

# **Eksploracja danych i wyszukiwanie informacji w mediach społecznościowych**

**Wykład 1 - sprawy organizacyjne, wstęp, przykłady**

**dr inż. Julian Sienkiewicz**

**8 października 2018**

## Kontakt

dr inż. Julian Sienkiewicz

Pracownia Fizyki w Ekonomii i Naukach Społecznych

Gmach Matematyki, pokój 529

tel. 22 234 5808, email: [julian.sienkiewicz@pw.edu.pl](mailto:julian.sienkiewicz@pw.edu.pl)

WWW: [www.fizyka.pw.edu.pl/~julas/TEXT](http://www.fizyka.pw.edu.pl/~julas/TEXT)

## Kontakt

dr inż. Julian Sienkiewicz

Pracownia Fizyki w Ekonomii i Naukach Społecznych

Gmach Matematyki, pokój 529

tel. 22 234 5808, email: [julian.sienkiewicz@pw.edu.pl](mailto:julian.sienkiewicz@pw.edu.pl)

WWW: [www.fizyka.pw.edu.pl/~julas/TEXT](http://www.fizyka.pw.edu.pl/~julas/TEXT)

## Organizacja przedmiotu

- wykład 15h (pierwsza połowa semestru),
- laboratorium 30h,
- 2 grupy laboratorium: 10<sup>15</sup>-11<sup>45</sup> oraz 16<sup>15</sup>-17<sup>45</sup>,
- wykład: ogólny opis,
- laboratorium: konkretne przykłady w pakiecie  $\mathbb{R}$

## Zasady zaliczania przedmiotu

### ● wykład:

- kolokwium na ostatnich zajęciach (**26 listopada**) - 45 min,
- **20 punktów** do zdobycia,
- dziesięć pytań zamkniętych (test wyboru) po 0.5 pkt każde + 3 pytania otwarte po 5 pkt. każde,
- przykładowe kolokwium na stronie najpóźniej 20 listopada

### ● laboratorium:

- 13 zajęć + zajęcia organizacyjne + wstęp do R,
- 8 punktowanych zadań po max. 10 punktów = **80 pkt**,
- brak kolokwium
- na ocenę składa się **suma punktów z wykładu i lab.**
- standardowa skala: 51-60 dst, 61-70 dst+, 71-80 db, 81-90 db+, 91-100 bdb
- brak warunków koniecznych uzyskania co najmniej połowy dostępnych punktów z wykładu lub laboratorium

Brakuje konkretnej literatury w języku polskim

- D. Spinczyk, M. Dzieciątko, *Text mining. Metody, narzędzia, zastosowania*, PWN (2016),

Polecam również poniższe pozycje w jęz. angielskim:

- Ch. Aggarwal, Ch-X Zhai, C. O'Neil *Mining Text Data*, Springer (2012).
- D. Robinson, J. Silge, *Text Mining with R*, O'Reilly (2017)

## Text mining wg Wikipedii (ang.)

**Text mining**, also referred to as **text data mining**, roughly equivalent to **text analytics**, is the process of deriving high-quality information from text.

### Text mining wg Wikipedii (ang.)

**Text mining**, also referred to as **text data mining**, roughly equivalent to **text analytics**, is the process of deriving high-quality information from text.

### Text mining wg Wikipedii (pol.)

Text mining (eksploracja tekstu) — ogólna nazwa metod eksploracji danych służących do wydobywania danych z tekstu i ich późniejszej obróbki.

## Text mining wg Wikipedii (ang.)

**Text mining**, also referred to as **text data mining**, roughly equivalent to **text analytics**, is the process of deriving high-quality information from text.

## Text mining wg Wikipedii (pol.)

Text mining (eksploracja tekstu) — ogólna nazwa metod eksploracji danych służących do wydobywania danych z tekstu i ich późniejszej obróbki.

## Text mining wg Marti Hearst

Another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known.



[Grafika pobrana z:  
<https://www.ischool.berkeley.edu>]



## Po co text mining?

Z drugiej strony, warto zadać sobie pytanie **po co potrzebujemy eksploracji tekstu?** lub **jakie jest zadanie eksploracji tekstu?**. Ogólną odpowiedzią jest oczywiście: **aby (w automatyczny sposób) zrozumieć zawartość danego tekstu...**

## Po co text mining?

Z drugiej strony, warto zadać sobie pytanie **po co potrzebujemy eksploracji tekstu?** lub **jakie jest zadanie eksploracji tekstu?**. Ogólną odpowiedzią jest oczywiście: **aby (w automatyczny sposób) zrozumieć zawartość danego tekstu...**

## Po co text mining?

... niestety to założenie wydaje się być zbyt trudne. Dlatego skupiamy się raczej pomniejszych zdaniach.

## Dlaczego analiza tekstu jest **trudna**?

Cieężko jest oddać abstrakcyjne pojęcia w postaci innych, dobrze zdefiniowanych pojęć



## Dlaczego analiza tekstu jest **trudna**?

Cieężko jest oddać abstrakcyjne pojęcia w postaci innych, dobrze zdefiniowanych pojęć



**Time** **flies** like an  
arrow.

Niezliczone kombinacje subtelnych i abstrakcyjnych relacji pomiędzy pojeciami

## Dlaczego analiza tekstu jest **trudna**?

Cieężko jest oddać abstrakcyjne pojęcia w postaci innych, dobrze zdefiniowanych pojęć



**Time** **flies** like an arrow.



Niezliczone kombinacje subtelnych i abstrakcyjnych relacji pomiędzy pojeciami

Wiele sposobów opisywania tych samych pojęć

## Dlaczego analiza tekstu jest **trudna**?

Cieężko jest oddać abstrakcyjne pojęcia w postaci innych, dobrze zdefiniowanych pojęć



**Time** **flies** like an arrow.



Niezliczone kombinacje subtelnych i abstrakcyjnych relacji pomiędzy pojeciami

Wiele sposobów opisywania tych samych pojęć

Wysoka wymiarowość problemu



## Dlaczego analiza tekstu jest **trudna**?

Cieężko jest oddać abstrakcyjne pojęcia w postaci innych, dobrze zdefiniowanych pojęć



**Time** flies like an  
arrow.



Niezliczone kombinacje subtelnych i abstrakcyjnych relacji pomiędzy pojeciami

Wiele sposobów opisywania tych samych pojęć

Wysoka wymiarowość problemu



Automatic Feature Analysis

Term: You found 2 distinct items  
which are highly similar

0.9999999999999999

Term	Rank	Value	Rank	Value	Rank	Value
1	1	1	1	1	1	1
2	2	1	2	1	2	1
3	3	1	3	1	3	1
4	4	1	4	1	4	1
5	5	1	5	1	5	1
6	6	1	6	1	6	1
7	7	1	7	1	7	1
8	8	1	8	1	8	1
9	9	1	9	1	9	1
10	10	1	10	1	10	1
11	11	1	11	1	11	1
12	12	1	12	1	12	1
13	13	1	13	1	13	1
14	14	1	14	1	14	1
15	15	1	15	1	15	1
16	16	1	16	1	16	1
17	17	1	17	1	17	1
18	18	1	18	1	18	1
19	19	1	19	1	19	1
20	20	1	20	1	20	1
21	21	1	21	1	21	1
22	22	1	22	1	22	1
23	23	1	23	1	23	1
24	24	1	24	1	24	1
25	25	1	25	1	25	1
26	26	1	26	1	26	1
27	27	1	27	1	27	1
28	28	1	28	1	28	1
29	29	1	29	1	29	1
30	30	1	30	1	30	1
31	31	1	31	1	31	1
32	32	1	32	1	32	1
33	33	1	33	1	33	1
34	34	1	34	1	34	1
35	35	1	35	1	35	1
36	36	1	36	1	36	1
37	37	1	37	1	37	1
38	38	1	38	1	38	1
39	39	1	39	1	39	1
40	40	1	40	1	40	1
41	41	1	41	1	41	1
42	42	1	42	1	42	1
43	43	1	43	1	43	1
44	44	1	44	1	44	1
45	45	1	45	1	45	1
46	46	1	46	1	46	1
47	47	1	47	1	47	1
48	48	1	48	1	48	1
49	49	1	49	1	49	1
50	50	1	50	1	50	1
51	51	1	51	1	51	1
52	52	1	52	1	52	1
53	53	1	53	1	53	1
54	54	1	54	1	54	1
55	55	1	55	1	55	1
56	56	1	56	1	56	1
57	57	1	57	1	57	1
58	58	1	58	1	58	1
59	59	1	59	1	59	1
60	60	1	60	1	60	1
61	61	1	61	1	61	1
62	62	1	62	1	62	1
63	63	1	63	1	63	1
64	64	1	64	1	64	1
65	65	1	65	1	65	1
66	66	1	66	1	66	1
67	67	1	67	1	67	1
68	68	1	68	1	68	1
69	69	1	69	1	69	1
70	70	1	70	1	70	1
71	71	1	71	1	71	1
72	72	1	72	1	72	1
73	73	1	73	1	73	1
74	74	1	74	1	74	1
75	75	1	75	1	75	1
76	76	1	76	1	76	1
77	77	1	77	1	77	1
78	78	1	78	1	78	1
79	79	1	79	1	79	1
80	80	1	80	1	80	1
81	81	1	81	1	81	1
82	82	1	82	1	82	1
83	83	1	83	1	83	1
84	84	1	84	1	84	1
85	85	1	85	1	85	1
86	86	1	86	1	86	1
87	87	1	87	1	87	1
88	88	1	88	1	88	1
89	89	1	89	1	89	1
90	90	1	90	1	90	1
91	91	1	91	1	91	1
92	92	1	92	1	92	1
93	93	1	93	1	93	1
94	94	1	94	1	94	1
95	95	1	95	1	95	1
96	96	1	96	1	96	1
97	97	1	97	1	97	1
98	98	1	98	1	98	1
99	99	1	99	1	99	1
100	100	1	100	1	100	1

Bardzo wiele cech (features)

Dlaczego analiza tekstu może być **łatwa**?



## Dlaczego analiza tekstu może być **łatwa**?

W tekście zwykle jest spora ilość nadmiarowych lub powtarzających się informacji.

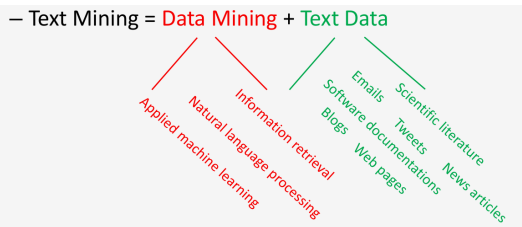
## Dlaczego analiza tekstu może być **łatwa**?

W tekście zwykle jest spora ilość nadmiarowych lub powtarzających się informacji.

W zasadzie większość prostych algorytmów może osiągnąć całkiem dobre wyniki przy wykonywaniu w następujących nieskomplikowanych zadań:

- wydobądź "istotne" wyrażenia,
- znajdź istotnie powiązane słowa,
- stwórz pewnego rodzaju podsumowanie dokumentów

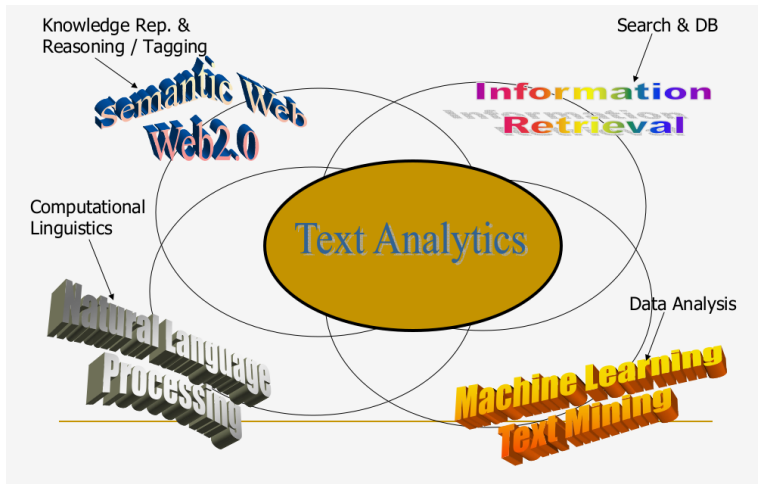
Można również próbować zilustrować powiązania pomiędzy eksploracją tekstu a innymi dziedzinami:



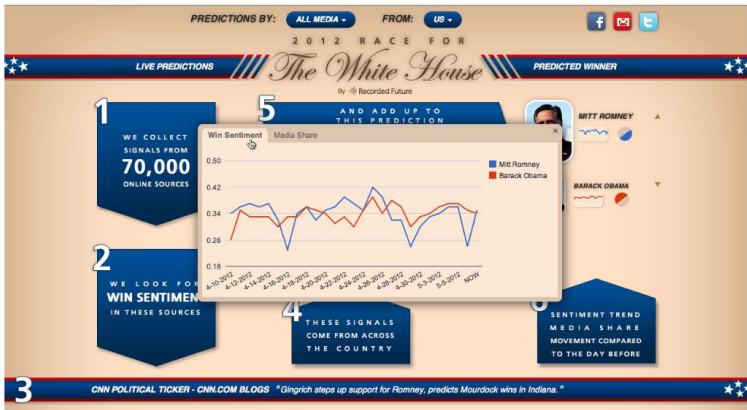
	Finding Patterns	Finding "Nuggets"	
		Novel	Non-Novel
Non-textual data	General data-mining	Exploratory analysis	Database queries
Textual data	Comp Ling		Information retrieval

**Text Mining**

Można również próbować zilustrować powiązania pomiędzy eksploracją tekstu a innymi dziedzinami:



## Przykłady: analiza sentymentu – wybory



## Przykłady: podsumowywanie dokumentów



## Przykłady: systemy rekomendujące

### FOREIGN SUGGESTIONS (about 104) [See all >](#)



#### Tell No One

Because you enjoyed:  
Memento  
Syriana  
Children of Men



#### Let the Right One In

Because you enjoyed:  
Seven Samurai  
This Is Spinal Tap  
The Big Lebowski



#### I've Loved You So Long

Because you enjoyed:  
The Queen  
Syriana  
Good Night, and Good Luck

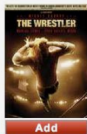


#### Downfall

Because you enjoyed:  
Das Boot  
The Killing Fields  
Seven Samurai

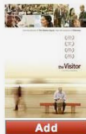


### DRAMA SUGGESTIONS (about 82) [See all >](#)



#### The Wrestler

Because you enjoyed:  
Sin City  
Reservoir Dogs  
The Big Lebowski



#### The Visitor

Because you enjoyed:  
Gandhi  
The Motorcycle Diaries  
The Queen



#### Brick

Because you enjoyed:  
The Big Lebowski  
Rushmore  
Fight Club

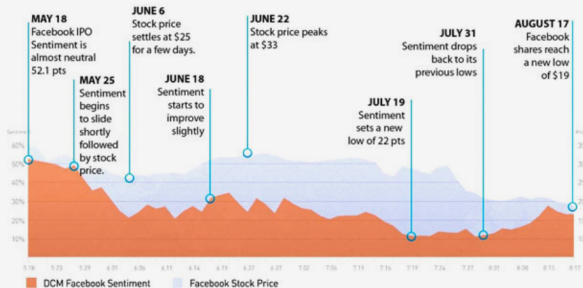


#### The Pianist

Because you enjoyed:  
Amadeus  
The Killing Fields  
Empire of the Sun



## Przykłady: analiza tekstu w serwisach finansowych





## Przykłady: analiza danych medycznych

REQUEST FOR MEDICAL/DENTAL RECORDS		DATE
1. PATIENT (Last, First, Middle Initial) (Asterisk Name)		0 December 20, 1987
2. PROVIDER (Last, First, Middle Initial) (Asterisk Name)		Dr. [redacted]
3. REQUESTOR (Last, First, Middle Initial) (Asterisk Name)		John [redacted]
4. REQUESTOR'S ADDRESS (Street, City, State, Zip)		100 [redacted] Street, [redacted] City, [redacted] State, [redacted] Zip
5. REQUESTOR'S PHONE NUMBER (Area Code, Number)		[redacted]
6. REQUESTOR'S FAX NUMBER (Area Code, Number)		[redacted]
7. REQUESTOR'S E-MAIL ADDRESS		[redacted]
8. REQUESTOR'S BUSINESS ADDRESS (Street, City, State, Zip)		[redacted]
9. REQUESTOR'S BUSINESS PHONE NUMBER (Area Code, Number)		[redacted]
10. REQUESTOR'S BUSINESS FAX NUMBER (Area Code, Number)		[redacted]
11. REQUESTOR'S BUSINESS E-MAIL ADDRESS		[redacted]
12. REQUESTOR'S BUSINESS WEBSITE		[redacted]
13. REQUESTOR'S BUSINESS SOCIAL MEDIA		[redacted]
14. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
15. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
16. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
17. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
18. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
19. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
20. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
21. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
22. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
23. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
24. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
25. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
26. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
27. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
28. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
29. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
30. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
31. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
32. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
33. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
34. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
35. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
36. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
37. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
38. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
39. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
40. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
41. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
42. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
43. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
44. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
45. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
46. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
47. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
48. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
49. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
50. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
51. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
52. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
53. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
54. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
55. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
56. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
57. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
58. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
59. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
60. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
61. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
62. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
63. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
64. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
65. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
66. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
67. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
68. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
69. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
70. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
71. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
72. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
73. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
74. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
75. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
76. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
77. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
78. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
79. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
80. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
81. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
82. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
83. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
84. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
85. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
86. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
87. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
88. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
89. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
90. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
91. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
92. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
93. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
94. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
95. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
96. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
97. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
98. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
99. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]
100. REQUESTOR'S BUSINESS OTHER CONTACT INFORMATION		[redacted]

WebMD moderated

### WebMD® Heart Disease Community

Home  
Discussions  
Tips  
Resources  
About This Community  
Staying Informed  
My Watchlist  
Related Men's Health Communities  
All Communities  
Community FAQs  
Chris Assistance  
Sign Up  
Sign Up

#### What's Happening Now

See All Discussions | Tips | Resources



**11 surprising ways to prevent a heart attack**  
http://news.bloomber.com/health/2012/01/10/11-surprising-ways-to-prevent-a-heart-attack/...  
Chances are you're still riding the New Year's high and you're motivated and committed to eating healthy...

Posted by cardiobus1

Was this helpful?

Yes No

2 of 2 found this Resource helpful

0 Replies

Post Now

1 day ago



**Reply: Angiogram**  
Consult with an interventional cardiologist and bring the disc of the angiogram video with you.

Posted by cardiobus1

3 Replies

INCLUDES EXPERT CONTENT

Report This

1 day ago



**Reply: Internal Bleeding after heart cath**  
Could be that there isn't enough in it for the surgery. My husband had this because a top video was suppressed.

Posted by cardiobus1

16 Replies

Report This

3 days ago



**Reply: Trouble Breathing**  
You need to consult with a doctor. If you don't have the money to pay for it, use the internet to find the...

Posted by cardiobus1

1 Reply

Report This

Search This Community

130

#### Popular Discussions

Start a Discussion | See All

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

1 day ago

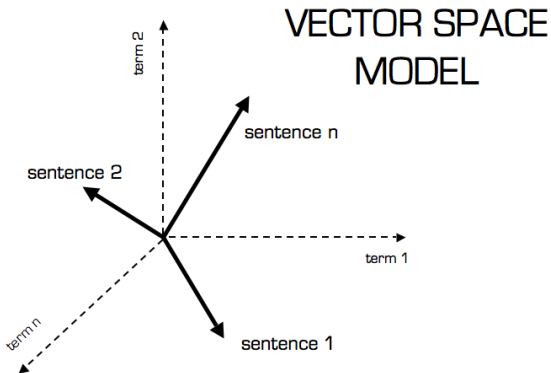
1 day ago

1 day ago

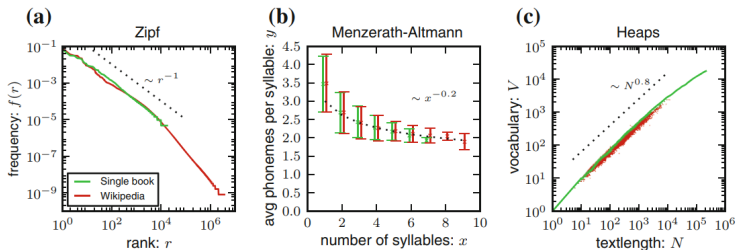
## Ogólny plan wykładu

- 1 reprezentacja tekstu
- 2 prawo Zipfa
- 3 przetwarzanie języka naturalnego (NLP)
- 4 analiza sentymentu
- 5 topic modeling
- 6 analiza mediów społecznościowych

## 2 Reprezentacja tekstu...



### 3 Prawo Zipfa i pokrewne...



[Altmann, Gerlach, Statistical laws in Linguistics, Creativity and Universality in Language, Springer (2017)]

## 4 przetwarzanie języka naturalnego (NLP)

### Part of speech:

NP NP RB VBD IN NP NP CC PRF VBZ RB VBG PRP IN PRP .  
Mrs. Clinton previously worked for Mr. Obama, but she is now distancing herself from him .

### Named entity recognition:

Person Date Person Date  
Mrs. Clinton previously worked for Mr. Obama, but she is now distancing herself from him.

### Co-reference:

Mention Ment M Mention M  
Mrs. Clinton previously worked for Mr. Obama, but she is now distancing herself from him.

### Basic dependencies:

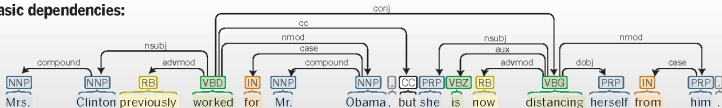
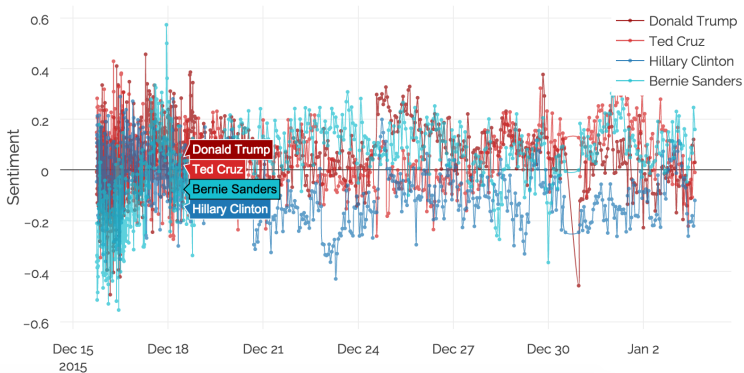


Fig. 1. Many language technology tools start by doing linguistic structure analysis. Here we show output from Stanford CoreNLP. As shown from top to

[Hirschberg, Manning, Advances in natural language processing, Science 349, 261 (2015)]

## 5 Analiza sentymentu

### How Twitter Feels About the 2016 Election Candidates



## 5 Analiza sentymentu: klasyfikatory słownikowe vs uczenie pod nadzorem

### Dictionary-based Approach

Create lists of **positive/negative** words (phrases).

#### Negative

suck  
terrible  
awful  
unwatchable  
hideous

#### Positive

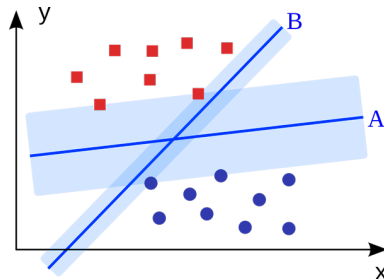
dazzling  
brilliant  
phenomenal  
excellent  
fantastic

Sentiment = |Positive words| - |Negative words|

Around 65% accuracy!



5



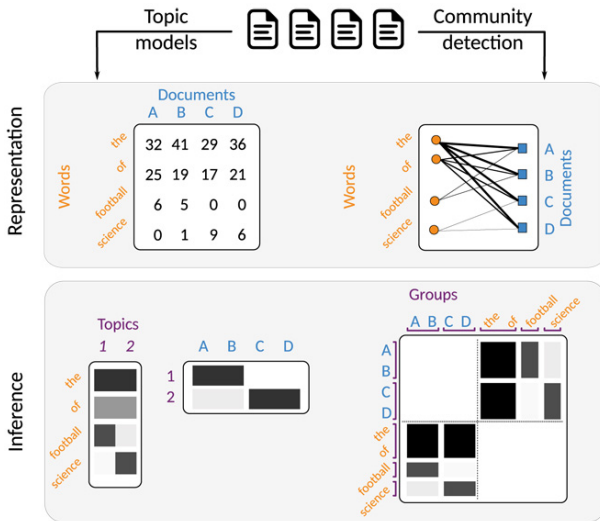
[<https://www.slideshare.net/jchoi7s/cs571-sentiment-analysis>]

[<https://medium.com/nlpython/sentiment-analysis-analysis-part-2-support-vector-machines-31f78baeee09>]

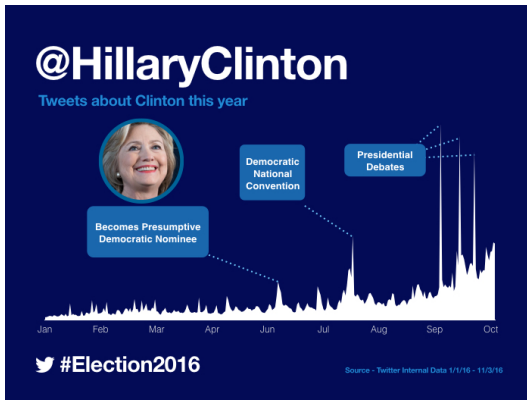




## 6 Topic modelling

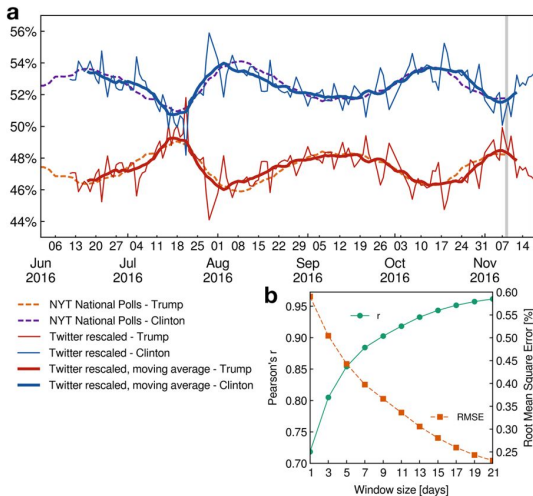


## 7 Analiza mediów społecznościowych



[Bovet, Morone, Makse, Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump, Scientific Reports (2018)]

## 7 Analiza mediów społecznościowych



[Bovet, Morone, Makse, Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump, Scientific Reports (2018)]