# 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
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: $10^{15}$-$11^{45}$ oraz $16^{15}$-$17^{45}$,
- wykład: ogólny opis,
- laboratorium: konkretne przykłady w pakiecie `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?

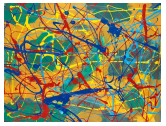
Z drugiej strony, warto zadać sobie pytanie **po co potrzebujemy eksploracji tekstu?** lub **jakie jest zadanie eksploracji tekstu?**. Ogólną odpowiedzią jest oczywiście: **aby (w automatyczny sposób) zrozumieć zawartość danego tekstu**...

## Po co text mining?

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

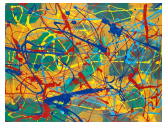
# Dlaczego analiza tekstu jest **trudna**?

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



# Dlaczego analiza tekstu jest **trudna**?

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

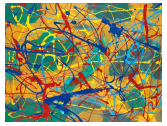


Time flies like an arrow.

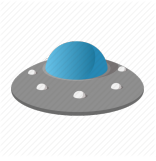
Niezliczone kombinacje subtelnych i abstrakcyjnych relacji pomiędzy pojeciami

# Dlaczego analiza tekstu jest **trudna**?

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



Time flies like an arrow.

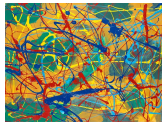


Niezliczone kombinacje subtelnych i abstrakcyjnych relacji pomiędzy pojeciami

Wiele sposobów opisywania tych samych pojęć

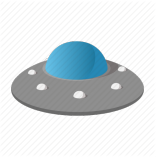
# Dlaczego analiza tekstu jest **trudna**?

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



Time flies like an arrow.

Niezliczone kombinacje subtelnych i abstrakcyjnych relacji pomiędzy pojeciami

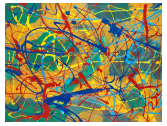


Wiele sposobów opisywania tych samych pojęć

Wysoka wymiarowość problemu

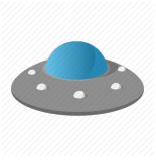
# Dlaczego analiza tekstu jest **trudna**?

Ciężko 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


Bardzo wiele cech (features)

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

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

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

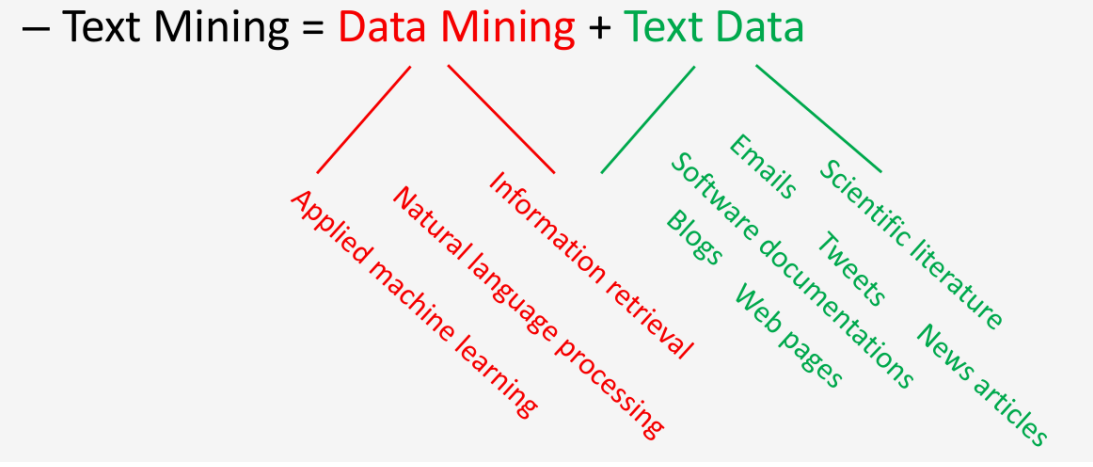
## Dlaczego analiza tekstu może być **łatwa**?

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

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

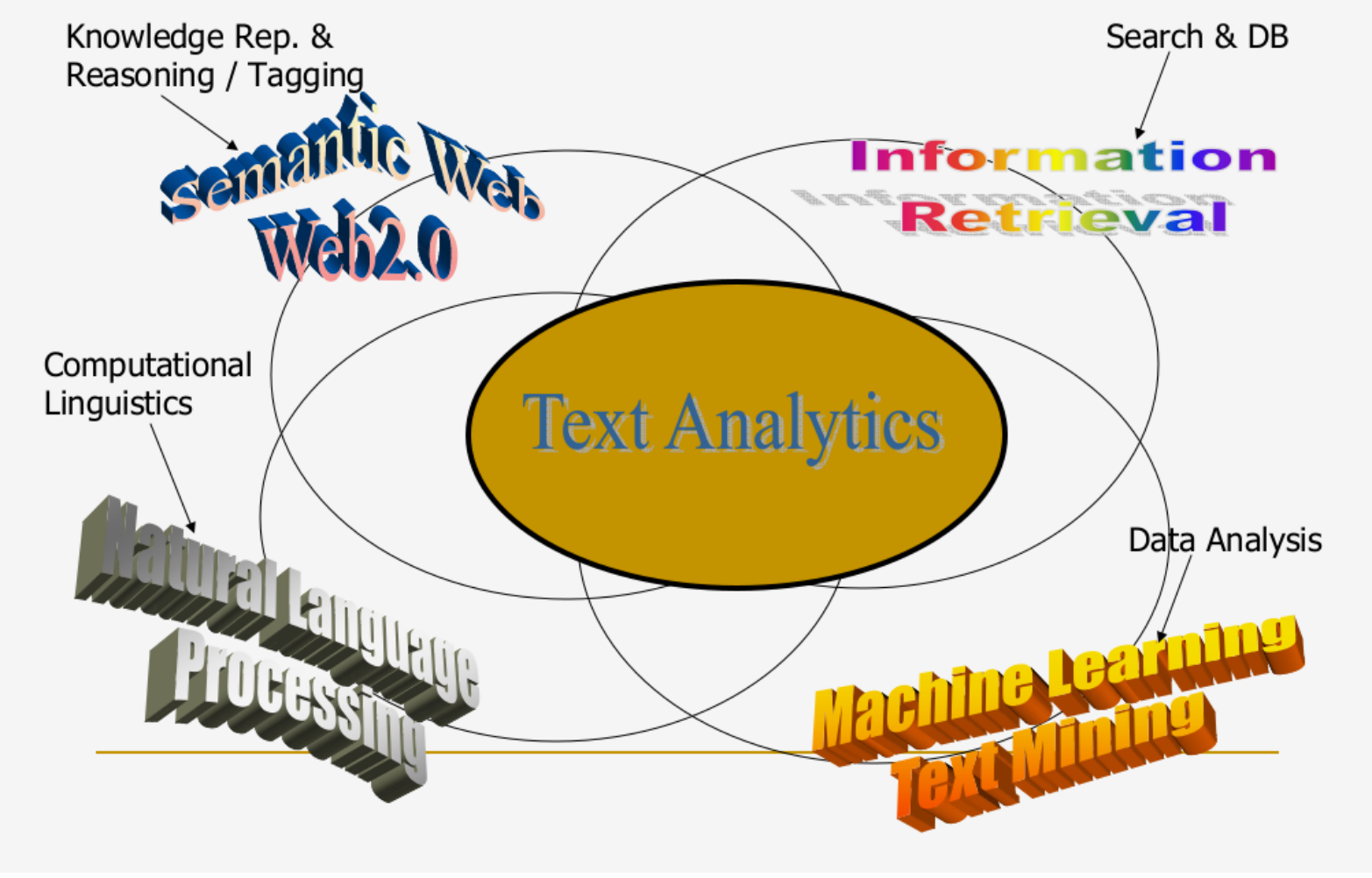
- wydobądź “istotne” wyrażenia,
- znajdź istotnie powiązane słowa,
- stwórz pewnego rodzaju podsumowanie dokumentów

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

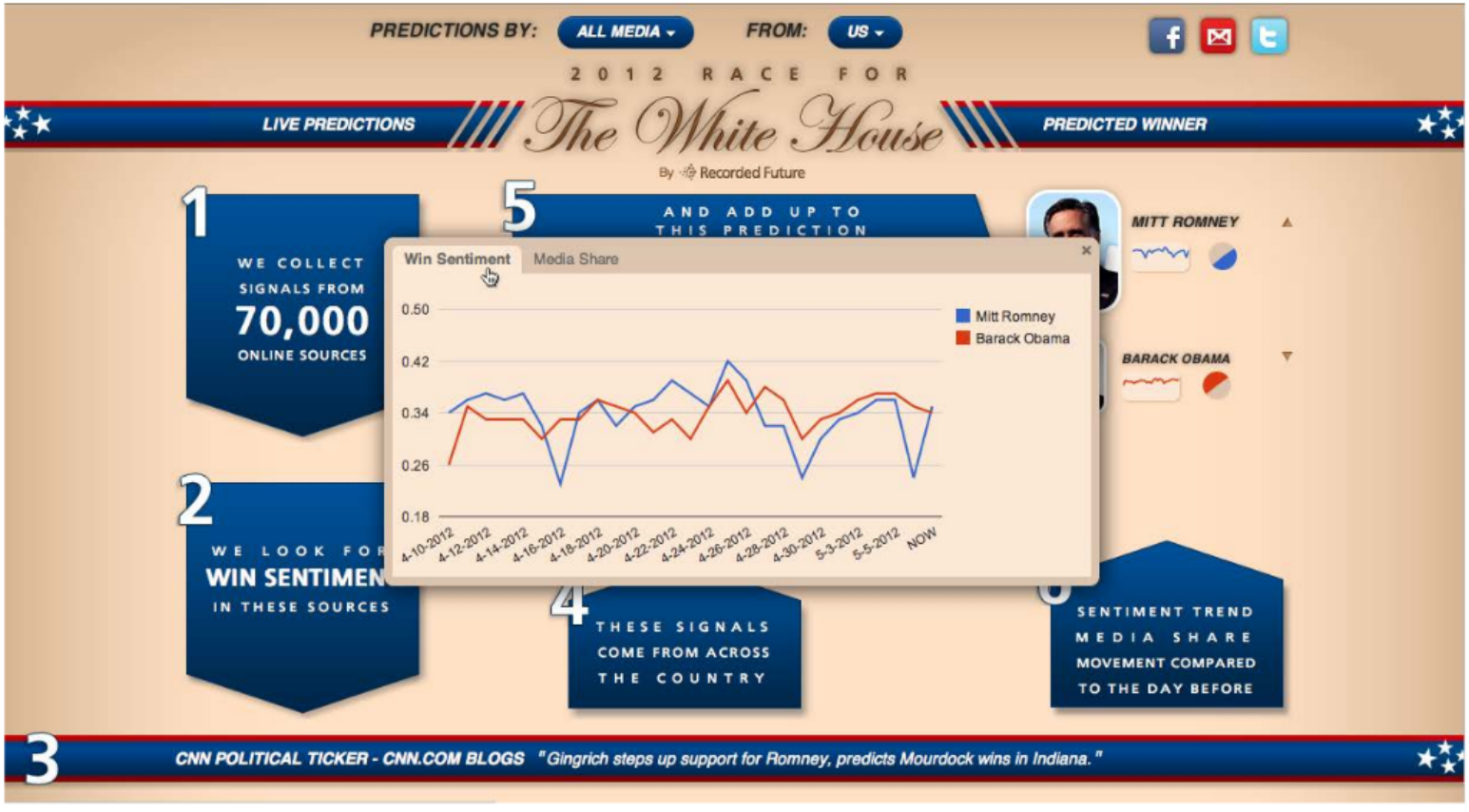



| | Finding Patterns | Finding “Nuggets” | |
|---|---|---|---|
| | | Novel | Non-Novel |
| Non-textual data | General data-mining | Exploratory | Database queries |
| Textual data | Comp Ling | Text Mining | Information retrieval |

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

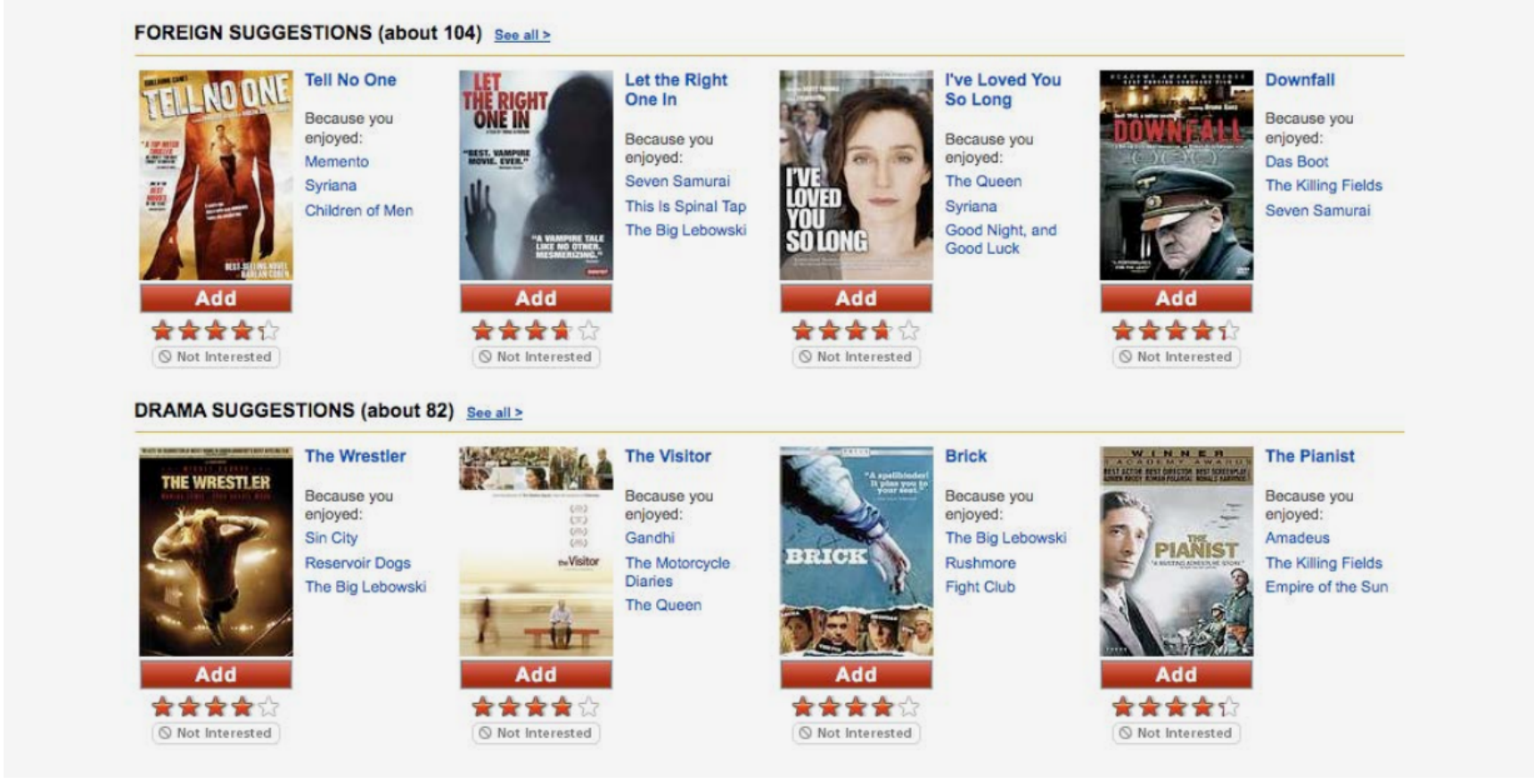


# Przykłady: analiza sentymentu – wybory

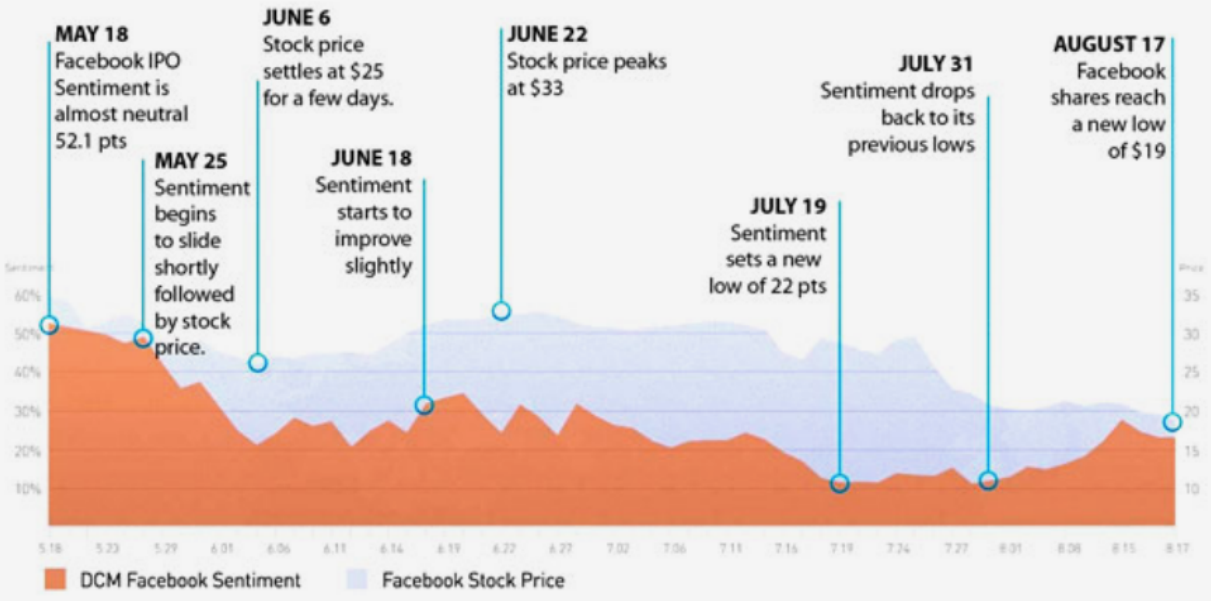


# Przykłady: podsumowywanie dokumentów

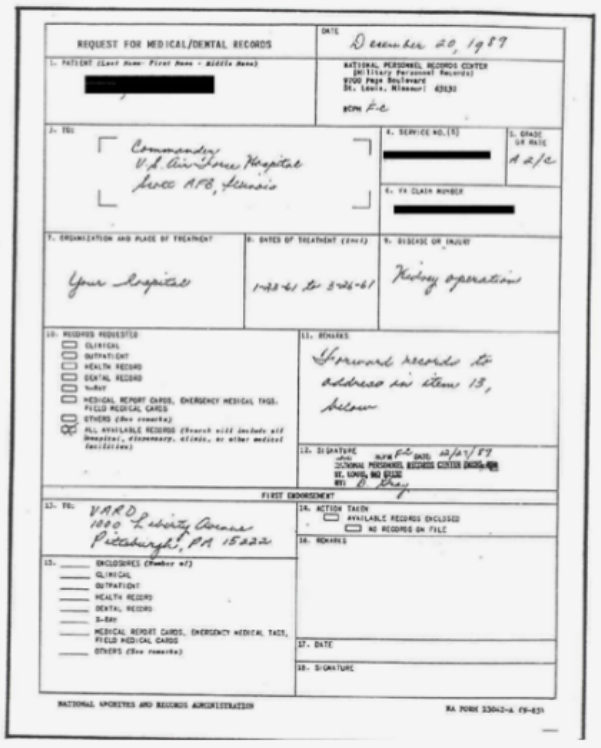
# Przykłady: systemy rekomendujące



# Przykłady: analiza tekstu w serwisach finansowych



# Przykłady: analiza danych medycznych



REQUEST FOR MEDICAL/DENTAL RECORDS

DATE December 20, 1989

1. PATIENT (Last Name - First Name - Middle Name) [redacted]

NATIONAL PERSONNEL RECORDS CENTER (Military Personnel Records) 9700 Page Boulevard St. Louis, Missouri 63132

SSN [redacted]

2. TO: Commanding V.A. Air Force Hospital Sioux Falls, Illinois

4. SERVICE NO. (S) [redacted]

5. GRADE OR RATE A A/C

6. VA CLAIM NUMBER [redacted]

7. ORGANIZATION AND PLACE OF TREATMENT: Your Hospital

8. DATES OF TREATMENT (Dates): 1-12-61 to 2-16-61

9. DISEASE OR INJURY: Kidney operation

10. RECORDS REQUESTED
- CLINICAL
- OUTPATIENT
- HEALTH RECORD
- DENTAL RECORD
- X-RAY
- MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS
- OTHERS (Give remarks)
- ALL AVAILABLE RECORDS (Search will include all Hospital, dispensary, clinic, or other medical facilities)

80

11. REMARKS: Forward records to address in item 13, below

12. SIGNATURE [redacted] DATE 2/27/89 TELEPHONE PERSONNEL RECORDS CENTER (MAIL) ST. LOUIS, MO 63132 RE: D. Ray

FIRST ENDORSEMENT

13. TO: VARO 1000 Liberty Avenue Pittsburgh, PA 15222

14. ACTION TAKEN
- AVAILABLE RECORDS ENCLOSED
- NO RECORDS ON FILE

16. REMARKS

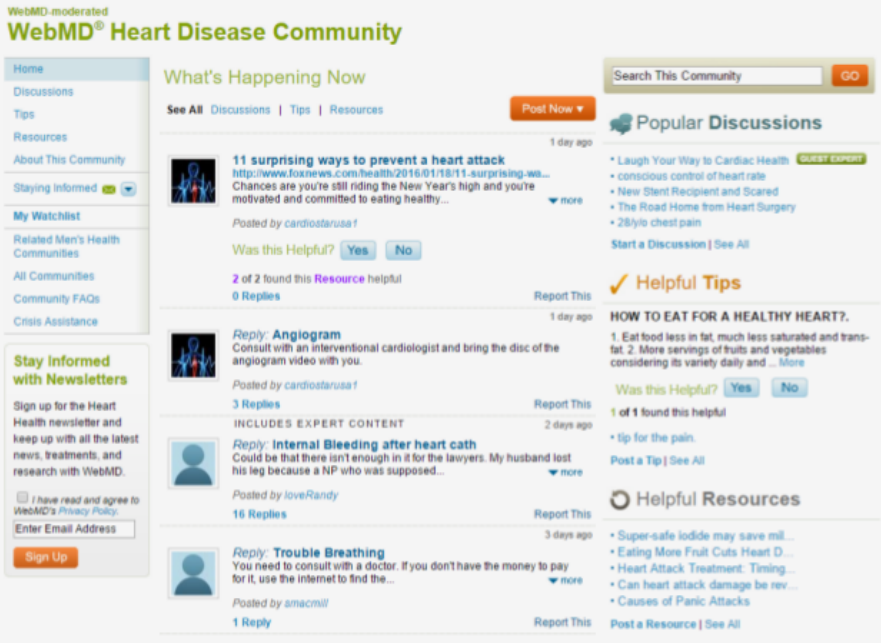
15. ENCLOSURES (Number of)
- CLINICAL
- OUTPATIENT
- HEALTH RECORD
- DENTAL RECORD
- X-RAY
- MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS
- OTHERS (Give remarks)

17. DATE

18. SIGNATURE

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION

NA FORM 13042-A (9-83)



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
Crisis Assistance

Stay informed with Newsletters

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

Post Now

1 day ago

11 surprising ways to prevent a heart attack
http://www.foxnews.com/health/2016/01/18/11-surprising-wa...
Chances are you're still riding the New Year's high and you're motivated and committed to eating healthy... more

Posted by cardiostarusa1

Was this Helpful? Yes No

2 of 2 found this Resource helpful

0 Replies Report This

1 day ago

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

Posted by cardiostarusa1

3 Replies Report This

INCLUDES EXPERT CONTENT 2 days ago

Reply: Internal Bleeding after heart cath
Could be that there isn't enough in it for the lawyers. My husband lost his leg because a NP who was supposed... more

Posted by loveRandy

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... more

Posted by amacmill

1 Reply Report This

Search This Community GO

Popular Discussions

- Laugh Your Way to Cardiac Health QUEST EXPERT
- conscious control of heart rate
- New Stent Recipient and Scared
- The Road Home from Heart Surgery
- 28/y/o 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? Yes No

1 of 1 found this helpful

- tip for the pain.

Post a Tip | See All

Helpful Resources

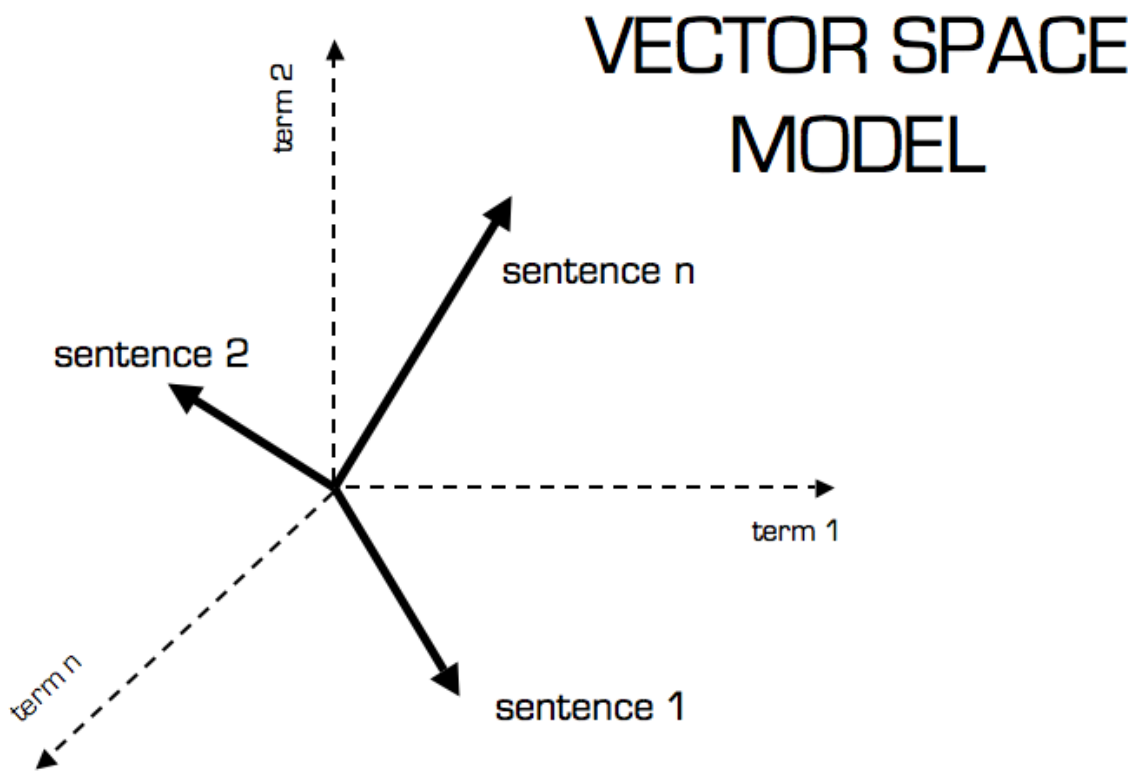
- Super-safe iodide may save mil...
- Eating More Fruit Cuts Heart D...
- Heart Attack Treatment: Timing...
- Can heart attack damage be rev...
- Causes of Panic Attacks

Post a Resource | See All

## Ogólny plan wykładu

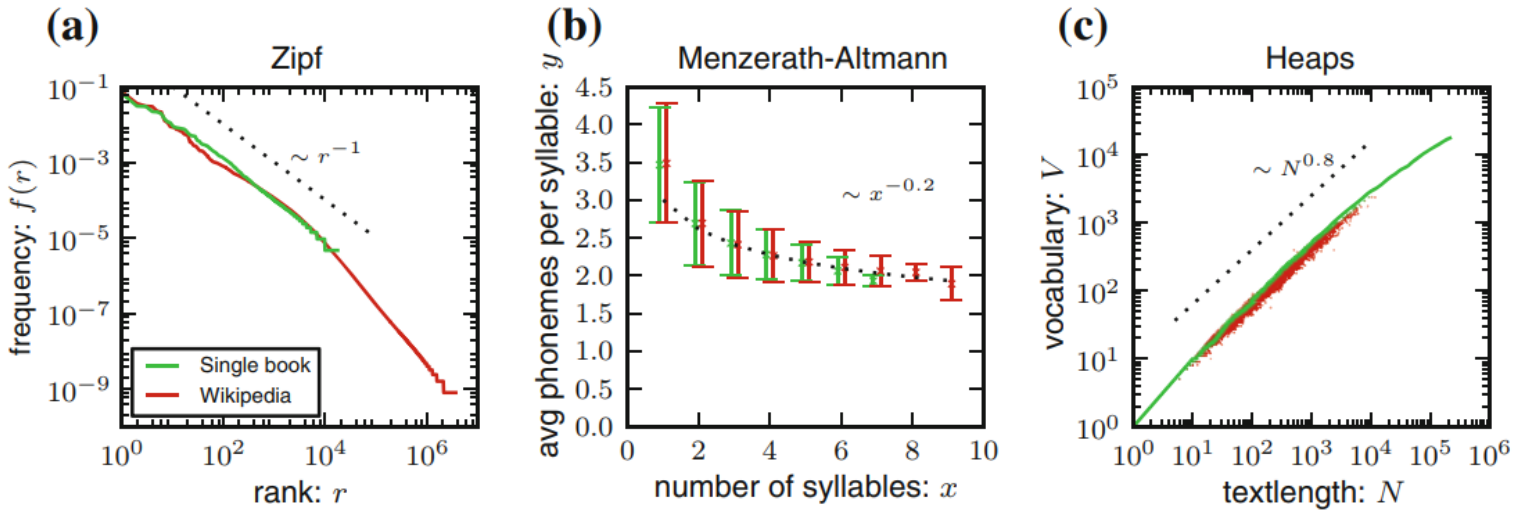
1. reprezentacja tekstu
2. prawo Zipfa
3. przetwarzanie języka naturalnego (NLP)
4. analiza sentymentu
5. topic modeling
6. analiza mediów społecznościowych

2. Reprezentacja tekstu...



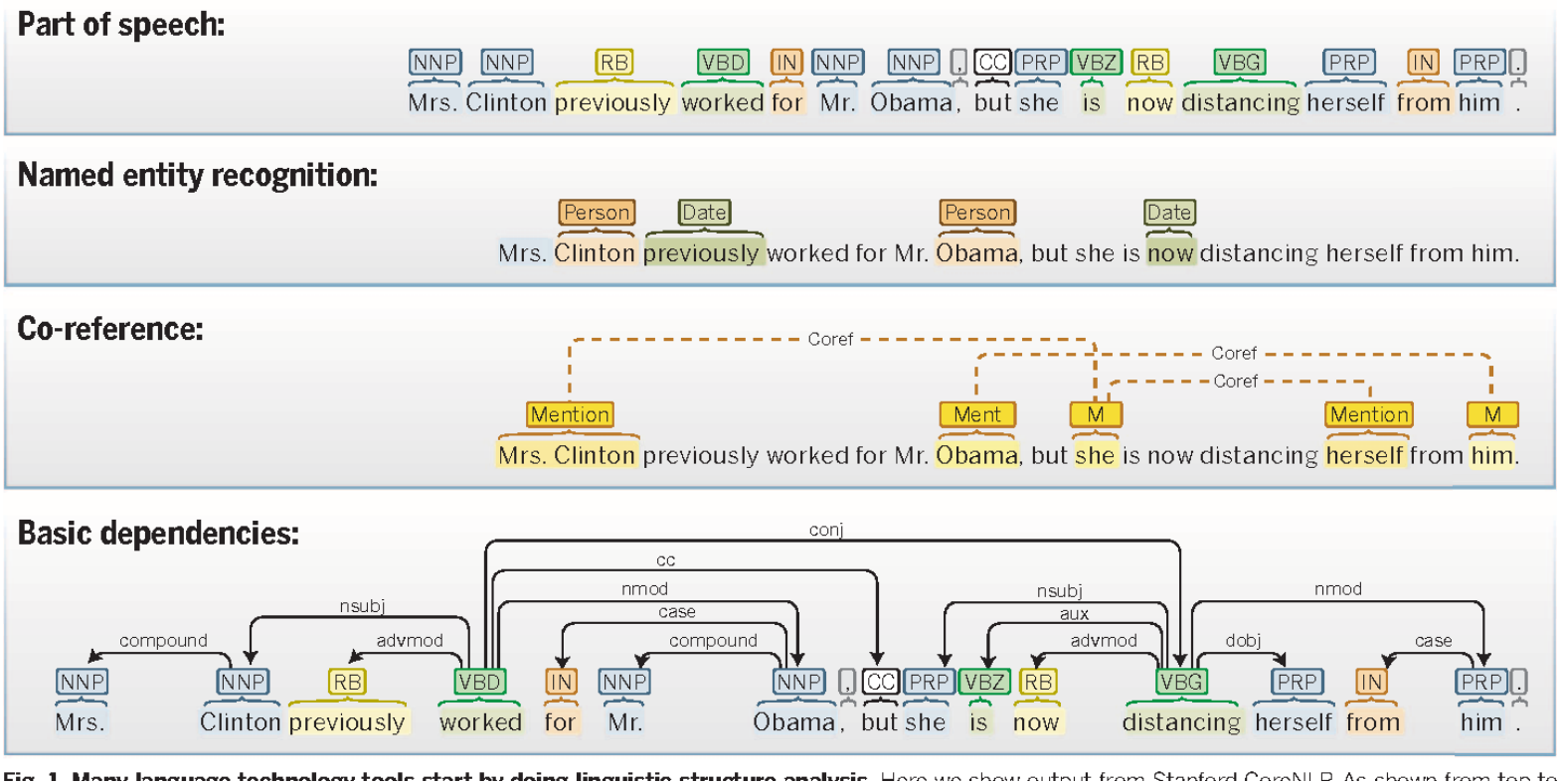

[http://bitsearch.blogspot.com/2011/01/vector-space-model-for-scoring.html]

## 3 Prawo Zipfa i pokrewne...



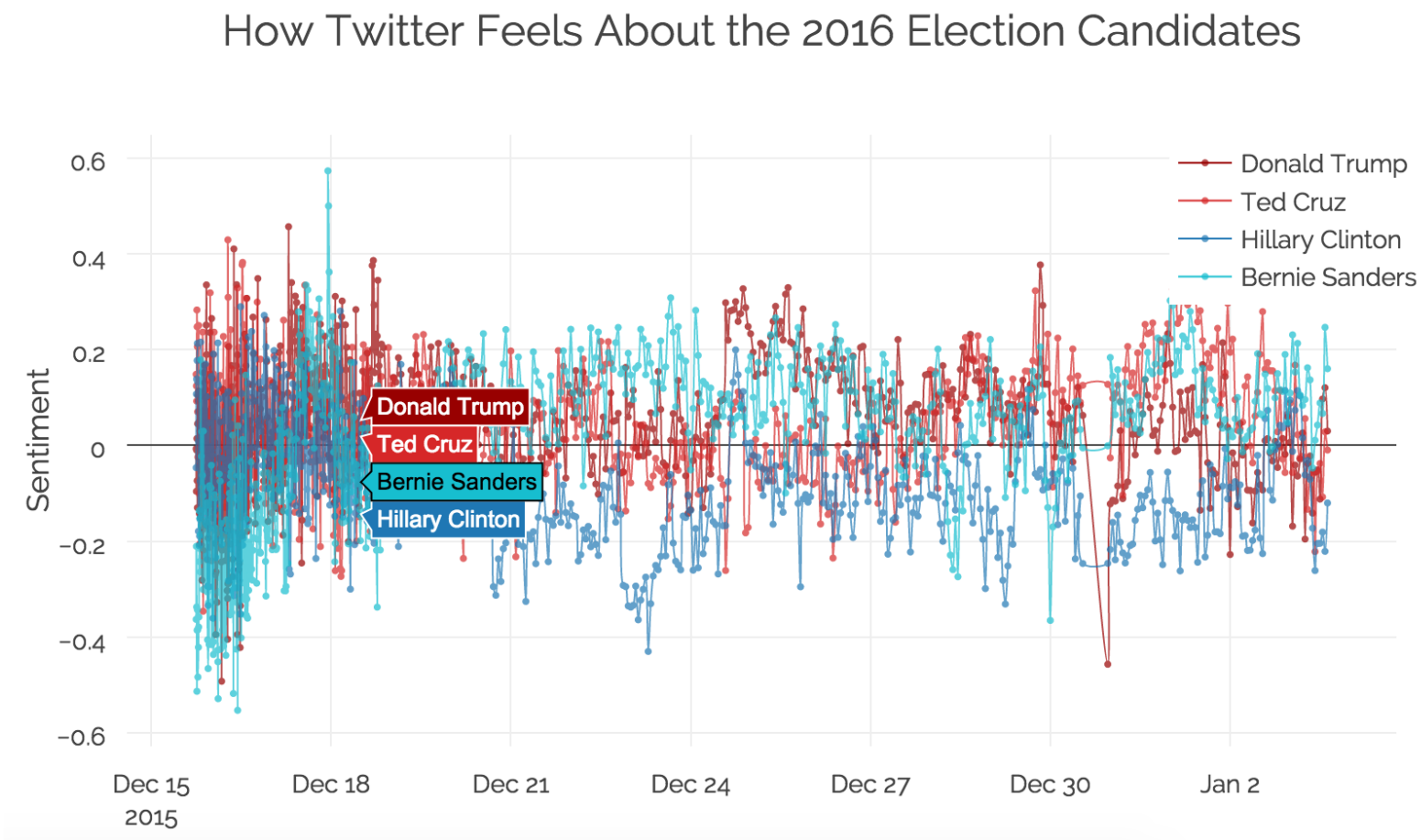

[Altmann, Gerlach, Statistical laws in Lingustics, Creativity and Universality in Language, Springer (2017)]

4 przetwarzanie języka naturalnego (NLP)




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

## 5 Analiza sentymentu




[Blei, Probabilistic topic models, Communications of the ACM 55 (4), Apr 2012]

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

### Dictionary-based Approach

Create lists of positive/negative words (phrases).

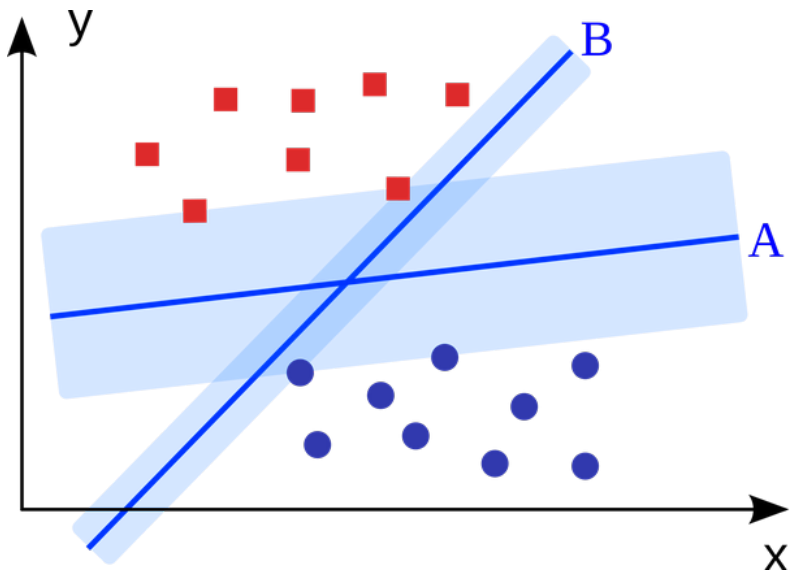
| Negative | Positive |
|---|---|
| suck | dazzling |
| terrible | brilliant |
| awful | phenomenal |
| unwatchable | excellent |
| hideous | fantastic |

Sentiment = |Positive words| - |Negative words|

Around 65% accuracy!

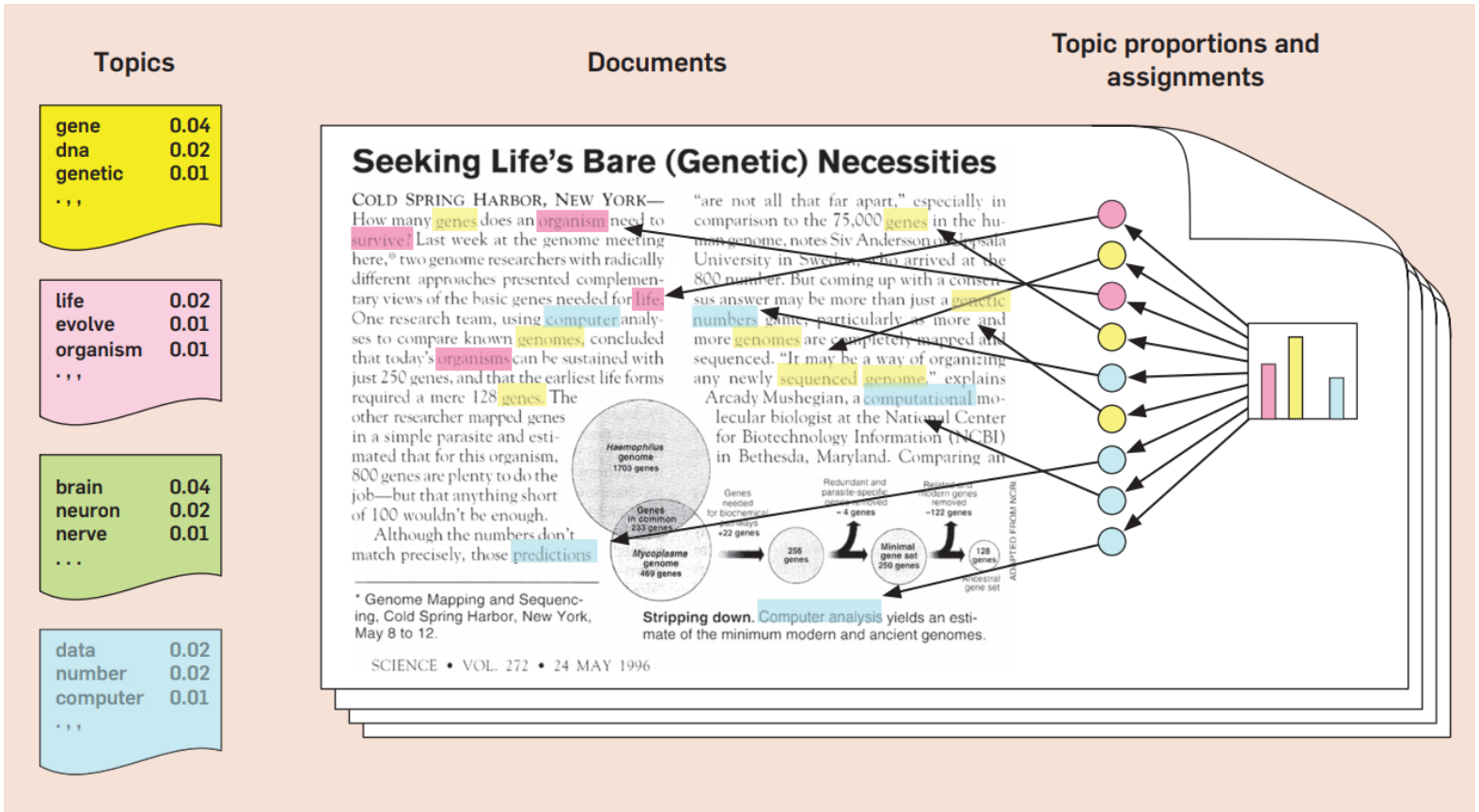
5

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



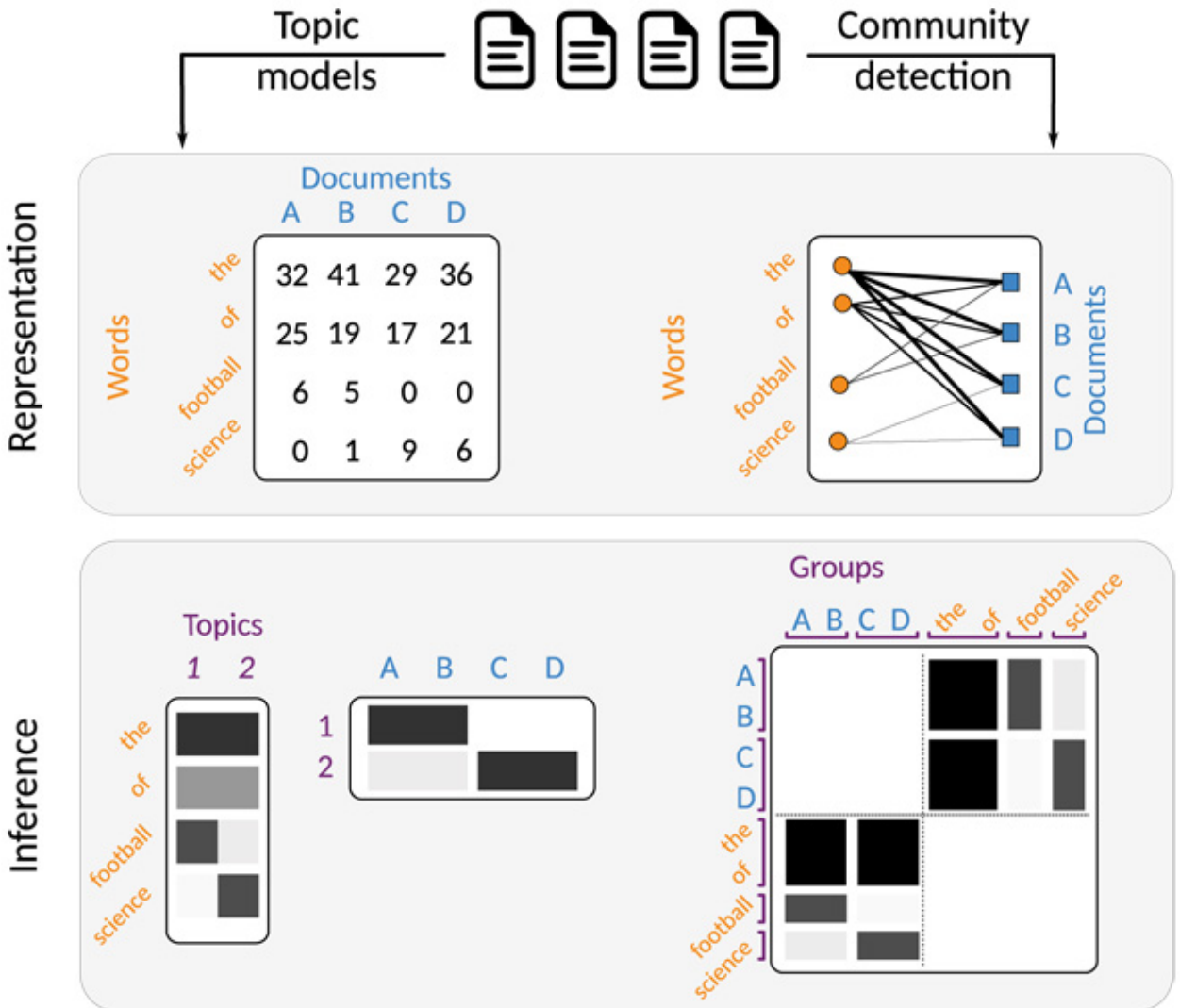

[https://medium.com/nlpython/sentiment-analysis-analysis-part-2-support-vector-machines-31f78baeee09]

# 6 Topic modelling



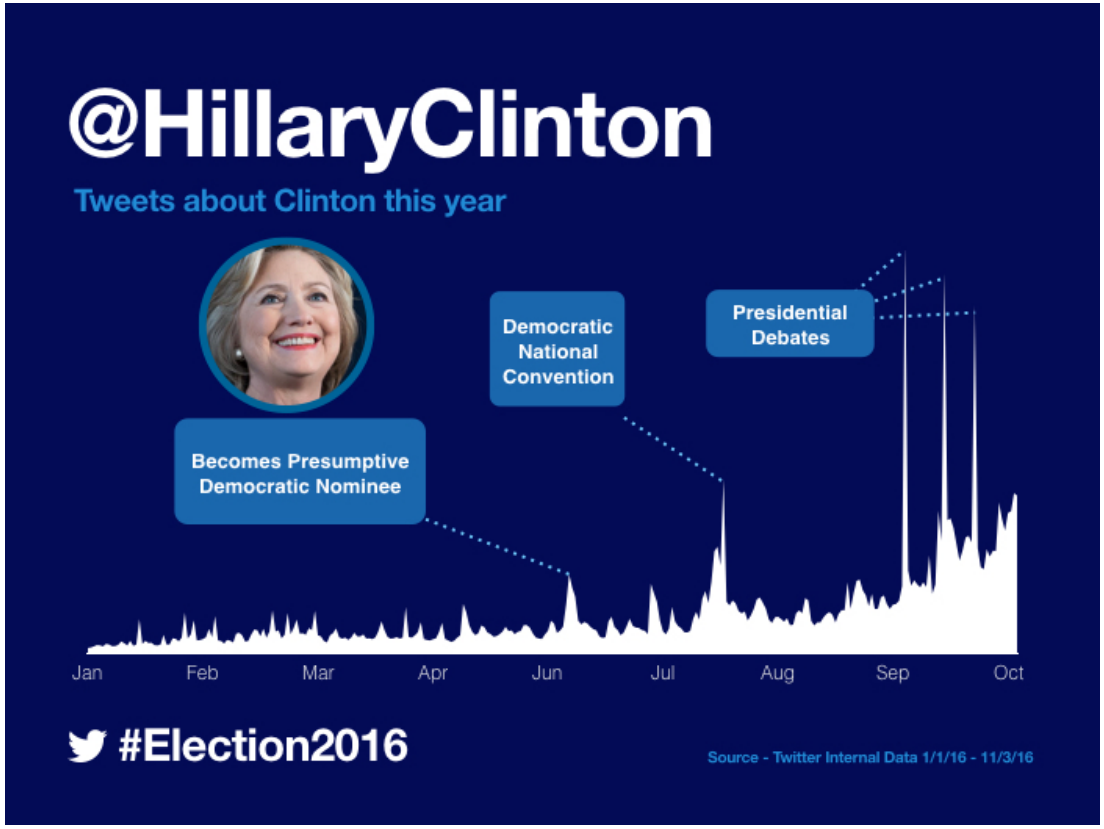

[Blei, Probabilistic topic models, Communications of the ACM 55 (4), Apr 2012]

# 6 Topic modelling



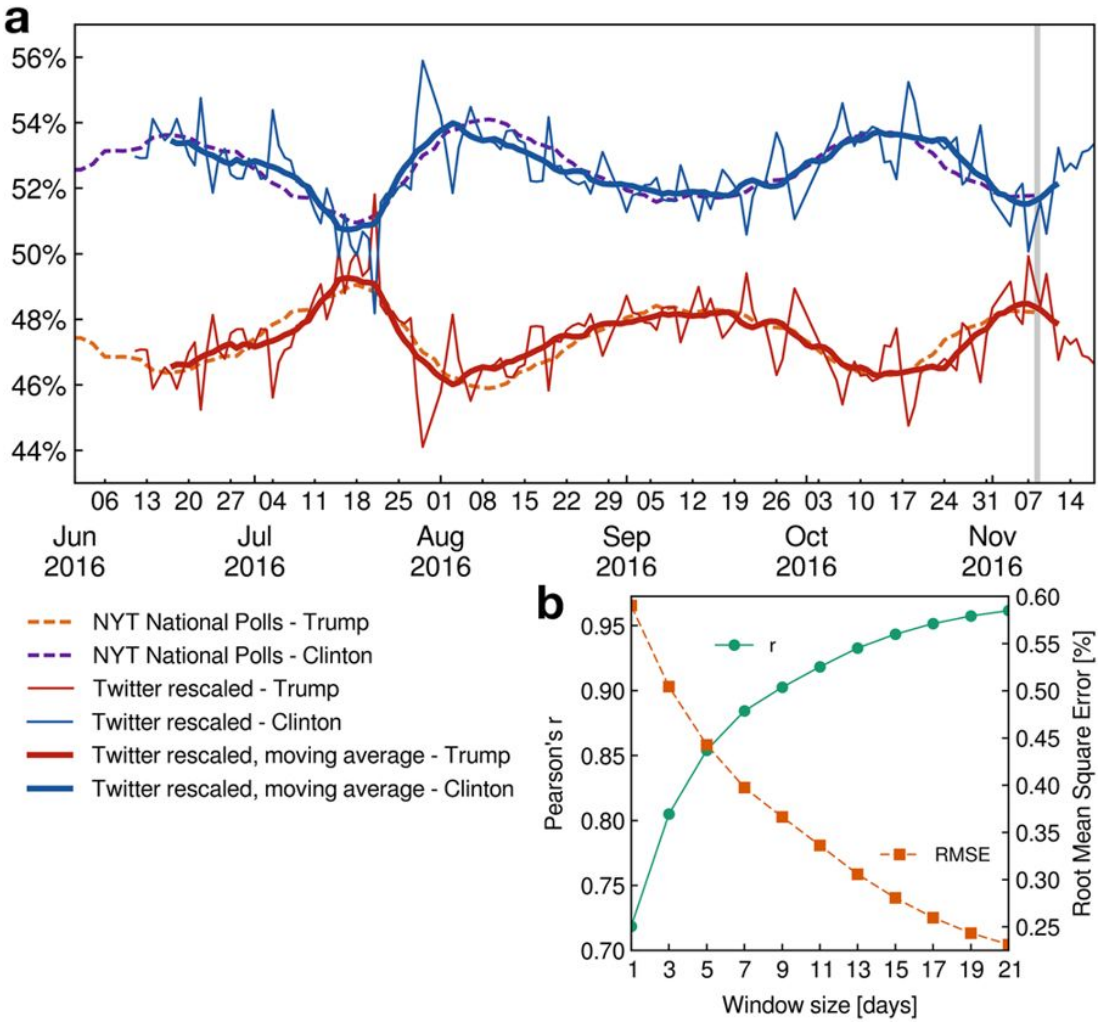

[Gerlach, Peixoto, Atlmann, A network approach to topic models, Science Advances (2018)]

7 Analiza mediów społecznościowych




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

# 7 Analiza mediów społecznościowych




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