

Topic Models als sinologisches Hilfsmittel: Möglichkeiten und Grenzen

Gesa Stupperich

Universität Heidelberg, Asia and Europe in a Global Context

25. Januar 2018

Hierüber möchte ich sprechen

- Dokumenten-Clustering mit Topic Models
- Projekt: Visualisierung von Topic Models von Verwaltungshandbüchern aus der Qing-Zeit (1661–1911)
- ⇒ Hilfreiches Werkzeug zur Erschließung größerer Quellensammlungen?
- Vorläufiges Fazit zu Möglichkeiten und Grenzen

Topic Models: Verfahren

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
organism 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 **genes**, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a geneticist at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a **concise** answer may be more than just a **simple** numbers game, particularly as **more** and **more** **genomes** are **sequenced** and **mapped** and **sequenced**. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing on



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



Topic Models: Input

- Sammlung von Texten (Korpus)

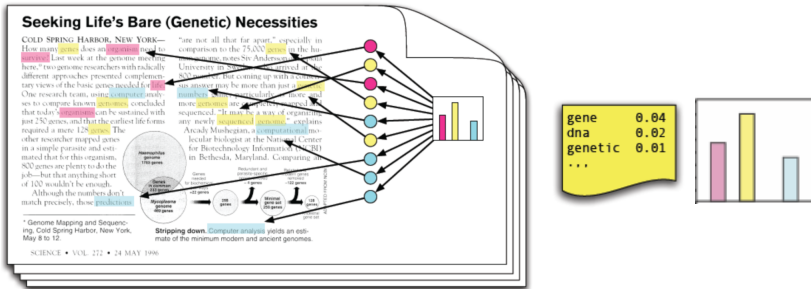


- Liste von Stoppwörtern

- das, zu, und, wie...
- that, to, and, how...
- 此、爲、與、安

- Anzahl der Topics: 20/50/100/150...

Topic Models: Output

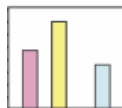


Output: Topic Assignments

- Wortverteilungen pro Topic: Wörter geordnet nach relativer Häufigkeit
- Topicanteile pro Dokument: Topics geordnet nach ihrem Anteil am Dokument

Topic Models: Interpretation

gene	0.04
dna	0.02
genetic	0.01
...	



Interpretation: Wovon "handeln" die Topics?

⇒ Wovon handeln die Texte?

Topic Models: Anwendung

- “Suchmaschine”: Finde Dokumente, die von einem bestimmten Thema handeln.
- Welche Themen bekommen Aufmerksamkeit und verändert sich das mit der Zeit?
- Wie verändert sich der Diskurs über bestimmte Phänomene?

Anstoß für das Projekt

- Topic Modeling deckt semantische Zusammenhänge auf (?)
- Keine weiteren Trainingsdaten nötig
- Möglicherweise hilfreich als Hilfsmittel bei der Erschließung großer Quellensammlungen in klassischem Chinesisch
 - ⇒ Eigenes einfaches Topic Modeling Tool zum Testen verschiedener Parametereinstellungen, Modellierungs- und Implementierungsentscheidungen und Visualisierungsoptionen

Die *Anthologien zur Staatskunst der Qing-Dynastie* (*Huangchao jingshi wenbian* 皇朝經世文編)

- Serie von Verwaltungshandbüchern aus der Qing-Dynastie
- Erschienen zwischen 1827 und 1903
- Thematische Struktur: Sektionen und Kapitel
- Texte zu technischen und theoretischen Verwaltungsfragen
- Offizieller und privater Schriftverkehr (Memoranden, Edikte, Briefe), Abhandlungen und Essays



Image downloaded from: <https://zhidao.baidu.com/question/458301589104584885.html>

Tool für die Generierung und Visualisierung von Topic Models

Input

- Korpus und Liste von Stoppwörtern
- Anzahl der Topics

Lda Lab

You can explore one of the following topic models or create a new model with one of the available corpora.

Available Models

- [Model with 10 topics inferred from corpus hcjswb_he_1827_small \(Jan-6-2018-12:06-PM\) \[delete\]](#)
- [Model with 10 topics inferred from corpus hcjsw_combined \(Jan-6-2018-12:18-PM\) \[delete\]](#)
- [Model with 20 topics inferred from corpus hcjsw_combined \(Jan-6-2018-12:42-PM\) \[delete\]](#)
- [Model with 25 topics inferred from corpus hcjsw_combined \(Jan-6-2018-01:10-PM\) \[delete\]](#)
- [Model with 20 topics inferred from corpus hcjswb_he_1827_filtered \(Jan-6-2018-05:11-PM\) \[delete\]](#)
- [Model with 50 topics inferred from corpus hcjsw_combined_filtered \(Jan-6-2018-05:49-PM\) \[delete\]](#)
- [Model with 50 topics inferred from corpus hcjsw_combined_filtered \(Jan-6-2018-07:58-PM\) \[delete\]](#)
- [Model with 35 topics inferred from corpus hcjsw_combined_filtered \(Jan-7-2018-04:09-PM\) \[delete\]](#)

Available Corpora

Corpus: 

Number of topics

Tool für die Generierung und Visualisierung von Topic Models

Output

■ Liste der generierten Topics und Liste der Dokumente

- 0 兵營軍將人用二長各戰
- 1 水利溝田民大地旱人成
- 2 盜捕民匪苗官賊獲土方
- 3 子父母禮服宗後祖人夫
- 4 錢銀鈔用行民法利重易
- 5 臣行皇上實奏督撫查等
- 6 民鄉長縣法官行保盜甲
- 7 人書文子言經古天乎謂
- 8 吏官縣州司事上下治民
- 9 役差人戶名查各官冊分
- 10 東河南西北流州水府合

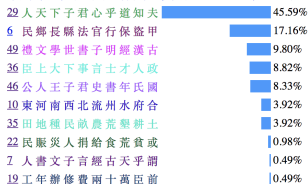
盛編1897·卷二十吏政三官制 >>

盛編1897·卷十九吏政二吏論下 >>

盛編1897·卷十八吏政一吏論上 >>

賀編1827·卷二十一吏政七守令上 <<

[書李綱薦所知於張徐州 \(賀編1827·卷二十一吏政七守令上\)](#)



[讀史縣令箚 \(賀編1827·卷二十一吏政七守令上\)](#) >>

■ Dokumentansicht: Topics und Text

■ Topicansicht: Wörter, Wortsequenzen und Dokumente

Tool für die Generierung und Visualisierung von Topic Models

■ Output

- Liste der generierten Topics und Liste der Dokumente
- Dokumentansicht: Topics und Text

Topics

47 倉米穀價石積年買州縣	60.06%
22 民賑災人捐給食荒貧或	13.21%
5 臣行皇上實奏督撫查等	12.31%
2 役差人戶名查各官冊分	6.31%
43 人事言見心得能知書生	2.70%



Content (Topic Ids in Brackets)

竊臣前以通裕民食之法。請於秋成確探上游豐收之處。差員領奉旨發議。惟是豐收地方。業經各處差員赴買。恐一時米價昂商大戶。恣意囤積所致。若必強之使賣。恐有不肖胥役。藉端。莫善於各省府州縣。遍開常平倉捐穀之例。以備積儲。以資

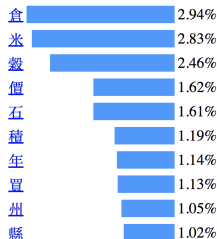
- Topicansicht: Wörter, Wortsequenzen und Dokumente

Tool für die Generierung und Visualisierung von Topic Models

Output

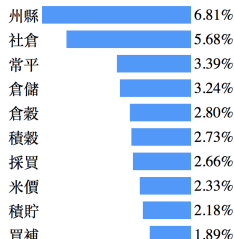
- Liste der generierten Topics und Liste der Dokumente
- Dokumentansicht: Topics und Text
- Topicansicht: Wörter, Wortsequenzen und Dokumente

Words



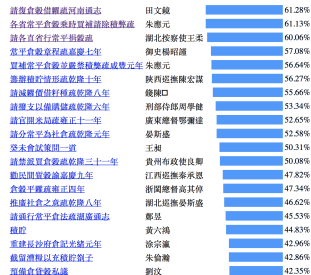
+ -

Word Sequences



+ -

Documents



+ -

Natur der generierten Topics (T=50)

- 30% Klar abgrenzbare Fachterminologie spezifischer Verwaltungszweige (Salzsteuer, Steuergetreidetransport, Rechtsprechung, Getreidespeicher, usw.)
- 30% Fachterminologie Wasserbau und Militär
- 20% Diskurse abstrakter Natur (Theorien guter Regierung, Kontrolle innerhalb der Verwaltung, Effizienz der Lokalverwaltung usw.)
- 10% Wortarten mit spezifischer Funktion (Zahlen und Zählseinheitswörter, Memorandenfloskeln, oft zitierte Autoritäten usw.)
- 10% Hintergrundrauschen

Beispiele

- Salzsteuer (Topic 27): Produktion, Transport, Handel mit Salz, Regulierung und Besteuerung
- Provinzverwaltung (Topic 8): Kontrolle von Korruption innerhalb der Verwaltung durch Aufsicht, Sanktionen, gute Personalführung

Fazit

- Überblick über Fachterminologie der einzelnen Ressorts
- Auffindung von Texten zu abstrakteren Diskursen mit charakteristischem Vokabular
- ⇒ Neue Perspektiven und Forschungsfragen
- Abhängig von der Verfügbarkeit und Zuverlässigkeit digitalisierter Quellen
- Bei spezifischem Forschungsinteresse/-fokus eventuell nicht hilfreich

Vielen Dank für Ihre Aufmerksamkeit!

<https://github.com/neinkeinkaffee/lda-lab>

Quellen

- Van Atteveldt, Wouter, et al. "LDA models topics... But what are 'topics'?", *Glasgow Big Data* (2014).
- Miller, Ian M. "Rebellion, crime and violence in Qing China, 1722—1911: a topic modeling approach." *Poetics* 41.6 (2013): 626-649.
- Blei, David M., "Probabilistic Topic Models (Survey)." *Communications of the ACM* 55 (2012): 77-84.
- Goldstone, Andrew: "dfr-browser: Take a MALLET to disciplinary history", <https://agoldst.github.io/dfr-browser/> (2013–2016).
- Mimno, David: "jsLDA: In-browser topic modeling", <https://mimno.infosci.cornell.edu/jsLDA/index.html> (2012–2017).
- Chang, Jonathan, et al. "Reading tea leaves: How humans interpret topic models." *Advances in neural information processing systems* (2009).
- Griffiths, Thomas L., and Mark Steyvers. "Finding scientific topics." *Proceedings of the National Academy of Sciences* 101 (2004): 5228-5235.