

Eksploracja Tekstu i Analiza Danych On-line

Wykład 1 - sprawy organizacyjne, wstęp, przykłady

dr inż. Julian Sienkiewicz

7 października 2019

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

WWW: 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

WWW: www.fizyka.pw.edu.pl/~julas/TEXT

Organizacja przedmiotu

- wykład 15h (pierwsza połowa semestru),
- laboratorium 30h,
- 2 grupy laboratorium: 10¹⁵-11⁴⁵ oraz 16¹⁵-17⁴⁵,
- wykład: ogólny opis,
- laboratorium: konkretne przykłady w pakiecie $\mathbb{R}$

Zasady zaliczania przedmiotu

● **wykład:**

- kolokwium na ostatnich zajęciach (**18 listopada**),
- **20 punktów** do zdobycia,
- dwadzieścia pytań zamkniętych (test wyboru) po 0.5 pkt każde + 4 pytania otwarte po 2.5 pkt. każde,
- przykładowe kolokwium na stronie najpóźniej 4 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

Text: You found a book about time
Author: unknown
ID: 123456789

Feature	Book	Time	Found	Author	ID
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					

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.

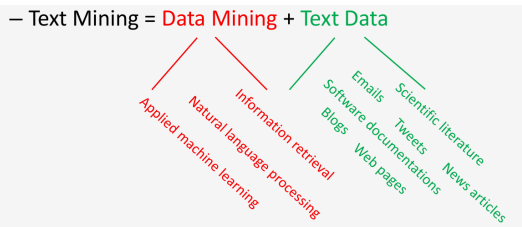
Dlaczego analiza tekstu może być **łatwa**?

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

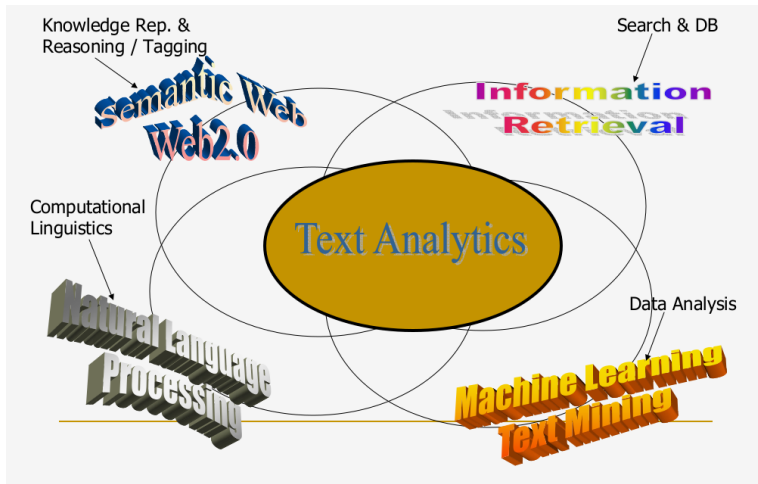
Można również próbować zilustrować powiązania pomiędzy eksploracją tekstu a innymi dziedzinami:



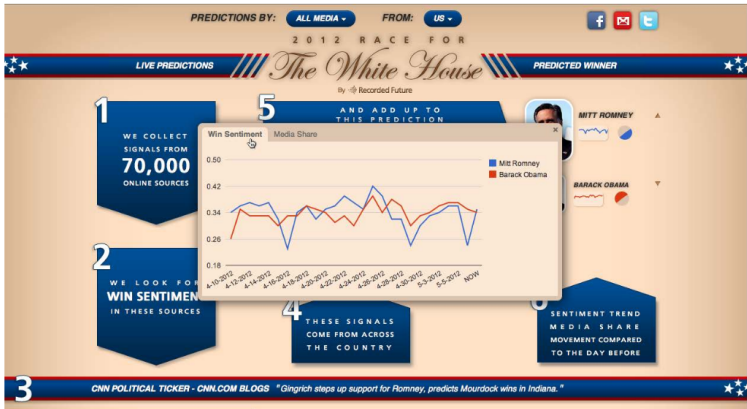
	Finding Patterns	Finding "Nuggets"	
		Novel	Non-Novel
Non-textual data	General data-mining	Exploratory analysis	Database queries
Textual data	Comp Ling		Information retrieval

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

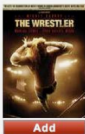


Downfall

Because you enjoyed:
Das Boot
The Killing Fields
Seven Samurai

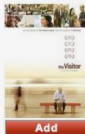


DRAMA SUGGESTIONS (about 82) [See all >](#)



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

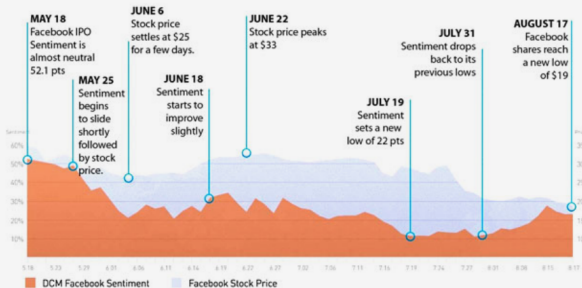


The Pianist

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 NAME, FIRST NAME, & AKA) NAME		0 December 20, 1977
2. PROVIDER (LAST NAME, FIRST NAME, & AKA) NAME		Dr. [redacted]
3. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
4. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
5. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
6. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
7. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
8. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
9. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
10. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
11. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
12. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
13. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
14. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
15. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
16. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
17. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
18. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
19. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
20. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
21. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
22. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
23. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
24. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
25. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
26. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
27. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
28. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
29. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
30. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
31. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
32. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
33. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
34. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
35. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
36. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
37. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
38. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
39. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
40. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
41. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
42. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
43. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
44. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
45. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
46. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
47. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
48. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
49. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
50. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
51. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
52. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
53. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
54. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
55. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
56. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
57. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
58. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
59. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
60. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
61. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
62. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
63. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
64. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
65. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
66. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
67. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
68. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
69. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
70. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
71. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
72. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
73. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
74. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
75. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
76. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
77. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
78. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
79. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
80. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
81. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
82. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
83. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
84. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
85. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
86. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
87. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
88. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
89. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
90. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
91. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
92. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
93. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
94. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
95. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
96. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
97. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
98. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
99. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [redacted]
100. REQUESTOR (LAST NAME, FIRST NAME, & AKA) NAME		John [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 for the Heart Health newsletter and keep up with all the latest news, treatments, and research with WebMD.

☐ I have read and agree to WebMD's Privacy Policy

Enter Email Address

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 cardiovascular1

Was this Helpful?

2 of 2 found this Resource helpful

0 Replies

Post New

Search This Community

Popular Discussions

- Laugh Your Way to Cardiac Health
- Continuous control of heart rate
- New Diet Required and Stated
- The Road Home from Heart Surgery
- 2012's chest pain

Start a Discussion | See All

Helpful Tips

HOW TO EAT FOR A HEALTHY HEART.

1. Eat food less in fat, much less saturated and trans-fat. 2. More servings of fruits and vegetables considering its variety daily and ... More

Was this Helpful?

1 of 1 found this helpful

Tip for the pain

Post a Tip | See All

Helpful Resources

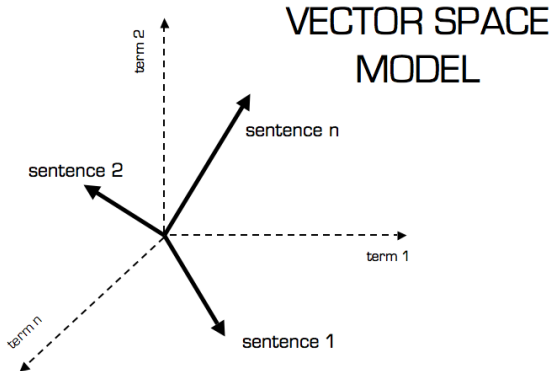
- Super safe iodine may save mil...
- Eating More Fruit Cuts Heart D...
- Heart Attack Treatment: Timing...
- Can heart attack damage be re...
- Causes of Plastic Attacks

Post a Resource | See All

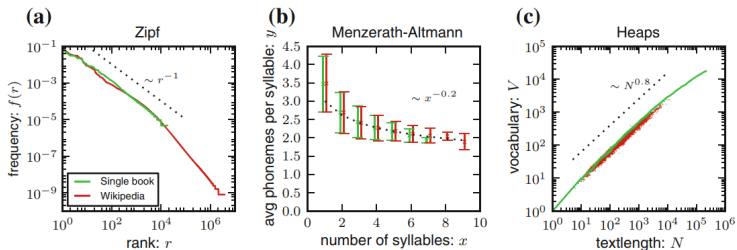
Ogólny plan wykładu

- 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:

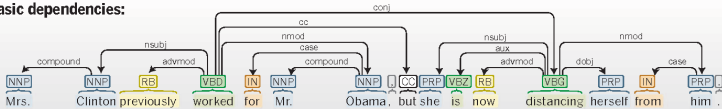
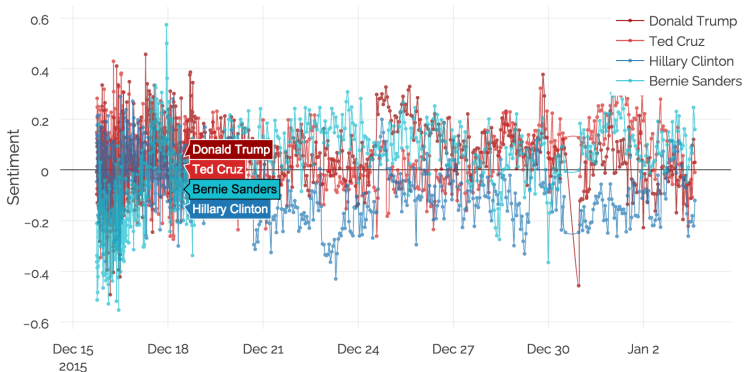


Fig. 1. Many language technology tools start by doing linguistic structure analysis. Here we show output from Stanford CoreNLP. As shown from top to

[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

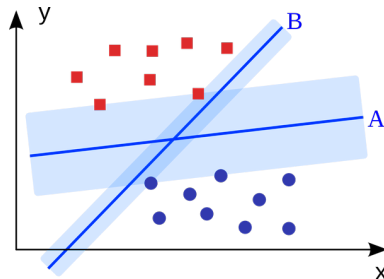
dazzling
brilliant
phenomenal
excellent
fantastic

Sentiment = |Positive words| - |Negative words|

Around 65% accuracy!



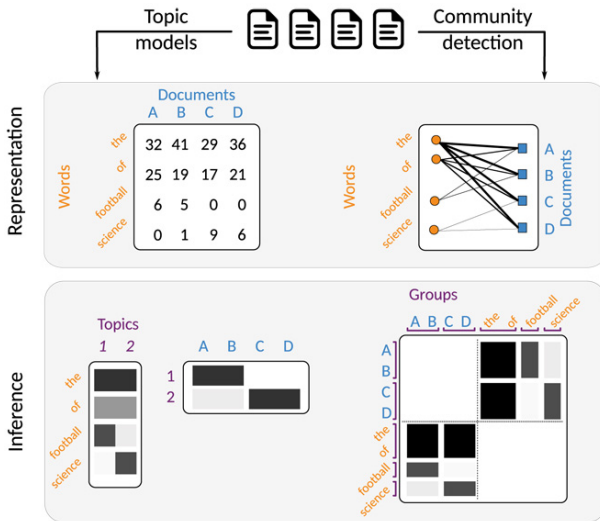
5



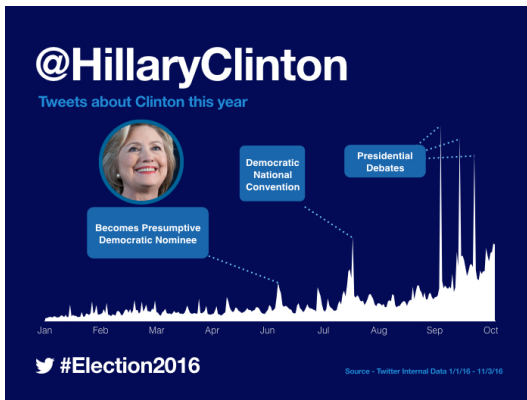
[<https://www.slideshare.net/jchoi7s/cs571-sentiment-analysis>]

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

6 Topic modelling

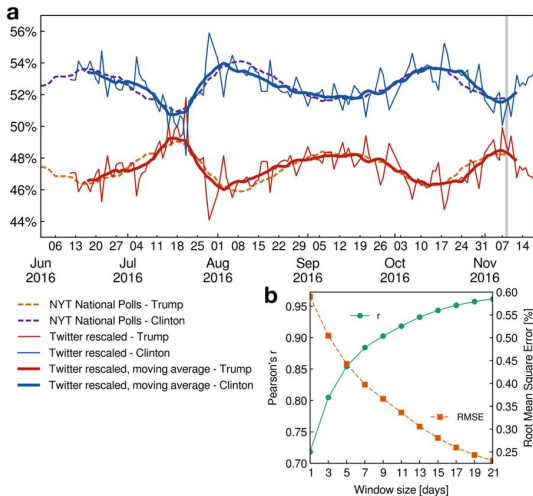


7 Analiza mediów społecznościowych



[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)]