# Topic Models als sinologisches Hilfsmittel: Möglichkeiten und Grenzen

Gesa Stupperich

Universität Heidelberg, Asia and Europe in a Global Context

25. Januar 2018

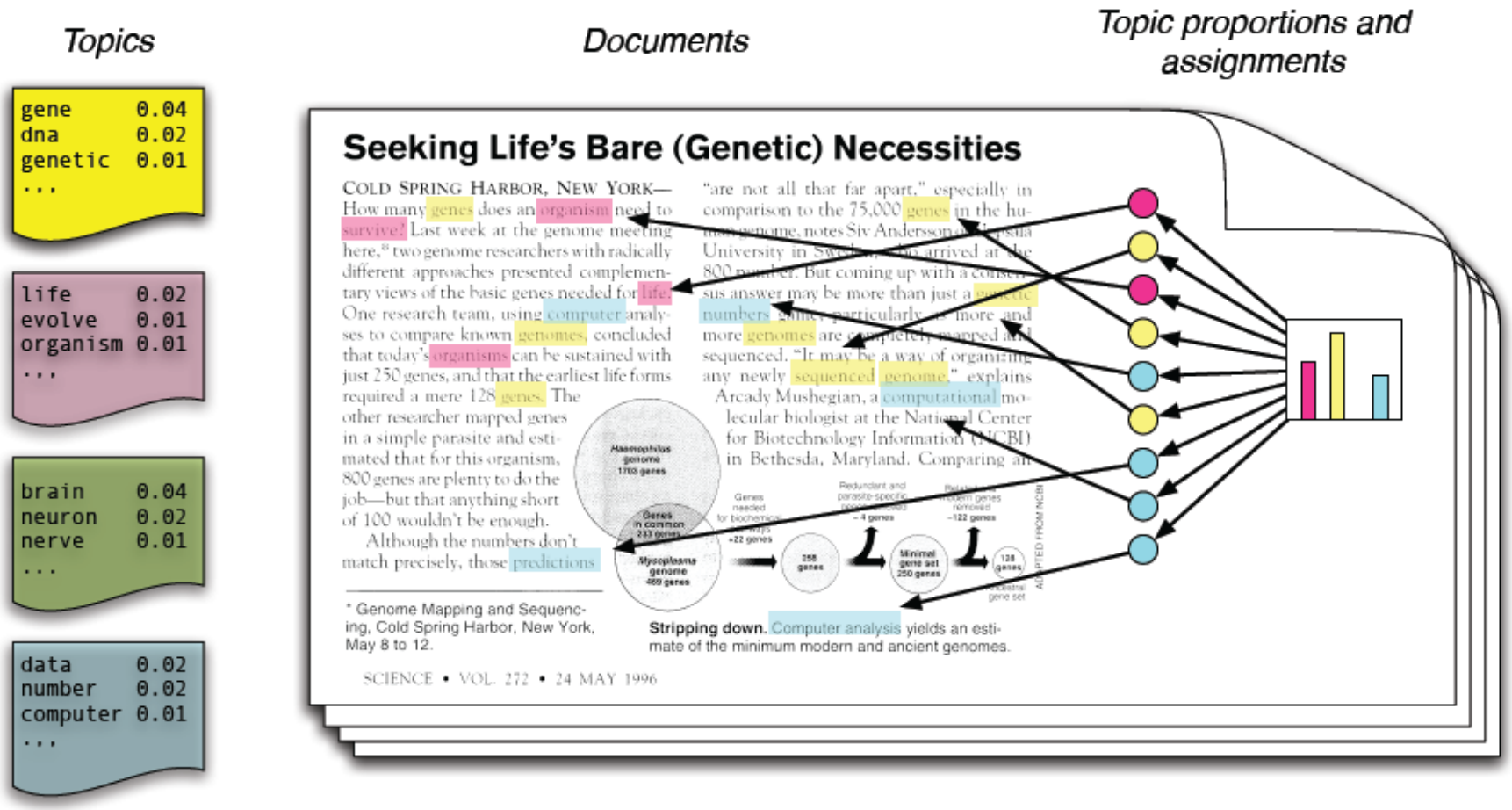
# Hierüber möchte ich sprechen

- 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

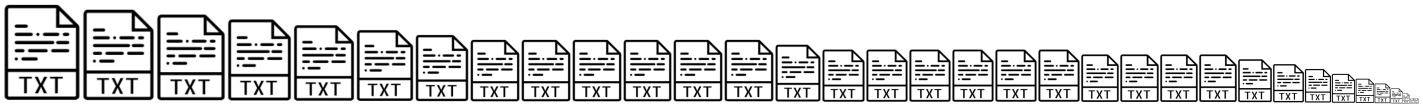
# Topic Models: Verfahren




David M. Blei (2012): "Probabilistic topic models: Surveying a suite of algorithms that offer a solution to managing large document archives", *Communications of the ACM*, Vol. 55, No. 4, S. 78.

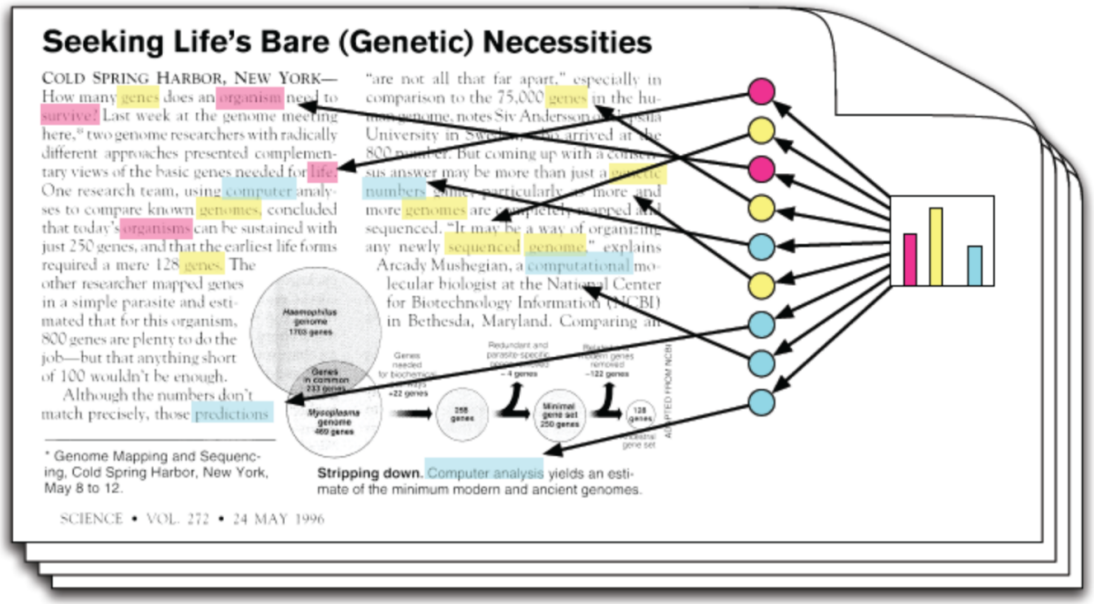
# Topic Models: Input

- 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

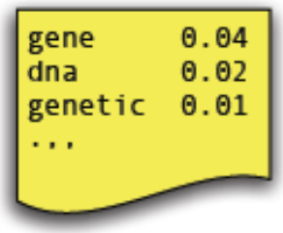


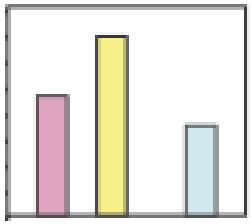


Output: Topic Assignments

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

# Topic Models: Interpretation



$\Rightarrow$


Interpretation: Wovon "handeln" die Topics?
$\Rightarrow$ 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

  $\Rightarrow$ 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)

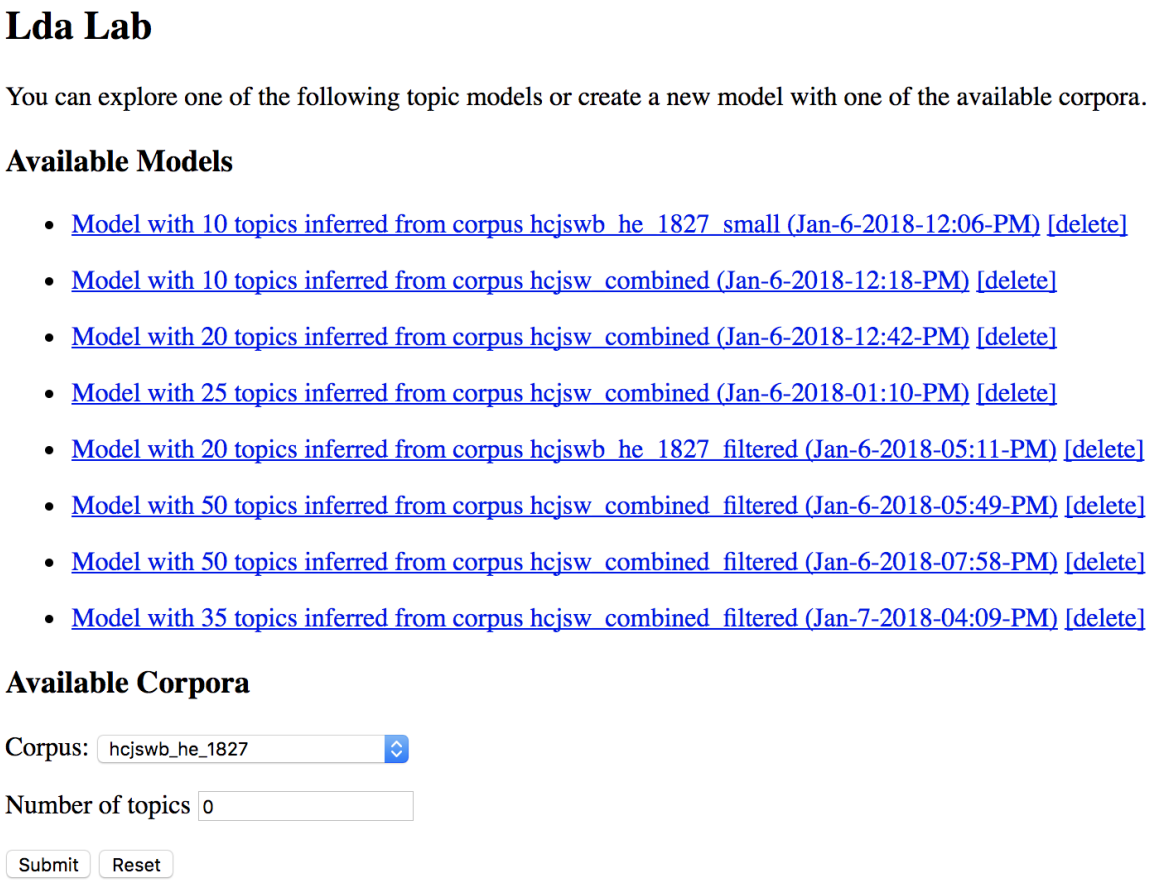
- Serie von Verwaltungshandbüchern aus der Qing-Dynastie
- Erschienen zwischen 1827 und 1903
- Thematische Struktur: Sektionen und Kapitel
- Texte zu technischen und theoretischen Verwaltungsfragen
- Offizieller und privater Schriftverkehr (Memoranden, Edikte, Briefe), Abhandlungen und Essays

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: hcjswb_he_1827

Number of topics 0

Submit Reset

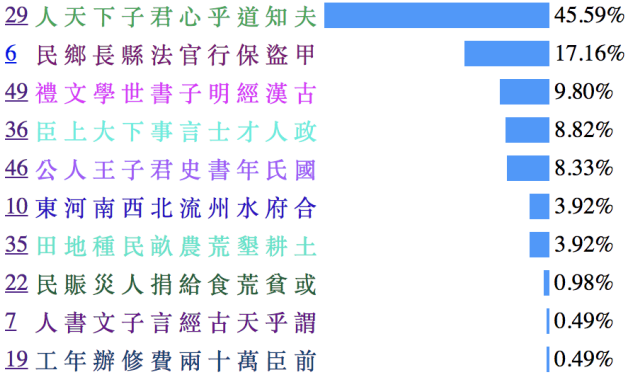
# Tool für die Generierung und Visualisierung von Topic Models

## 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 <<
書李翹薦所知於張徐州（賀編1827·卷二十一吏政七守令上）



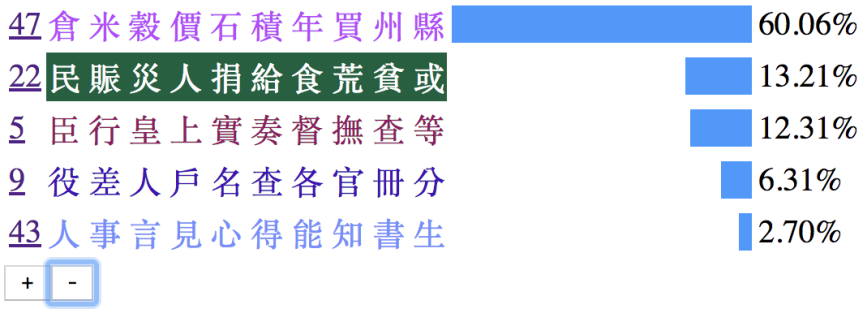

讀史縣令箋（賀編1827·卷二十一吏政七守令上）>>

- 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**




**Content (Topic Ids in Brackets)**

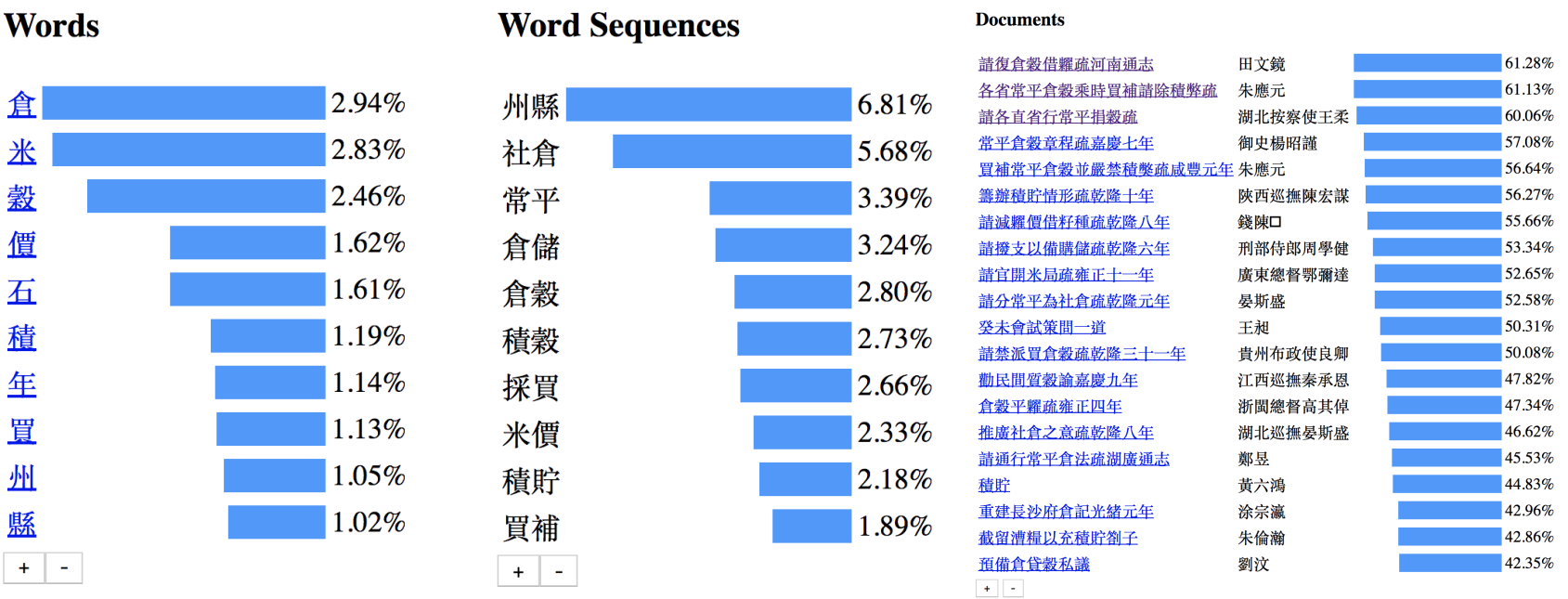
竊臣前以通裕民食之法。請於秋成確探上游豐收之處。差員領奉旨發議。惟是豐收地方。業經各處差員赴買。恐一時米價昂商大戶。恣意囤積所致。若必強之使賣。恐有不肖胥役。藉端。莫善於各省府州縣。遍開常平倉捐穀之例。以備積儲。以資

  - 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



# 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 Zähleinheitswörter, 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.