

Cyfrowe narzędzia w przekładoznawstwie

Co to jest stylometria?

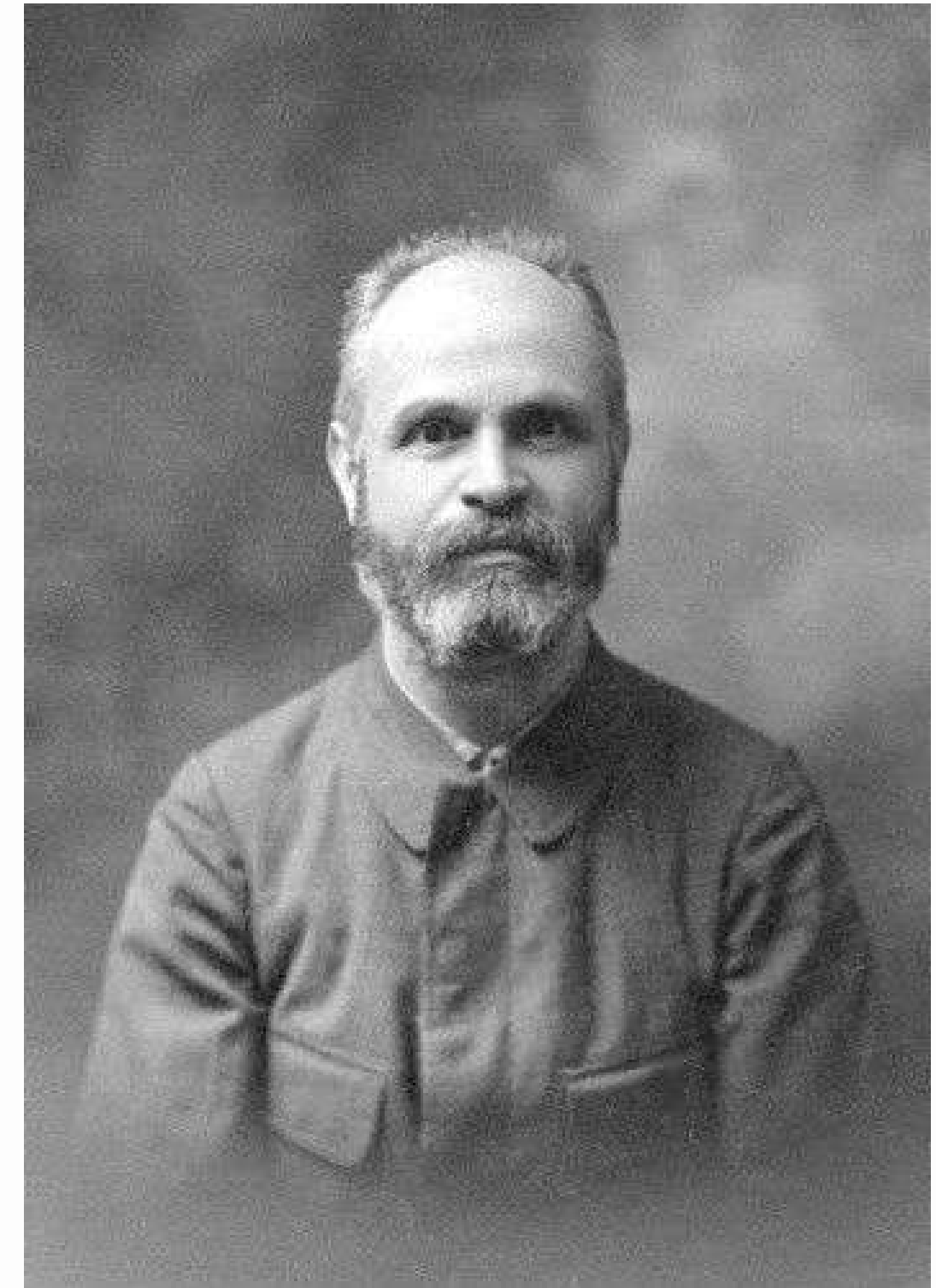
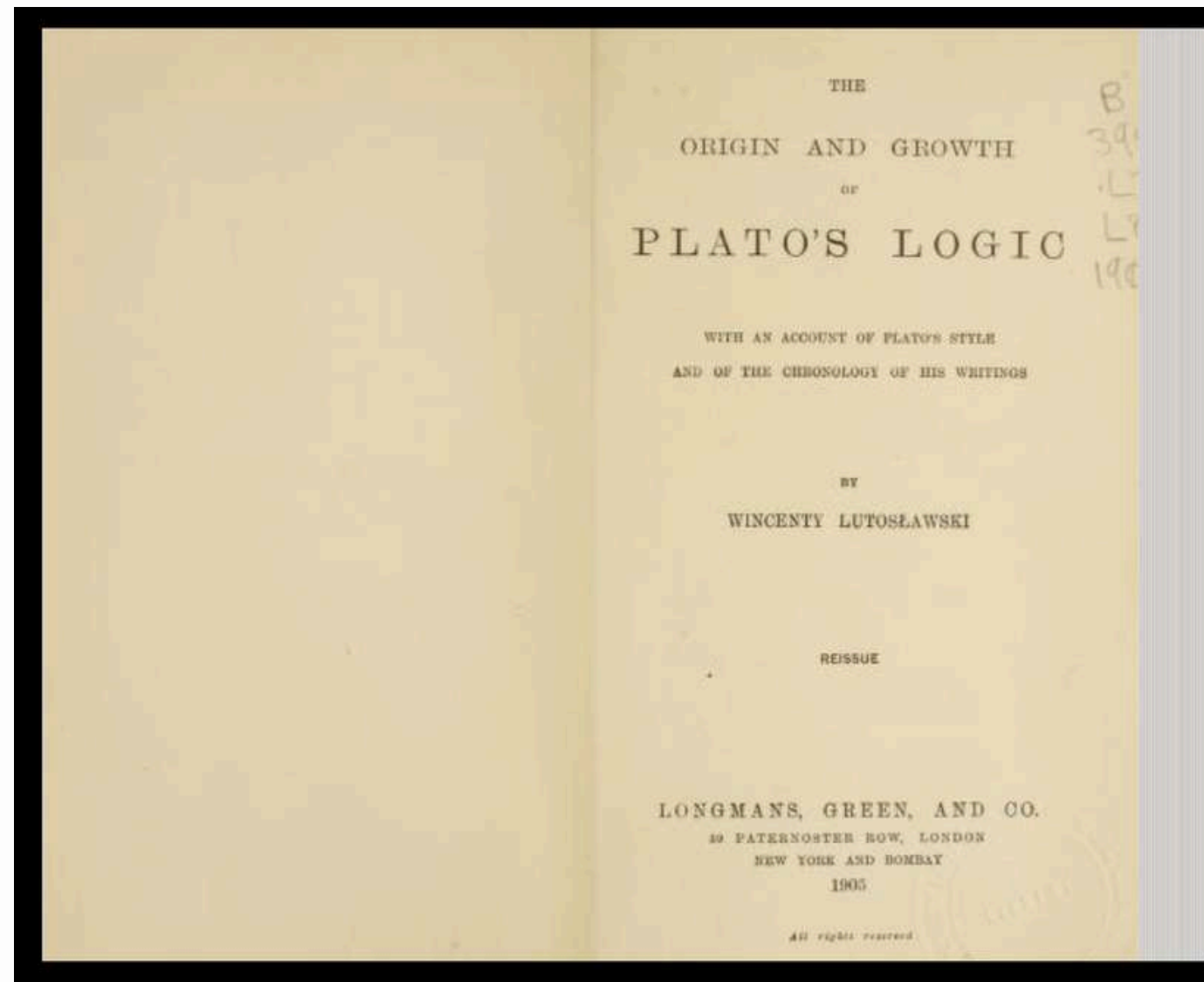
Sposób badania tekstów, w których **liczymy wszelkie jednostki językowe, które jesteśmy w stanie policzyć**

Metoda ilościowej analizy tekstów

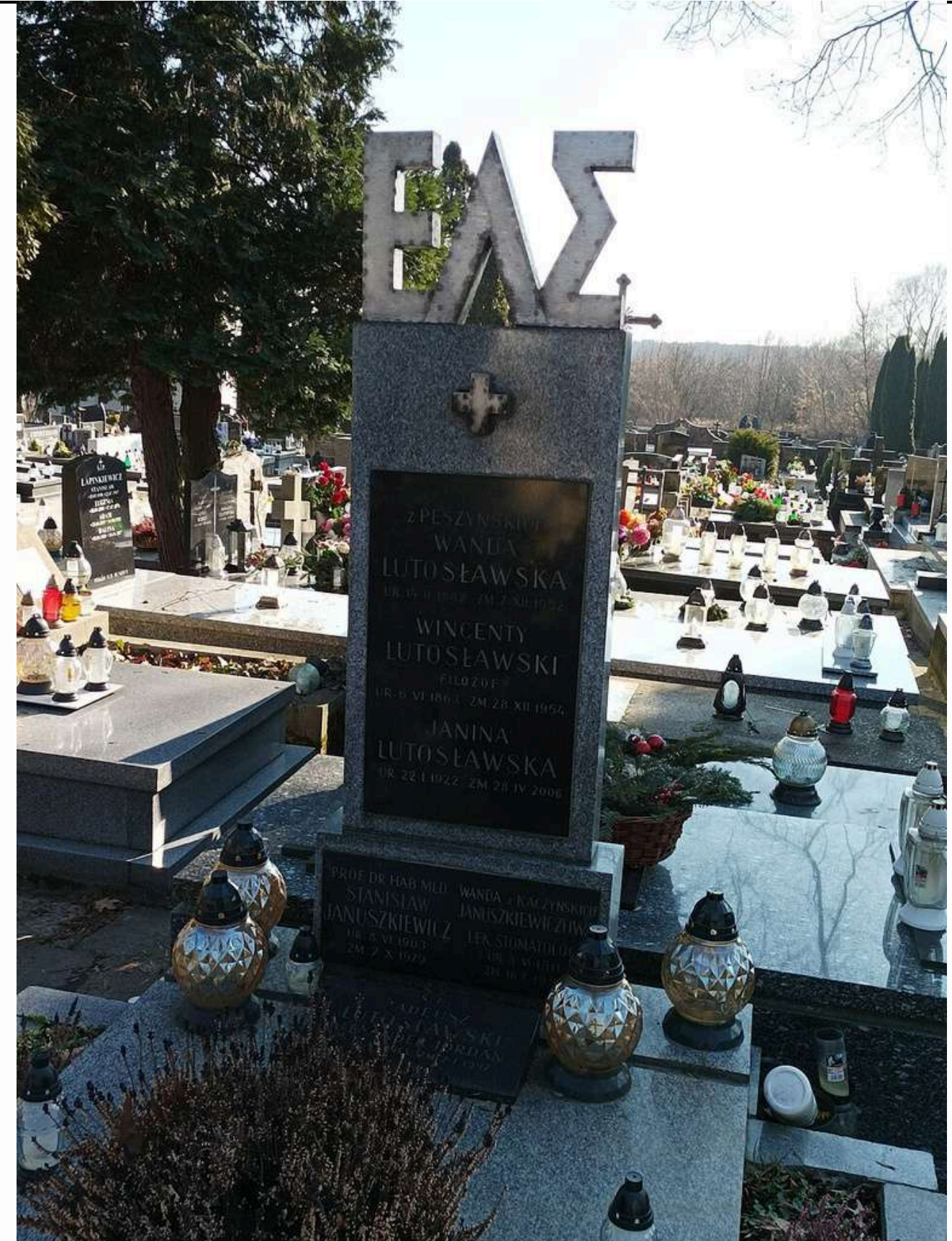
Czerpanie wiedzy niekoniecznie bezpośrednio z tekstów, a raczej z **obserwacji relacji między tekstami**

Metoda obserwacji **podobieństw i różnic** między tekstami

Trochę historii...



Trochę historii...



Distant reading (F. Moretti)

Wyjście „poza”
(pojedynczy)
tekst

Badanie „wszystkich”
a nie tylko
„reprezentatywnych”
tekstów

Pozyskiwanie i
analiza danych z
bardzo wielu,
różnych tekstów

Jeśli w ciągu 2 dni
jesteśmy w stanie
przeczytać 1
książkę, to przez 50
lat przeczytamy
tylko 9125 książek

Distant reading (F. Moretti)

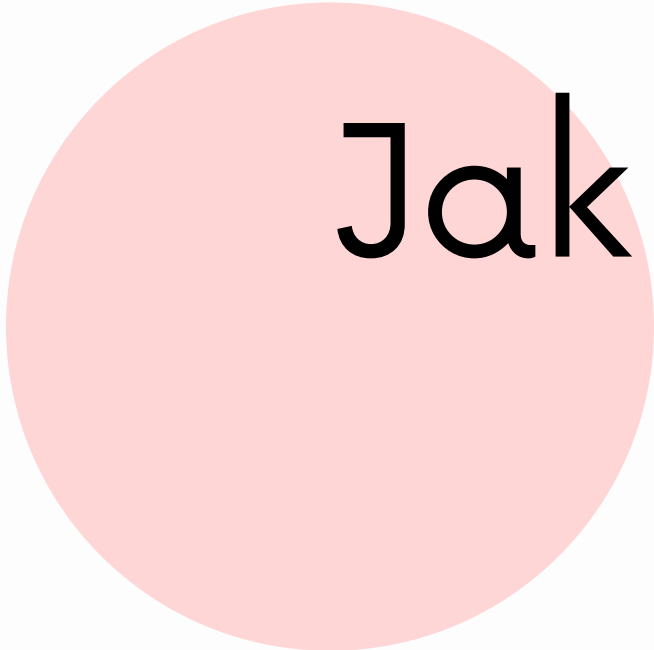
Wyjście „poza”
(pojedynczy)
tekst

Badanie „wszystkich”
a nie tylko
„reprezentatywnych”
tekstów

Pozyskiwanie i
analiza danych z
bardzo wielu,
różnych tekstów

Jeśli w ciągu 2 dni
jesteśmy w stanie
przeczytać 1
książkę, to przez 50
lat przeczytamy
tylko 9125 książek

W Polsce co roku wydawanych jest 3000-4000 nowych książek.



Jak więc czytać, żeby nie czytać?

Jak więc czytać, żeby nie czytać? Stylometrycznie

- Wystarczy policzyć wartości najczęściej występujących słów, żeby wykryć autorstwo tekstu



Jest prawdą powszechnie mało znaną, że w większości rozważań nad dziełami literackimi **zachowujemy się tak, jakby jednej trzeciej, dwóch piątych, połowy materiału – w ogóle nie było**. A tymczasem ta jedna trzecia, dwie piąte czy połowa materiału to odpowiednio dwadzieścia, trzydzieści i pięćdziesiąt najczęściej występujących słów. Co ciekawe, **są to zwykle te same słowa niezależnie od powieści czy autora: zaimki osobowe, czasowniki posiłkowe, parę przysłówków, spójników, przedimków...** Wszystkie one spełniają więcej niż jedną funkcję gramatyczną i zwykle plasują się – w dodatku na mniej więcej tej samej pozycji – wśród najczęstszych słów każdej powieści (Burrows 1987: 1)

The basic assumptions behind stylometric methods are that the **style is, to a large extent, transmitted by frequent, systematic elements of language, which are unconsciously applied**, with authors not controlling their use of these less meaning-making units, which in turn **allows for examining the frequency distributions of these features and combining them into authorial profiles**, that can be compared by means of traditional statistical methods as well as contemporary machine learning algorithms (Byszuk 2024).

No ale jak to działa?

1. Bag-of-words

2. Lista
frekwencyjna
jakichś jednostek
(np. słów)

3. Zamiana
wartości
liczbowych na
wektory

4. Analiza
statystyczna
danych

Bag-of-words

Nie chodzi tutaj — u kaduka! — o herb ani o szeregi przodków podgolonych, z sarmackimi wąsami i przy karabelach — ani wydekoltowane prababki w fiokach. Ojciec i matka — otóż i cały rodowód, jak to jest u nas, w dziejach nowoczesnych ludzi bez wczoraj. Z konieczności wzmianka o jednym dziadku, z musu notatka o jednym jedynym pradziadku. Chcemy uszanować nasyconą do pełna duchem i upodobaniem semickim awersję ludzi nowoczesnych do obciążania sobie pamięci wiadomościami, w którym kościele czy na jakim cmentarzu dany dziadek spoczywa.

Bag-of-words

Nie chodzi tutaj — u kaduka! — o herb ani o szeregi przodków podgolonych, z sarmackimi wąsami i przy karabelach — ani wydekoltowane prababki w fiokach. Ojciec i matka — otóż i cały rodowód, jak to jest u nas, w dziejach nowoczesnych ludzi bez wczoraj. Z konieczności wzmianka o jednym dziadku, z musu notatka o jednym jedynym pradziadku. Chcemy uszanować nasyconą do pełna duchem i upodobaniem semickim awersję ludzi nowoczesnych do obciążania sobie pamięci wiadomościami, w którym kościele czy na jakim cmentarzu dany dziadek spoczywa.

Stefan Żeromski, *Przedwiośnie*

Bag-of-words

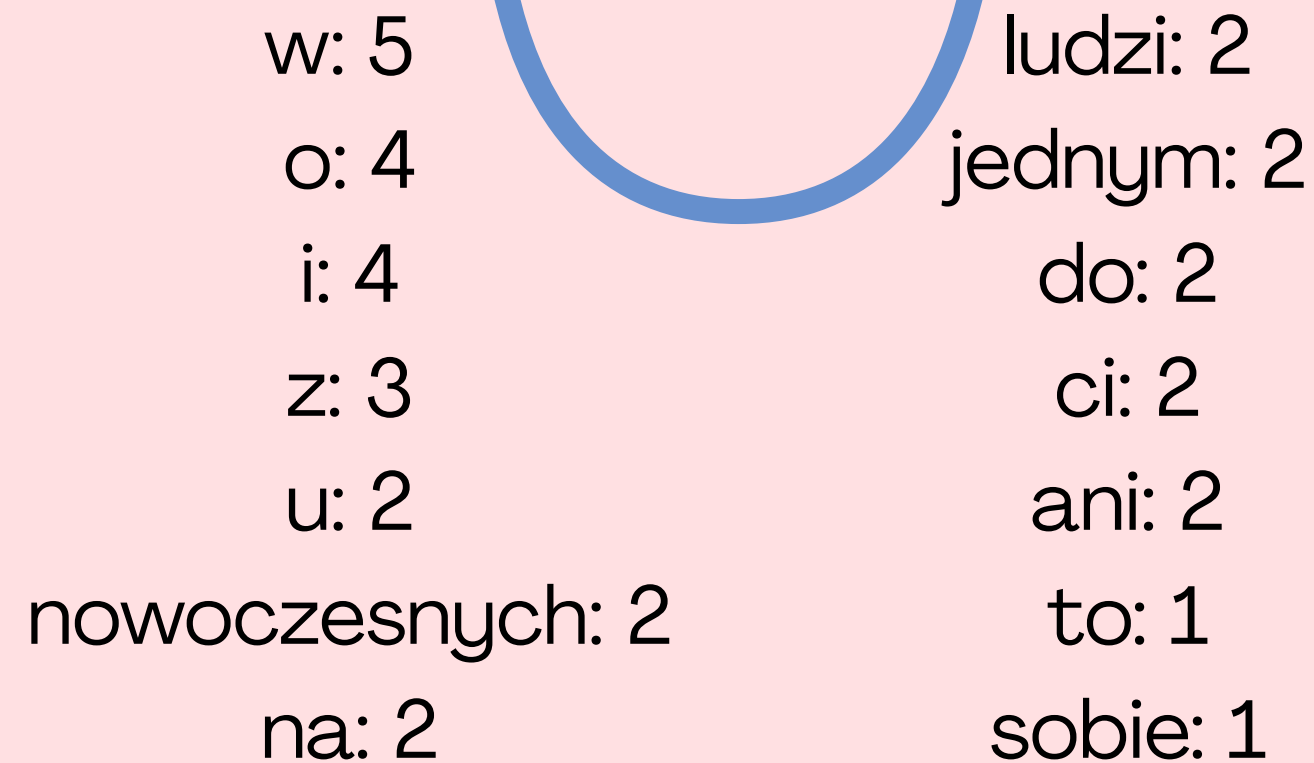
Nie chodzi tutaj —
sarmackimi wąsami
Ojciec i matka — otóż
bez wczoraj. Z koni
jedynym pradziadku
semickim awersję
w którym kościele



rodków podgolonych, z
ne prababki w fiokach.
ach nowoczesnych ludzi
nusu notatka o jednym
duchem i upodobaniem
amięci wiadomościami,
wa.

Bag-of-words

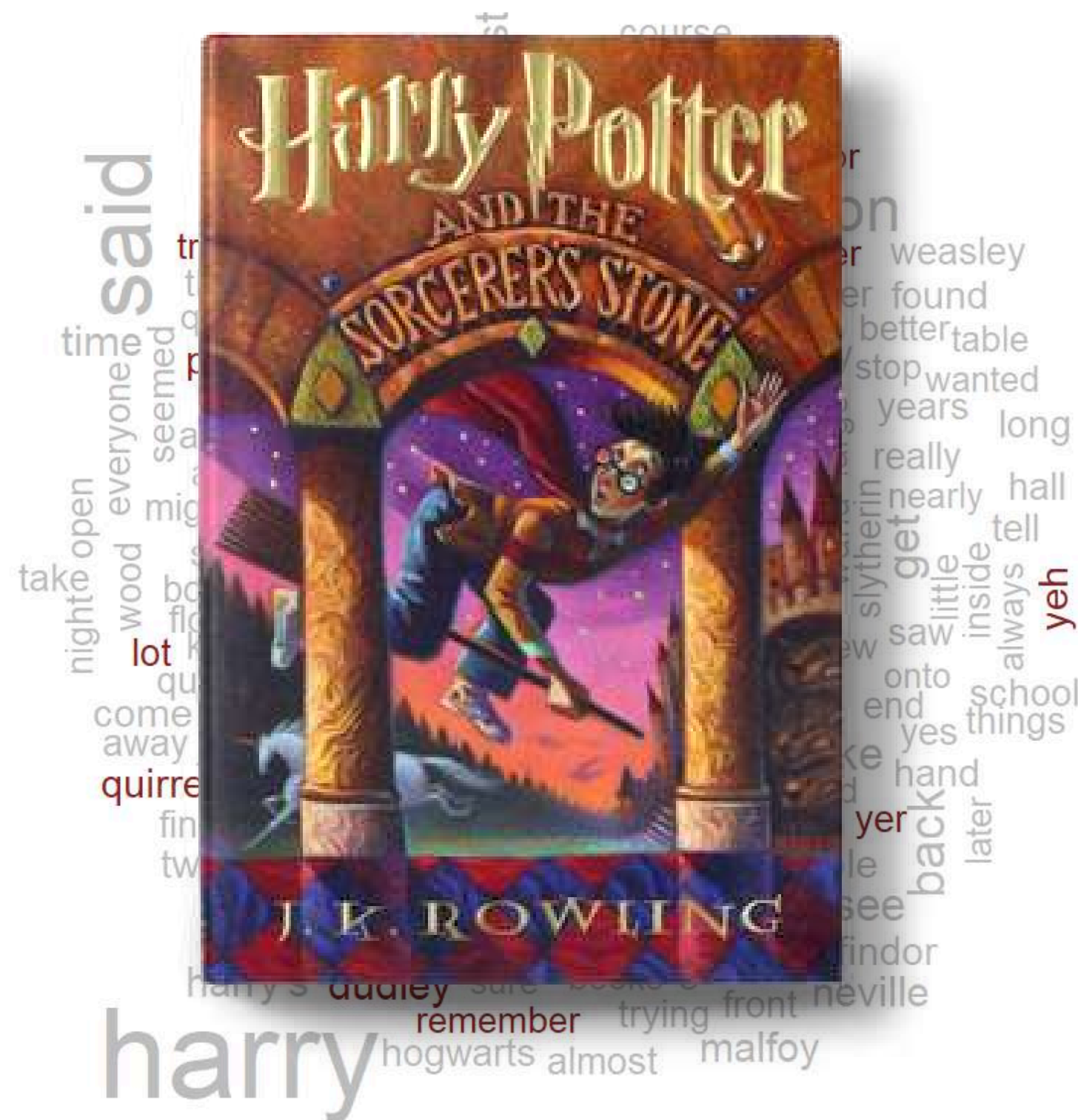
Nie chodzi tutaj —
sarmackimi wąsami
Ojciec i matka — otó
bez wczoraj. Z koni
jedynym pradziadku
semickim awersję l
w którym kościele c

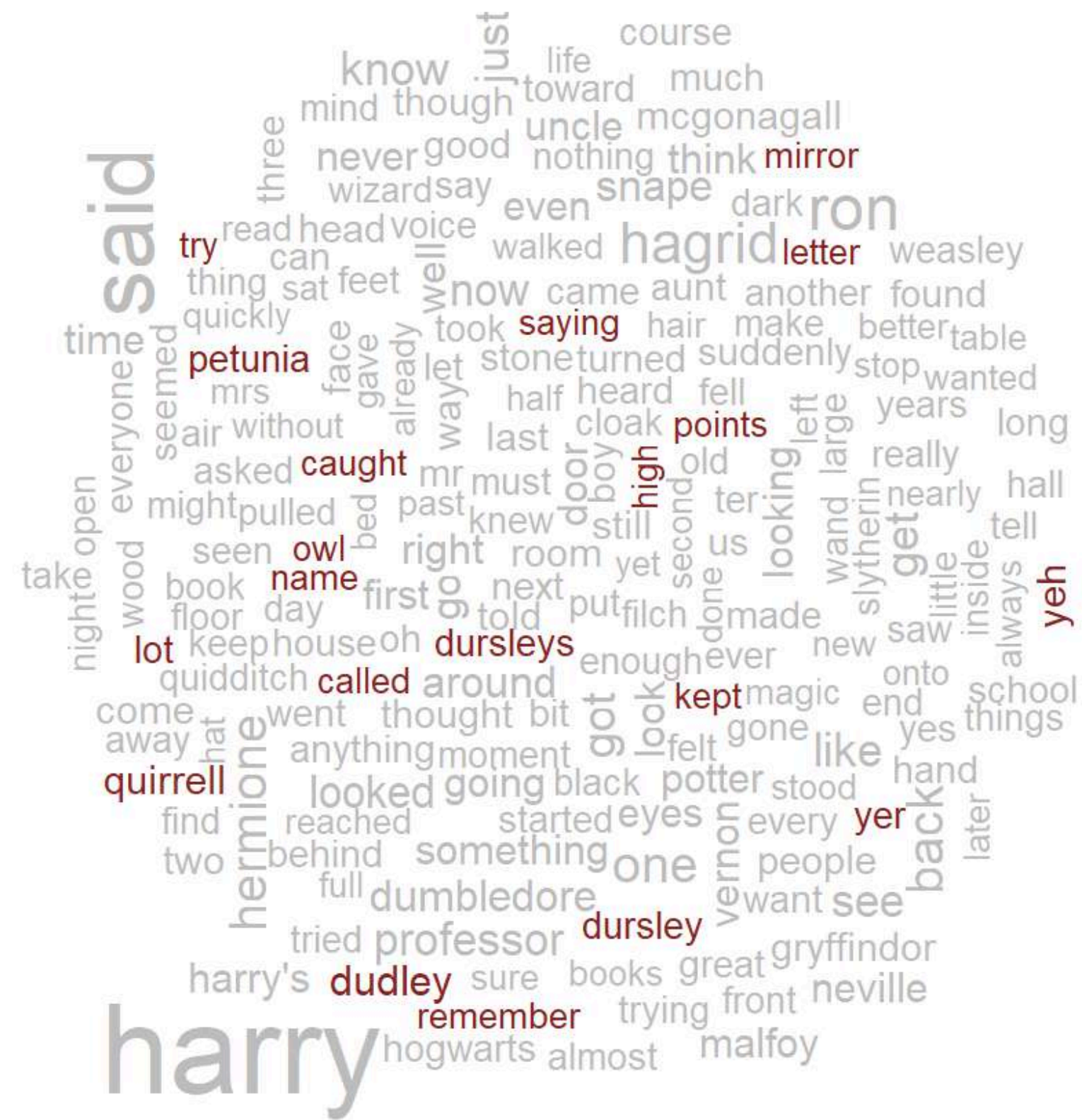


w: 5	ludzi: 2
o: 4	jednym: 2
i: 4	do: 2
z: 3	ci: 2
u: 2	ani: 2
nowoczesnych: 2	to: 1
na: 2	sobie: 1

odków podgolonych, z
ne prababki w fiokach.
ach nowoczesnych ludzi
nusu notatka o jednym
dlichem i upodobaniem
amięci wiadomościami,
/wa.

Czy możemy z tej metody coś wywnioskować?

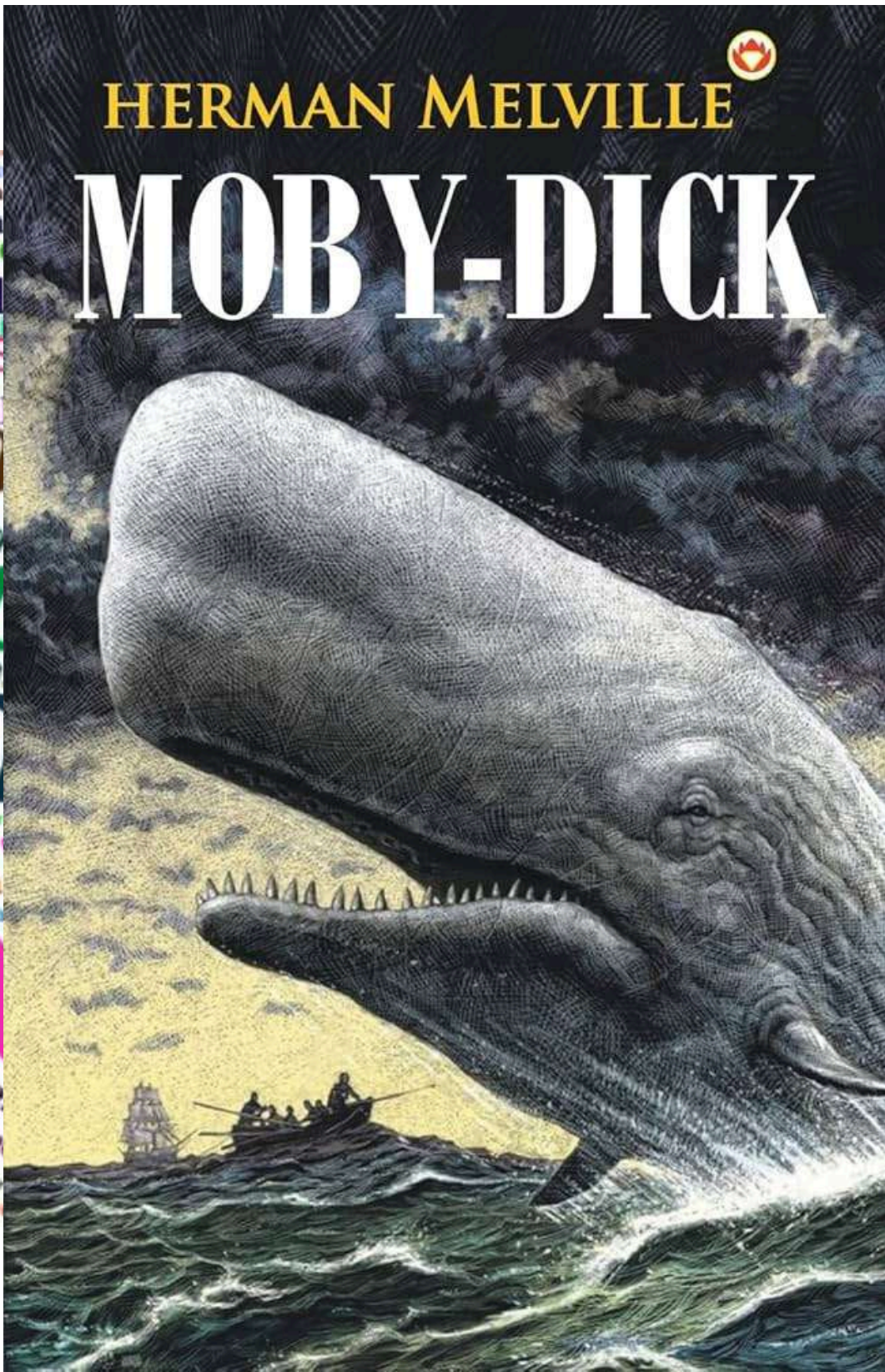






HERMAN MELVILLE

MOBY-DICK

A dramatic illustration of the whale Moby-Dick breaching the ocean's surface. The whale's massive, textured head and open mouth, filled with sharp teeth, dominate the right side of the frame. In the lower-left background, a small wooden boat with several crew members is visible on the dark, choppy water, looking up at the whale. A larger sailing ship is also visible in the distance. The sky is dark and stormy, with a small red circular logo in the top right corner.



Jak zamienić wartości liczbowe
wystąpień najczęstszych słów w
miarę podobieństwa tekstów?

- czyli trochę statystyki, ale przystępnej
-

Wróćmy do Johna Burrowsa i jego delty

$$\Delta = \sum_{i=1}^n \frac{|z(x_i) - z(y_i)|}{n}$$



text 1



text 2



$\Delta (T1, T2)$

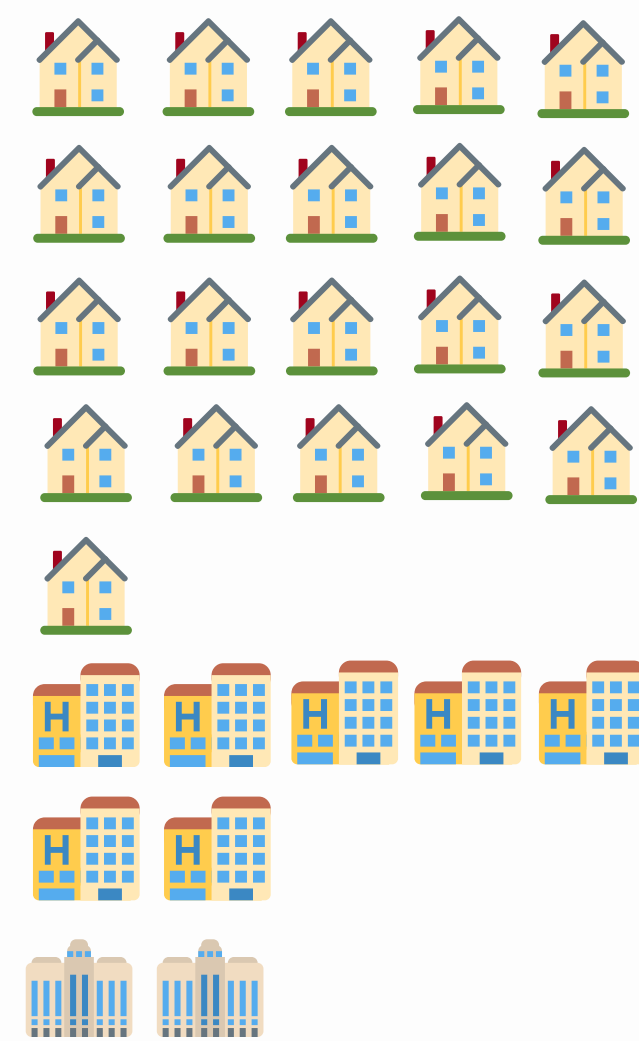
text 1



text 2



$$\Delta (T1, T2)$$



text 1



text 2

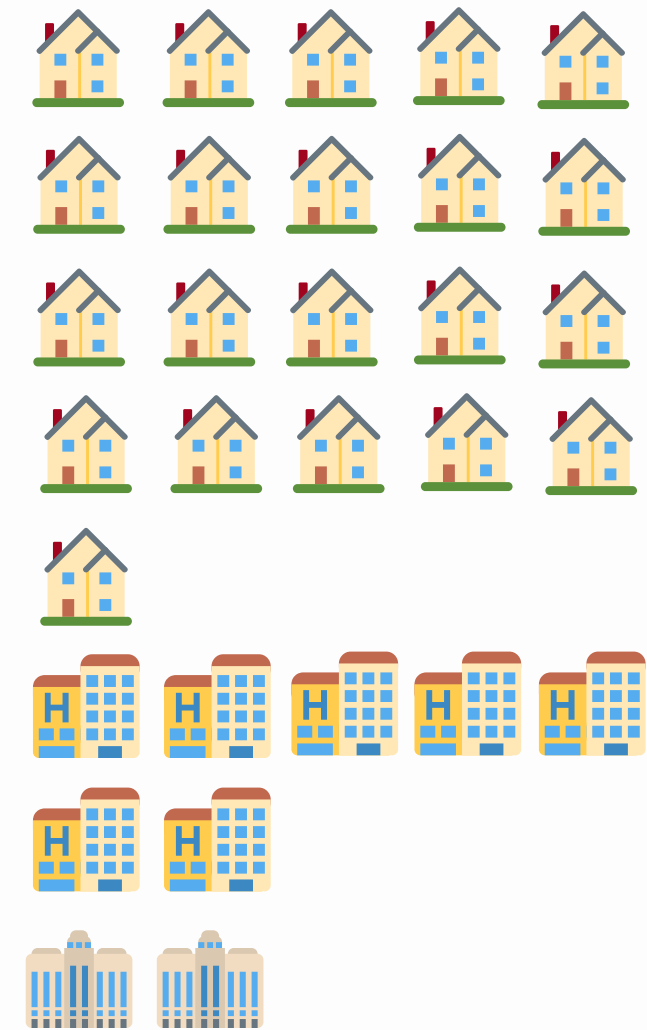


$\Delta (T1, T2)$

T1(6, 15, 10)



T1(21, 7, 2)



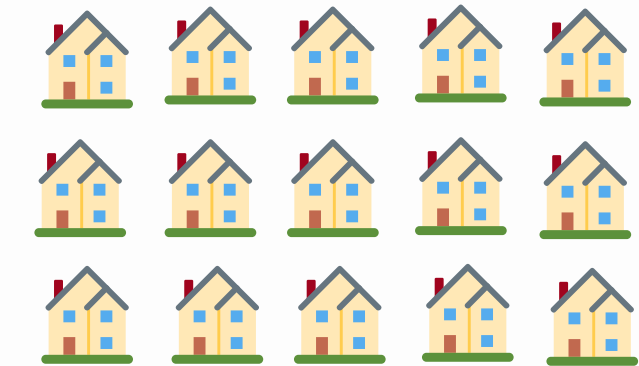
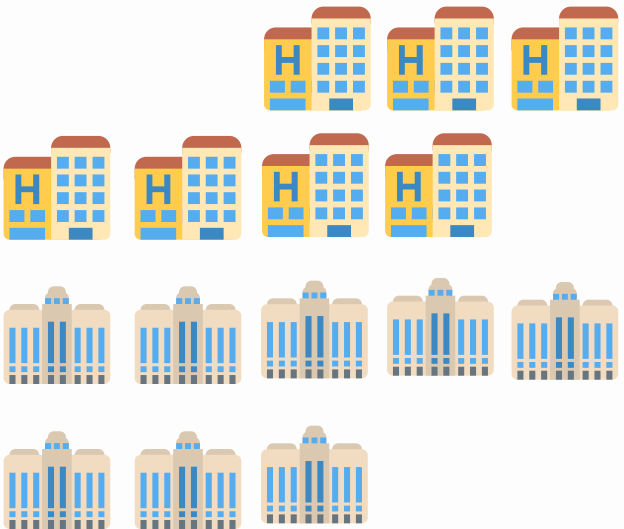
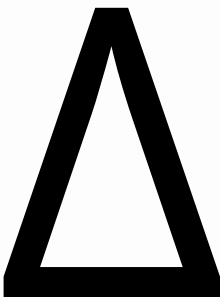
text 1



text 2



$$\Delta (T1, T2)$$



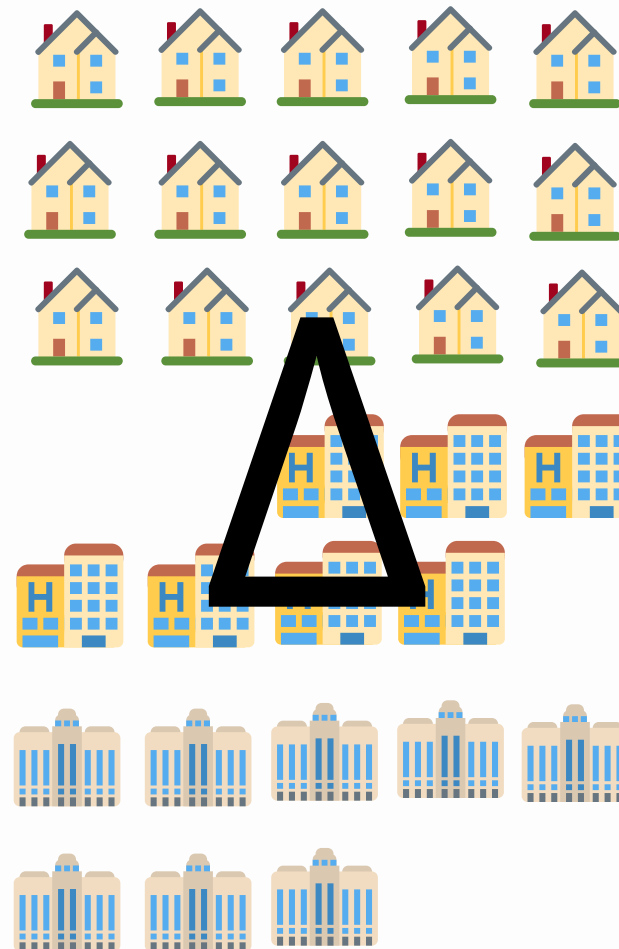
text 1



text 2



$\Delta (T1, T2)$



$\Delta (T1, T2) = (15, 7, 8)$

$$\Delta(T1, T2) = 15 + 7 + 8 = 30$$

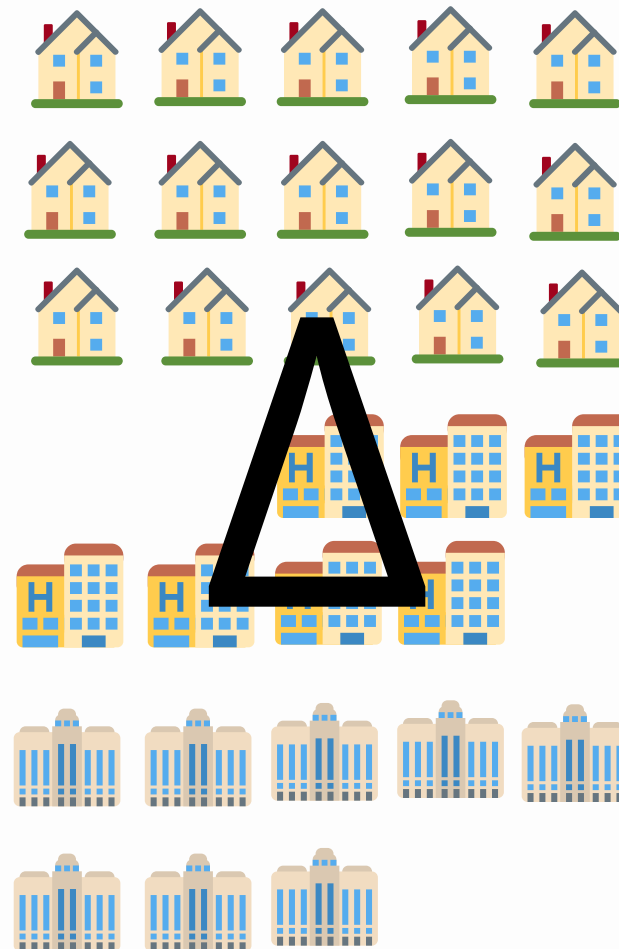
text 1



text 2



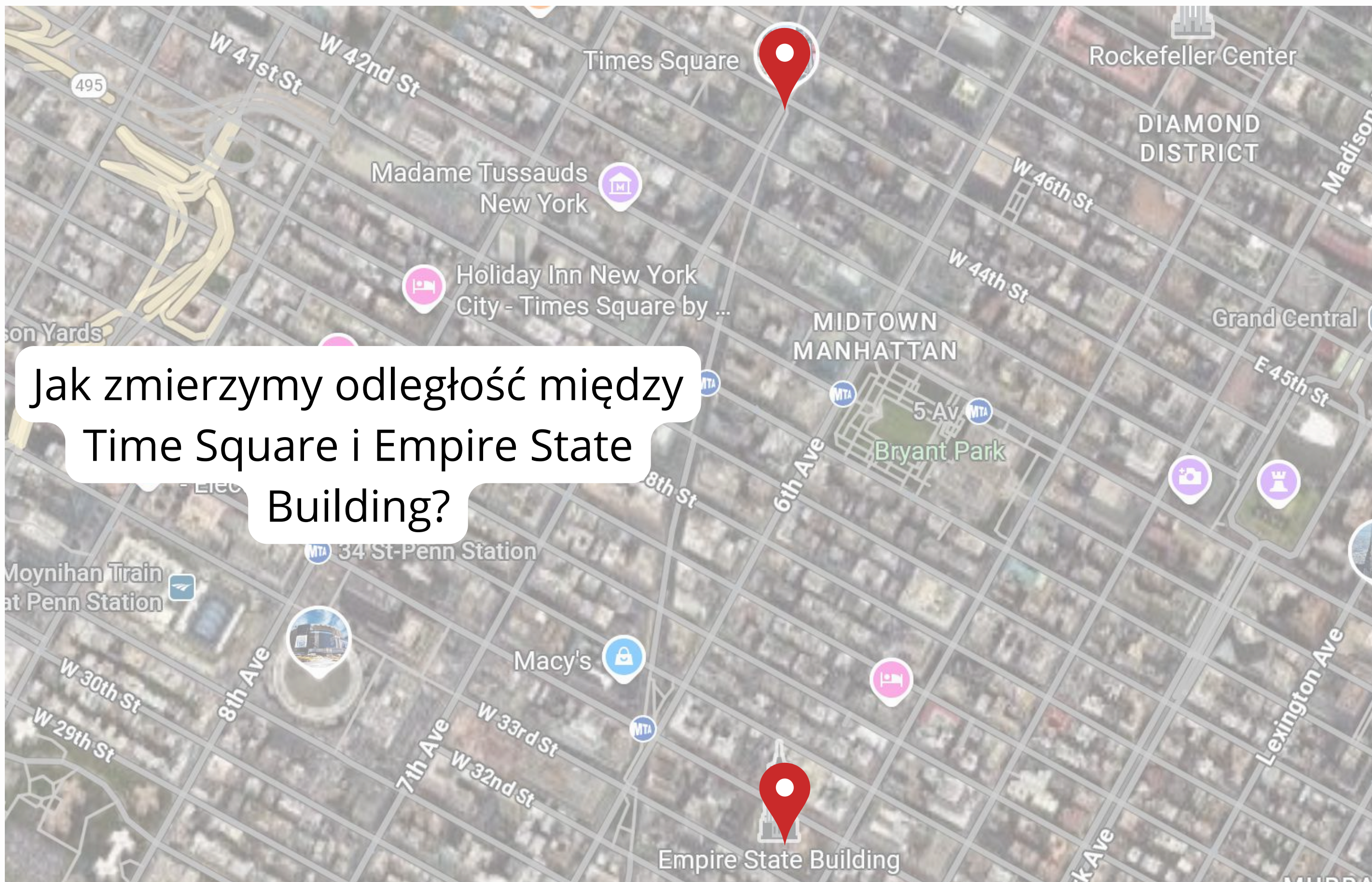
$\Delta (T1, T2)$

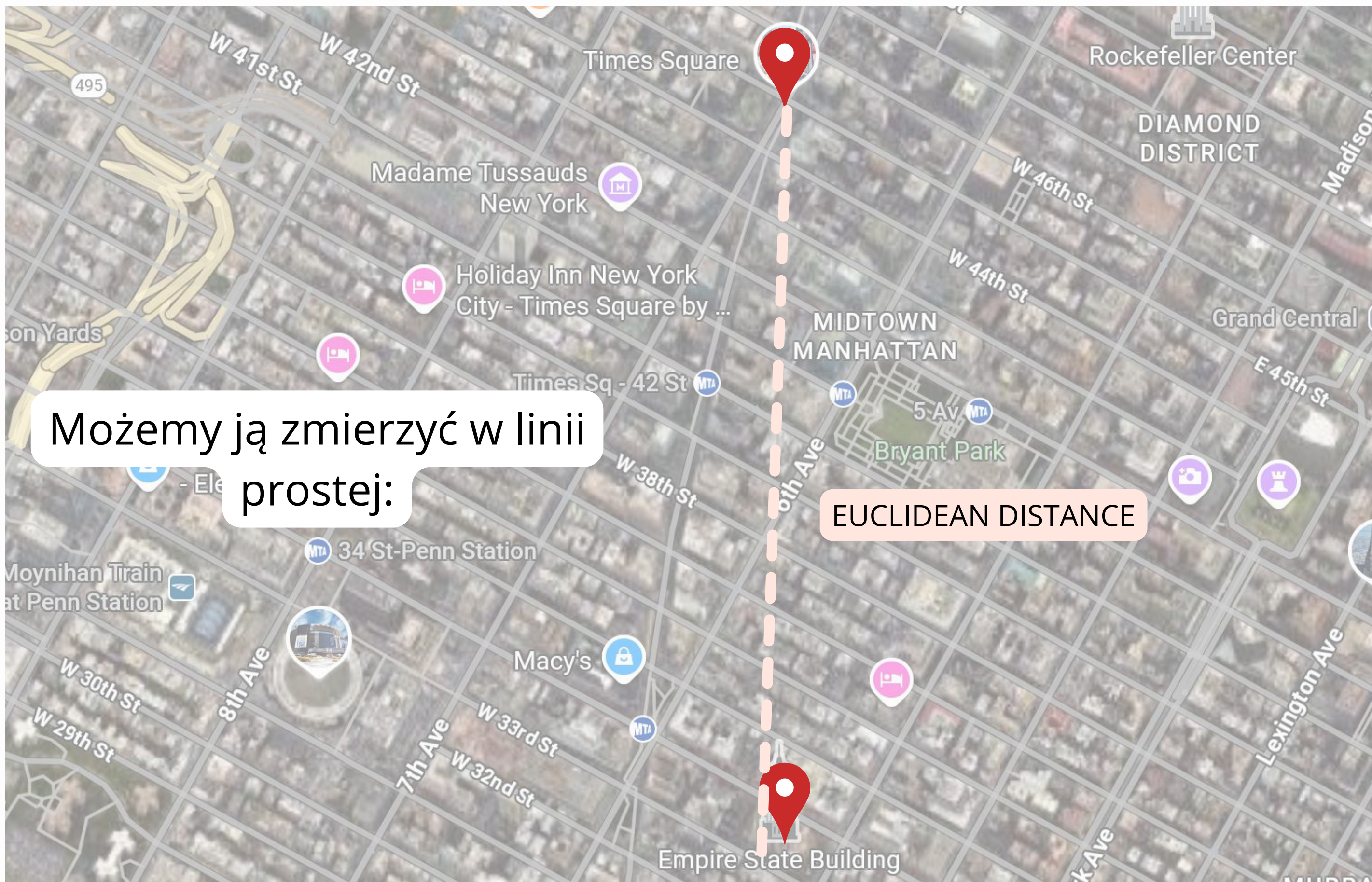


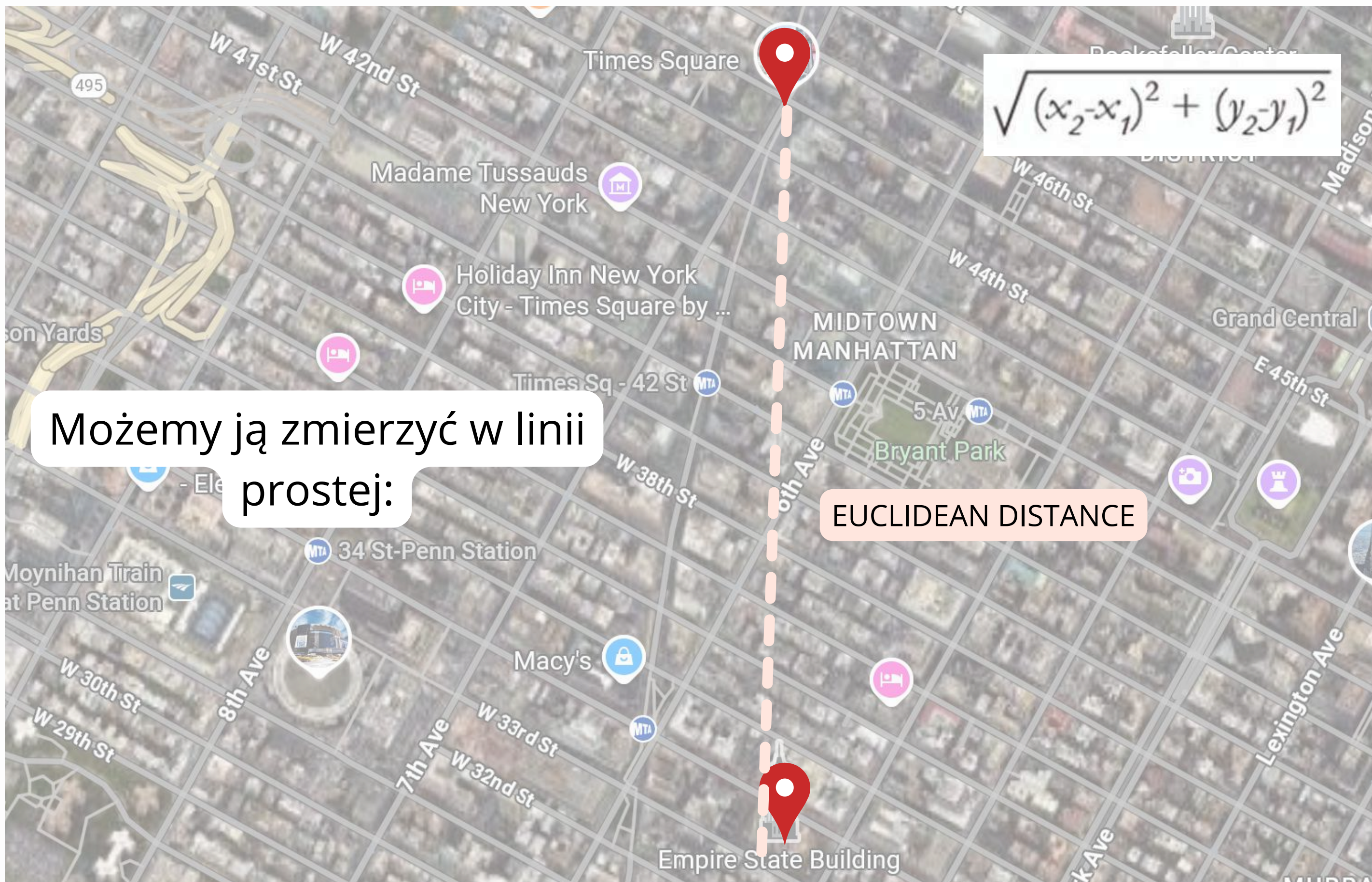
$$\Delta (T1, T2) = (15, 7, 8)$$

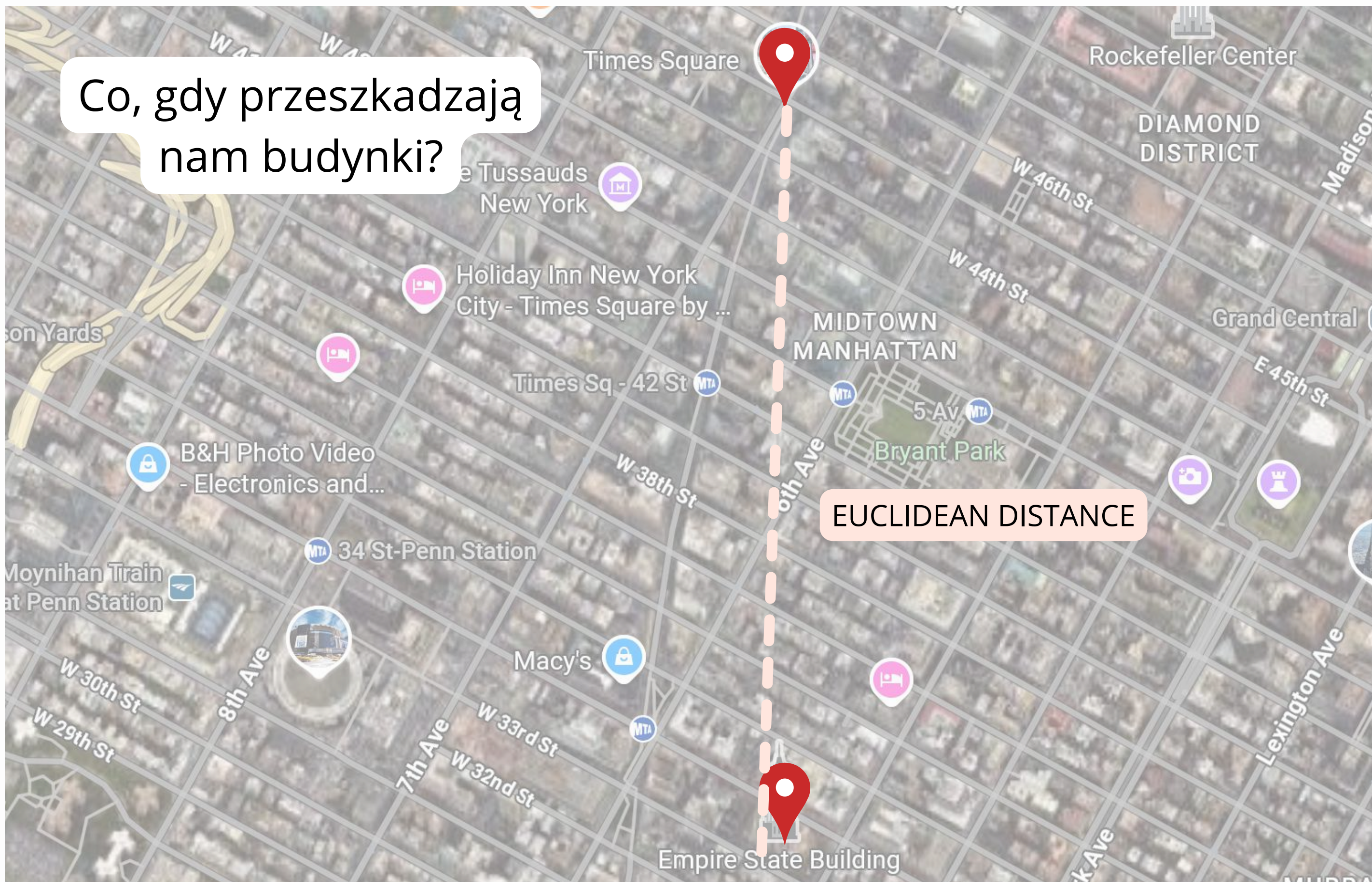
$$\Delta (T1, T2) = 15 + 7 + 8 = 30$$

tzw. Manhattan distance (city block distance), z małymi poprawkami Burrowsa



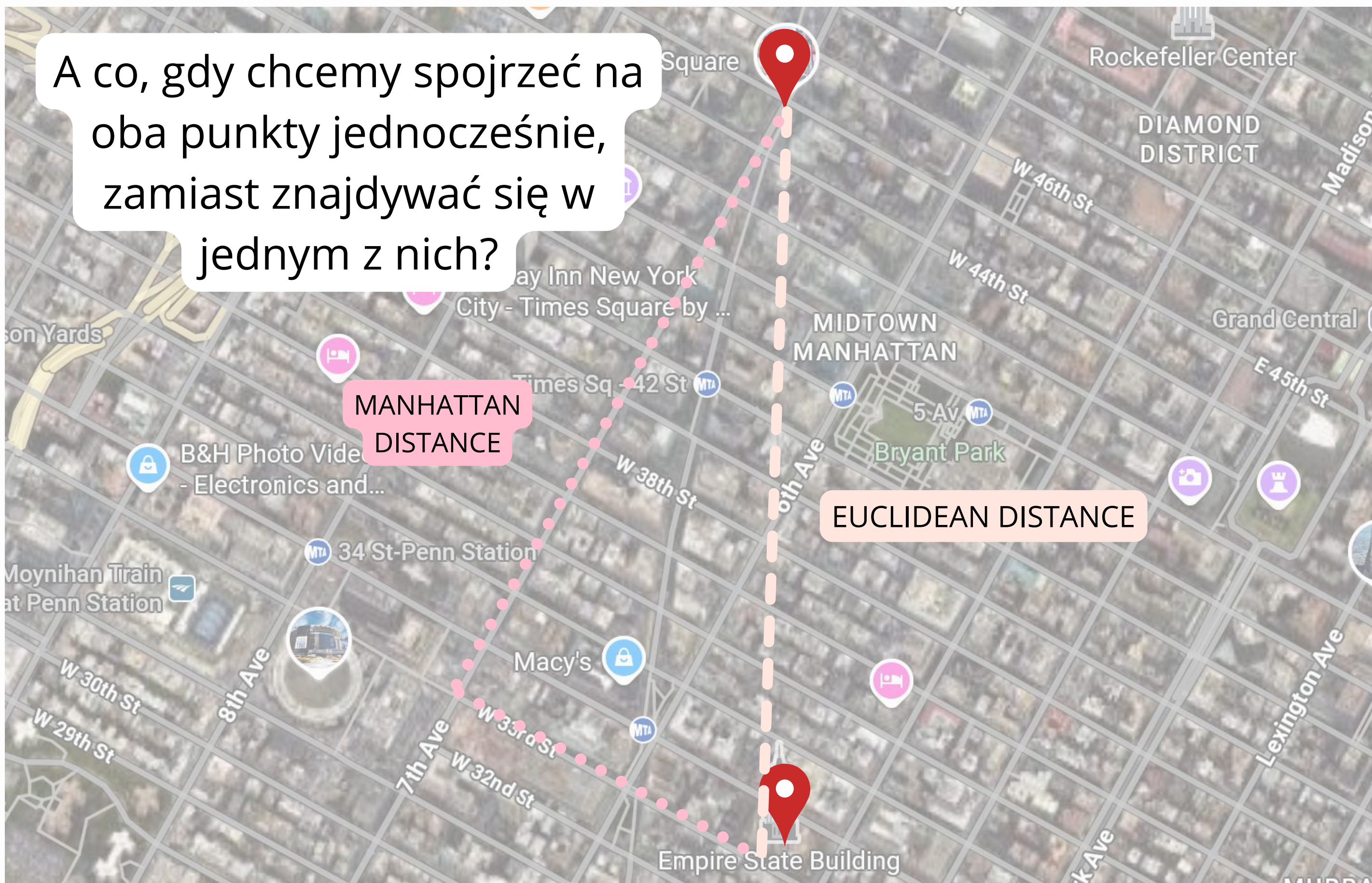








A co, gdy chcemy spojrzeć na oba punkty jednocześnie, zamiast znajdować się w jednym z nich?





A co, gdy chcemy spojrzeć na oba punkty jednocześnie, zamiast znajdować się w jednym z nich?



$$\cos \alpha = \frac{\sum_{i=1}^{n_s} x_i y_i}{\sqrt{(\sum_{i=1}^{n_s} x_i^2)} \sqrt{(\sum_{i=1}^{n_s} y_i^2)}},$$

where $x = z(T)$ and $y = z(T_1)$

$$\text{and } z(T) = \frac{f_s(T) - \mu_s}{\sigma_s}$$

Dobrze, ale jak to przełożyć na badania tekstów, literatury?

- frekwencję wystąpień każdego ze słów możemy zamienić na **wektory**
 - wektory słów umieszczamy w **wielowymiarowej przestrzeni**
 - przestrzeń ma tyle wymiarów, ile mamy wektorów słów dla danego tekstu
 - między wektorami tworzą się **kąty**, których miary możemy analizować
 - w obrębie tekstu wykonujemy dokładnie te same czynności, co przed chwilą na mapie Nowego Jorku
-

	<u>Agnes Grey</u>	Pride and prejudice	Jane Eyre	David Copperfield	The Mill on the Floss	Tom Jones
the	2511	4330	7835	13693	8690	16690
and	2733	3577	6618	12275	6182	8652
to	2366	4136	5152	10456	6082	11425
of	1602	3609	4359	8683	5042	10467
I	2204	2064	7165	13185	2760	6382
a	1296	1948	4467	7879	4806	6785
in	911	1866	2762	6218	3101	5680
that	776	1577	1655	5233	2946	4037
he	659	1338	1902	3526	2289	5034
was	1000	1847	2525	5314	3045	3997
it	795	1532	2403	4749	2401	3343



Co nam mówi ta tabelka?

	Agnes Grey	Pride and prejudice	Jane Eyre	David Copperfield	The Mill on the Floss	Tom Jones
the	2511	4330	7835	13693	8690	16690
and	2733	3577	6618	12275	6182	8652
to	2366	4136	5152	10456	6082	11425
of	1602	3609	4359	8683	5042	10467
I	2204	2064	7165	13185	2760	6382
a	1296	1948	4467	7879	4806	6785
in	911	1866	2762	6218	3101	5680
that	776	1577	1655	5233	2946	4037
he	659	1338	1902	3526	2289	5034
was	1000	1847	2525	5314	3045	3997
it	795	1532	2403	4749	2401	3343



Co nam mówi ta tabelka?

- w sumie to niewiele
-



Co nam mówi ta tabelka?

- w sumie to niewiele
- może jedynie, że *Tom Jones* jest prawdopodobnie czterokrotnie dłuższą książką niż *Duma i uprzedzenie*



Co nam mówi ta tabelka?

- w sumie to niewiele
 - może jedynie, że *Tom Jones* jest prawdopodobnie czterokrotnie dłuższą książką niż *Duma i uprzedzenie*
 - albo że Fielding pisał dość mało zdań współrzędnie złożonych
-

	Agnes Grey	Pride and prejudice	Jane Eyre	David Copperfield	The Mill on the Floss	Tom Jones
Tenant...	0.81	1.07	0.88	0.92	0.98	1.16
Emma	1.12	0.78	1.28	1.15	1.2	1.25
Sense and sensibility	1.14	0.69	1.24	1.16	1.25	1.13
The Professor	1.06	1.21	0.69	0.94	1	1.27
Villette	1.07	1.26	0.65	0.91	0.96	1.28
Bleak House	1.09	1.18	0.92	0.55	0.87	1.21
Hard Times	1.16	1.25	0.96	0.65	0.91	1.26
Wuthering Heights	1.13	1.31	0.81	0.94	1.01	1.32
Middlemarch	1.01	1.1	0.99	0.87	0.65	1.17
Adam Bede	1.2	1.37	0.95	0.9	0.66	1.42
Joseph Andrews	1.15	1.19	1.24	1.18	1.29	0.64



Co nam mówi ta tabelka?

- w sumie to niewiele
-

	Agnes Grey	Pride and prejudice	Jane Eyre	David Copperfield	The Mill on the Floss	Tom Jones
Tenant...	0.81	1.07	0.88	0.92	0.98	1.16
Emma	1.12	0.78	1.28	1.15	1.2	1.25
Sense and sensibility	1.14	0.69	1.24	1.16	1.25	1.13
The Professor	1.06	1.21	0.69	0.94	1	1.27
Villette	1.07	1.26	0.65	0.91	0.96	1.28
Bleak House	1.09	1.18	0.92	0.55	0.87	1.21
Hard Times	1.16	1.25	0.96	0.65	0.91	1.26
Wuthering Heights	1.13	1.31	0.81	0.94	1.01	1.32
Middlemarch	1.01	1.1	0.99	0.87	0.65	1.17
Adam Bede	1.2	1.37	0.95	0.9	0.66	1.42
Joseph Andrews	1.15	1.19	1.24	1.18	1.29	0.64



Co nam mówi ta tabelka?

- w sumie to już wiele



Co nam mówi ta tabelka?

- w sumie to już wiele
 - możemy wyznaczyć, które teksty są do siebie podobne stylistycznie i najprawdopodobniej zostały napisane przez tę samą osobę
-



Co nam mówi ta tabelka?

- w sumie to już wiele
 - możemy wyznaczyć, które teksty są do siebie podobne stylistycznie i najprawdopodobniej zostały napisane przez tę samą osobę
 - możemy ustalić relacje rodzinne między autorami, którzy są spokrewnieni (z czego może wynikać podobieństwo stylistyczne np. sióstr Brontë?)
-

Co nam mówi ta tabelka?

- w sumie to już wiele
 - możemy wyznaczyć, które teksty są do siebie podobne stylistycznie i najprawdopodobniej zostały napisane przez tę samą osobę
 - możemy ustalić relacje rodzinne między autorami, którzy są spokrewnieni (z czego może wynikać podobieństwo stylistyczne np. sióstr Brontë?)
 - możemy ustalić podobieństwo gatunkowe, które jest mniej wyraźne niż sygnał autorski, ale jest dalej widoczne
-

	Agnes Grey	Pride and prejudice	Jane Eyre	David Copperfield	The Mill on the Floss	Tom Jones
Tenant...	0.81	1.07	0.88	0.92	0.98	1.16
Emma	1.12	0.78	1.28	1.15	1.2	1.25
Sense and sensibility	1.14	0.69	1.24	1.16	1.25	1.13
The Professor	1.06	1.21	0.69	0.94	1	1.27
Villette	1.07	1.26	0.65	0.91	0.96	1.28
Bleak House	1.09	1.18	0.92	0.55	0.87	1.21
Hard Times	1.16	1.25	0.96	0.65	0.91	1.26
Wuthering Heights	1.13	1.31	0.81	0.94	1.01	1.32
Middlemarch	1.01	1.1	0.99	0.87	0.65	1.17
Adam Bede	1.2	1.37	0.95	0.9	0.66	1.42
Joseph Andrews	1.15	1.19	1.24	1.18	1.29	0.64



R jest oprogramowaniem darmowym i dostarczany jest BEZ JAKIEJKOLWIEK GWARANCJI.
Możesz go rozpowszechniać pod pewnymi warunkami.
Wpisz 'license()' lub 'licence()' aby uzyskać szczegóły dystrybucji.

R jest projektem kolaboracyjnym z wieloma uczestnikami.
Wpisz 'contributors()' aby uzyskać więcej informacji oraz
'citation()' aby dowiedzieć się jak cytować R lub pakiety R w publikacjach.

Wpisz 'demo()' aby zobaczyć demo, 'help()' aby uzyskać pomoc on-line, lub
'help.start()' aby uzyskać pomoc w przeglądarce HTML.
Wpisz 'q()' aby wyjść z R.

```
> library(stylo)
```

```
### stylo version: 0.7.5 ###
```


If you plan to cite this software (please do!), use the following reference:
Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R:
a package for computational text analysis. R Journal 8(1): 107-121.
<<https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>>

To get full BibTeX entry, type: citation("stylo")
> stylo()

	INPUT & LANGUAGE	FEATURES	STATISTICS	SAMPLING	OUTPUT
INPUT:	plain text <input checked="" type="radio"/>	xml <input type="radio"/>	xml (plays) <input type="radio"/>	xml (no titles) <input type="radio"/>	html <input type="radio"/>
LANGUAGE:	English <input type="radio"/>	English (contr.) <input type="radio"/>	English (ALL) <input checked="" type="radio"/>	Latin <input type="radio"/>	Latin (u/v > u) <input type="radio"/>
	Polish <input type="radio"/>	Hungarian <input type="radio"/>	French <input type="radio"/>	Italian <input type="radio"/>	Spanish <input type="radio"/>
	Dutch <input type="radio"/>	German <input type="radio"/>	CJK <input type="radio"/>	Other <input type="radio"/>	Native encoding <input type="checkbox"/>
OK					

Ważne informacje, pojęcia przed wykonaniem analiz

- pracujemy na „surowych” plikach (***raw files***), najlepiej w formacie **txt**
 - sprawdzamy sposób kodowania plików, żeby się upewnić, że analiza przebiegnie poprawnie - sprawdzamy, czy jest format **UTF-8**
 - pracujemy na **niezlematyzowanych słowach**, ale możemy wykonywać analizę **n-gramów** poszczególnych znaków (przydatne, gdyby ktoś chciał badać np. chiński)
-



Jak dowiedzieć się w jakim formacie tekstowym zapisujemy plik?

Formaty tekstowe plików:

TXT	RTF	DOC/DOCX	ODT	XML	CSS	HTML	WPS
-----	-----	----------	-----	-----	-----	------	-----

Jak dowiedzieć się w jakim formacie tekstowym zapisujemy plik?

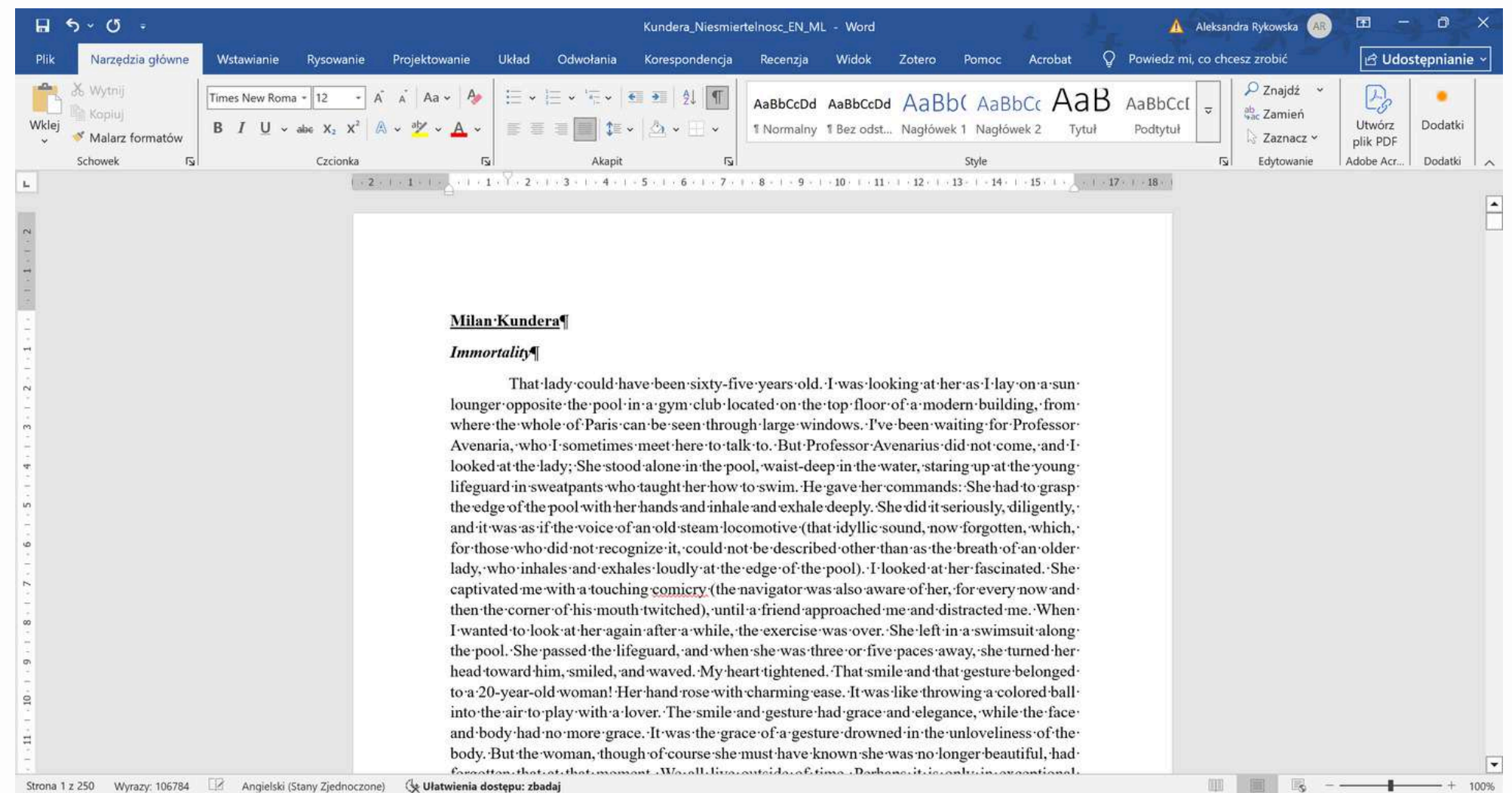
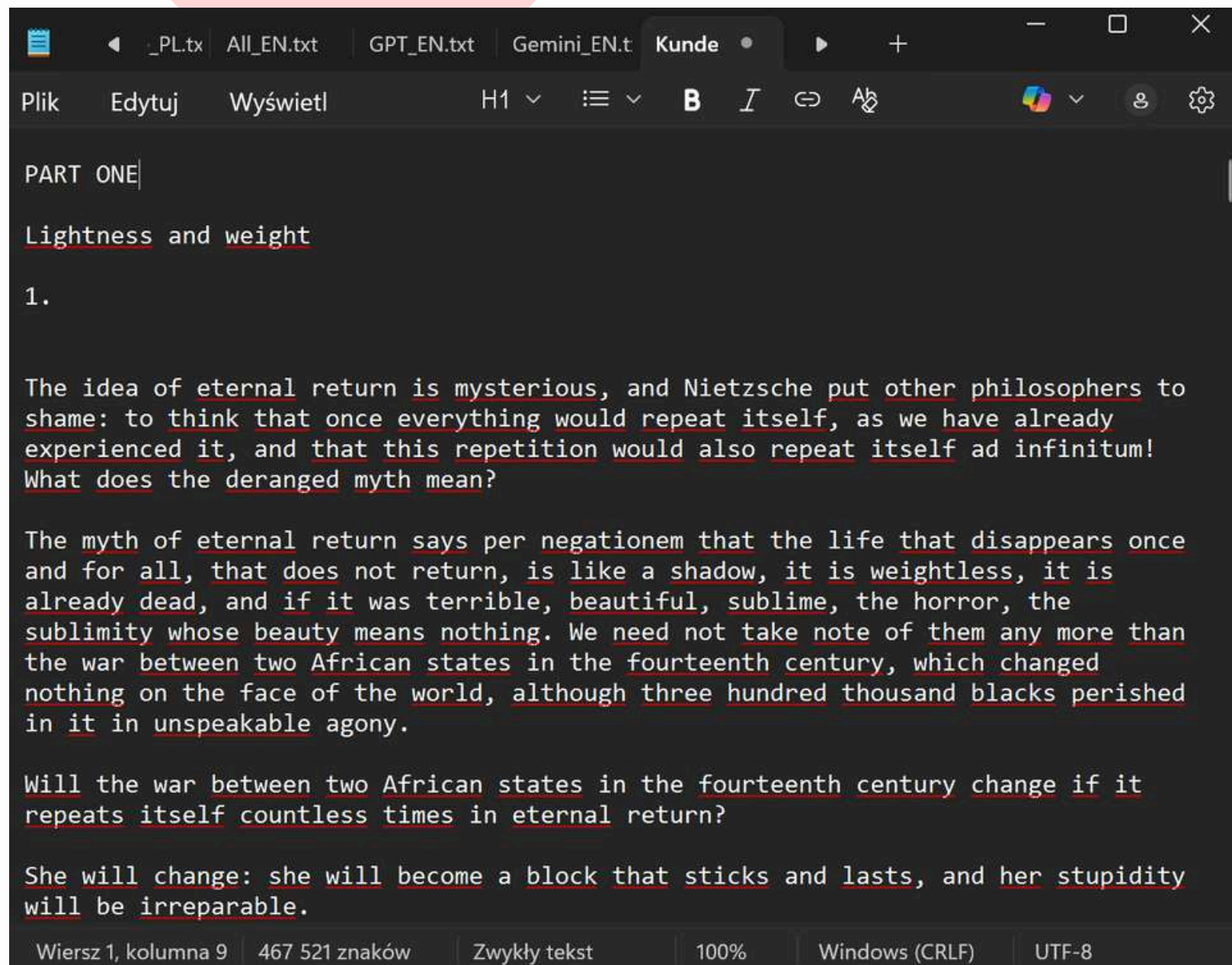
Formaty tekstowe plików:

TXT	RTF	DOC/DOCX	ODT	XML	CSS	HTML	WPS
-----	-----	----------	-----	-----	-----	------	-----

Formaty akceptowane przez *stylo*:

INPUT:	plain text	xml	xml (plays)	xml (no titles)	html
	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Jak dowiedzieć się w jakim formacie tekstowym zapisujemy plik?



Jak dowiedzieć się w jakim formacie tekstowym zapisujemy plik?



Kundera_Niesmiertelnosc_MT - w
corpus




Kundera_Niesmiertelnosc_MT - w
nowe gt

Jak dowiedzieć się w jakim formacie tekstowym zapisujemy plik?

Nazwa	Data modyfikacji	Typ	Rozmiar
Kundera_Niesmiertelnosc_MT	09.07.2025 08:51	Dokument programu Microsoft Word	345 KB
Lackberg_Kaznodzieja_MT	09.07.2025 08:52	Dokument programu Microsoft Word	436 KB
Lackberg_SrebrneSkrzydla_MT	09.07.2025 08:52	Dokument programu Microsoft Word	312 KB
Lackberg_ZlotaKlatka_MT	09.07.2025 08:52	Dokument programu Microsoft Word	431 KB
Lagerlof_CudownaPodroz_MT	09.07.2025 08:52	Dokument programu Microsoft Word	110 KB
Lagerlof_GostaBerling_MT	09.07.2025 08:53	Dokument programu Microsoft Word	498 KB
Lagerlof_TetniaceSerce_MT	09.07.2025 08:53	Dokument programu Microsoft Word	216 KB
Link_Oszukana_MT	09.07.2025 08:53	Dokument programu Microsoft Word	620 KB
Link_Poszukiwanie_MT	09.07.2025 08:53	Dokument programu Microsoft Word	691 KB
Lispector_GodzinaGwiazdy_MT	09.07.2025 08:53	Dokument programu Microsoft Word	116 KB
Lispector_PasjaWedlugGH_MT	09.07.2025 08:54	Dokument programu Microsoft Word	148 KB
Lispector_WPoblizuDzikiegoSerca_MT	09.07.2025 08:54	Dokument programu Microsoft Word	180 KB
Llosa_BurzliweCzasy_MT	09.07.2025 08:54	Dokument programu Microsoft Word	303 KB
Llosa_CiotkaJuliaSkryba_MT	09.07.2025 08:54	Dokument programu Microsoft Word	407 KB
Llosa_DzielnicaWystepku_MT	09.07.2025 08:54	Dokument programu Microsoft Word	246 KB
Llosa_SzelmostwaNiegrzecznejDziewczynki_MT	09.07.2025 08:55	Dokument programu Microsoft Word	366 KB
Llosa_ZielonyDom_MT	09.07.2025 08:55	Dokument programu Microsoft Word	450 KB
Mankell_Chinczyk_MT	09.07.2025 08:55	Dokument programu Microsoft Word	596 KB
Mankell_FalszywyTrop_MT	09.07.2025 08:56	Dokument programu Microsoft Word	566 KB
Mankell_MordercaBezTwarzy_MT	09.07.2025 08:56	Dokument programu Microsoft Word	369 KB


Nazwa	Data modyfikacji	Typ	Rozmiar
Koch_DomLetniZBasenem_FR_HT	08.07.2025 19:29	Dokument tekstowy	680 KB
Koch_Kolacja_FR_HT	08.07.2025 19:04	Dokument tekstowy	490 KB
Kundera_Niesmiertelnosc_FR_HT	08.07.2025 19:16	Dokument tekstowy	689 KB
Kundera_Zart_FR_HT	08.07.2025 18:55	Dokument tekstowy	703 KB
Lackberg_Kaznodzieja_FR_HT	08.07.2025 19:10	Dokument tekstowy	789 KB
Lackberg_SrebrneSkrzydla_FR_HT	08.07.2025 18:46	Dokument tekstowy	471 KB
Lackberg_ZlotaKlatka_FR_HT	08.07.2025 18:48	Dokument tekstowy	587 KB
Lagerlof_CudownaPodroz_FR_HT	08.07.2025 19:07	Dokument tekstowy	1 230 KB
Lagerlof_GostaBerling_FR_HT	08.07.2025 18:53	Dokument tekstowy	443 KB
Lagerlof_TetniaceSerce_FR_HT	08.07.2025 19:12	Dokument tekstowy	371 KB
Link_Oszukana_FR_HT	08.07.2025 18:47	Dokument tekstowy	772 KB
Link_Poszukiwanie_FR_HT	08.07.2025 19:13	Dokument tekstowy	777 KB
Lispector_PasjaWedlugGH_GodzinaGwiazdy_FR_...	08.07.2025 19:17	Dokument tekstowy	515 KB
Lispector_WPoblizuDzikiegoSerca_FR_HT	08.07.2025 19:23	Dokument tekstowy	17 KB
Llosa_BurzliweCzasy_FR_HT	08.07.2025 19:26	Dokument tekstowy	624 KB
Llosa_CiotkaJuliaSkryba_FR_HT	08.07.2025 18:56	Dokument tekstowy	828 KB
Llosa_DzielnicaWystepku_FR_HT	08.07.2025 19:30	Dokument tekstowy	471 KB
Llosa_SzelmostwaNiegrzecznejDziewczynki_FR_...	08.07.2025 19:27	Dokument tekstowy	725 KB
Llosa_ZielonyDom_FR_HT	08.07.2025 18:54	Dokument tekstowy	1 034 KB
Mankell_Chinczyk_FR_HT	08.07.2025 19:02	Dokument tekstowy	918 KB


Jak dowiedzieć się w jakim formacie tekstowym zapisujemy plik?



Kundera_Niesmiertelnosc_MT


Typ pliku: Dokument programu Microsoft Word (.docx)

Otwierany za pomocą:  Word [Zmień...](#)



Kundera_Niesmiertelnosc_FR_HT

Typ pliku: Dokument tekstowy (.txt)

Otwierany za pomocą:  Notepad++ [Zmień...](#)

Jak dowiedzieć się w jakim formacie tekstowym zapisujemy plik?

Directory: C:\Users\aleks\Desktop\studia\doktorat\korpus\oryginały

Mode	LastWriteTime		Length	Name
----	-----		-----	----
d-----	09.07.2025	09:26		nowe
-a-----	05.12.2024	11:27	616931	Andric_MostNaDrinie.docx
-a-----	04.12.2024	14:50	1195521	Andric_MostNaDrinie.txt
-a-----	05.12.2024	11:58	330633	Blixen_PozegnanieZAfryka.docx
-a-----	04.12.2024	14:49	702547	Blixen_PozegnanieZAfryka.txt
-a-----	05.12.2024	12:38	534510	Bulhakov_MistrzIMalgorzata.docx
-a-----	04.12.2024	15:06	1431645	Bulhakov_MistrzIMalgorzata.txt
-a-----	05.12.2024	11:20	285199	Calvino_BaronDrzewolaz.docx
-a-----	04.12.2024	17:47	536059	Calvino_BaronDrzewolaz.txt
-a-----	05.12.2024	11:23	236376	Calvino_JesliZimowaNocaPodrozny.docx
-a-----	04.12.2024	17:21	461636	Calvino_JesliZimowaNocaPodrozny.txt
-a-----	05.12.2024	10:52	620366	Cortazar_GrawKlasy.docx
-a-----	04.12.2024	14:50	1047278	Cortazar_GrawKlasy.txt
-a-----	05.12.2024	12:37	1173200	Dostojewski_BraciaKaramazov.docx
-a-----	04.12.2024	14:49	3281542	Dostojewski_BraciaKaramazov.txt
-a-----	05.12.2024	13:27	1061760	Dostojewski_Idiota.docx
-a-----	04.12.2024	17:48	2423300	Dostojewski_Idiota.txt
-a-----	05.12.2024	12:33	1575729	Dostojewski_ZbrodniaIKara.docx
-a-----	04.12.2024	14:49	4324305	Dostojewski_ZbrodniaIKara.txt
-a-----	05.12.2024	11:25	624544	Eco_ImieRozy.docx
-a-----	04.12.2024	17:21	1219346	Eco_ImieRozy.txt
-a-----	04.12.2024	17:21	1428592	Eco_WagadloFoucaulta.txt
-a-----	05.12.2024	11:26	236378	Eco_WahadloFoucaulta.docx

Jak dowiedzieć się w jakim formacie tekstowym zapisujemy plik?

```

2  <html lang="pl">
28  <body>
29    <div id="wrapper">
49      <nav>
50        <div class="nav">
63          </li>
64          <li><a href="kontakt.html">Kontakt</a></li>
65        </ol>
66      </div>
67    </nav>
68    <div class="content">
69      
70      <div class="content2">
71        <div class="text">
72          <h2>Tłumaczenia pisemne</h2>
73          <p>W biurze tłumaczeń oferujemy usługi tłumaczeń tekstu na język polski</p>
74          <p>Wykonujemy tłumaczenia tekstów użytkowych, specjalistycznych oraz</p>
75          <p>Gwarantujemy najwyższą jakość wykonanej pracy, zachęcamy do zapoznania</p>
76          <h2>Tłumaczenia ustne</h2>
77          <p>Biuro oferuje również tłumaczenia ustne [ ] symultaniczne oraz</p>
78          <p><strong><a href="kontakt.html" class="internal-link">Zachęcamy</a></strong></p>
79          <h2>Korekta tekstu</h2>
80          <p>Wykonujemy korektę tekstów literackich, specjalistycznych, użytkowych</p>
81          <h2>MTPE (Machine-Translation Post-Editing)</h2>
82          <p>zaoferować Państwu redakcję, edycję oraz korektę tłumaczeń maszynowych</p>
83          <p>W przypadku specyficznych potrzeb, które nie zostały opisane powyżej</p>
84        </div>
85      </div>
86    </div>
87    <footer>

```

```

[Wersja zapoznawcza] README.md  persons.xml  laks-correspondence-008.xml X
data > letters > laks-correspondence-008.xml
1  <TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="SL-008">
116    <text>
118      <div>
119        <div type="letter">
120          <opener>
134          </opener>
135          <salute>Wielce Szanowna Pani,</salute>
136          <p>Przepraszam, że znów pozwalam sobie zaniepokoić Panią moimi małymi problemami,
137            które z pewnego jednak punktu widzenia mogą się okazać nie tak bardzo małe.</p>
138          <p>Przypuszczam, że nie uszło Pani uwagi „przeproszenie” mnie przez prof. <persName ref="gnd-171150694">Pragiera</persName> na łamach „Wiadomości” (nr
139            1544 z 2 [XI [19]75)<anchor n="1" corresp="SL-008-note-2"/>. Na wszelki wypadek
140            przypominam, o co poszło: prof. <persName ref="gnd-171150694">Pragier</persName>, bez mojej zgody
141            1971<anchor n="2" corresp="SL-008-note-3"/>, drugi raz w swej książce
142            <title key="bibl_SL_062">Czas teraźniejszy</title> w r. 1975<anchor n="3" corresp="SL-008-note-4">
143            - mój prywatny do niego list tuż obok listu prof. <persName ref="gnd-107498413">Romana Karsta</persName>
144            - mojego prywatnego „collage’u” wynikało dla czytelnika, że epitet „skończony
145            idiota”, użyty przez prof. <persName ref="gnd-107498413">Karsta</persName>, odnosił się do mojej
146            1971)<anchor n="5" corresp="SL-008-note-7"/> był antysemitą, a prof.
147            <persName ref="gnd-107498413">Karst</persName> - niezależnie ode mnie - napisał do Pragiera, że
148            idiotą, żeby dopatrzeć się (w tym artykule) choćby antysemitki nutki.</p>
149          <p>Sprawa ta rozeszłaby się „po kościach”, gdyby <persName ref="gnd-171150694">Pragier</persName> nie
150            jej ponownie publiczny rozgłos przez opublikowanie obu tych listów obok siebie,
151            tym razem w książce, po czterech latach.</p>
152          <p>„Przeproszenie” mnie na łamach „Wiadomości” dało mi pewną satysfakcję zabarwioną
153            niesmakiem. Ale jądro sprawy - czy artykuł był czy nie był antysemitki - gdzieś
154            się w tym wszystkim zawieruszyło. Toteż uważałem za stosowne wysłać załączony
155            przy niniejszym tekście, przedtem do „Wiadomości”, potem do „Tygodnia Polskiego” z
156            przypisem: „Przepraszam, że znów pozwalam sobie zaniepokoić Panią moimi małymi problemami,
157            które z pewnego jednak punktu widzenia mogą się okazać nie tak bardzo małe.”</p>

```

Ważne informacje, pojęcia przed wykonaniem analiz

- ~~pracujemy na „surowych” plikach (**raw files**), najlepiej w formacie **txt**~~
 - sprawdzamy sposób kodowania plików, żeby się upewnić, że analiza przebiegnie poprawnie - sprawdzamy, czy jest format **UTF-8**
 - pracujemy na **niezlematyzowanych słowach**, ale możemy wykonywać analizę **n-gramów** poszczególnych znaków (przydatne, gdyby ktoś chciał badać np. chiński)
-

Co to UTF-8 i dlaczego jest ważne?

The screenshot shows the homepage of the Unicode Consortium. The browser's address bar displays 'home.unicode.org'. The page features a grid of various Unicode characters, each with its corresponding code point (e.g., U+025C, U+271E, U+0939, U+02C6, U+055E, U+3145, U+263C, U+063A, U+03B5, U+1F9F8, U+10DB, U+0298, U+03A6, U+00F0, U+0EC2, U+FF65, U+FF4D, U+203A, U+269B, U+1F64C, U+0573, U+3008, U+0C39, U+309D, U+2669, U+30C9, U+266F, U+D392, U+13EA, U+53F8, U+3002, U+26A2, U+0920, U+1F603, U+1D54, U+01D0, U+0422, U+FF0C, U+26AA, U+2039, U+8DD1, U+0D1A, U+104D, U+1F91E). On the left, there is a sidebar with links: 'About Unicode', 'Technical Quick Start Guide', 'Support Unicode', 'Adopt a Character', 'Membership', 'News and Events', 'Emoji', and 'Newsletter Signup'. In the center, a message states: 'Everyone in the world should be able to use their own language on phones and computers.' Below this is a link: 'LEARN MORE ABOUT UNICODE'. On the right, there is a 'TM' logo and a button that says 'ADOPT A CHARACTER'.

Co to UTF-8 i dlaczego jest ważne?

Do najpowszechniejszych metod bajtowego kodowania znaków należą:

- UTF-32/UCS-4
- UTF-16
- UTF-8.

Mniej popularnymi kodowaniami Unicode są:

- UTF-7
- UCS-2.

Istnieją również inne kodowania, stanowiące margines lub pozostające na etapie propozycji, na przykład:

- UTF-9 i UTF-18
- UTF-EBCDIC
- UTF-6
- UTF-5.

Co to UTF-8 i dlaczego jest ważne?

A Chinese character: 汉
its Unicode value: U+6C49
convert 6C49 to binary: 01101100 01001001

Binary format of bytes in sequence

1st Byte	2nd Byte	3rd Byte	4th Byte	Number of Free Bits	Maximum Expressible Unicode Value
0xxxxxxx				7	007F hex (127)
110xxxxx	10xxxxxx			(5+6)=11	07FF hex (2047)
1110xxxx	10xxxxxx	10xxxxxx		(4+6+6)=16	FFFF hex (65535)
11110xxx	10xxxxxx	10xxxxxx	10xxxxxx	(3+6+6+6)=21	10FFFF hex (1,114,111)

Header	Place holder	Fill in our Binary	Result
1110	xxxx	0110	11100110
10	xxxxxx	110001	10110001
10	xxxxxx	001001	10001001

PREMIÈRE PARTIE

LUDVIK

Ainsi, après bien des années, je me retrouvais chez moi. Debout sur la grande place (quand enfant, puis gamin, puis jeune homme, j'avais mille fois traversée), je ne r
Des années durant, rien ne m'avait attiré vers ma ville natale ; je me disais qu'elle m'était devenue indifférente, et cela me paraissait naturel : depuis quinze
Une fois encore je parcourus d'un l'œil narquois la place disgracieuse avant de lui tourner le dos pour prendre la rue de l'hôtel où ma chambre était retenue pour la
Je m'assis sur la chaise, le regard perdu vers les rideaux éclairés en transparence, et je réfléchis. À cet instant, des pas et des voix se firent entendre du corri
Je me levai, ma résolution était prise ; je me lavai encore les mains dans le lavabo, les essuyai avec la serviette et quittai l'hôtel sans bien savoir d'abord où
Cet hôpital est un ensemble de bâtiments et de pavillons semés et là sur un vaste espace de jardins ; je pénétrai dans la petite guérite qui jouxte le portail
Je lui expliquai que j'étais arrivé moins d'une heure plus tôt pour une affaire sans importance qui me retiendrait ici environ deux jours, et il manifesta tout de s
« Je cours les filles, répondis-je. « Ce n'est pas pour des femmes, c'est pour moi qu'il me faut ma liberté », dit-il, et il ajouta : « Écoutez, venez un mo
Sortis de l'enceinte de l'hôpital, nous parvînmes bientôt à un groupe d'immeubles neufs qui, l'un à côté de l'autre, jaillissaient sans harmonie d'un sc
Je fis à Kostka l'alloc de sa chambre et lui demandai comment était sa salle de bains. « Rien de luxueux », dit-il, content de l'intérêt que je marquais, et il
Ma question l'étonna, mais sur-le-champ (comme s'il craignait que je ne le soupçonnasse de manquer d'empressement) il me dit : « Bien volontiers, il est à vous » E
Après quoi nous prîmes place autour de la petite table (Kostka avait préparé du café) et bavardâmes un moment (assis sur le divan, j'en constatais avec plaisir la
Kostka exprima en partant le vœu que son studio me procure « vraiment quelque chose de beau ». « Oui, lui dis-je, il va me permettre d'effectuer une belle destructio
Nous nous retrouvions là où nous nous étions séparés la dernière fois (peut-être quelque neuf ans plus tôt) ; notre différend revêtait à présent une allure m
Tandis que j'accompagnais Kostka pour regagner l'hôpital à l'autre bout de la ville, je jouais avec les clés au fond de ma poche et je me sentais bien au côté d
Je ne déclinai pas les bons offices de Kostka et me laissai emmener dans un petit salon où devant trois glaces étaient plantés trois grands fauteuils pivotants dont de
Je maintins mes yeux au plafond même après avoir senti sur mon cou les doigts de la coiffeuse qui glissaient sous le col de ma chemise le bord d'un linge blanc. Puis l
Puis les caresses cessèrent et j'entendis la coiffeuse s'écarter afin cette fois de vraiment saisir le rasoir et je me dis à ce moment (car les pensées contin

46
47 PREMIÈRE PARTIE
48
49 LUDVIK
50
51 Ainsi, après bien des années, je me retrouvais chez moi. Debout sur la grande place (qu'enfant, puis gamin, puis jeune homme, j'avais mille fois traversée), je ne ressentais
52
53 Des années durant, rien ne m'avait attiré vers ma ville natale ; je me disais qu'elle m'était devenue indifférente, et cela me paraissait naturel : depuis quinze ans déjà
54
55 Une fois encore je parcourus d'un œil narquois la place disgracieuse avant de lui tourner le dos pour prendre la rue de l'hôtel où ma chambre était retenue pour la nuit. I
56
57 Je m'assis sur la chaise, le regard perdu vers les rideaux éclairés en transparence, et je réfléchis. À cet instant, des pas et des voix se firent entendre du corridor ; c
58
59 Je me levai, ma résolution était prise ; je me lavai encore les mains dans le lavabo, les essuyai avec la serviette et quittai l'hôtel sans bien savoir d'abord où j'irais
60
61 Cet hôpital est un ensemble de bâtiments et de pavillons semés çà et là sur un vaste espace de jardins ; je pénétrai dans la petite guérite qui jouxte le portail et je pri
62
63 Je lui expliquai que j'étais arrivé moins d'une heure plus tôt pour une affaire sans importance qui me retiendrait ici environ deux jours, et il manifesta tout de suite un
64
65 – Je cours les filles, répondis-je. – Ce n'est pas pour des femmes, c'est pour moi qu'il me faut ma liberté », dit-il, et il ajouta : « Écoutez, venez un moment chez moi,
66
67 Sortis de l'enceinte de l'hôpital, nous parvînmes bientôt à un groupe d'immeubles neufs qui, l'un à côté de l'autre, jaillissaient sans harmonie d'un sol poussiéreux non a
68
69 Je fis à Kostka l'éloge de sa chambre et lui demandai comment était sa salle de bains. « Rien de luxueux », dit-il, content de l'intérêt que je marquais, et il me fit pass
70
71 Ma question l'étonna, mais sur-le-champ (comme s'il craignait que je ne le soupçonne de manquer d'empressement) il me dit : « Bien volontiers, il est à vous » Et de pours
72
73 Après quoi nous prîmes place autour de la petite table (Kostka avait préparé du café) et bavardâmes un moment (assis sur le divan, j'en constatais avec plaisir la fermeté,
74
75 Kostka exprima en partant le vœu que son studio me procure « vraiment quelque chose de beau ». « Oui, lui dis-je, il va me permettre d'effectuer une belle destruction. – V
76
77 Nous nous retrouvions là où nous nous étions séparés la dernière fois (peut-être quelque neuf ans plus tôt) ; notre différend revêtait à présent une allure métaphorique pe
78
79 Tandis que j'accompagnais Kostka pour regagner l'hôpital à l'autre bout de la ville, je jouais avec les clés au fond de ma poche et je me sentais bien au côté de l'ami de
80
81 Je ne déclinai pas les bons offices de Kostka et me laissai emmener dans un petit salon où devant trois glaces étaient plantés trois grands fauteuils pivotants dont deux é
82
83 Je maintins mes yeux au plafond même après avoir senti sur mon cou les doigts de la coiffeuse qui glissaient sous le col de ma chemise le bord d'un linge blanc. Puis la co
84
85 Puis les caresses cessèrent et j'entendis la coiffeuse s'écarter afin, cette fois, de vraiment saisir le rasoir et je me mis à ce moment (car les pensées continuaient le

Ważne informacje, pojęcia przed wykonaniem analiz

- ~~pracujemy na „surowych” plikach (**raw files**), najlepiej w formacie **txt**~~
 - ~~sprawdzamy sposób kodowania plików, żeby się upewnić, że analiza przebiegnie poprawnie - sprawdzamy, czy jest format **UTF-8**~~
 - pracujemy na **niezlematyzowanych słowach**, ale możemy wykonywać analizę **n-gramów** poszczególnych znaków (przydatne, gdyby ktoś chciał badać np. chiński)
-

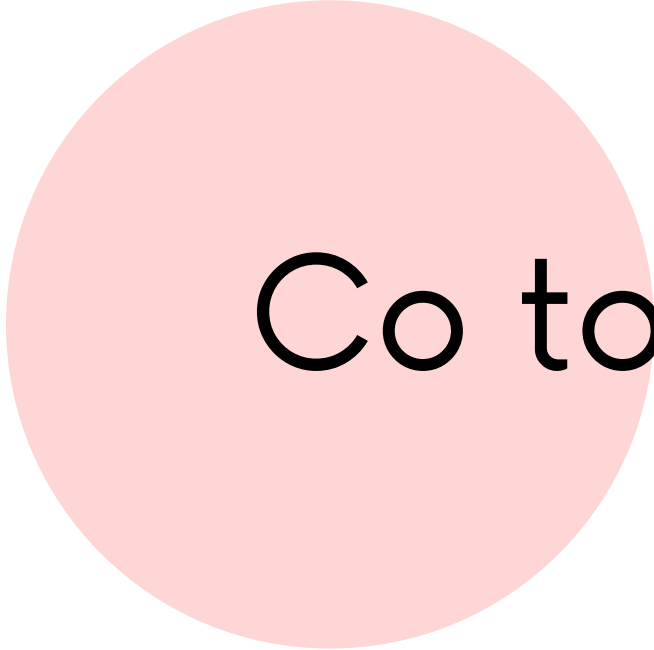


Czym jest lematyzacja?

Wszyscy mamy bardzo fajne mamy.



Wszyscy mieć bardzo fajny mama.



Co to n-gramy?

To be or not to be.



2-gramy

Co to n-gramy?

To be or not to be.



2-gramy

to-o_-b-be-e_-o-or-r_-n-no-ot-t_-t-to-o_-b-be-e.

Co to n-gramy?

To be or not to be.



3-gramy

to_-o_b-be-be_-e_o-or-or_-r_n-no-not-ot_-t_t-to-to_-o_b-be-be.

Co to n-gramy?

To be or not to be.



2-gramy słów

to be-be or-or not-not to-to be

Co to n-gramy?

To be or not to be.



3-gramy słów

to be or-be or not-or not to-not to be

