Conceptual Topic Aggregation (FAT-CAT)

A Case Study on the ETYNTKE Dataset

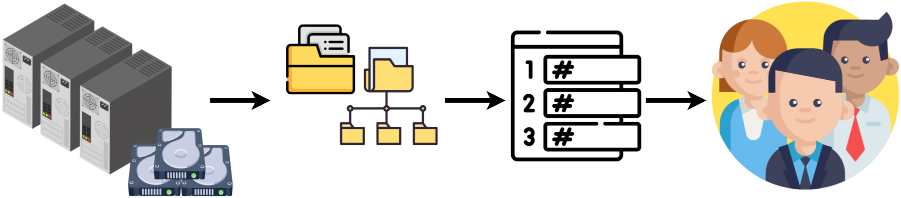
Klara M. Gutekunst Dominik Dürrschnabel Johannes Hirth Gerd Stumme

University of Kassel

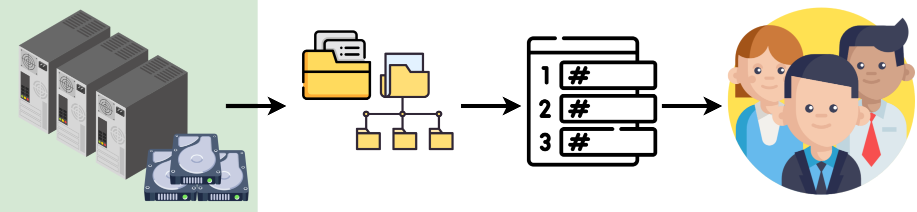
September 12, 2025

Contents

Motivation



Motivation



Motivation

Motivation

Motivation

Motivation

Problem

Employees at the tax office often lack prior knowledge about the contents of the data before starting exploration.

Motivation

Problem

Employees at the tax office often lack prior knowledge about the contents of the data before starting exploration.

What is our objective?

Motivation

Problem

Employees at the tax office often lack prior knowledge about the contents of the data before starting exploration.

What is our objective?
→ **Gain an initial understanding of the data.**

Motivation

Objective

Gain an initial understanding of the data.

Motivation

Objective

Gain an initial understanding of the data.

Challenges:

Motivation

Objective

Gain an initial understanding of the data.

Challenges:

- ✗ Lack of labeled data for supervised training

Motivation

Objective

Gain an initial understanding of the data.

Challenges:

- ✗ Lack of labeled data for supervised training
- ✗ Diverse data formats and structures

Motivation

Objective

Gain an initial understanding of the data.

Challenges:

- ✗ Lack of labeled data for supervised training
- ✗ Diverse data formats and structures

How did we tackle these problems?

Motivation

Objective

Gain an initial understanding of the data.

Challenges:

- ✗ Lack of labeled data for supervised training
- ✗ Diverse data formats and structures

How did we tackle these problems?

- ✓ Fully unsupervised analysis pipeline

Motivation

Objective

Gain an initial understanding of the data.

Challenges:

- ✗ Lack of labeled data for supervised training
- ✗ Diverse data formats and structures

How did we tackle these problems?

- ✓ Fully unsupervised analysis pipeline
- ✓ Handles multiple languages and file types

Dataset

Direct use of tax data is restricted due to privacy concerns, so we rely on a proxy dataset for our case study.

Dataset

Direct use of tax data is restricted due to privacy concerns, so we rely on a proxy dataset for our case study.

Requirements for the proxy dataset:

Dataset

Direct use of tax data is restricted due to privacy concerns, so we rely on a proxy dataset for our case study.

Requirements for the proxy dataset:

- Dataset heterogeneity

Dataset

Direct use of tax data is restricted due to privacy concerns, so we rely on a proxy dataset for our case study.

Requirements for the proxy dataset:

- Dataset heterogeneity
- Structured hierarchy

Everything You Need To Know Ever (ETYNTKE)[1]

- ≈ 101 GB
- Images, text documents, and other file types
- Malformed files

```
'Interrogation, Psychology'
 Language
'Law and Law Enforcement'
 Literature
'Magical Library (UNORGANIZED)'
'Marksmanship and Sniping'
'Martial Arts'
 Math
 Medical
 Military
 Music
 Nanotechnology
 NBC
 Nonfiction
 Paranormal
'PDF Library'
'Periodic tables'
```

Figure: Subset of top-level directories.

[1] https://archive.org/details/ETYNTKE (29.01.2025)

ETYNTKE[1]

- $\approx$ 101 GB
- Images, text documents, and other file types
- Malformed files
- Uploaded in 2015

```
'Interrogation, Psychology'
 Language
'Law and Law Enforcement'
 Literature
'Magical Library (UNORGANIZED)'
'Marksmanship and Sniping'
'Martial Arts'
 Math
 Medical
 Military
 Music
 Nanotechnology
 NBC
 Nonfiction
 Paranormal
'PDF Library'
'Periodic tables'
```

Figure: Subset of top-level directories.

[1] https://archive.org/details/ETYNTKE (29.01.2025)

ETYNTKE[1]

- ≈ 101 GB
- Images, text documents, and other file types
- Malformed files
- Uploaded in 2015
- Topics include politics, weaponry, and computer science

```
'Interrogation, Psychology'
 Language
'Law and Law Enforcement'
 Literature
'Magical Library (UNORGANIZED)'
'Marksmanship and Sniping'
'Martial Arts'
 Math
 Medical
 Military
 Music
 Nanotechnology
 NBC
 Nonfiction
 Paranormal
'PDF Library'
'Periodic tables'
```

Figure: Subset of top-level directories.

[1] https://archive.org/details/ETYNTKE (29.01.2025)

ETYNTKE[1]

- ≈ 101 GB
- Images, text documents, and other file types
- Malformed files
- Uploaded in 2015
- Topics include politics, weaponry, and computer science
- Hierarchical organization into directories

```
'Interrogation, Psychology'
 Language
'Law and Law Enforcement'
 Literature
'Magical Library (UNORGANIZED)'
'Marksmanship and Sniping'
'Martial Arts'
 Math
 Medical
 Military
 Music
 Nanotechnology
 NBC
 Nonfiction
 Paranormal
'PDF Library'
'Periodic tables'
```

Figure: Subset of top-level directories.

[1] https://archive.org/details/ETYNTKE (29.01.2025)

Formal Concept Analysis (FCA)

Formal context $\mathbb{K} := (G, M, I)$

- G object set
- M attribute set
- $I \subseteq G \times M$ incidence relation

Derivation operations

- For $A \subseteq G$: $A' := \{m \in M \mid (g, m) \in I \ \forall g \in A\}$
- For $B \subseteq M$: $B' := \{g \in G \mid (g, m) \in I \ \forall m \in B\}$

Formal concepts $\mathfrak{B}(\mathbb{K}) := \{(A, B) | A' = B \text{and} B' = A\}$

- $A \subseteq G$ and $B \subseteq M$

Concept lattice $\underline{\mathfrak{B}}(\mathbb{K}) := (\mathfrak{B}(\mathbb{K}), \leq)$ with $(A_1, B_1) \leq (A_2, B_2) \iff A_1 \subseteq A_2$.

- $A_i \subseteq G$ and $B_i \subseteq M$

Selection of Related Work

Assuming we work on data structured in directories.

Selection of Related Work

Assuming we work on data structured in directories.

- Topic models (LDA, BERTopic, and Top2Vec)

Selection of Related Work

Assuming we work on data structured in directories.

- Topic models (LDA, BERTopic, and Top2Vec)
  - ✓ Semantic topics
  - × Do not leverage intrinsic hierarchical structure of documents

Selection of Related Work

Assuming we work on data structured in directories.

- Topic models (LDA, BERTopic, and Top2Vec)
  - ✓ Semantic topics
  - × Do not leverage intrinsic hierarchical structure of documents
- FCA on Twitter data (Cigarrán, Castellanos, and García-Serrano 2016)

Selection of Related Work

Assuming we work on data structured in directories.

- Topic models (LDA, BERTopic, and Top2Vec)
  - ✓ Semantic topics
  - × Do not leverage intrinsic hierarchical structure of documents
- FCA on Twitter data (Cigarrán, Castellanos, and García-Serrano 2016)
  - ✓ Discard high- and low-frequency terms
  - × Concepts are topics, but based on one-hot BoW encoding (not semantics)

Related Work

- Representation of document-topic relations using a concept lattice (Hirth and Hanika 2024)

Related Work

- Representation of document-topic relations using a concept lattice (Hirth and Hanika 2024)

Table: Document-topic relation

Documents \ Topics	t_1	t_2	$\cdots$	t_n
d_1	$w_{1,1}$	$w_{1,2}$	$\cdots$	$w_{1,n}$
d_2	$w_{2,1}$	$w_{2,2}$	$\cdots$	$w_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_l	$w_{l,1}$	$w_{l,2}$	$\cdots$	$w_{l,n}$

Table: Term-topic relation

Terms \ Topics	t_1	t_2	$\cdots$	t_n
s_1	$\hat{w}_{1,1}$	$\hat{w}_{1,2}$	$\cdots$	$\hat{w}_{1,n}$
s_2	$\hat{w}_{2,1}$	$\hat{w}_{2,2}$	$\cdots$	$\hat{w}_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_l	$\hat{w}_{l,1}$	$\hat{w}_{l,2}$	$\cdots$	$\hat{w}_{l,n}$

Related Work

- Representation of document-topic relations using a concept lattice (Hirth and Hanika 2024)
  - ✓ Hierarchical representation
  - ✓ Excluding low support items
  - × Homogeneous dataset (academic papers on Machine Learning)
  - × Exclusively textual data
  - × Lacks ability to aggregate information

Table: Document-topic relation

Documents \ Topics	t_1	t_2	$\cdots$	t_n
d_1	$w_{1,1}$	$w_{1,2}$	$\cdots$	$w_{1,n}$
d_2	$w_{2,1}$	$w_{2,2}$	$\cdots$	$w_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_l	$w_{l,1}$	$w_{l,2}$	$\cdots$	$w_{l,n}$

Table: Term-topic relation

Terms \ Topics	t_1	t_2	$\cdots$	t_n
s_1	$\hat{w}_{1,1}$	$\hat{w}_{1,2}$	$\cdots$	$\hat{w}_{1,n}$
s_2	$\hat{w}_{2,1}$	$\hat{w}_{2,2}$	$\cdots$	$\hat{w}_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_l	$\hat{w}_{l,1}$	$\hat{w}_{l,2}$	$\cdots$	$\hat{w}_{l,n}$

FAT-CAT

FAT-CAT FCA-based Aggregation for Topics using Conceptual Analysis and Taxonomies

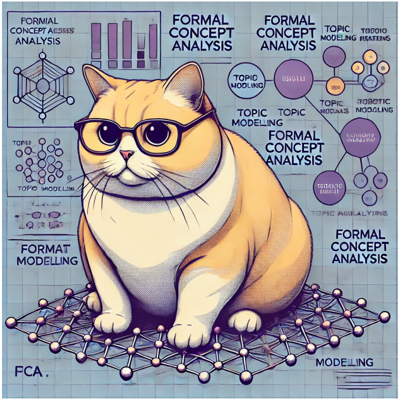
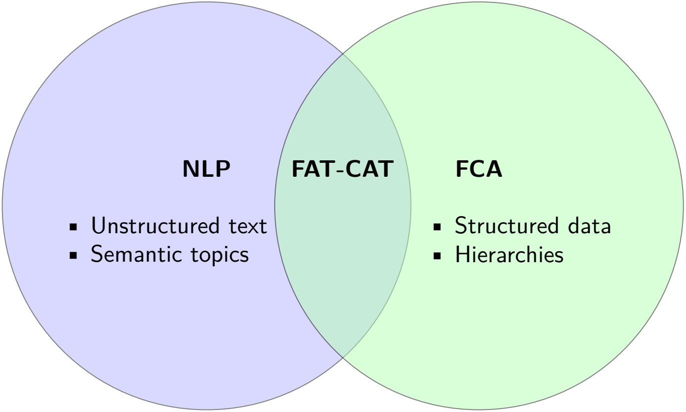
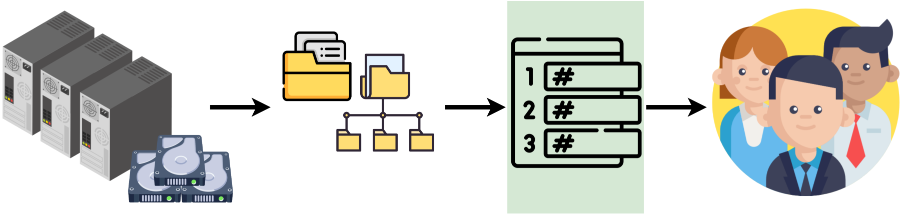


Image generated by DALLE (February 21, 2025)

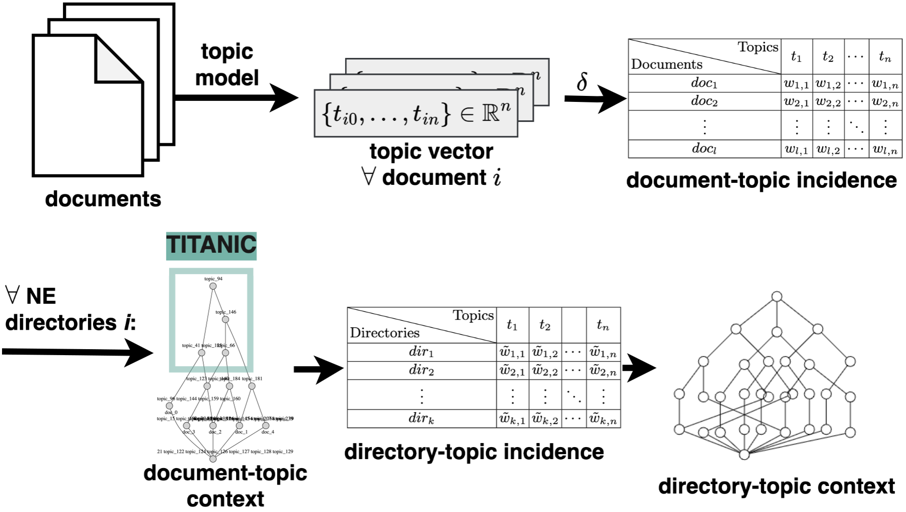
FAT-CAT

FAT-CAT FCA-based Aggregation for Topics using Conceptual Analysis and Taxonomies




1 #
2 #
3 #

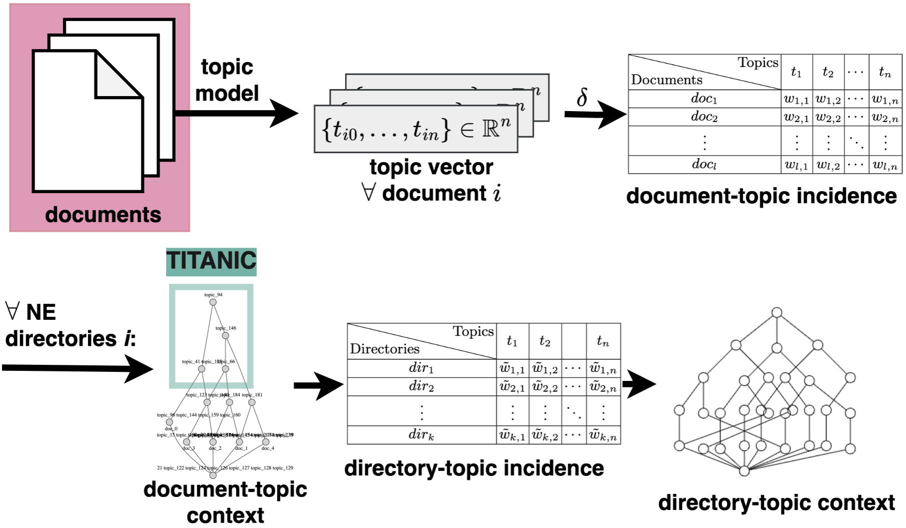
FAT-CAT



Documents \ Topics	t_1	t_2	$\cdots$	t_n
doc_1	$w_{1,1}$	$w_{1,2}$	$\cdots$	$w_{1,n}$
doc_2	$w_{2,1}$	$w_{2,2}$	$\cdots$	$w_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
doc_l	$w_{l,1}$	$w_{l,2}$	$\cdots$	$w_{l,n}$

Directories \ Topics	t_1	t_2		t_n
dir_1	$\tilde{w}_{1,1}$	$\tilde{w}_{1,2}$	$\cdots$	$\tilde{w}_{1,n}$
dir_2	$\tilde{w}_{2,1}$	$\tilde{w}_{2,2}$	$\cdots$	$\tilde{w}_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
dir_k	$\tilde{w}_{k,1}$	$\tilde{w}_{k,2}$	$\cdots$	$\tilde{w}_{k,n}$

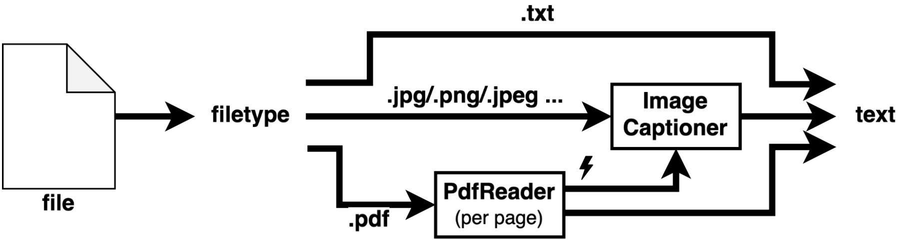
Text Extraction



Documents \ Topics	t_1	t_2	$\cdots$	t_n
doc_1	$w_{1,1}$	$w_{1,2}$	$\cdots$	$w_{1,n}$
doc_2	$w_{2,1}$	$w_{2,2}$	$\cdots$	$w_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
doc_l	$w_{l,1}$	$w_{l,2}$	$\cdots$	$w_{l,n}$

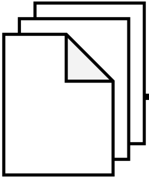
Directories \ Topics	t_1	t_2		t_n
dir_1	$\tilde{w}_{1,1}$	$\tilde{w}_{1,2}$	$\cdots$	$\tilde{w}_{1,n}$
dir_2	$\tilde{w}_{2,1}$	$\tilde{w}_{2,2}$	$\cdots$	$\tilde{w}_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
dir_k	$\tilde{w}_{k,1}$	$\tilde{w}_{k,2}$	$\cdots$	$\tilde{w}_{k,n}$

Text Extraction



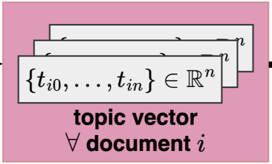
Appendix details: Image Captioner

FAT-CAT



documents

topic model

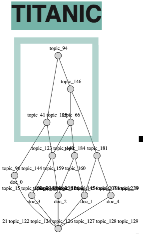


δ

Documents \ Topics	t_1	t_2	$\cdots$	t_n
doc_1	$w_{1,1}$	$w_{1,2}$	$\cdots$	$w_{1,n}$
doc_2	$w_{2,1}$	$w_{2,2}$	$\cdots$	$w_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
doc_l	$w_{l,1}$	$w_{l,2}$	$\cdots$	$w_{l,n}$

document-topic incidence

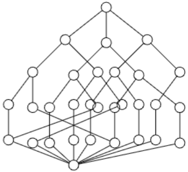
$\forall$ NE directories *i*:



document-topic context

Directories \ Topics	t_1	t_2		t_n
dir_1	$\tilde{w}_{1,1}$	$\tilde{w}_{1,2}$	$\cdots$	$\tilde{w}_{1,n}$
dir_2	$\tilde{w}_{2,1}$	$\tilde{w}_{2,2}$	$\cdots$	$\tilde{w}_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
dir_k	$\tilde{w}_{k,1}$	$\tilde{w}_{k,2}$	$\cdots$	$\tilde{w}_{k,n}$

directory-topic incidence



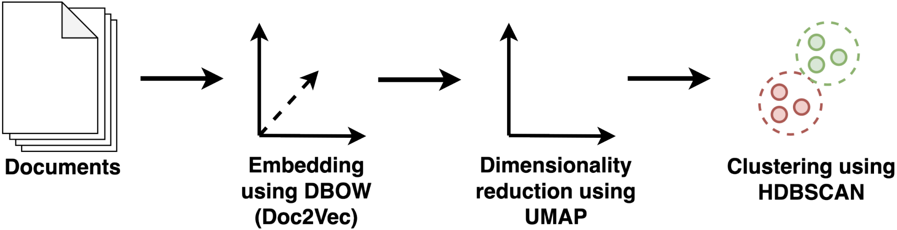
directory-topic context

Topic Modeling

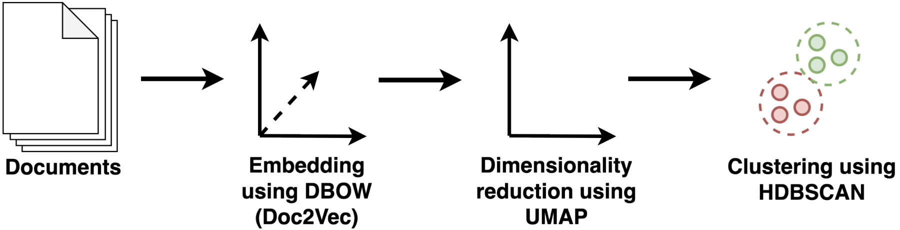
Definition

Topic modeling is used for discovering latent semantic structure, usually referred to as topics, in a large collection of documents (Angelov 2020).

Top2Vec (Angelov 2020)

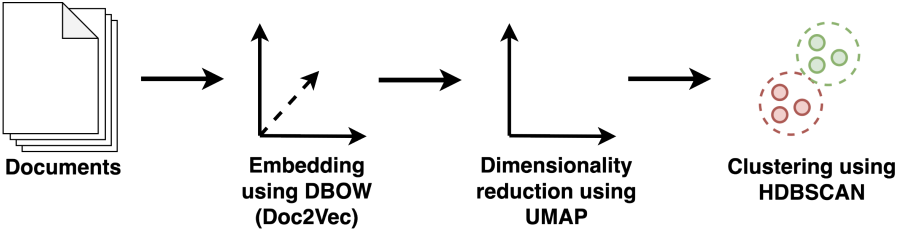


Top2Vec (Angelov 2020)



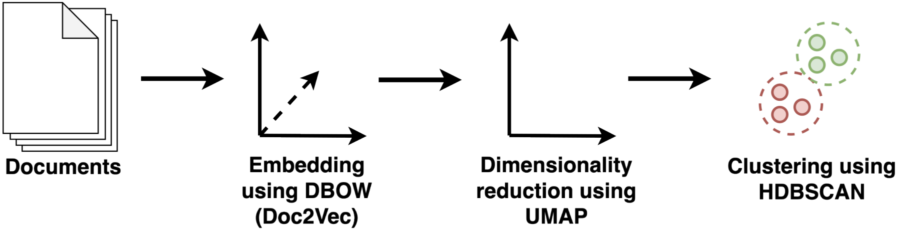
- Words, documents and topics share vector space
- Multilingual USE
  - 16 languages
  - 300-dimensional vector space

Top2Vec (Angelov 2020)



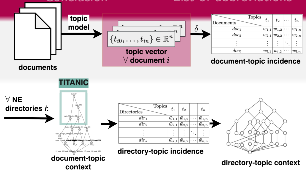
- Words, documents and topics share vector space
- Multilingual USE
  - 16 languages
  - 300-dimensional vector space
- Dimensionality reduction to five dimensions

Top2Vec (Angelov 2020)



- Words, documents and topics share vector space
- Multilingual USE
  - 16 languages
  - 300-dimensional vector space
- Dimensionality reduction to five dimensions
- Embeddings are clustered into topics

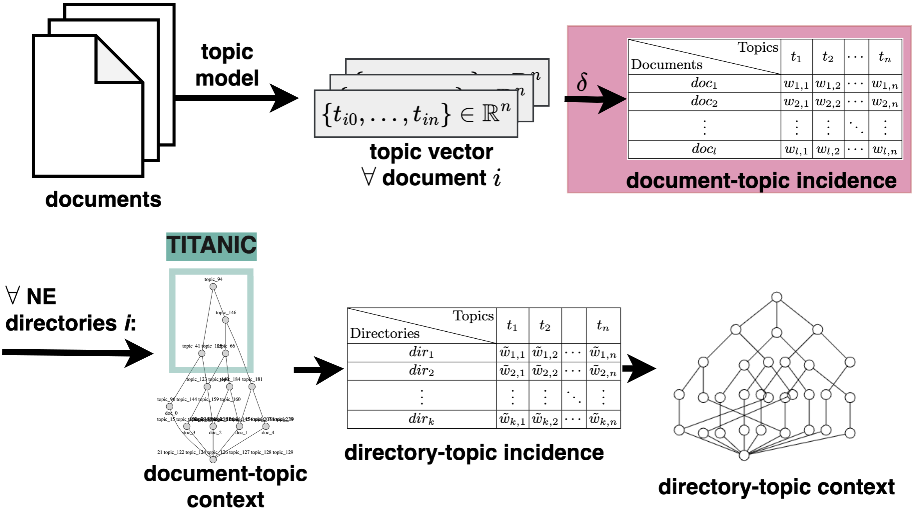
Directory-Topic Concept Lattice



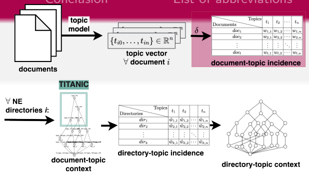
Assuming all documents are organized in directories.

1. Top 10 topics per document incl. relevance scores

Directory-Topic Concept Lattice

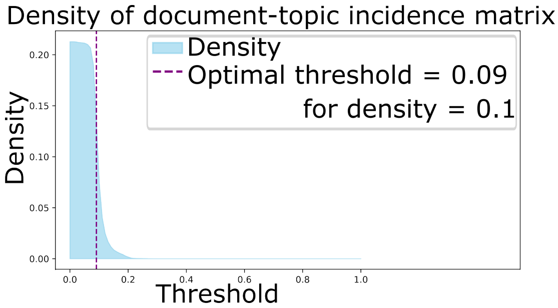


Directory-Topic Concept Lattice

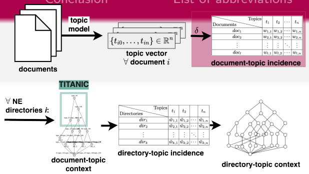


Assuming all documents are organized in directories.

1. Top 10 topics per document incl. relevance scores
2. Compute threshold δ s.t. document-topic matrix density < 0.1



Directory-Topic Concept Lattice



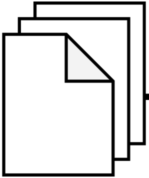
Assuming all documents are organized in directories.

1. Top 10 topics per document incl. relevance scores
2. Compute threshold δ s.t. document-topic matrix density < 0.1
3. $\mathcal{D}_\delta(i,j) = \begin{cases} 1 & \text{if } w_{i,j} \geq \delta, \\ 0 & \text{else.} \end{cases}$

Table: Document-topic relation $\mathcal{D}_\delta$ (Hirth and Hanika 2024)

Documents \ Topics	t_1	t_2	$\cdots$	t_n
d_1	$w_{1,1}$	$w_{1,2}$	$\cdots$	$w_{1,n}$
d_2	$w_{2,1}$	$w_{2,2}$	$\cdots$	$w_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_l	$w_{l,1}$	$w_{l,2}$	$\cdots$	$w_{l,n}$

Directory-Topic Concept Lattice



documents

topic model

$\{t_{i0},\dots,t_{in}\} \in \mathbb{R}^n$

topic vector
$\forall$ document i

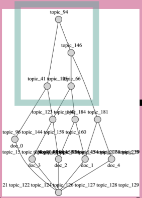
δ

Topics / Documents	t_1	t_2	$\cdots$	t_n
doc_1	$w_{1,1}$	$w_{1,2}$	$\cdots$	$w_{1,n}$
doc_2	$w_{2,1}$	$w_{2,2}$	$\cdots$	$w_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
doc_l	$w_{l,1}$	$w_{l,2}$	$\cdots$	$w_{l,n}$

document-topic incidence

$\forall$ NE directories *i*:

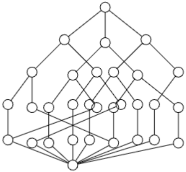
TITANIC



document-topic context

Topics / Directories	t_1	t_2		t_n
dir_1	$\tilde{w}_{1,1}$	$\tilde{w}_{1,2}$	$\cdots$	$\tilde{w}_{1,n}$
dir_2	$\tilde{w}_{2,1}$	$\tilde{w}_{2,2}$	$\cdots$	$\tilde{w}_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
dir_k	$\tilde{w}_{k,1}$	$\tilde{w}_{k,2}$	$\cdots$	$\tilde{w}_{k,n}$

directory-topic incidence



directory-topic context

Iceberg Lattice (Stumme et al. 2002)

Problem

The size of a concept lattice can grow exponentially with respect to the size of the context.

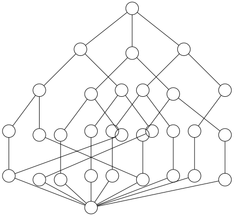


Figure: Concept lattice.

Iceberg Lattice (Stumme et al. 2002)

Problem

The size of a concept lattice can grow exponentially with respect to the size of the context.

Solution

Consider only top-level concepts.

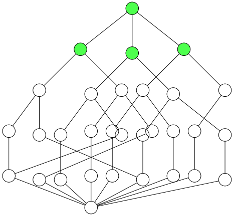


Figure: Concept lattice.

Iceberg Lattice (Stumme et al. 2002)

Let $\mathbb{K} := (G, M, I)$ be a formal context, $A \subseteq G$ and $B \subseteq M$.

Definition

Support $supp(B) = \frac{|B'|}{|G|}$

Iceberg Lattice (Stumme et al. 2002)

Let $\mathbb{K} := (G, M, I)$ be a formal context, $A \subseteq G$ and $B \subseteq M$.

Definition

Support $supp(B) = \frac{|B'|}{|G|}$

minimum support $minsupp \in [0, 1]$: Threshold below which attribute sets are not considered frequent

Iceberg Lattice (Stumme et al. 2002)

Let $\mathbb{K} := (G, M, I)$ be a formal context, $A \subseteq G$ and $B \subseteq M$.

Definition

Support $supp(B) = \frac{|B'|}{|G|}$

minimum support $minsupp \in [0, 1]$: Threshold below which attribute sets are not considered frequent

frequent attribute set $B \subseteq M$: $supp(B) \geq minsupp$

Iceberg Lattice (Stumme et al. 2002)

Let $\mathbb{K} := (G, M, I)$ be a formal context, $A \subseteq G$ and $B \subseteq M$.

Definition

Support $supp(B) = \frac{|B'|}{|G|}$

minimum support $minsupp \in [0, 1]$: Threshold below which attribute sets are not considered frequent

frequent attribute set $B \subseteq M$: $supp(B) \geq minsupp$

frequent concept (A, B): B is frequent attribute set

Iceberg Lattice (Stumme et al. 2002)

Definition

The iceberg concept lattice of a context $\mathbb{K}$ comprises all frequent concepts.

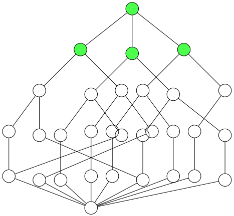


Figure: Iceberg concept lattice coloured.

Iceberg Lattice (Stumme et al. 2002)

Definition

The iceberg concept lattice of a context $\mathbb{K}$ comprises all frequent concepts.

TITANIC algorithm computes iceberg concept lattices.

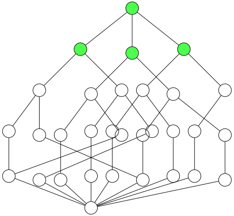
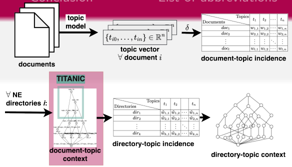


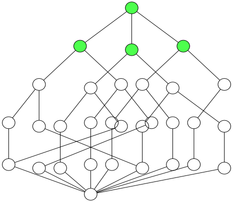
Figure: Iceberg concept lattice coloured.

Directory-Topic Concept Lattice



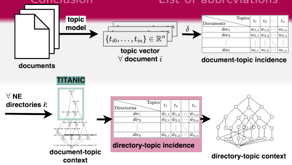
Assuming all documents are organized in directories.

1. Top 10 topics per document incl. relevance scores
2. Compute threshold δ s.t. document-topic matrix density < 0.1
3. $\mathcal{D}_\delta(i,j) = \begin{cases} 1 & \text{if } \ w_{i,j} \geq \delta, \\ 0 & \text{else.} \end{cases}$
4. Compute iceberg lattice for each directory



Iceberg concept lattice coloured.

Directory-Topic Concept Lattice



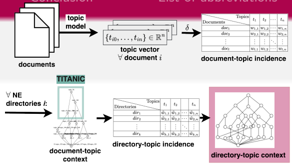
Assuming all documents are organized in directories.

1. Top 10 topics per document incl. relevance scores
2. Compute threshold δ s.t. document-topic matrix density < 0.1
3. $\mathcal{D}_\delta(i,j) = \begin{cases} 1 & \text{if } w_{i,j} \geq \delta, \\ 0 & \text{else.} \end{cases}$
4. Compute iceberg lattice for each directory
5. Combine frequent topics across directories into directory-topic matrix

Table: Weighted directory-topic relation. The weights $\tilde{w}_{i,j} \in \{0,1\}$ represent whether a topic t_j is present in a directory $\hat{d}_i$.

Directories \ Topics	t_1	t_2		t_n
$\hat{d}_1$	$\tilde{w}_{1,1}$	$\tilde{w}_{1,2}$	$\cdots$	$\tilde{w}_{1,n}$
$\hat{d}_2$	$\tilde{w}_{2,1}$	$\tilde{w}_{2,2}$	$\cdots$	$\tilde{w}_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\hat{d}_k$	$\tilde{w}_{k,1}$	$\tilde{w}_{k,2}$	$\cdots$	$\tilde{w}_{k,n}$

Directory-Topic Concept Lattice



Assuming all documents are organized in directories.

1. Top 10 topics per document incl. relevance scores
2. Compute threshold δ s.t. document-topic matrix density < 0.1
3. $\mathcal{D}_\delta(i,j) = \begin{cases} 1 & \text{if } w_{i,j} \geq \delta, \\ 0 & \text{else.} \end{cases}$
4. Compute iceberg lattice for each directory
5. Combine frequent topics across directories into directory-topic matrix

$\rightarrow$ Directory-topic context

Topics

- 50 words per topic
- Multi-lingual

Topics

- 50 words per topic
- Multi-lingual

Table: Subset of translated topic IDs and their top 5 descriptive words for the directory-topic concept lattice.

Topic ID	Topic Words				
34	gun	weapons	rifles	pistol	firearm
36	terminological	nonverbal	linguist	codeword	longtext
54	microelectron	microkernels	microparticle	microphysics	microcosmic
56	pythonpath	pythonmac	python_version	python_object	pythonpowered
58	nuclear	hazardous	explosive	nanotoxicity	chernobyl
94	project	fundraiser	donation	volunteering	charity

Directory-Topic Concept Lattice

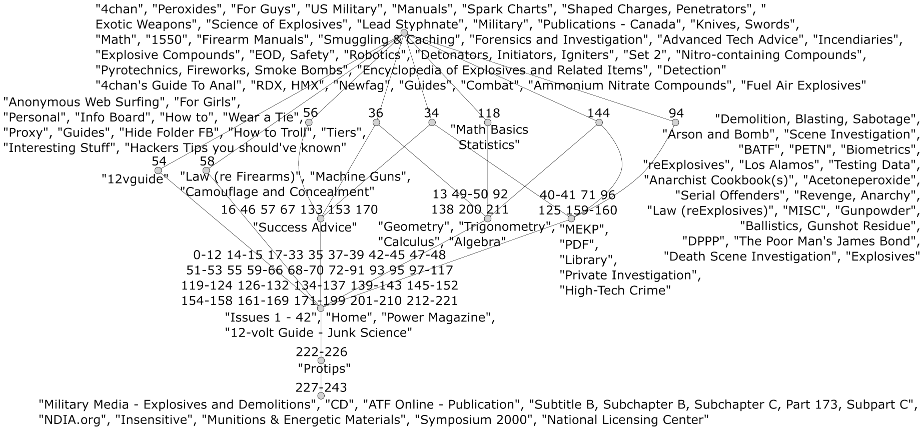


Figure: ETYNTKE directory-topic concept lattice. The text entries below the nodes represent directory names, number above nodes denote topic IDs.

Directory-Topic Concept Lattice

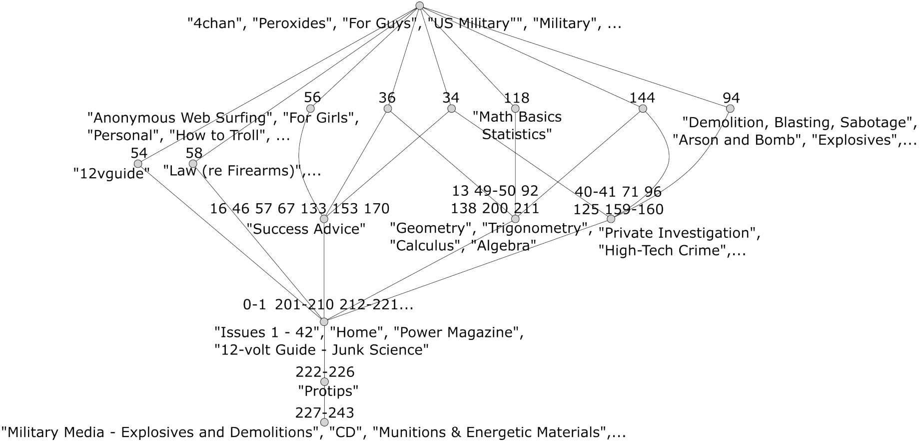


Figure: Incomplete ETYNTKE directory-topic concept lattice. The text entries below the nodes represent directory names, number above nodes denote topic IDs.

What I've learnt about ETYNTKE

- Majority of bottom nodes are subdirectories

What I've learnt about ETYNTKE

- Majority of bottom nodes are subdirectories
- Lower nodes introduce philosophical topics

What I've learnt about ETYNTKE

- Majority of bottom nodes are subdirectories
- Lower nodes introduce philosophical topics
- Intriguing thematic associations
  - `High-Tech Crime` contains explosives related topics
  - `Success Advice` contains chemical related topics

Conclusion

Objective

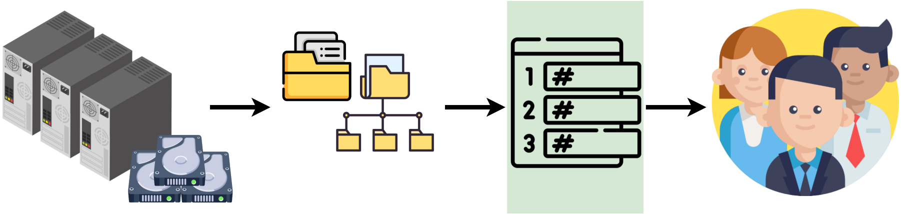
Gain an initial understanding of the data.

Conclusion

Objective

Gain an initial understanding of the data.

+ Dataset exploration without manual inspection

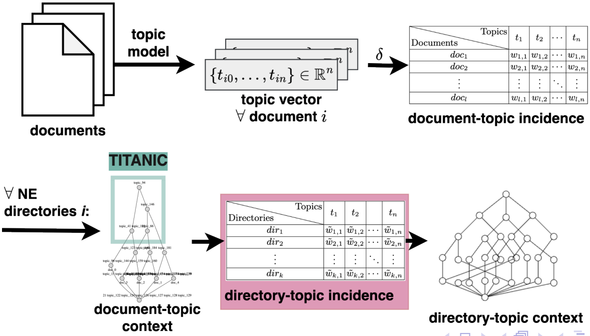


Conclusion

Objective

Gain an initial understanding of the data.

- + Dataset exploration without manual inspection
- + Semantic relationship of directories

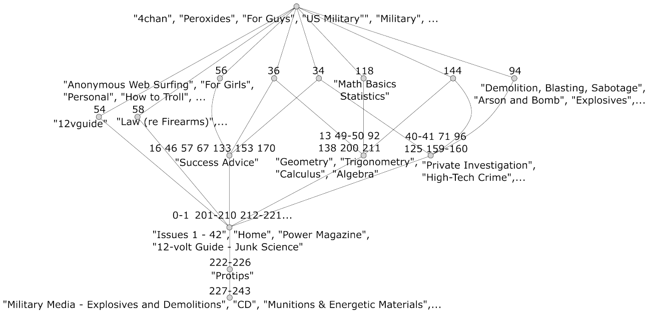


Conclusion

Objective

Gain an initial understanding of the data.

- \+ Dataset exploration without manual inspection
- \+ Semantic relationship of directories
- \- Unintuitive for inexperienced users



Conclusion

Objective

Gain an initial understanding of the data.

+ Dataset exploration without manual inspection
+ Semantic relationship of directories
- Unintuitive for inexperienced users
- Numerical topic representation

Topic ID	Topic Words				
34	gun	weapons	rifles	pistol	firearm
36	terminological	nonverbal	linguist	codeword	longtext
54	microelectron	microkernels	microparticle	microphysics	microcosmic
56	pythonpath	pythonmac	python_version	python_object	pythonpowered
58	nuclear	hazardous	explosive	nanotoxicity	chernobyl
94	project	fundraiser	donation	volunteering	charity

Conclusion

Objective

Gain an initial understanding of the data.

+ Dataset exploration without manual inspection
+ Semantic relationship of directories
- Unintuitive for inexperienced users
- Numerical topic representation

Thank you for your attention!

BoW	Bag of Words
ETYNTKE	Everything You Need To Know Ever
FAT-CAT	FCA-based Aggregation for Topics using Conceptual Analysis and Taxonomies
FCA	Formal Concept Analysis
GIT	Generative Image-to-text Transformer
NLP	Natural Language Processing
SBERT	Sentence-BERT
USE	Universal Sentence Encoder

FAT-CAT

FAT-CAT FCA-based Aggregation for Topics using Conceptual Analysis and Taxonomies

Our contributions:

- Open-source[2]
- Minimal training requirements
- Non-textual data
- Information aggregation across directories

[2] https://github.com/KlaraGtknst/text_topic (29.01.2025) & https://github.com/KlaraGtknst/clj_exploration_leaks (29.01.2025)

Content of `Military` directory

```
[kgu@watzmann:/norgay/bigstore/kgu/data/ETYNTKE/Military$ ls
'American & NVA Weapons & Tactics In Vietnam.pdf'
"Army Non-Commissioned Officer's Guide - FM 7-22-7.pdf"
'Attacks - Erwin Rommel.pdf'
'Company Officer'\''s Handbook of the German Army 1944 (MID Special Series No. 22).pdf'
'Corp Operations (1998) - B-GL-321-001.pdf'
'Decisive Force - The Army in Theater of Operations - FM 100-7.pdf'
'Encyclopedia of Modern US Military Weapons.pdf'
'German Infantry Weapons 1943 (MIS Special Series No. 14).pdf'
'German Methods of Warfare in the Libyan Desert (Information Bulletin No. 20).pdf'
'German Military Training 1942 (MIS Special Series No. 03).pdf'
'German Mountain Warfare (MID Special Series No. 21).pdf'
'German Rifle Company-For Study and Translation (MIS Special Series No. 15).pdf'
'German Squad in Combat 1943 (MIS Special Series No. 09).pdf'
'German Tactical Doctrine (MIS Special Series No. 08).pdf'
'German Winter Warfare (MID Special Series No. 18).pdf'
"Handbook of the Chinese People's Liberation Army.pdf"
'History of US Army Lessons Learned.pdf'
'Land Force Command (1996) - B-GL-300-003.pdf'
'Land Force Information Operations (1999) - B-GL-300-00.pdf'
'Land Force Sustainment (1999) - B-GL-300-004.pdf'
'Land Force Tactical Doctrine (1997) - B-GL-300-002.pdf'
'Logistics Operations - MCWP 4-1.pdf'
 Mercs_-_True_Stories_of_Mercenaries_In_Action_-_Bill_Fawcett.rar
'Military Almanac 2001-2002.pdf'
'Military Manuals'
'Military Training'
'Nazi Robot Soldiers.pdf'
'Operational Level Doctrine for the Canadian Army (1998) - B-GL-300-001.pdf'
'Professional Knowledge Gained from Operational Experience in Vietnam, 1965-1966 - FMFRP 12-40.pdf'
'Professional Knowledge Gained from Operational Experience in Vietnam, 1968 - FMFRP 12-42.pdf'
'Spetsnaz - FMFRM 3-201.pdf'
'SWAT glossary.rar'
'The Army - FM 1.pdf'
'The Encyclopedia of Weapons of World War II.pdf'
'US Army History Of Lessons Learned.pdf'
'Vietnam Primer - Lessons Learned.pdf'
'Wartime Exploits of a Mindanao Guerrilla - Nick Minecci.pdf'
```

Figure: Screenshot of content of `Military` directory.

Content of Business directory

```
kgu@watzmann:/norgay/bigstore/kgu/data/ETYNTKE/Business$ ls
 5.04.FinancialFitness.pdf
 7HiddenPsychological.pdf
'Benjamin Graham - The Intelligent Investor_investing_+_found_at_redsamara.com.pdf'
 Brands.That.Rock.What.Business.Leaders.Can.Learn.from.the.World.of.Rock.and.Roll.eBook-EEn.pdf
'Business Blogging Toolset--100 Internet Resources for Entrepreneur-Writers.pdf'
'Consulting-Ten Commandments of Small Business [Lesson].doc'
'CRM McGraw Hill - Customer Relationship Management (2002).pdf'
'David Frey - Business - Six Deadly Small Business Marketing Secrets.pdf'
'Donald Trump - Fortune Without Fear Real Estate Riches.pdf'
 dropshipreport.pdf
'(Ebook) - David Frey - The Small Business Marketing Bible 2003 (312 Pages).pdf'
'Ebook -en- Cooking- Jamie Oliver - The Naked Chef 2 Cook Book.pdf'
'(Ebook) Harper Business - Stock Market Wizards (Jack Schwager).pdf'
'[eBook] How To Write A Business Plan - Nolo.pdf'
'Economics - Freakonomics - A Rogue Economist Explores the Hidden Side of Everything - S D Levitt & S J D~0.pdf'
'Father to Son.pdf'
'Finance Investments 5th Ed - Z Bodie, A Kane & A J Marcus (McGraw-Hill) - 2001.pdf'
'Finance, Investment, Stock, Trading - Soros Unauthorized Biography.pdf'
'Finance Investment - The Global Money Markets.pdf'
'Finance Investment - Trump Strategies for Real Estate.pdf'
'Finance Management Project Strategic - Handbook for small business.pdf'
'Five-Minute Mba In Corporate Finance.pdf'
 Harvard.Business.Press.Blue.Ocean.Strategy.2005.pdf
'Harvard Business Review - How information gives you competitive advantage - Michael Porter.pdf'
'Harvard Business Review - The 5 Stages Of Small Business Growth.pdf'
'Harvard Business Review - What Is Strategy - Michael Porter.pdf'
'Harvard Business Review - Why Good Companies Go Bad.pdf'
'Harvard Business School Press - Blue Ocean Strategy (2005).pdf'
'Helping Small Business Owners and Self-Employed Professionals Do More With Less.pdf'
'IDIOTS GUIDE - Complete Idiots Guide to Small Business.pdf'
"Insider's Guide to Trading the World's Stock Markets.pdf"
'Inside The Guru Mind - Peter Drucker.pdf'
'John Wiley & Sons - The 7 Irrefutable Rules of Small Business Growth - 2005.pdf'
'Just 1% The Power of Microtrends.pdf'
'Kaplan Professional - Financing Secrets of a Millionaire Real Estate Investor - 2003.pdf'
'Kaplan - The Business Start Up Kit.pdf'
'Katz & McCormick 2000 The Encyclopedia of Trading Strategies.pdf'
'Learn the Principles of Business Writing.pdf'
 List.txt
'Making A Living From Your eBay Business, 2nd Edition.chm'
'Managing Your Business with Outlook 2003 for Dummies.pdf'
'Marketing Branding - Defending the Brand - Aggressive Strategies for Protecting Your Brand in the Online~0.pdf'
'Marketing, Planning And Strategy.pdf'
'Mba In Finance.pdf'
'Mcgraw-Hill - Fundamentals Of Managerial Economics.pdf'
'McGraw-Hill - The Law of Financial Success.pdf'
'Menshikov - Millionaires and Managers - Structure of US Financial Oligarchy (1969).pdf'
'Powerful Sleep Package'
'Real Estate - Mortgage & Finance.pdf'
'Ron Paul --  Pillars of Prosperity - Free Markets, Honest Money, Private Property.pdf'
'smart strategies for small business.pdf'
 Strategic_Marketing_Handbook.pdf
'Strategy and the Internet - Michael Porter (Harvard Business Review).pdf'
'Tao Te Ching.pdf'
'The Creation of Conscious Culture through Educational Innovation.pdf'
'The Financial Analyst'\''s Handbook (Stock Market).pdf'
'The Secrets of Market-Driven Leaders A Pragmatic Approach.pdf'
'The Turnover Dilemma A Question to Keep Employees.pdf'
'The World Is Flat A Brief History Of The 21St Century.pdf'
'think and grow rich.pdf'
 Trader
'Trading For A Living In The Forex Market.pdf'
'What You Would Learn At Top-Tier Business Schools - Wiley 2004.pdf'
'Wiley Intermarket Technical Analysis Trading Strategies For The Global S Vorschau Commodity, And Curre~0.pdf'
```

Figure: Screenshot of content of Business directory.

Content of Business directory

Wordcloud of directory: Business

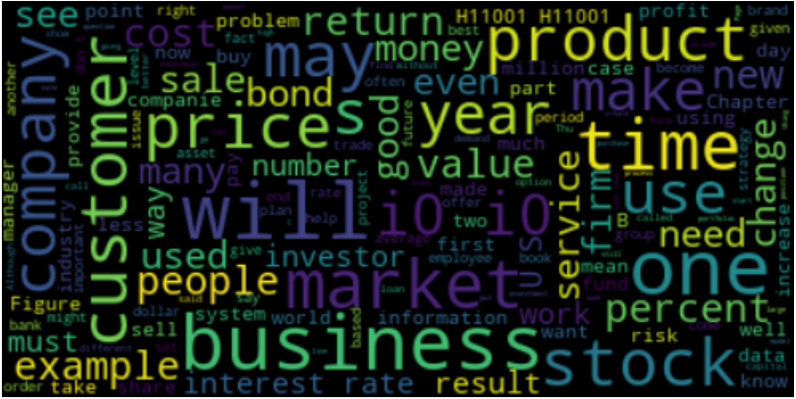
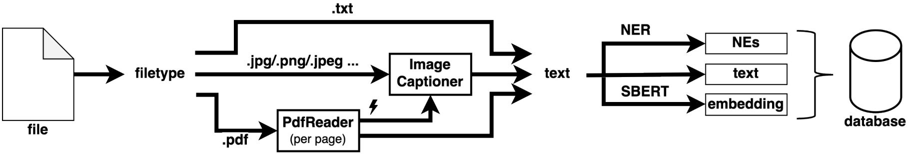


Figure: Wordcloud[3] of content of Business directory.

[3] https://github.com/amueller/word_cloud (01.02.2025)

Text Processing



Appendix details: Image Captioner SBERT

Image Captioner (Wang et al. 2022)

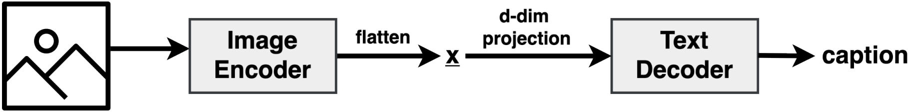
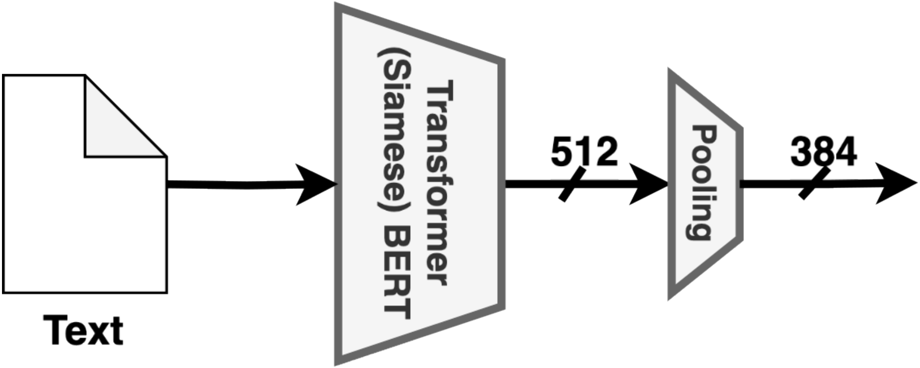


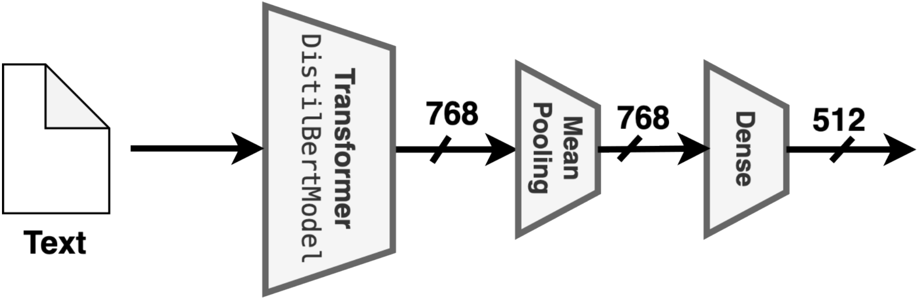
Figure: Generative Image-to-text Transformer (GIT)[4]

[4] https://huggingface.co/microsoft/git-base (29.01.2025)

Sentence-BERT (SBERT) (Reimers and Gurevych 2019)



USE (Yang et al. 2019)



References I

Angelov, Dimo (2020). "Top2vec: Distributed representations of topics". In: *arXiv preprint arXiv:2008.09470*.

Cigarrán, Juan, Ángel Castellanos, and Ana García-Serrano (2016). "A step forward for Topic Detection in Twitter: An FCA-based approach". In: *Expert Systems with Applications* 57, pp. 21–36.

Hirth, Johannes and Tom Hanika (2024). "The Geometric Structure of Topic Models". In: *CoRR* abs/2403.03607. DOI: 10.48550/ARXIV.2403.03607. arXiv: 2403.03607.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics*.

Stumme, Gerd et al. (2002). "Computing iceberg concept lattices with titanic". In: *Data & knowledge engineering* 42.2, pp. 189–222.

Wang, Jianfeng et al. (2022). "Git: A generative image-to-text transformer for vision and language". In: *arXiv preprint arXiv:2205.14100*.

References II

Yang, Yinfei et al. (2019). "Multilingual universal sentence encoder for semantic retrieval". In: *arXiv preprint arXiv:1907.04307*.