Interdyscyplinarny blog badaczy z Instytutu Językoznawstwa i Pracowni Systemów Informacyjnych Uniwersytetu Adama Mickiewicza.

RE-RESEARCH.PL

re-research

.pl .en .de .ru rss atom

Interdyscyplinarny blog badawczy pracowników i UAM

wyszukaj

Start Publikacje Zasoby O stronie Kontakt Archiwum

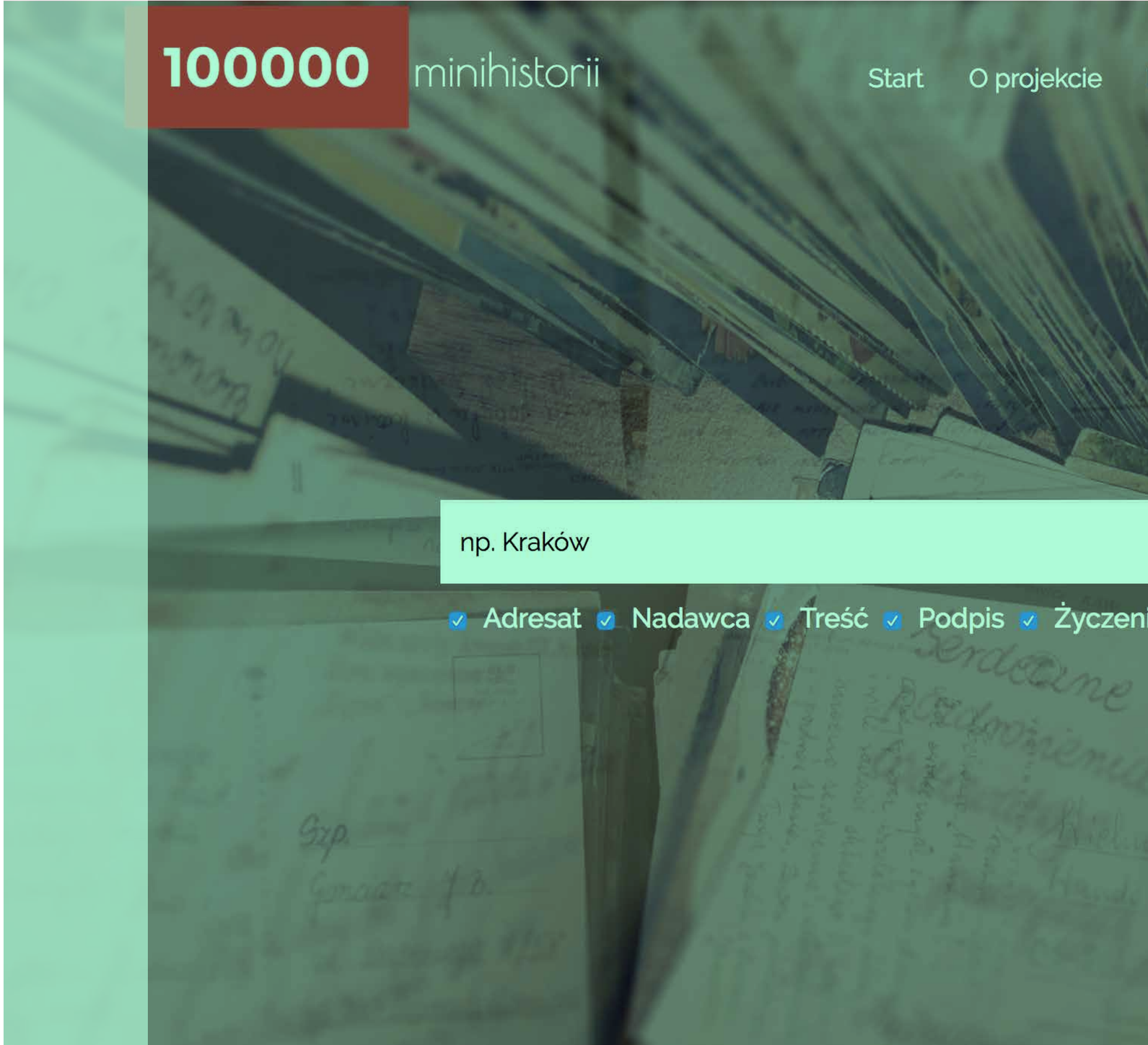
Szynobus tygodnia 9

Piszący te słowa spotkał się kiedyś z opinią, że słuchanie Slayera o poranku jest lepsze nawet od porannej kawy. Być to wszystko może, trzeba jednak wziąć pod uwagę, że niektórym udaje się wytrzymać bez obu i całkiem nieźle sobie radzą. Nie znaczy to wcale, że nie czerpią

Zespół, czyli kto?

100 000 minihistorii

Celem projektu jest językoznawczy i realioznawczy opis rzeczywistości, jaka wyłania się z analizy korpusu pocztówek polskich, które zostały wysłane do adresatów w latach 1945–1989.



Biogram Narodu

Sztuczna inteligencja czyta dostępne elektronicznie polskie teksty historyczne i wyławia z masy 12,3 mld słów wszystkie odniesienia do osób.

Automatyczny biograf

Sztuczna inteligencja czyta dostępne elektronicznie polskie teksty historyczne i wyławia z masy 12,3 mld słów wszystkie odniesienia do osób.

Pobierz konspekt

1901

Ołoniec, miasto pow. gub. Ołonieckiej, nad rz. Megregą i Ołonką, w pobliżu granic Finlandji, liczy 1,303 miesz. (1897).— *Ołoniecki powiat,* ma na przestrzeni 8,113 w. kw. 41,239 miesz.— *Ołoniecka gubernja,* zajmuje 112,322 w. kw. i graniczy na północ i wschód z gubernją Archangielską, na południe z gub. Wołogodzką i Petersburską, na zachód z jeziorem Ładoskiem i Finlandją; pod wzglę-

Encyklopedja Powszechna

Z ILUSTRACJAMI I MAPAMI.

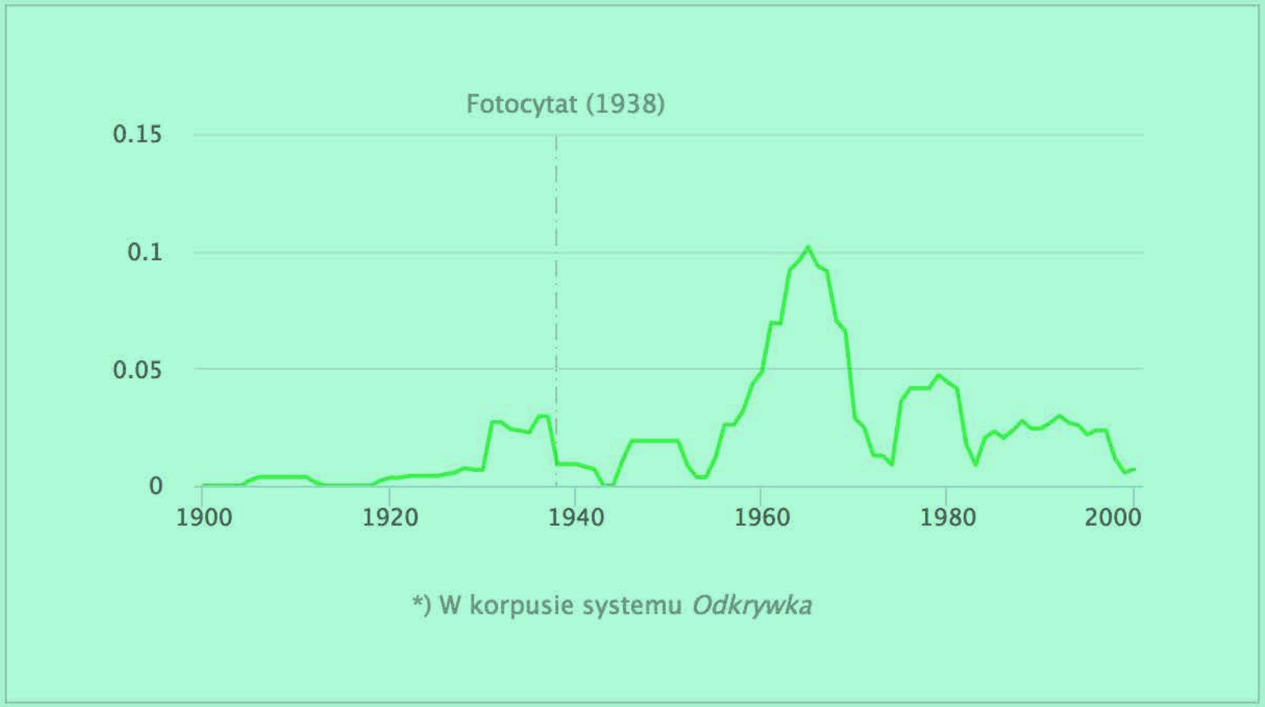
1901
1914
1918
1923
1937
1938
1939
1940
1943
1944

Narodowy Fotokorpus Języka Polskiego

Największy zbiór leksykalny polszczyzny XX w. wraz z poświadczeniem cytatowym. Jego naczelną dystynktywną zasadą jest każdorazowe udokumentowanie ekscerptu w postaci fotodokumentacyjnej,

Dodatkowe informacje

Diachroniczna częstość użycia słowa (wystąpień na milion wyrazów):



*) W korpusie systemu *Odkrywka*

Lokalizacja ekscerptu na stronie:

Adres bibliograficzny:

Doroszewski, Witold 1938. Język polski w Stanach Zjednoczonych AP, Warszawa : Nakł. TNW

Redatacje

Nieregularnik prezentujący bieżące odkrycia źródłowe, które pozwalają przesunąć wstecz, choćby o rok tylko, datę pojawienia się w druku poszczególnych wyrazów czy wyrażeń – w stosunku do datacji znanej z istniejącej literatury przedmiotu.

Miłośnikom polskiego słowa do lotu!

REDATACJE

Nieregularnik Leksykograficzny
Jana Wawrzyńczyka i Piotra Wierzchonia

Nr 7 Warszawa, 6 lipca 2016 r. Druk bezpłatny

1 **ateizacja** 1988 → 1896

Jednostka opracowana przez Teresę Smółkową pośpiesznie[1]. Wyraz znany, np. w suplemencie do Słownika Szymczaka, wcześniej opisany przez Jana Miodka[2]. Dysponuję starszym cytatem: [...] wystawia ona [demokracja] społeczność na niebezpieczeństwo ateizacji państwa i społeczeństwa.[3]

Jan Wawrzyńczyk

2 **komisarka** 2004 → 1930

Chodzi o formę odpowiadającą masculinum **komisarz** w pracy Smółkowej[4] udokumentowaną przez tę ekscerptorkę cytatem z tygodnika „Wprost". Piotr Wierzchoń znalazł ją w popularnym (niegdyś) „Ikacu" sporo wcześniejszym[5].

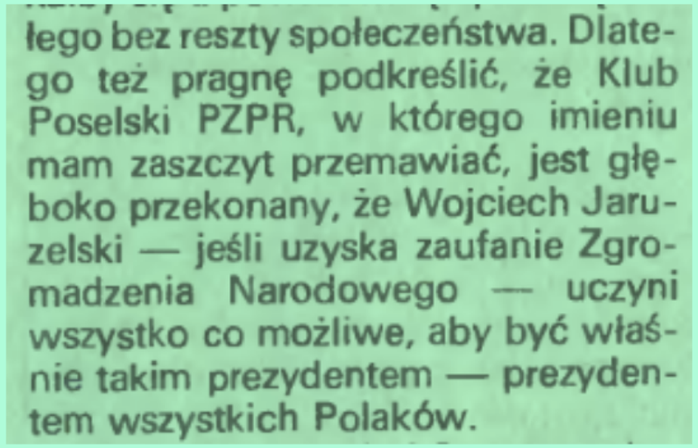
Jan Wawrzyńczyk

Pierwszy prezydent wszystkich Polaków?

Daniel Dzienisiewicz

Podczas kampanii prezydenckich często słyszymy, że zwycięzca wyścigu po ten urząd powinien być prezydentem wszystkich Polaków, wznoszącym się ponad podziałami partyjnymi i ideologicznymi. Zainteresowało nas więc, kto mógł być pierwszym *prezydentem wszystkich Polaków*. Może określano kogoś w taki sposób już w burzliwych czasach dwudziestolecia międzywojennego?

Otóż nie! Moda na jednoczenie narodu rozpoczęła się w roku 1989, a palma pierwszeństwa należy się Wojciechowi Jaruzelskiemu, który został zarekomendowany jako prezydent wszystkich Polaków z ramienia PZPR:


łego bez reszty społeczeństwa. Dlatego też pragnę podkreślić, że Klub Poselski PZPR, w którego imieniu mam zaszczyt przemawiać, jest głęboko przekonany, że Wojciech Jaruzelski — jeśli uzyska zaufanie Zgromadzenia Narodowego — uczyni wszystko co możliwe, aby być właśnie takim prezydentem — prezydentem wszystkich Polaków.


łego bez reszty społeczeństwa. Dlatego też pragnę podkreślić, że Klub Poselski PZPR, w którego imieniu mam zaszczyt przemawiać, jest głęboko przekonany, że Wojciech Jaruzelski — jeśli uzyska zaufanie Zgromadzenia Narodowego — uczyni wszystko co możliwe, aby być właśnie takim prezydentem — prezydentem wszystkich Polaków.

GŁOS POMORZA

Proletariusze wszystkich krajów łączcie się!

KOSZALIN

SŁUPSK

Czwartek, 20 lipca 1989r.

SPECJALISTA
Władysław Sikorski garbuje skórę dzikich zwierząt. — Specjalność: iberalesy. — Kraków, Veit-Stoss Str. 34
1343k

GONIEC KRAKOWSKI

Kraków, piątek 14 czerwca 1940 r.

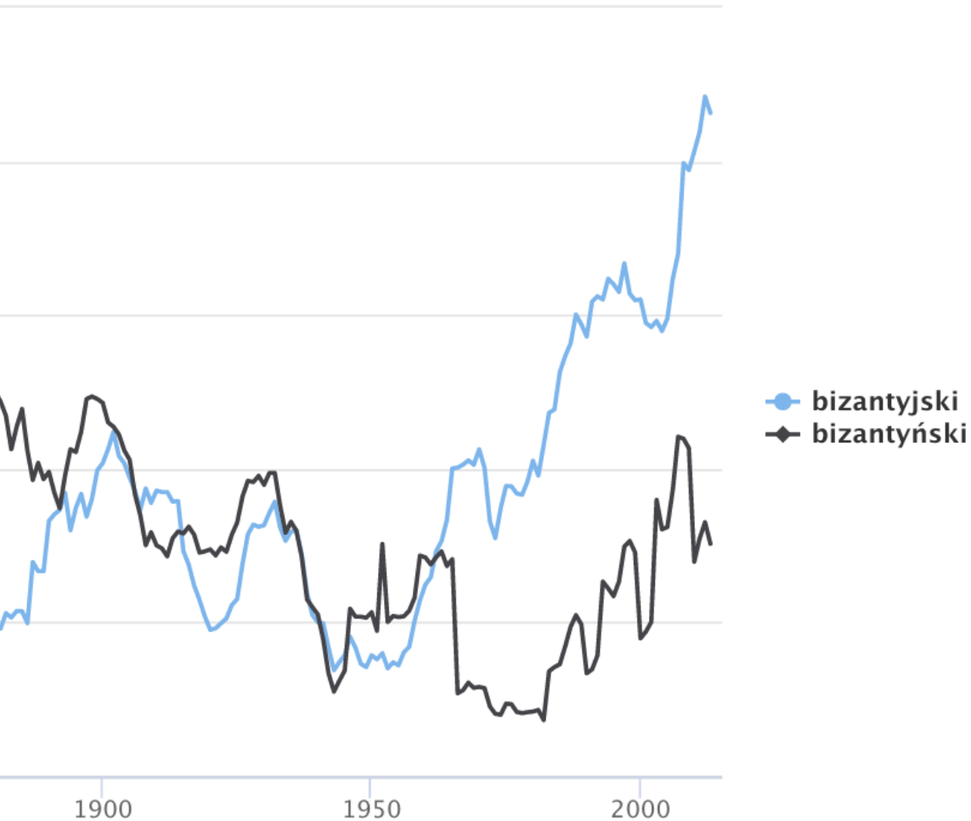
Garbowanie skór iberalesów

Filip Graliński

W książce „Smutek anegdot. Etniczne dygresje do wspomnień i pomysły refleksji” (skądinąd bardzo inspirujące usypisko opowieści) Krzysztof Kwaśniewski wspomina:

> Ja jednak na własne oczy widziałem drobne ogłoszenie w zupełnie oficjalnym wydaniu owego „Gońca…” [polskojęzycznej gadzinówki wydawanej przez niemieckiego okupanta], które brzmiało: „Garbuję skóry dzikich zwierząt. Specjalność iberalesy”. Po czym następował jakiś adres, zapewne fikcyjny (dziś niektórzy mówią, że był to adres „Kraków, Veit Stoss Str. 34”, a nawet, że zaczynało się to ogłoszenie od słów „Specjalista Władysław Sikorski…”, ale z tym stanowczo się nie zetknąłem). [s. 229]

Można by myśleć, że to tylko okupacyjna legenda miejska ku pokrzepieniu serc. Skąd, to najprawdziwsza prawda!



Bizantyjski czy bizantyński?

Daniel Dzienisiewicz

10/05

Nasze dotychczasc
wszystkim na celu
wystąpień określor
przedstawienie ich
przez nas tekstów.
ortograficznymi.

Jeśli ktoś zapytałb
bizantyjski czy *biz*
pierwsza. Tak bow
otoczeniu oraz tylk
szkole, np. na lekcj
spotkałem wyrazu

Łatwo więc wyobra
zbadałem częstość
tekstach. Ogółem w
naszych źródłach 1
niewiele więcej, bo
Należy zaznaczyć,
bizantyński. Na pr
względnym równo
drugiej połowie XX
przeważać. Obecni
częstszą. Frekwen
została zilustrowar

Konkursy redatacyjne

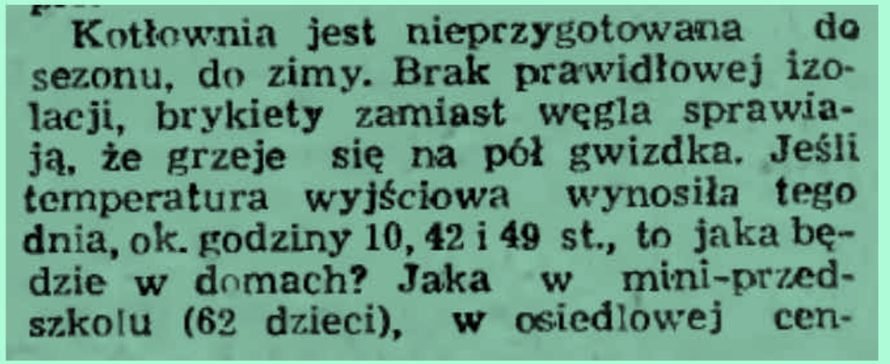
Na pół gwizdka — kto da wcześniej?

Filip Graliński

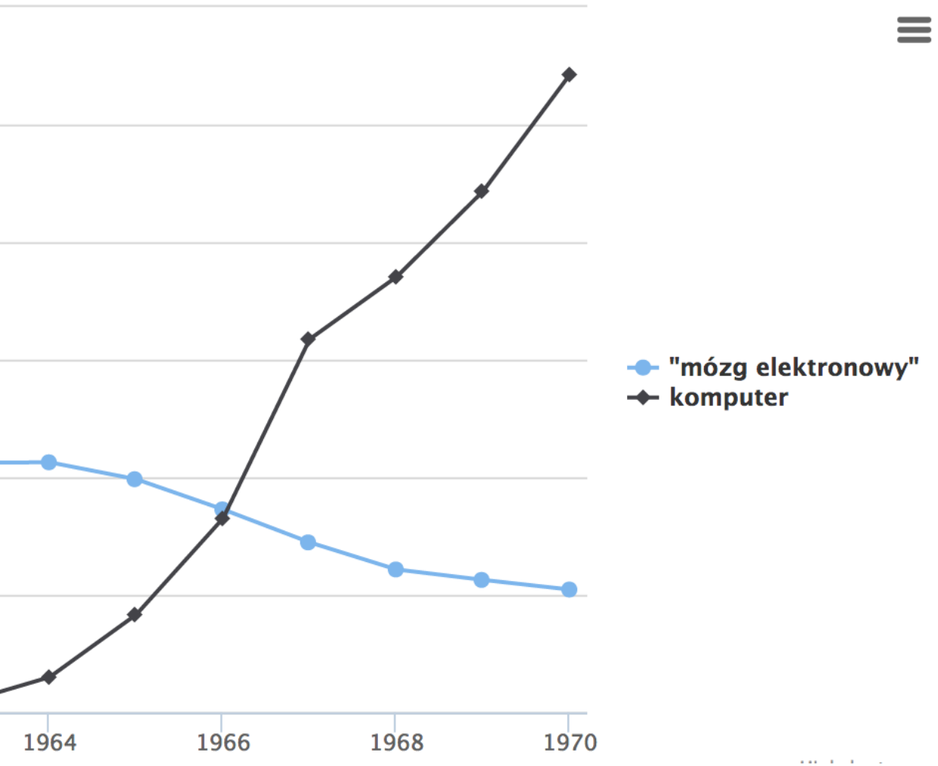
Uwaga! Pierwszy konkurs chronologizacyjny! Za każdy rok poniżej znanego najwcześniejszego datowania — 5 zł.

Zaczynamy od frazeologizmu „na pół gwizdka” i roku 1980:

na pół gwizdka



Kotłownia jest nieprzygotowana do sezonu, do zimy. Brak prawidłowej izolacji, brykiety zamiast węgla sprawiają, że grzeje się na pół gwizdka. Jeśli temperatura wyjściowa wynosiła tego dnia, ok. godziny 10, 42 i 49 st., to jaka będzie w domach? Jaka w mini-przedszkolu (62 dzieci), w osiedlowej cen-



Komputer – kto da wcześniej?

Daniel Dzienisiewicz

Kolejny konkurs chronologizacyjny! Tym razem na tapet bierzemy wyraz *komputer*.

W naszych zbiorach rzeczownik ten ujawnia się najwcześniej w 1965 r. Kontekst odsyła nas jednak do tekstu starszego o rok:

[4] Na terenie wyższych uczelni, instytutów badawczych PAN, przemysłu oraz ekonomiki, używanych jest obecnie dla celów prognoz meteorologicznych, podliczania wyników egzaminów wstępnych, do obliczeń naukowo-technicznych i do zarządzania — 6 typów komputerów, produkowanych w Polsce (zob. artykuł *Maszyn matematycznych rozmowy*. Tryb. Ludu nr 194 z 16 VII 1964).

PRZEGLĄD
BIBLIOTECZNY

ROCZNIK XXXIII — ZESZYT 4
PAŹDZIERNIK — GRUDZIEŃ 1965

Józef Stalin był Polakiem i szewcem

Marcin Pigulak

Dziś 17 września, więc postaraliśmy się o „radziecki” temat. Ciekawe, czy (nie)sławny sowiecki dyktator wiedział, że w Polsce żył człowiek posługujący się identycznym co on imieniem i nazwiskiem?

SZEWSKICH czeladników
na stała targową pracę —
dwóch uczni — przyjmę za-
raz. Mieszkanie, wikt — w
miejscu. Józef Stalin, Cha-
bówka. 8155g

Cena numeru w Krakowie 20 gr. na prowincji

ILUSTROWANY

KURYER CODZIENNY

PRENUMERATA WYNOSI

Kraków, piątek 17 września 1926.

SZEWSKICH czeladników
na stała targową pracę —
dwóch uczni — przyjmę za-
raz. Mieszkanie, wikt — w
miejscu. Józef Stalin, Cha-
bówka. 8155g

Cena numeru w Krakowie 20 gr. na prowincji

ILUSTROWANY

KURYER CODZIENNY

PRENUMERATA WYNOSI

33) Ten Tytus Albucyusz przed swem jeszcze wygnaniem za zdzierstwa popełnione w Sardynii, osiadł w Atenach, i zarzuciwszy język, zwyczaje i obyczaje swego kraju, zupełnie zgreczał. Lucyliusz w żartobliwych na Albucyusza wierszach powiada, że Scewola przybywszy do Aten w urzędzie pretora, spotkał go i rzekł do niego: „Wolałeś udawać Greka, niżeli pozostać Rzymianinem, Sabinem, spółziomkiem Poncyusza, Trytanniusza, dzielnym setnikiem, chorążym w wojsku krajowem. Więc ja pretor Rzymski, witam cię, mój Tytusie, w Atenach po grecku: *Χαῖρε, Tite.* Liktorowie, żołnierze, wszyscy obecni zawołali z śmiechem: *Χαῖρε, Tite.* I toć to jest, dodaje Scewola, czego mi Albucyusz nie przebaczył za co jest moim nieprzyjacielem.“ Cicero, *de Finibus*, I, 3.

PISMA

KRASOMOWCZE I POLITYCZNE

MARKA TULIUSZA

CYCERONA

ROZMOWA O MOWCY
BRUTUS CZYLI O SŁAWNYCH MOWCACH
MOWCA BRUTUSOWI POŚWIĘCONY
O DOSKONAŁYCH MOWCACH
O RZECZYPOSPOLITEJ
O PRAWACH

PRZEŁOŻONE

PRZEZ

E. RYKACZEWSKIEGO.

POZNAŃ.

NAKŁADEM BIBLIOTEKI KÓRNICKIEJ.

1873.

Tytus Albucjusz udawał Greka

Daniel Dzienisiewicz

Profesor Jerz
pochodzenie

> Nierozumie
> nieznajom
> *tureckim k*
> chcemy po
> według teg
> o kimś mó
> udaje kogo
> dobrze wie
> dla mnie ja

Do tej pory pr
mnie niejasn
pewne domys
hipotezy etyn

Cyceron pisał
mówca Tytus
w Atenach m
władać greką,
prawie za Gre
kpin, co znala
cytowanej prz
niewłaściwe a

W niedawno otwartym w Katowicach barze pod szyldem „hot-dog" – o każdej porze dnia pełno zajadających. Specjalności: gorące parówki, podane na wąskiej bułeczce, przyprawione dla smaku pikantym ketchupem, albo musztardą, kanapki z twarożkiem, jajkiem, serem, śledzikiem, posypane cebulką, pietruszką. Tu odkrywamy, że kanapka jest dobra zawsze, na śniadanie i kolację, podtrzymuje siły, gdy nie ma czasu zjeść obiadu i pokrzepia w służbowej podróży.

Psiokrwisty międzywojenny hot dog

Daniel Dzienisiewicz

09/19

Dziś bę
kilka sł
się pora

Poniew
zawsze
rytualn
temat n
food" za
napisal
słownic
tekst si
tym fak

Ale ad r

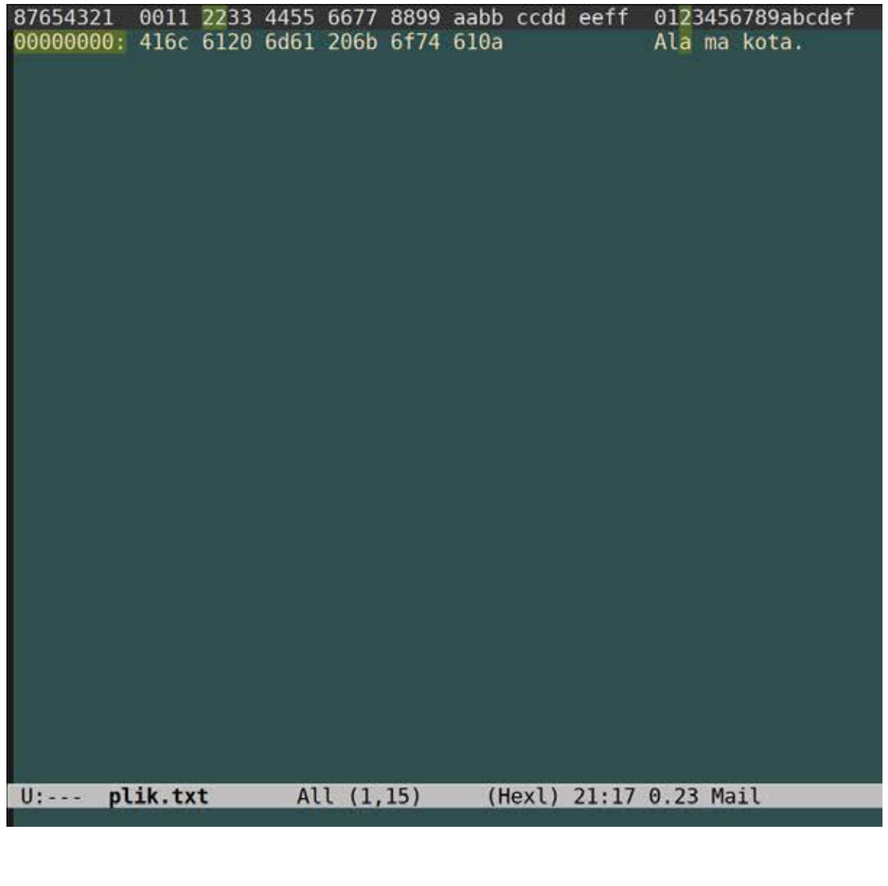
Wyraz t
tekstac

1.5

1.25

1

wyrazów)



Anatomia pliku tekstowego — część I (wstęp)

Filip Graliński

Plik tekstowy? tekstowy?

```
Ala ma kota.
```

Widziałem ludz
Widziałem zma
tysiące dolaró
pracowników k

Wszystko dlate
plik tekstowy t

W tym cyklu w
plików tekstow
czyhają między

Czy jesteś prog
informacje tuta
Czytelniku, prz
poślesz mi bom
że przeczytane
czas i pieniądz

Zera i

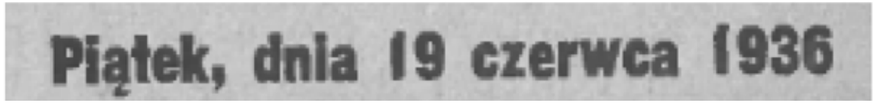
Tak naprawdę,
— w przeciwie

W r. 2139 będą na świecie sami warjaci

Jeden z Anglików obliczył, że w roku 2139 cały świat będzie składał się tylko z samych warjatów. Nie są to żadne przepowiednie, ani też jakieś dociekania, tylko ściśle matematyczne obliczenia. W roku 1859 wypadał 1 warjat na 535 zdrowych, w roku 1897 jeden warjat już tylko na 312 zdrowych; dalej rozwija się stosunek następująco: w r. 1926 — 1:150, w r .1977 będzie 1:100, idąc zaś tym sposobem liczenia, dochodzi się do wniosku, że w roku 2139 cały świat będzie się składać z samych warjatów. Te obliczenia są dokonane ściśle matematycznie. Jak można więc nie wierzyć!

Egzempl. 10 groszy Miesięcznie 1,95 zł (dostawa poczt. 39 gr)

Orędownik



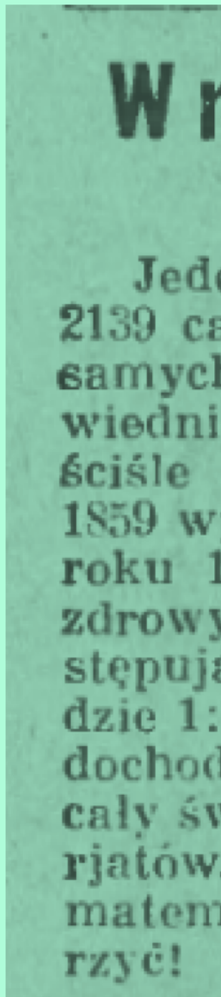
Piątek, dnia 19 czerwca 1936

W 2139 „będą na świecie sami warjaci”, czyli historia pewnych ekstrapolacji

Łukasz Borchmann

10/29

Czytelnicy „
zaznajomić
obliczeniam
przynajmnie
w roku 2139
tylko *warjat*



W
Jed
2139 ca
samych
wiedni
ściśle
1859 w
roku 1
zdrowy
stępuj
dzie 1:
dochod
cały św
rjatów
matem
rzyć!

Vive la petite différence! Exploiting small differences for gender attribution of short texts

Author hidden for review

Affiliation hidden for review

Abstract. This article describes a series of experiments on gender attribution of Polish texts. The research was conducted on the publicly available corpus called "He Said She Said", consisting of a large number of short texts from the Polish version of Common Crawl. As opposed to other experiments on gender attribution, this research takes on a task of classifying relatively short texts, authored by many different people.

For the sake of this work, the original "He Said She Said" corpus was filtered in order to eliminate noise and apparent errors in the training data. In the next step, various machine learning algorithms were developed in order to achieve better classification accuracy.

Interestingly, the results of the experiments presented in this paper are fully reproducible, as all the source codes were deposited in the open platform *Gonito.net*. Gonito.net allows for defining machine learning tasks to be tackled by multiple researchers and provides the researchers with easy access to each other's results.

Keywords: gender attribution, text classification, corpus, Common Crawl, re-seach reproducibility

1 Introduction

Gender classification of written language has been a subject of linguistic studies for decades now. Note, for instance, a ground-breaking book [8], describing characteristic features of women's language. In more recent years, linguists and socio-linguists researching this subject have been aided by statisticians, see for instance: [11] and [10]. Furthermore, development of social media opened a possibility of building large-scale text corpora, annotated with meta information regarding the author's gender. This resulted in numerous projects aimed at automatic gender classification based on training data acquired from the Web. The annotated text resources used to build the corpora were often taken from blogs, e.g. [10].

However, we believe that gender annotated corpora scraped from the selected Web sources are prone to some flaws. Firstly, they may suffer from thematic bias – women tend to write about different subjects than men, see [11]. Secondly, there might be significantly more text written by authors of either gender. This is due to the fact that if a corpus is a collection of gender annotated items such as blog entries, these items might differ significantly in length. And lastly – the volume of the corpus may not be sufficient for reliable statistical analysis. All these flaws result in a situation, where gender

Brneńskie południe i konferencja TSD 2016

Rafał Jaworski

09/20

12 wrześni
konferenc
miejscowa
stolicy Mo
jedenastej

Dokładnie
Jest to zw
tego miast
Lennarta T
wraz ze sw
miesięcy, a
południa n
czym przy
miasta bro
gdyby nie
we wszyst
o jedenast
uczynił —

Na szczęś
przeszkod
czasie i wy
prezentacj
lingwistyc
przewodni

Różne, różne treści...

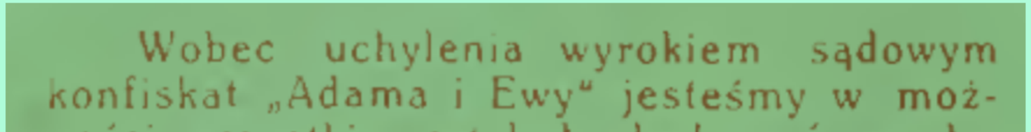
Adam i Ewa: świerszczyk lat trzydziestych? (18+)

Daniel Dzienisiewicz

Wraz z odzyskaniem upragnionej niepodległości, w międzywojennej Polsce nastała wolność prasy, ogłoszona przez rząd Jędrzeja Moraczewskiego 20 grudnia 1918 r. Rzecz jasna, w tym przypadku możemy operować pojęciem wolności wyłącznie w kategoriach teoretycznych, gdyż już od 1919 władze próbowały wprowadzać różne formy kontroli nad publikowanymi treściami. Ta tendencja pogłębiała się przez lata, a jej kwintesencją był zamordystyczny przepis z kwietnia 1938 r. ustanawiający karę pozbawienia wolności do lat pięciu za pośmiertne znieważanie Józefa Piłsudskiego.

Wydaje się jednak, że prawo prasowe dotyczyło w przeważającej mierze treści politycznych, gdyż w latach 30. w Łodzi w najlepsze wydawano świerszczyka pt. „Adam i Ewa”.

Owszem, czasopismo miewało pewne problemy w tych mimo wszystko dość konserwatywnych czasach, ale jakoś udawało mu się przetrwać.


Wobec uchylenia wyrokiem sądowym
konfiskat „Adama i Ewy“ jesteśmy w moż-

Dziękuję za uwagę i zapraszam

re-research.pl borch@amu.edu.pl