Modeling Health-Related Topics in an Online Forum Designed for the Deaf & Hard of Hearing

1st XJTLU Research Symposium on Healthy Ageing & Society,
Xi'an Jiaotong-Liverpool University, Suzhou, 14 Dec, 2015

Hang Dong, CSSE, Xi'an Jiaotong-Liverpool University
Biyang Yu, School of Information, Florida State University

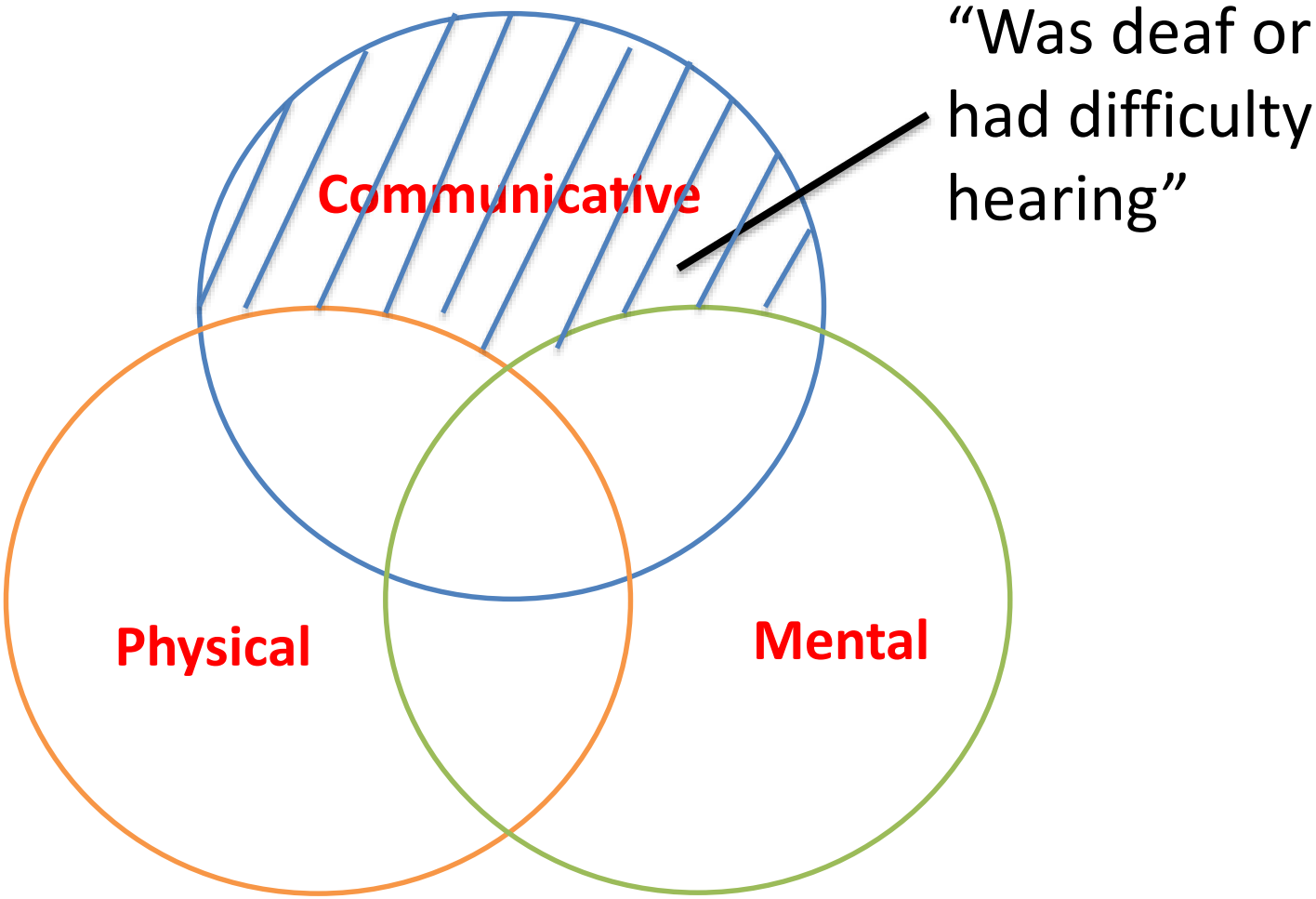
Xi'an Jiaotong-Liverpool University
西交利物浦大学

Email: Hang.Dong@xjtlu.edu.cn
Last edited on 9th Jan 2016 by Hang Dong.

Content

- Background and research question
- Topic modeling as a computational or quantitative method to analyze social media discourse
- Topic modeling vs. human coding (thematic analysis)

Definition for disabled people (15+) [Disability Status]



This graph is illustrated according to the definition of disability status proposed in the report by U.S. CENSUS BUREAU. Americans With Disabilities: 2010, issued on 2012. http://www.census.gov/prod/2012pubs/p70-131.pdf

Online communities for deaf & hard of hearing

- "Active users" of online communities (Snunith & Meital, 2012).

- Motivated because of (Snunith & Meital, 2012):
  - Easy communication;
  - Equality and empowerment;
  - Social Support.

Online communities for health related issues

- "A social life of health information" (Pew Research Center, 2011)

The Social Life of Health Information, 2011. The Pew Research Center. http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/

"…there is a social life of health information, as well as peer-to-peer support, as people exchange stories about their own health issues to help each other understand what might lie ahead." (Pew Research Center, 2013)

Health Online 2013, Susannah & Maeve, in Pew Research Center. http://www.pewinternet.org/2013/01/15/health-online-2013/

Image from © Columbia Business Times 2016 http://columbiabusinesstimes.com/2011/05/27/health-nonprofits-effectively-using-social-media/

Question

- What health issues are concerned/shared by deaf & hard-of-hearing in an online forum?

- Quantitative (computational) vs. qualitative (manual) methods, which one is more suitable? or both?

Dataset

- Dataset:
  - Alldeaf, the leading US online community for deaf and hard-of-hearing.
  - All threads the section "*Lifestyle, Health, Fitness & Food*".

  - 80650 posts in 3772 threads created by 1829 users , 2003-2015, 2.3m words. [Quantitative method: Topic Modeling]
  - Manually selected 559 threads, 450k words, related to health inquiries (Biyang, Jongwook & Hang, 2015). [Quantitative + Qualitative: Human Coding]

New Thread

Threads in Forum : Lifestyle, Health, Fitness & Food

	Thread / Thread Starter
	What's Your Dinner Tonight -- Part II (1 2 3 4 5 6 7 8 9 10 11 12 13) [redacted]
	What favorite childhood ice cream [redacted]
	Brain shock [redacted]
	Birth control packaging error leads to lawsuit [redacted]
	15 Of the fattest USA states [redacted]
	Georgia officer helps injured runner cross finish line [redacted]
	Fighting back against Parkinson's [redacted]
	How young adults brains develops [redacted]
	Pork chopped from federal prison menus [redacted]

07-30-2015, 09:55 PM #1

[redacted]
Registered User

Join Date: Nov 2004
Location: [redacted]
Posts: 681
Likes: 0
Liked 13 Times in 8 Posts

Seizure Disorder

A few months back, I had an EEG done to check my brainwaves. The EEG came back saying that I have a seizure disorder when I never had visual symptoms. This news shocked me because I thought that all seizures come with convulsions. I started anti-seizure medication because there is a family history of seizures. The first few days after the diagnosis, I was very upset. I have come to terms with it.

RIP Kyle Jean-Baptist (died August 29, 2015)

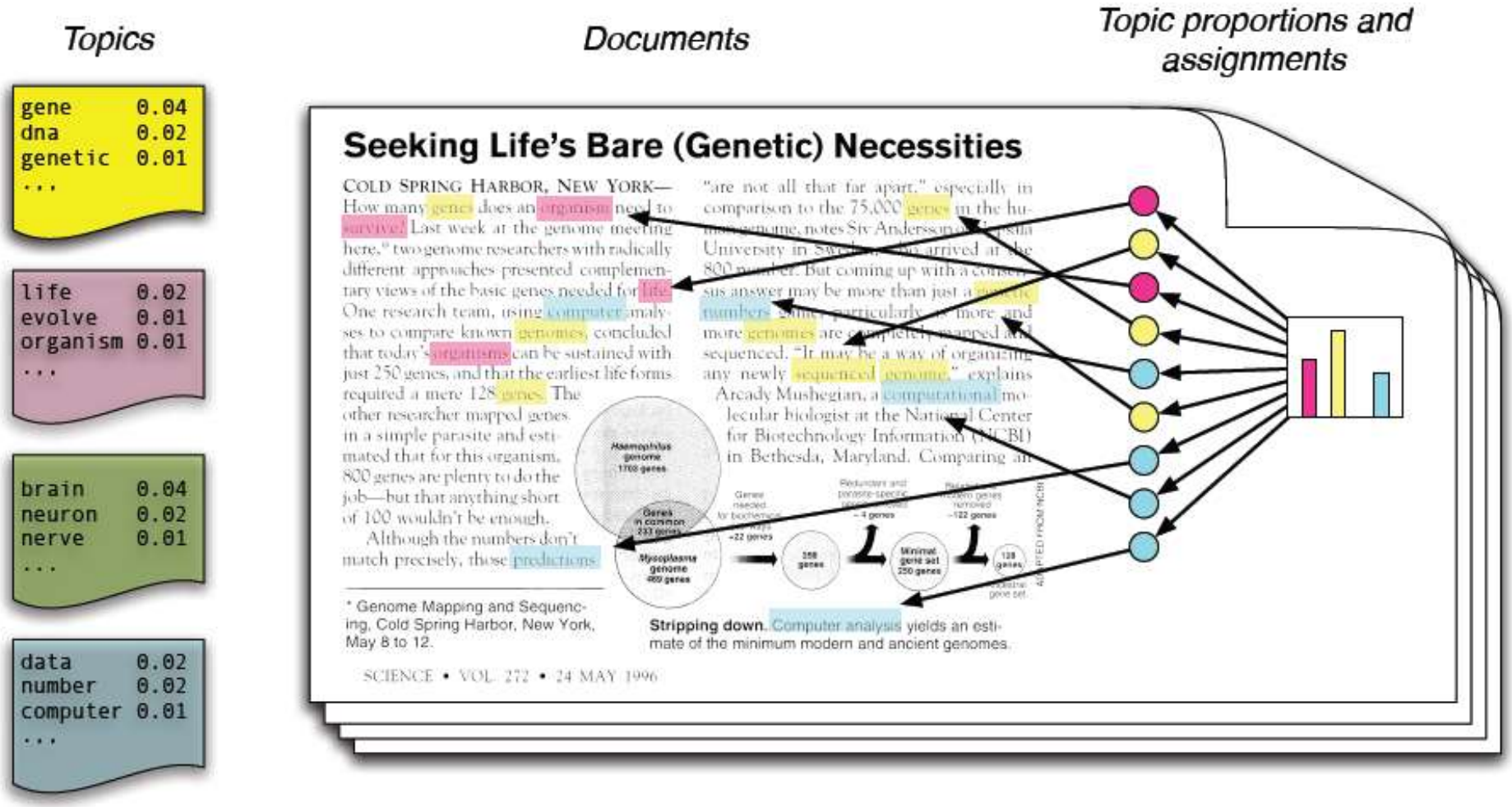
Quantitative Method: Topic modelling

Topic models uncover the hidden thematic structure in document collections; can help develop new ways to search, browse and summarize large archives of texts.
(David M. Blei)

Latent Dirichlet Allocation (David, Andrew & Michael, 2003):

- (1) Input: documents → weighted word-document matrix. [not limited to documents, e.g. genetic data]
- (2) Output: p (words | topics), p (topics | docs).
- (3) Unsupervised learning, no need to specify the meaning of the topics first;
- (4) Based on word co-occurrences, but also can handle <u>polysemy</u> and <u>synonymy</u>.

Generative model: Latent Dirichlet Allocation



Acknowledgement to the image on Scott Weingart's blog, http://www.scottbot.net/HIAL/?p=221

Why use a topic model for classification?

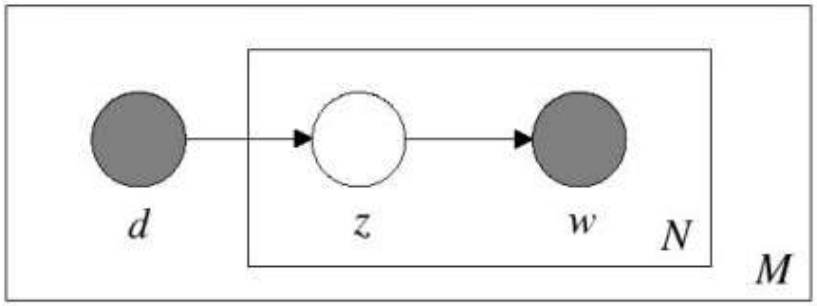
- Topic models help handle <u>polysemy</u> and <u>synonymy</u>
  - The count for a topic in a document can be much more informative than the count of individual words belonging to that topic.
- Topic models help combat <u>data sparsity</u>
  - You can control the number of topics
  - At a reasonable choice for this number, you'll observe the topics many times in training data

    (unlike individual words, which may be very sparse)

This slide is **only for your further reference**: **copied** from page 113 in "LSA, pLSA, and LDA Acronyms, oh my!" by Thomas et al., 2011. http://knight.cis.temple.edu/~yates/cis8538/sp11/slides/intro-to-lsa-lda.ppt

From pLSA to LDA

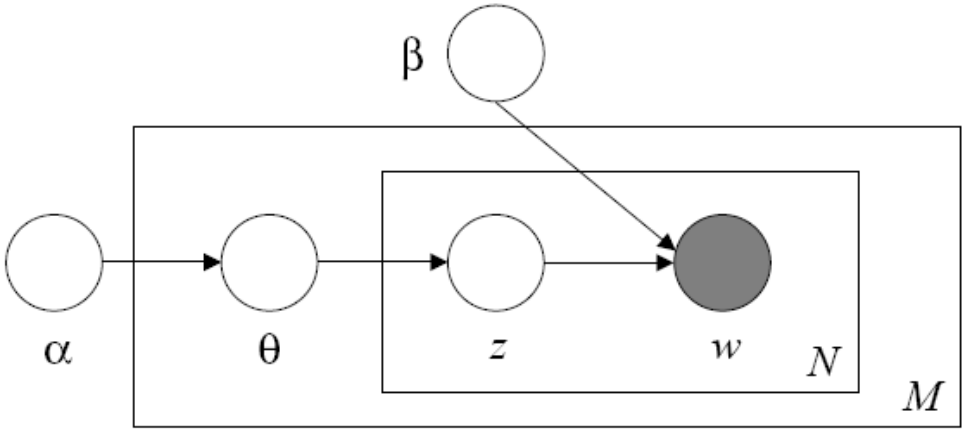
pLSA



$$P(d, z, w) = P(d)P(w|z)P(z|d)$$

$$P(d, w) = \sum_z P(z)P(d|z)P(w|z)$$

LDA



Steps for topic modeling

- Data extraction: web crawling
- Data **preprocessing**:
  - Lemmatization

    (illnesses → illness, recovered → recover)
  - Remove stop words (too high frequent "the", "is"; too low frequent <= 5)
  - [2.2m words after preprocessing, 17,050 distinct words]
- Running the algorithm ($K = 100$; $\alpha = 5/K$; $\beta = 0.1$; $Tw = 20$)
- <u>Manual labelling of the topics</u>

Original text:

Have any of you heard Waardenburg Syndrome before?? Explain?? Or how did u know about it or have u see anyone have them? I do have one since I born.. Many deafies never heard it before... I'll say 95% of deaf people have them and 5% of hearing people have them. I am curious everyone's saying in this thread.. If u never heard it before.. The link u can check out at: http://www.nidcd.nih.gov/health/hearing/waard.asp

After preprocessing:

waardenburg syndrome explain bear deafies deaf people people curious everyone say thread link check

Tools for LDA topic modeling

- **JgibbLDA** (we used this)
- **GibbsLDA++**
- **MALLET Toolkit from UMass**
- Matlab Topic Modeling Toolbox 1.4
- R package

Topics & manual labelling

(with all threads: 39% health, 61% food, lifestyle, others)

```
Topic 5th:
    cancer 0.17974747087409101
    woman 0.03186942668498003
    breast 0.029986962190352025
    skin 0.016112501655871546
    tan 0.011580642687322647
    risk 0.011162317244071198
    hpv 0.011162317244071198
    mammogram 0.010813712708029756
    sun 0.009419294563860864
    test 0.009000969120610196
    cell 0.008303760048525751
    prostate 0.007745992790858194
    men 0.00767627188364975
    die 0.007467109162024416
    cervical 0.007118504625982193 5
    pap 0.006839620997148414 4
    tumor 0.006630458275523081
    treatment 0.006351574646689302
    age 0.006142411925063968 5
    gardasil 0.004747993780895077
```

cancers and treatment

```
Topic 46th:
    ear 0.08430161669819217
    hear 0.04218705718748558
    tinnitus 0.028939754430596657
    vertigo 0.013253893454777926
    infection 0.012924358560327953
    doctor 0.012331195750318002
    sound 0.011408498045858076
    dizzy 0.009365381700268242
    loss 0.009365381700268242
    aid 0.008508590974698 31
    ring 0.007849521185798364
    head 0.007519986291348391
    help 0.006597288586888465 5
    meniere 0.006399567650218482
    ent 0.006399567650218482
    wax 0.006333660671328486
    leave 0.006004125776878514
    day 0.005872311819098524
    deaf 0.005674590882428 54
    loud 0.005674590882428 54
```

conditions of ear and head

Topics & manual labelling (with 559 threads)

```
Topic 0th:
    cancer 0.05629627507984018
    old 0.01662680624973148
    drive 0.016197173013304308
    woman 0.011327996333796383
    age 0.011327996333796383
    family 0.010182307703323928
    breast 0.009609463388087701
    live 0.007461297205951853
    mammogram 0.007461297205951853
    mother 0.007031663969524682
    mom 0.006888452890715626
    brother 0.006602030733097512 4
    sperm 0.006315608575479 4
    handicap 0.005313131023816003
    test 0.004597075629770719
    aid 0.004597075629770719
    elderly 0.004597075629770719
    father 0.004453864550961663
    die 0.004453864550961663
    aunt 0.004310653472152606
```

mammary/breast cancer for old people

```
Topic 2th:
    shot 0.06024455217830138
    flu 0.04982047421281666
    child 0.018382778761354798
    vaccine 0.016562701656270166
    school 0.01093700878600857
    vaccination 0.010275162565977795
    sick 0.00961331634594702
    virus 0.008124162350877774
    require 0.006965931465823917
    health 0.006635008355808528
    immunization 0.006469546800800834
    hepatitis 0.006469546800800834
    food 0.00630408524579314
    die 0.004980392805731588
    hand 0.0048149312507238945
    law 0.0048149312507238945
    department 0.004649469695716201
    parent 0.004484008140708507
    vaccinate 0.004484008140708507
    h1n1 0.004484008140708507
```

vaccination and illness

```
Topic 7th:
    deaf 0.059621975635413486
    people 0.024443765765154342
    deafness 0.01629961428393679
    read 0.014489802843666223
    disease 0.012453764973361836
    lip 0.009739047812955986
    loss 0.008494802447769971
    understand 0.007589896727634688
    sign 0.007363670297600868
    mental 0.006798104222516316
    call 0.006684991007499406
    language 0.006684991007499406
    sound 0.005780085287364123
    hoh 0.005440745642313391
    phone 0.005214519212279571
    gp 0.00510140599726266
    group 0.004875179567228839
    themselves 0.004535839922178108
    asl 0.004535839922178108
    speech 0.004535839922178108
```

Conversation issues for deaf & HoH

```
Topic 10th:
    kidney 0.03986341412939442
    stone 0.035465762992472255
    drink 0.028739943606591302
    water 0.02020332669374 24
    pain 0.012960136585870606
    soda 0.012960136585870606
    juice 0.008562485448948444
    urine 0.007786429365962181
    cranberry 0.007269058643971338
    uti 0.006751687921980 4956
    belly 0.006234317199989653
    operation 0.005975631838994232
    infection 0.005199575756007968
    button 0.005199575756007968
    painful 0.004940890395012547
    bladder 0.004682205034017126
    gp 0.004423519673021704
    yeast 0.004164834312026283
    foxrac 0.004164834312026283
    pop 0.003647463590035 44
```

Kidney stones and treatment

A selected list of manual labelled topics: from TM

mammary cancer / breast cancer

exercise and pain treatment

vaccination and illnesses

diabetes and medication

smoking and second hand smoking

sinus, noise related illnesses and organ transplantation

migraine and medication

deafness and family doctor

skin diseases

mental illnesses

What do the topics mean?

- Essence of the text (Carina, Wouter & Kaspar, 2015)
- Issues or “voice”, important things (DiMaggio, Nag, and Blei, 2013)
- A categorization or “frame” (DiMaggio, Nag, and Blei, 2013; Carina, Wouter & Kaspar, 2015)
- Evidence (Andrew & Ted, 2012)
- Events during a certain period (for journalism, politics, bibliometrics)

Qualitative Method: human coding

- Using a theoretical model, or a framework, to manually categorize the discourse on social media.

- Framework: **sixteen categories suggested by MedlinePlus** (http://www.nlm.nih.gov/medlineplus/healthtopics.html)
- Select 559/3772 threads manually which are related to health inquiries.

Results of human coding

Health Concerns	Numbers of Questions
Ear, Nose, and Throat	91 (16.3%)
Mental Health	76 (13.6%)
Female Reproductive System	47 (8.4%)
Digestive System	36 (6.4%)
Eyes and Vision	28 (5.0%)
Skin, Hair and Nails	22 (3.9%)
Substance Abuses	17 (3.0%)
Lungs and Breathing	13 (2.3%)
Mouth and Teeth	13 (2.3%)
Endocrine System	13 (2.3%)
Immune System	10 (1.8%)
Kidneys and Urinary System	8 (1.4%)
Nutrition	8 (1.4%)
Male Reproductive System	2 (0.4%)
Others	67 (12.0%)
Total	559

Back to the result of TM method

mammary cancer / breast cancer

exercise and pain treatment

vaccination and illnesses

diabetes and medication

smoking and second hand smoking

sinus, noise related illnesses and organ transplantation

migraine and medication

deafness and family doctor

skin diseases

mental illnesses

Compared to professional topics

WebMD Health A-Z Directory Common Topics

- de Lange Syndrome
- De Morsier Syndrome
- De Santis Cacchione Syndrome
- Deafness
- Deafness and Pili Torti, Bjornstad Type
- Deafness-Dwarfism-Retinal Atrophy
- Deafness-Functional Heart Disease
- Dealing With Emergencies
- Dealing with Gestational Diabetes
- Dealing with Low Blood Sugar (Hypoglycemia)
- Debrancher Deficiency
- Debre's Syndrome
- Deciduous Skin
- Decubitis Ulcers

http://www.webmd.com/a-to-z-guides/health-topics/d.htm

Can TM substitute HC?

- No, it can only assist human/manual coding.
- Human coding: flexible, dynamic, can use specific coding schemes, easier to make sense.
- Topic modeling: objective, immediate, suitable for huge volume of data, reproducible.

Take home message

- Topic modeling:
  - a useful method to analyze social media discourse
  - can be used to get the essence (issues and categorization) of a large volume of data
  - unsupervised, based on probability, detecting the co-occurrence of words

- TM cannot substitute qualitative methods.
- Carry out both to attain a fuller image.

Key References

- Blei. "Topic Modeling", https://www.cs.princeton.edu/~blei/topicmodeling.html Retrieved Dec 11, 2015
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993-1022.
- DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, *41*(6), 570-606.
- Goldstone, A., & Underwood, T. (2012). What can topic models of PMLA teach us about the history of literary scholarship. *Journal of Digital Humanities, 2*(1), 39-48.
- Jacobi, C., van Atteveldt, W., & Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 1-18.
- Shoham, S., & Heber, M. (2012). Characteristics of a virtual community for individuals who are d/deaf and hard of hearing. *American Annals of the deaf*,*157*(3), 251-263.
- S. Weingart, *"Topic Modeling and Network Analysis"*, http://www.scottbot.net/HIAL/?p=221 Retrieved Dec 7, 2015
- Yu, B., Lee, J. & Dong, H. (2016). Health Information Seeking Behavior of Individuals with Hearing Loss in an Online Community. *In iConference 2016 Proceedings.* (accepted)

Q & A

Thank you for your attention