ANALYSIS OF UNSTRUCTURED TEXT DATA WITH TOPIC MODELS

WK RECH 2019, Frankfurt School of Finance & Management
Since October 2018: Professor at the Department of Management Information Systems, Paderborn University

2016-2018: Associate Professor at the Business IT Department, IT University of Copenhagen

2011-2016: Assistant Professor at the Institute of Information Systems, University of Liechtenstein

2007-2011: PhD at European Research Center for Information Systems (ERCIS), Westfälische Wilhelms-Universität Münster
1. Using big data and machine learning to solve relevant business and societal problems
2. Analysis of unstructured data (e.g., text, images)
3. Acceptance and value of big data analytics
I. Big Text Data
II. Fundamentals of Topic Modeling
III. Topic Modeling Walkthrough
BIG TEXT DATA
"data sets that are too large to fit into main memory or even local disks" (Cox and Ellsworth 1997)

The story of how data became big starts many years before the current buzz around big data. Already seventy years ago we encounter the first attempts to quantify the growth rate in the volume of data or what has popularly been known as the "information explosion" (a term first used in 1941, according to the

Source: http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#41b6f62055da
The 3 "V"s of Big Data (Laney, 2001)

A Very Short History Of Big Data

The history of how data became big starts many years before the current buzz and big data. Already seventy years ago we encounter the first attempts to quantify the growth rate in the volume of data or what has popularly been known as the "Information Explosion" (a term coined by von Neumann in 1948), according to the third English Dictionary. Von Neumann was correct in his story of sizing data volumes plus other "firsts" in the evolution of the idea of "data" and observations pertaining to data or information explosion.

Update: December 21, 2013

Fremont Rider, Wesleyan University Librarian, publishes The Scholar and the Future of the Research Library. He estimates that American university libraries were doubling in size every sixteen years. Given this growth...

Source: http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#41b6f62055da
BIG TEXT DATA

Source: Laney (2001), IBM (2012)
Volume and Variety

![Graph showing total archived capacity by content type worldwide from 2008 to 2015 (petabytes)].

Source: Dhar (2013)
**Velocity**

- Business-relevant event occurs
- Event data stored
- Analysis information delivered
- Action initiated
- Action completed

Value lost through latency

Data latency

Analysis latency

Decision latency

Implementation latency

Reaction time

Big or not?

WIKIPEDIA
The Free Encyclopedia

English
5,550,000+ articles

Deutsch
2,142,000+ Artikel

Español
1,381,000+ artículos

Italiano
1,409,000+ voci

Português
988,000+ artigos

Français
1,947,000+ articles

Polski
1,260,000+ hasi

日本語
1,091,000+ 記事

Русский
1,447,000+ статей

中文
986,000+ 修目
Big or not?
Why is Text Analytics Difficult?

- **Text is Messy**
  - Cannot easily be represented in rows and columns of tables
  - Has complex linguistics structures that differ across languages

- **Text is Dirty**
  - Lots of words that are in no dictionary (e.g., spelling mistakes, slang, abbreviations, technical terms)

- **Text is Ambiguous**
  - Meaning of words depends on context
Text is Messy

Lorem Ipsum

What is Lorem Ipsum?
Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry’s standard dummy text ever since the 1960s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

Why do we use it?
It is a long established fact that a reader will be distracted by the readable content of a page when looking at its layout. The point of using Lorem Ipsum is that it has a normal distribution of letters, as opposed to using Content here, content here, making it look like readable English. Many desktop publishing packages and web page editors now use Lorem Ipsum as their default model text, and a search for "lorem ipsum" will uncover many web sites still in their infancy. Various versions have also evolved over the years, sometimes by accident, sometimes on purpose (ejected humour and the like).

sentence

noun phrase

verb phrase

noun phrase

article

noun

verb

noun

the

rat

ate

cheese
The Cool Parent’s Guide to
Internet Slang and Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Meaning</th>
<th>Abbreviation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFK</td>
<td>Away from Keyboard</td>
<td>msg</td>
<td>Message</td>
</tr>
<tr>
<td>ASL</td>
<td>Age/say/location?</td>
<td>MYOB</td>
<td>Mind Your Own Business</td>
</tr>
<tr>
<td>ATM</td>
<td>At the Moment</td>
<td>N/A</td>
<td>Not Available</td>
</tr>
<tr>
<td>b/c</td>
<td>Because</td>
<td>NC</td>
<td>No Comment</td>
</tr>
<tr>
<td>b/w</td>
<td>Between</td>
<td>net</td>
<td>Anyone</td>
</tr>
<tr>
<td>b4</td>
<td>Before</td>
<td>NM</td>
<td>Not much</td>
</tr>
<tr>
<td>BBL</td>
<td>Be back later</td>
<td>ncb</td>
<td>Newbie</td>
</tr>
<tr>
<td>BFF</td>
<td>Best Friends Forever</td>
<td>NTP</td>
<td>No Thanks Needed</td>
</tr>
<tr>
<td>BRB</td>
<td>Be Right Back</td>
<td>OMG</td>
<td>Oh My Gosh</td>
</tr>
<tr>
<td>BTW</td>
<td>By The Way</td>
<td>OMW</td>
<td>On My Way</td>
</tr>
<tr>
<td>CTFN</td>
<td>Can’t Talk Now</td>
<td>OT</td>
<td>Off Topic</td>
</tr>
<tr>
<td>CVE</td>
<td>Check Your E-Mail</td>
<td>PC</td>
<td>Personal Computer</td>
</tr>
<tr>
<td>dl</td>
<td>Download</td>
<td>pls</td>
<td>Please</td>
</tr>
<tr>
<td>ETA</td>
<td>Estimated Time of Arrival</td>
<td>POS</td>
<td>Parent Over Shoulder</td>
</tr>
<tr>
<td>FAWM</td>
<td>For What’s Worth</td>
<td>ppl</td>
<td>People</td>
</tr>
<tr>
<td>FYI</td>
<td>For Your Information</td>
<td>qt</td>
<td>Cute</td>
</tr>
<tr>
<td>GG</td>
<td>Good Game</td>
<td>re</td>
<td>Regarding</td>
</tr>
<tr>
<td>GJ</td>
<td>Good Job</td>
<td>SMH</td>
<td>Shaking my head</td>
</tr>
<tr>
<td>GL</td>
<td>Good Luck</td>
<td>Sry</td>
<td>Sorry</td>
</tr>
<tr>
<td>GSB</td>
<td>Great</td>
<td>TBA</td>
<td>To Be Announced</td>
</tr>
<tr>
<td>GTG</td>
<td>Got To Go</td>
<td>TBC</td>
<td>To Be Continued</td>
</tr>
<tr>
<td>GMV</td>
<td>Got My Vote</td>
<td>TC</td>
<td>Take Care</td>
</tr>
<tr>
<td>HTH</td>
<td>Hope this helps</td>
<td>txs</td>
<td>Thanks</td>
</tr>
<tr>
<td>HWm</td>
<td>Homework</td>
<td>TIA</td>
<td>Thanks in Advance</td>
</tr>
<tr>
<td>IAC</td>
<td>In Any Case</td>
<td>TLC</td>
<td>Tender Loving Care</td>
</tr>
<tr>
<td>IC</td>
<td>I see</td>
<td>TMI</td>
<td>Too Much Information</td>
</tr>
<tr>
<td>IDK</td>
<td>I Don’t Know</td>
<td>TTFN</td>
<td>Ta-Ta For Now</td>
</tr>
<tr>
<td>IIRC</td>
<td>If I Remember Correctly</td>
<td>TTCL</td>
<td>Talk To You Later</td>
</tr>
<tr>
<td>IKR</td>
<td>I Know, Right?</td>
<td>txt</td>
<td>Text</td>
</tr>
<tr>
<td>IM</td>
<td>Instant Message</td>
<td>TY</td>
<td>Thank You</td>
</tr>
<tr>
<td>IMO</td>
<td>In My Opinion</td>
<td>w/e</td>
<td>Whatever</td>
</tr>
<tr>
<td>IMHO</td>
<td>In My Humble Opinion</td>
<td>w/o</td>
<td>Without</td>
</tr>
<tr>
<td>IRL</td>
<td>In Real Life</td>
<td>WB</td>
<td>Wait</td>
</tr>
<tr>
<td>J/K</td>
<td>Just kidding</td>
<td>XOXO</td>
<td>Hugs and kisses</td>
</tr>
<tr>
<td>K</td>
<td>OK</td>
<td>Y</td>
<td>Why</td>
</tr>
<tr>
<td>LB</td>
<td>Late</td>
<td>YN</td>
<td>Why Not</td>
</tr>
<tr>
<td>LBR</td>
<td>Later</td>
<td>YOLO</td>
<td>You Only Live Once</td>
</tr>
<tr>
<td>LMK</td>
<td>Let Me Know</td>
<td>YW</td>
<td>You’re Welcome</td>
</tr>
<tr>
<td>LOL</td>
<td>Laughing Out Loud</td>
<td>ZZZ</td>
<td>Sleeping</td>
</tr>
</tbody>
</table>
Text is Ambiguous

Homophone
Same pronunciation, different meaning

Homograph
Same spelling, different meaning

Heterograph
Different spelling and meaning e.g. too/two

Homonym
Different meaning e.g. lie (untruth)/lie (recline)

Heteronym
Different pronunciation and meaning e.g. desert (arid region)/desert (leave)

Synonym
Different spelling and pronunciation e.g. settee/sofa

Same pronunciation

Identical words

Same spelling

Words with different spelling, pronunciation and meaning

FUNDAMENTALS OF TOPIC MODELING
From Counting to Categorizing

http://www.cse.buffalo.edu/~rapaport/575/categories.html
# FUNDAMENTALS OF TOPIC MODELING

## Text Categorization

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assumptions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Categories are predefined</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Relevant text features are known</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Mapping between text features and categories is known</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Costs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-analysis costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person-hours spent conceptualizing</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Level of substantive knowledge</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Analysis costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person-hours spent per text</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Level of substantive knowledge</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Post-analysis costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person hours spent interpreting</td>
<td>Moderate</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>Level of substantive knowledge</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

Source: Debortoli et al. (2016)
### Text Categorization

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Assumptions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Categories are predefined</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Relevant text features are known</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Mapping between text features and categories is known</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-analysis costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person-hours spent conceptualizing</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Level of substantive knowledge</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Analysis costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person-hours spent per text</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Level of substantive knowledge</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Post-analysis costs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person-hours spent interpreting</td>
<td>Moderate</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Level of substantive knowledge</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

Source: Debortoli et al. (2016)
## FUNDAMENTALS OF TOPIC MODELING

### Text Categorization

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories are predefined</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Relevant text features are known</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Mapping between text features and categories is known</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

### Costs

#### Pre-analysis costs

| Person-hours spent conceptualizing        | Low                       | High                      | High         | High                        | Low                         |
| Level of substantive knowledge           | Low                       | High                      | High         | High                        | Low                         |

#### Analysis costs

| Person-hours spent per text               | High                      | High                      | Low          | Low                         | Low                         |
| Level of substantive knowledge           | Moderate                  | Moderate                  | Low          | Low                         | Low                         |

#### Post-analysis costs

| Person hours spent interpreting          | Moderate                  | Low                       | Low          | Low                         | Moderate                    |
| Level of substantive knowledge           | High                      | High                      | High         | High                        | High                        |

Source: Debortoli et al. (2016)
# FUNDAMENTALS OF TOPIC MODELING

## Text Categorization

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories are predefined</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Relevant text features are known</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Mapping between text features and categories is known</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

**Costs**

<table>
<thead>
<tr>
<th>Analysis costs</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Person-hours spent conceptualizing</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Level of substantive knowledge</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Post-analysis costs</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Person-hours spent interpreting</td>
<td>Moderate</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>Level of substantive knowledge</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

Source: Debortoli et al. (2016)
### Text Categorization

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories are predefined</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Relevant text features are known</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Mapping between text features and categories is known</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

**Costs**

**Pre-analysis costs**

<table>
<thead>
<tr>
<th>Person-hours spent conceptualizing</th>
<th>Low</th>
<th>High</th>
<th>High</th>
<th>High</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of substantive knowledge</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

**Analysis costs**

<table>
<thead>
<tr>
<th>Person-hours spent per text</th>
<th>High</th>
<th>High</th>
<th>Low</th>
<th>Low</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of substantive knowledge</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
</tbody>
</table>

**Post-analysis costs**

<table>
<thead>
<tr>
<th>Person hours spent interpreting</th>
<th>Moderate</th>
<th>Low</th>
<th>Low</th>
<th>Low</th>
<th>Moderate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of substantive knowledge</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

Source: Debortoli et al. (2016)
# FUNDAMENTALS OF TOPIC MODELING

## Text Categorization

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories are predefined</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Relevant text features are known</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Mapping between text features and categories is known</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

### Costs

#### Pre-analysis costs

| Person-hours spent conceptualizing       | Low                       | High                     | High         | High                        | Low                           |
| Level of substantive knowledge           | Low                       | High                     | High         | High                        | Low                           |

#### Analysis costs

| Person-hours spent per text              | High                      | High                     | Low          | Low                         | Low                           |
| Level of substantive knowledge           | Moderate                  | Moderate                 | Low          | Low                         | Low                           |

#### Post-analysis costs

| Person hours spent interpreting          | Moderate                  | Low                      | Low          | Low                         | Moderate                      |
| Level of substantive knowledge           | High                      | High                     | High         | High                        | High                          |

Source: Debortoli et al. (2016)
What are Topic Models?

• **Unsupervised** machine learning methods for text mining (e.g., Latent Semantic Analysis, Latent Dirichlet Allocation)

• Theoretical grounding: *Distributional hypothesis* of linguistics
  • Words that co-occur together in similar contexts (e.g., ball, goal, offside) tend to have similar meanings
  • Co-occurrence patterns can be interpreted as topics (e.g., football) and used to cluster documents
Schematic Overview of Probabilistic Topic Modeling with LDA

Source: Debortoli et al. (2016)
Illustrative Example of Probabilistic Topic Modeling with LDA

Exemplary Customer Review about a Fitbit Flex

I bought this for my 14 year old daughter as a gift. She received it in July. It works great - she lost 6 pounds in 2 weeks. The Fitbit makes staying in shape easy. The iPhone app works fine.

Topics

- Birthday present
- Loosing weight
- Mobile app
FUNDAMENTALS OF TOPIC MODELING

Illustrative Example of Probabilistic Topic Modeling with LDA

Source: Debortoli et al. (2016)
TOPIC MODELING WALKTHROUGH
TOPIC MODELING WALKTHROUGH

Cross Industry Standard Process for Data Mining

Source: Shearer et al. (2000)
Research Question

“What factors drive employees’ company ratings?”

Company reviews and ratings. Get the whole story.

Search ratings and reviews of over 600,000 companies worldwide. Get the inside scoop and find out what it’s really like from people who’ve actually worked there.

Write a Review
"shortName": "Bank of America",
"date": "2010-10-03",
"stars": 3,
"text": "Great benefits for associates, Paid maternity/paternity leave, most associates receive 3 weeks of vacation leave per year (SSS, PB, AM and four weeks for BCM). Micro-management, poor leadership, lack of recognition, extremely under staffed. Do not forget the human aspect. Micro-management is not the answer to every situation. Put more people in the branches."
Data Understanding

- Sub-sample
  - Finance industry only
- Number of documents
  - 57,765
- Number or words (tokens)
  - 1,608,259
- Number of unique words
  - 1,740
TOPIC MODELING WALKTHROUGH

Data Understanding
TOPIC MODELING WALKTHROUGH

Data Understanding

- Citi
- Capital One
TOPIC MODELING WALKTHROUGH

Data Preparation

Tokenization → Part-Of-Speech Tagging → Stemming/Lemmatization → Stopword removal → Co-occurrence counting → Doc-term matrix
Modeling and Evaluation – Iteration #1

• Estimate 30 most prevalent topics
  • Takes approx. 10 minutes on a MacBook Pro
  • Wait for the “aha” effect
Modeling and Evaluation – Iteration #1

• Estimate 30 most prevalent topics
  • Takes approx. 10 minutes on a MacBook Pro
  • Wait for the “aha” effect

Top-3 Documents for Topic 6

| Health insurance, Vacation, sick pay, paid maternity leave 12 weeks. Every year perks decrease and are eliminated. Uneducated people with their nose in the air. |
| Decent benefits, decent bonuses, decent vacation/off time. Inadequate pay and inadequate coverage. |
| Huge Annual Bonus amount, Paid Overtime amount can be earned, 1 Time free meal & free transport facility. Less On paper CTC offered. Should include the approximate Annual Bonus & Gratuity in the offered CTC on-paper. |
Modeling and Evaluation – Iteration #2

- Find the right number of topics
  - Manual investigation
    - Are topics coherent? No duplicate topics? No fused topics?
    - Increase or reduce number of topics

LDavis: A R package for interactive topic model visualization.
Modeling and Evaluation – Iteration #2

- Find the right number of topics
  - **Automated** search
    - e.g.: From 10 to 100 topics, in steps of 10
    - Takes several hours on a MacBook Pro
    - Evaluate models with regards to **Semantic Coherence** and **Exclusivity**
Modeling and Evaluation – Iteration #3

- Final experimental evaluation through human coders
  - Word intrusion task
  - Topic intrusion task

http://etc.ch/bRsu
TOPIC MODELING WALKTHROUGH

Modeling and Evaluation – Iteration #4

• Modeling the relationship between topics and stars
TOPIC MODELING WALKTHROUGH

Modeling and Evaluation – Iteration #4

Topic Effects on Outcome

Topic 14: great, grow, alway, amaz, will, awesom, wonder ●
Topic 13: team, system, member, softwar, friend, atmospher, access ●
Topic 7: learn, lot, mobil, network, red, exposur, upward ●
Topic 15: innov, organ, global, slow, decis, tape, risk ●
Topic 3: balanc, life, flexibl, good, environ, hour, work ●
Topic 19: career, growth, compett, limit, develop, advanc, path ●
Topic 12: move, big, easi, around, larg, name, huge ●
Topic 10: think, treat, realli, bad, far, noth, none ●
Topic 9: turnov, hire, degre, rate, colleg, expect, requir ●
Topic 6: vacat, health, decent, pay, rais, low, bonus ●
Topic 16: goal, teller, custom, sale, banker, branch, pressur ●
Topic 2: outsourc, cut, india, american, job, cost, express ●
Topic 17: tell, ask, fire, told, talk, someth, wont ●
Topic 4: upper, manag, middl, poor, favorit, micro, lack ●

R²: 0.239
TOPIC MODELING WALKTHROUGH

Communication
TOPIC MODELING WALKTHROUGH

www.MineMyText.com


• Schmiedel, T., Müller, O., & vom Brocke, J. (2018). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. Organizational Research Methods, 1094428118773858.

• Icons: questions by Depb Dew, Jemis Mali from the Noun Project