Can you Polish your Dutch?

Jacek Pardyak

2017-08-10

Wprowadzenie

Popatrzmy na dwa słowniki (zbiory słów):

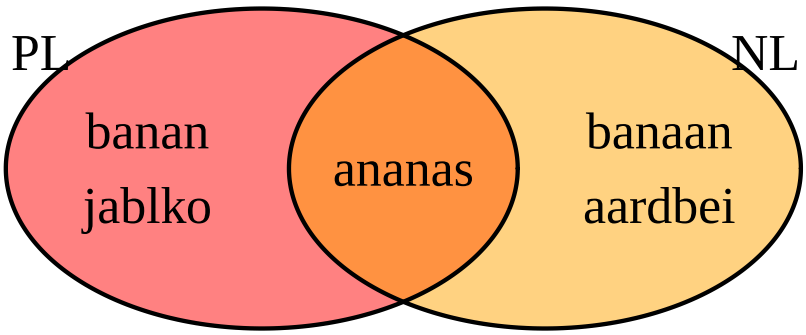
Table 1: Słowa w słowniku polskim

ananas
banan
jabłko

Table 2: Słowa w słowniku niderlandzkim

ananas
banaan
aardbei

Interesuje nas podobieństwo słów w obu słownikach, które może być zaprezentowane:



Będą nas interesować słowa identyczne i podobne (przy zdefiniowaniu, co to znaczy “podobne”).

Dane

Alfabety

Słowa to ciągi elementów (liter) należacych do pewnego zbioru (alfabetu). Oba języki posługują się innymi alfabetami. Możemy porównać oba alfabety:[1]

[1] Źródło: https://en.wikipedia.org/wiki/Polish_orthography , oraz https://en.wikipedia.org/wiki/Dutch_orthography

Table 3: Alfabety

Letter	NL	PL
a	TRUE	TRUE
ą	FALSE	TRUE
b	TRUE	TRUE
c	TRUE	TRUE
ć	FALSE	TRUE
d	TRUE	TRUE
e	TRUE	TRUE
ę	FALSE	TRUE
f	TRUE	TRUE
g	TRUE	TRUE
h	TRUE	TRUE
i	TRUE	TRUE
j	TRUE	TRUE
k	TRUE	TRUE
l	TRUE	TRUE
ł	FALSE	TRUE
m	TRUE	TRUE
n	TRUE	TRUE
ń	FALSE	TRUE
o	TRUE	TRUE
ó	FALSE	TRUE
p	TRUE	TRUE
q	TRUE	FALSE
r	TRUE	TRUE
s	TRUE	TRUE
ś	FALSE	TRUE
t	TRUE	TRUE
u	TRUE	TRUE
v	TRUE	FALSE
w	TRUE	TRUE
x	TRUE	FALSE
y	TRUE	TRUE
z	TRUE	TRUE

Letter	NL	PL
ź	FALSE	TRUE
ż	FALSE	TRUE

Dodatkowo oba języki stosują dwuznaki.

Słowa

Słowa powstają z liter alfabetu.

Słowniki

Słowniki to zbiory słów.[2]

[2] słowniki czynne, branżowe, słownictwo ….

Słowniki Aspell

Wybrałem słowniki **Aspell**, bo są tam oba języki (porównujemy jabłka do jabłek). Nie mógł być np SJP. **Aspell**[3] to standardowy w systemach GNU program do sprawdzania pisowni. Słowniki pochodzą z tej strony .

[3] `http://aspell.net/`

Słowa słownika Aspell

Jak wyglądają zgromadzone tam dane?

Table 4: Pierwsze słowa w słowniku polskim

Word
A
AA
AAN
AAP
ABA
ABB

Table 5: Ostatnie słowa w słowniku polskim

	Word
289835	ów
289836	ówczesność/MN
289837	ówczesny/bXxYy
289838	ówdzie
289839	ówże

	Word
289840	ósmi

Table 6: Pierwsze słowa w słowniku niderlandzkim

Word
A
A-attest
A-attesten
A-biljet
A-biljetten
A-bom

Table 7: Ostatnie słowa w słowniku niderlandzkim

	Word
341456	öres
341457	über
341458	über-ich
341459	überhaupt
341460	übermensch
341461	übermenschen

Polski słownik zawiera: 289840 słów, zaś niderlandzki 341461 słów.

Słowniki a listy słów

Niektóre języki, w tym polski stosują fleksję (z łac. przeistoczenie słów, przemiana) dla nadania słowom nowej funkcji gramatycznej.

Na przykład:

Table 8: Forma podstawowa

ówczesny

Table 9: Możliwe odmiany

nieówczesny
ówczesnymi
ówczesnych
ówczesnym

ówczesnemu
ówczesnego
ówczesną
ówczesnej
ówczesne
ówczesna
ówcześni
ówcześnie
nieówczesnymi
nieówczesnych
nieówczesnym
nieówczesnemu
nieówczesnego
nieówczesną
nieówczesnej
nieówczesne
nieówczesna

W Aspell zaimplementowano 7101 reguł odmiany końcówek słów w zależności od ich znaczenia. W języku niderlandzkim tych reguł niemal nie ma.

Po ich zastosowaniu lista polskich słów zawiera: 3761314 słów, zaś niderlandzkich tyle samo, bo 341461 słów. Spowodowało to 13 - krotny wzrost liczby słów polskich, gdy uwzględni się ich różne formy gramatyczne.

Listy słów pozyskano z programu Aspell w Ubuntu wykonując komendy:

```
aspell --lang=pl dump master | aspell --lang=pl expand | tr
' ' '\n' > wordsPL.dict
```

oraz

```
aspell --lang=nl dump master | aspell --lang=nl expand | tr
' ' '\n' > wordsNL.dict
```

Słowniki aspell –lang=en dump master > dictEN.dict

Przygotowanie danych

Pracować będziemy na słownikach, a nie listach słów. Zatem informacje o przypisanym słowom regułach możemy usunąć i posługiwać się formą podstawową słowa.

Opis statystyczny słów słowników polskiego i niderlandzkiego

Rozkład długości słów

Zaczynamy od zmierzenia długości słów.

Minimalna, maksymalna i średnia długość polskich słów:

Table 10: Podsumowanie długości słów polskich

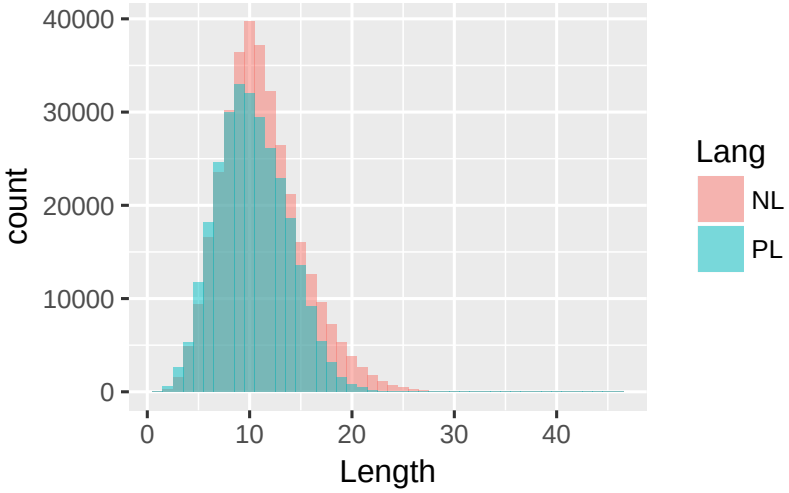
	V1
Min.	1.00000
1st Qu.	8.00000
Median	10.00000
Mean	10.31133
3rd Qu.	13.00000
Max.	33.00000

Minimalna, maksymalna i średnia długość niderlandzkich słów:

Table 11: Podsumowanie długości słów niderlandzkich

	V1
Min.	1.00000
1st Qu.	8.00000
Median	11.00000
Mean	11.18185
3rd Qu.	13.00000
Max.	46.00000

Wyniki możemy porównać graficznie:



Najdłuższe polskie słowo:
dziewięćdziesięciopięcioipółletni[4]

[4] Ninety-five-half-year-old

Najdłuższe niderlandzkie słowo:
arbeidsongeschiktheidsverzekeringsmaatschappij[5]

[5] Disability Insurance Society

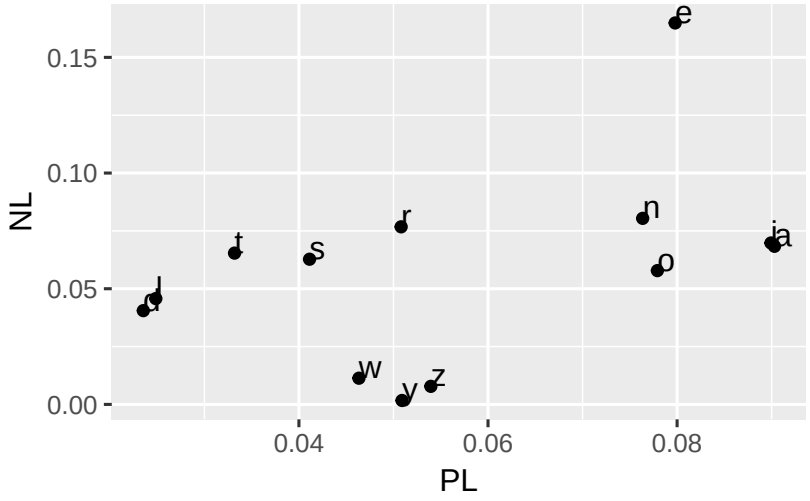
Częstotliwość występowania liter w obu językach

Porównujemy względną częstotliwość.

Table 12: Względna częstotliwość występowania liter w obu językach (top 10)

Letter	NL	PL
e	0.1648690	0.0797926
n	0.0804287	0.0763589
r	0.0767599	0.0508108
i	0.0697762	0.0899286
a	0.0683425	0.0902994
t	0.0653945	0.0331907
s	0.0627574	0.0411251
o	0.0578341	0.0779135
l	0.0457366	0.0248665
d	0.0405318	0.0235418
z	0.0078108	0.0539537
y	0.0017058	0.0509062
w	0.0113298	0.0463456

Wyniki przedstawia poniższy wykres.

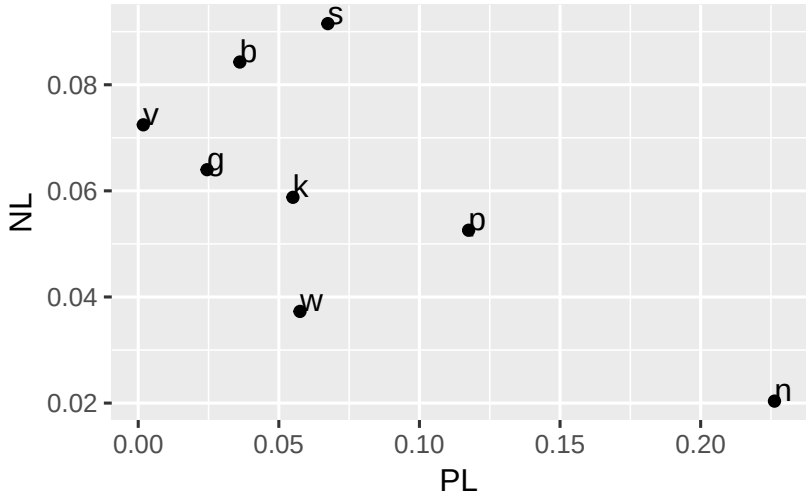


Częstotliwość występowania początkowych liter w obu językach

Największa przewidywalność początkowch liter słów.

Table 13: Względna częstotliwość występowania początkowych liter w obu językach (top 5)

Initial	NL	PL
s	0.0915214	0.0674269
b	0.0842790	0.0361303
v	0.0724504	0.0018044
g	0.0639897	0.0244756
k	0.0587856	0.0550166
n	0.0203742	0.2262559
p	0.0525770	0.1174924
w	0.0372810	0.0575248

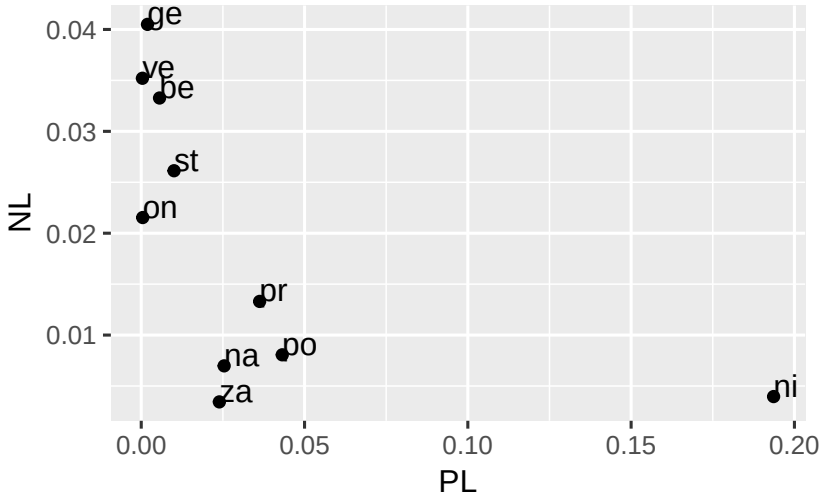


Częstotliwość występowania początkowych digrafów w obu językach

Table 14: Względna częstotliwość występowania początkowych digrafów w obu językach (top 5)

Digraph	NL	PL
ge	0.0404995	0.0019183
ve	0.0352105	0.0003657
be	0.0332717	0.0056065
st	0.0261318	0.0100400
on	0.0215369	0.0004485
ni	0.0039624	0.1936275
po	0.0080624	0.0431755
pr	0.0133046	0.0362476

Digraph	NL	PL
na	0.0069788	0.0253864
za	0.0034411	0.0239201



W niderlandzkim `ge-` to przedrostek określający abstrakcyjne koncepcje pochodzące od czasownika.

Table 15: Niderlandzki przedrostek ge-

Przedrostek	Grondwoord	Resultat
ge-	zeuren	gezeur
	piekeren	gepieker
	fluiten	gefluit

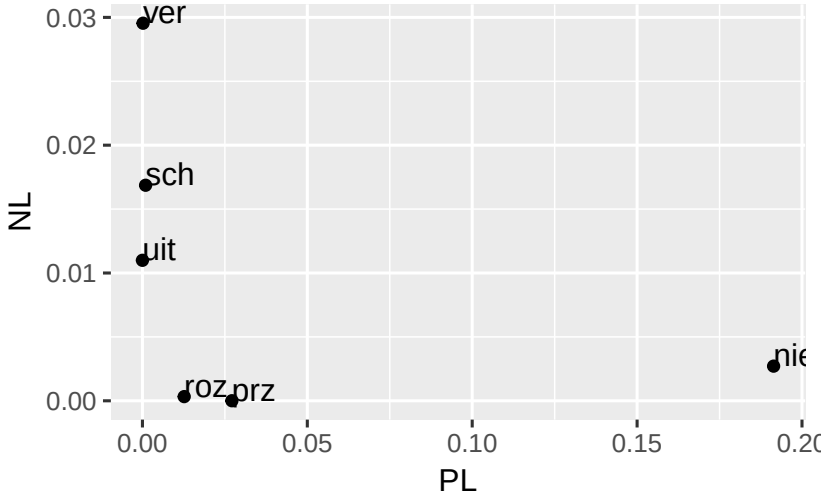
Table 16: Tłumaczenie słów z niderlandzkim przedrostkiem ge-

Przedrostek	Grondwoord	Resultat
ge-	zeuren	gezeur
	piekeren	gepieker
	fluiten	gefluit

Częstotliwość występowania początkowych trigrafów w obu językach

Table 17: Względna częstotliwość występowania początkowych trigrafów w obu językach (top 3)

Trigraph	NL	PL
ver	0.0295436	0.0001898
sch	0.0168658	0.0009591
uit	0.0109939	0.0000000
nie	0.0027119	0.1913918
prz	0.0000059	0.0271253
roz	0.0003280	0.0126587



W polskim `nie-` to przedrostek określający zaprzeczenie, które nie musi mieć negatywnego ładunku.

Table 18: Polski przedrostek nie-

Przedrostek	Rdzen	Rezultat
nie-	spokojny	niespokojny
	zwykły	niezwykły
	winny	niewinny

Probabilistyczny model tworzenia tekstu

Tworzenie pisanego tekstu języka naturalnego polega na tworzeniu określonych sekwencji liter. W procesie tworzenia tekstu dużą rolę odgrywa struktura probabilistyczna elementów tekstu - liter.

Modele wybudowano używając Łancuchy Markowa[6]

Za pomocą wybudowanych modeli możemy dokonać predykcji.

Otrzymujemy w ten sposób najbardziej prawdopodobne słowa:

[6] `https://pl.wikipedia.org/wiki/%C5%81a%C5%84cuch_Markowa`

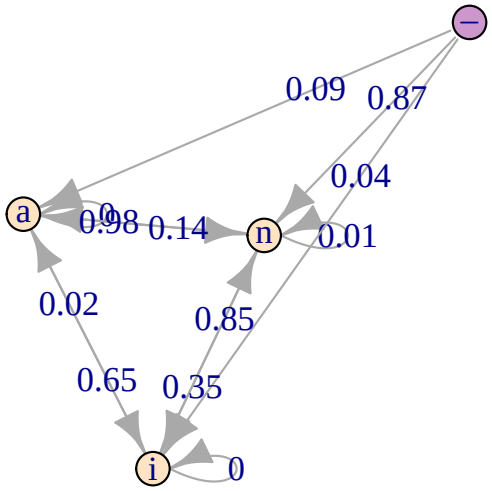


Figure 1: Fragment polskiego modelu zaprezentowany graficznie.

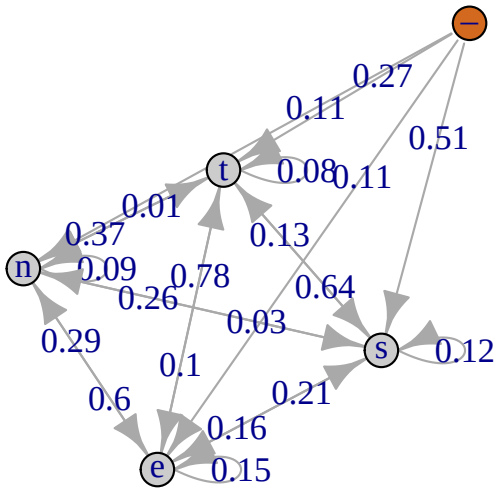


Figure 2: Fragment niderlandzkiego modelu zaprezentowany graficznie.

Polskie : **nie**[7]

i niderlandzkie: **sten**[8]

Ciekawe są "nowe słowa" wygenerowane za pomocą modeli - słowa mające cechy języka naturalnego, jednak ("póki co") w nim nie istniejące.

[7] nee, niet

[8] ?????

Table 19: Przykłady nowych słów polskich

skuka
prysy
mebiny
dęteka
donąc
rzyki
miezać
pozet
urtać
henesy

Table 20: Przykłady nowych słów niderlandzkich

hede
zorm
mideun
stuuine
aaste
hevé
notijs
gers
oubeel
smenge

Póki co widać różnice między językami, a gdzie są podobieństwa?

Podobieństwo słów z różnych słowników

Definicja podobieństwa dwóch słów

Istnieje wiele sposobów mierzenia podobieństwa dwóch łańcuchów znaków (pojedyńczych słów, wyrażeń, pełnych zdań, czy też tekstów)[9]. Do poszukiwania podobnych łańcuchów stosuje się **Przybliżone dopasowanie łańcuchów**[10]

[9] `https://en.wikipedia.org/wiki/String_metric`

[10] `https://en.wikipedia.org/wiki/Approximate_string_matching`

Identyczne słowa

Słowa identyczne są pisane dokładnie tak samo w obu językach. Z analiz wykluczono słowa zaczynające się wielką literą (skróty, imiona, nazwiska, nazwy geograficzne). Oraz słowa zbyt krótkie lub zbyt długie. Przykłady znalezionych identycznych słów:

Table 21: Przykłady identycznych słów w językach polskim i niderlandzkim

Word
ananas
balkon
chaos
duet
echo
filet
gratis
handel
impotent
jacht
kapsel
legenda
wiek

3352 słów w słowniku niderlandzkim występuje w polskim słowniku. Większość z tych słów pochodzi z angielskiego bądź francuskiego. Są też fałszywi przyjaciele[11]! Fałszywi przyjaciele to słowa w dwóch językach, które wyglądają i brzmią podobnie, ale znacząco różnią się w znaczeniu.

[11] `https://en.wikipedia.org/wiki/False_friend`

Z jednej strony mogą być źródłem pomyłek, a z drugiej śmiesznych skojarzeń ułatwiajacyh ich zapamiętanie:

Ania nosi buty na **hak**u[12]

[12] obcas

Kasia ma nowy **kapsel**[13] na głowie

[13] fryzura

Ja[14] mówię tak, a ptak na **tak**[15]u pyta jak?

[14] tak
[15] gałąź

Ten ptak ma dwa **wiek**[16]i.

[16] skrzydło

Pani z **pan**[17]em

[17] patelnia

Słowa podobne

Do wyszukiwania słów podobnych napisałem skrypt w Python zobacz **Załącznik**.

Table 22: Przykłady słów polskich i ich niderlandzkich przyjaciół

Word	Friend	Score
abiturient	abituriënt	95
banan	banaan	91
bestseler	bestseller	95
dermatolog	dermatoloog	95
fortepian	fortepiano	95
wachta	wacht	91

Takich słów jest 2629

Pełna lista jest dostępna pod adresem: `https://docs.google.com/spreadsheets/d/1rJojwRpEp0dHCa077zlWxW547PUHEIVzK-9vfIUOsG0/`

Ta metoda zawiedzie w przypadkach, gdy słowa są podobne, ale mają różne początkowe digrafy. wyjątek: wirus - virus kryzys - crisis

Załącznik

```
#from fuzzywuzzy import fuzz
from fuzzywuzzy import process
import pandas as pd

# Reading the datasets in a dataframe using Pandas
nl = 'C:\\Users\\A599131\\Documents\\PolishYourDutch\\dics\\nl.wl'
nl = pd.read_csv(nl, header = None)
nl.columns   = ['Word']
nl['Language'] = "NL"

pl = 'C:\\Users\\A599131\\Documents\\PolishYourDutch\\dics\\pl.wl'
pl = pd.read_csv(pl, header = None)
pl.columns = ['Word']
pl['Language'] = "PL"

dics = pd.concat([pl, nl])

dics['Word']  = dics['Word'].apply(lambda x: x.split('/', 1)[0])
dics['Length'] = dics['Word'].apply(lambda x: len(x))
dics['Upper']  = dics['Word'].apply(lambda x: x[0].isupper())
dics['Initial']  = dics['Word'].apply(lambda x: x[0:2])

# filter words with 3 < Length < 7
dics = dics.loc[dics['Length'] > 3]
dics = dics.loc[dics['Length'] < 12]
```

```
dics = dics.loc[dics['Upper'] == False]

# split back
pl = dics.loc[dics['Language'] == 'PL']
nl = dics.loc[dics['Language'] == 'NL']

# optionally reindex
pl = pl.reset_index(drop=True)
nl = nl.reset_index(drop=True)
dics = dics.reset_index(drop=True)

# save to file

pl.to_csv('C:\\Users\\A599131\\Documents\\PolishYourDutch\\dics\\pl_clean.csv',
          columns = ["Word"],
          index = True,
          encoding = 'utf-8')

nl.to_csv('C:\\Users\\A599131\\Documents\\PolishYourDutch\\dics\\nl_clean.csv',
          columns = ["Word"],
          index = True,
          encoding = 'utf-8')


def my_fun(x,y):
  query = x
  if y == 'PL':
    language = 'NL'
  else:
    language = 'PL'
  table = dics.loc[dics['Language'] == language]
  table = table.loc[table['Initial'] == query[0:2]] # the same inital
  table = table.loc[table['Length'] > len(query)-1]
  table = table.loc[table['Length'] < len(query)+2]
  if len(table.index) == 0:
    res = ('','','')
  else:
    table = table['Word']
    res = process.extractOne(query, table)
  return(res)


query = dics['Word'][5455] # banan
query = dics['Word'][2384] # ananas - ananas / good
```

```
query = dics['Word'][4499] # auto - auto / good
query = dics['Word'][3151] # apartament - appartement / bad (AM)
query = dics['Word'][3196] # aperitif - aperitief / good
query = dics['Word'][116913] # truskawka - trustakte
query = dics['Word'][144549] # ćmawy - ''
print(query)
print(my_fun(query, 'PL'))

dics['Friend'] = ''
dics['Score'] = ''

# [1: range(0, 3)
for index in range(0, 314271) :
  temp = my_fun(dics['Word'][index], dics['Language'][index])
  dics['Friend'][index] = temp[0]
  dics['Score'][index]  = temp[1]

dics.to_csv('C:\\Users\\A599131\\Documents\\PolishYourDutch\\dics\\dics_clean.csv',
          index = True,
          encoding = 'utf-8')

odmienność: tak
X   antytetycznego, antytetycznemu, antytetycznych, antytetycznym, antytetycznymi, nieantytetycznego, ni
Y   antytetyczni, nieantytetyczni
b   nieantytetyczny
x   antytetyczna, antytetyczną, antytetyczne, antytetycznej, nieantytetyczna, nieantytetyczną, nieantyte
y
```

References