachowaniem prawidłowym i będzie generowało dalsze błędy.

Wnioski

System sugerowanych złych odpowiedzi do wylosowanego zdania jest zadowalający, czasami nawet zaskakująco zadowalający w przypadku nazw własnych:

```
W 1967 władze zakazały wszelkich praktyk religijnych.
'W 1968 Albania skrytykowała interwencję _____ w Czechosłowacji.'
```

```
write_bad_words_for_single(chosen_word)
```

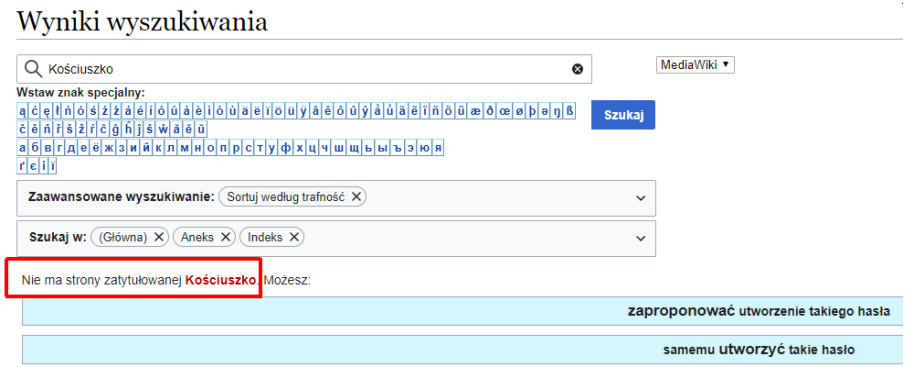
```
{'Układu Warszawskiego': ['Unii na rzecz Regionu Morza Śródziemnego',
  'Secret Intelligence Service',
  'MIR']}
```

Wnioski

Jednak budowanie systemu uwzględniającego polską deklinację wyrazów, którymi są nazwy własne nie jest wystarczająco efektywne, gdyż nie wszystkie nazwy własne (nawet jednowyrazowe, a szczególnie wielowyrazowe) występują w Wikisłowniku i o ile dany podmiot nie występuje w tekście odmieniony w odpowiednim przypadku, algorytm go nie odmieni poprawnie.

Tadeusz Kościuszko [edytuj]

Andrzej Tadeusz Bonawentura Kościuszko herbu Roch III (ur. 4 lutego 1746 w Mereczowszczyźnie, zm. 15 października 1817 w Solurze) – inżynier wojskowy[1], Najwyższy Naczelnik Siły Zbrojnej Narodowej w czasie insurekcji kościuszkowskiej, generał lejtnant wojska Rzeczypospolitej Obojga Narodów, generał major komenderujący w Dywizji Wielkopolskiej w 1792 roku[2], *brevet* generał brygady Armii Kontynentalnej w czasie wojny o niepodległość Stanów Zjednoczonych.



Ewentualne dalsze plany

Zwiększenie prawdopodobieństwa wylosowania pytań zawierających słowa znaczące, a nie głównie wygenerowane podmioty

Zastosowanie zapytań o deklinację również do Słownika gramatycznego języka polskiego
http://sgjp.pl/

Przyspieszenie działania algorytmu podobieństwa wyrazów wyszukującego niepoprawne odpowiedzi, gdy główny algorytm nic nie znalazł

Zawartość lokalnych plików JSON z przymiotnikami i rzeczownikami powinna się powiększać po każdorazowym odpytaniu Wiktionary

Wzbogacenie pytań o system punktowy

Formatowanie odpowiedzi np. dat za pomocą jednego stylu

Powiązanie pytań z sobą i stworzenie swego rodzaju automatu do generowania testów tak, aby np. poprawna odpowiedź w pytaniu pierwszym nie była taka sama w pytaniu drugim.