LLM-powered topic modeling

Washington University in St. Louis

Agenda

What is topic modeling?

How can it be applied to your research

Traditional topic models

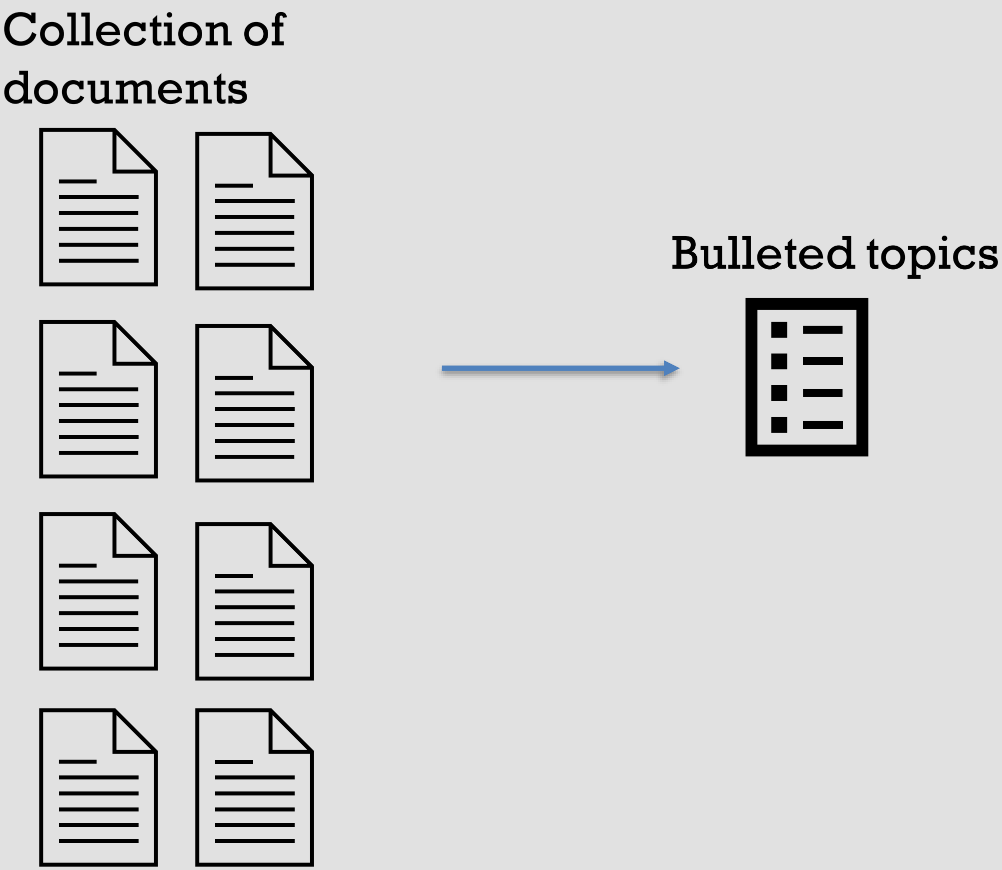
LDA

LLM-powered topic models

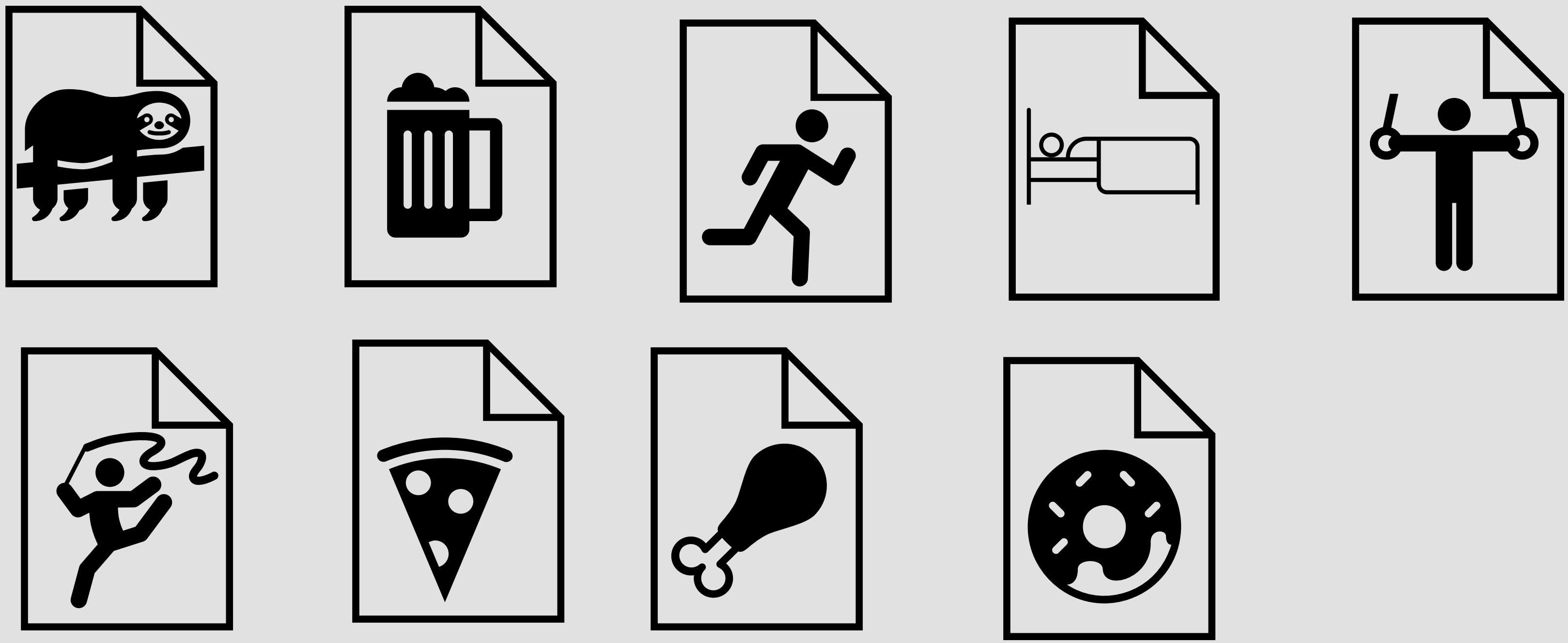
1. BERTopic

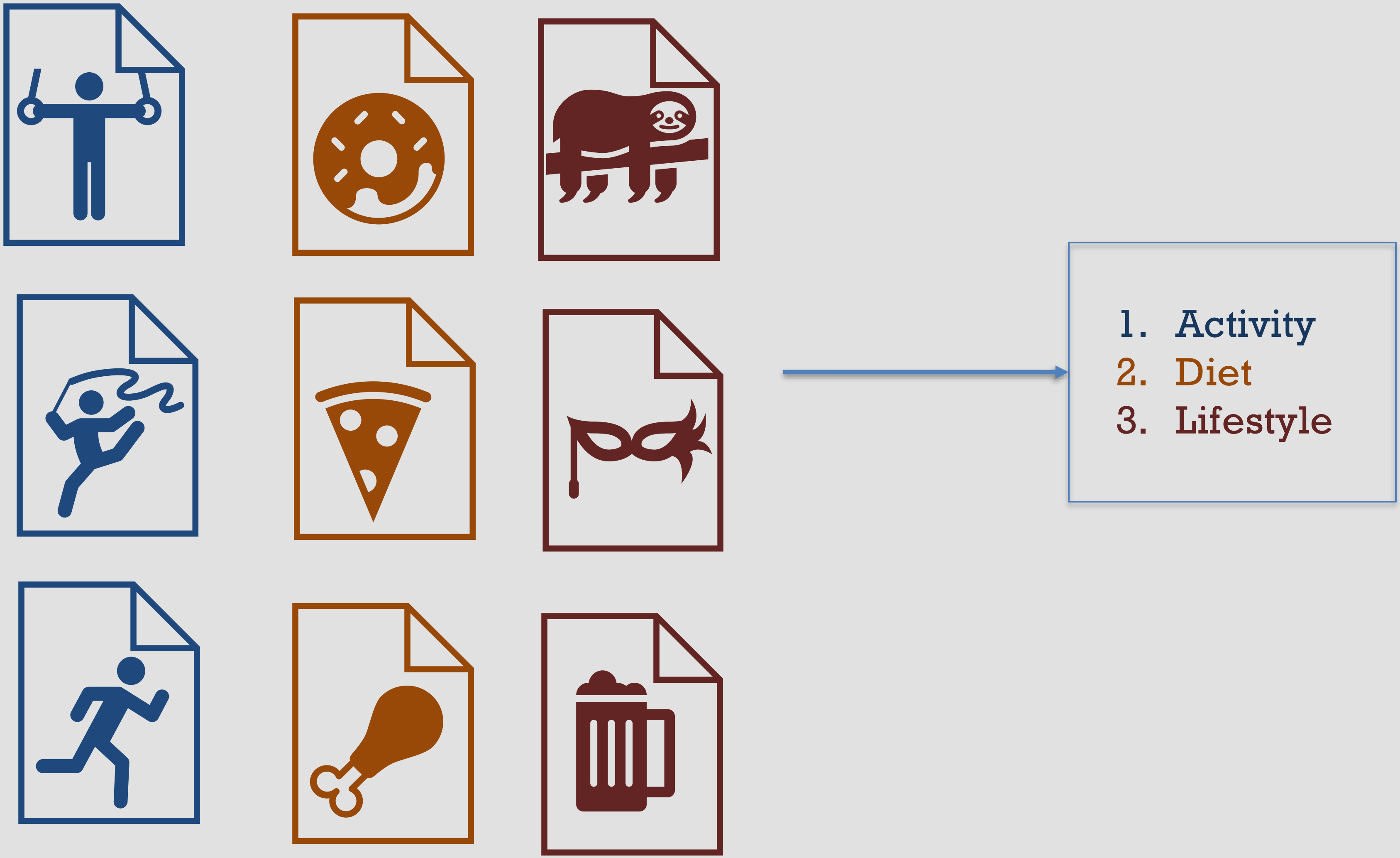
2. TopClus (separate presentation)

Topic modeling



RQ: What are the causes of obesity?



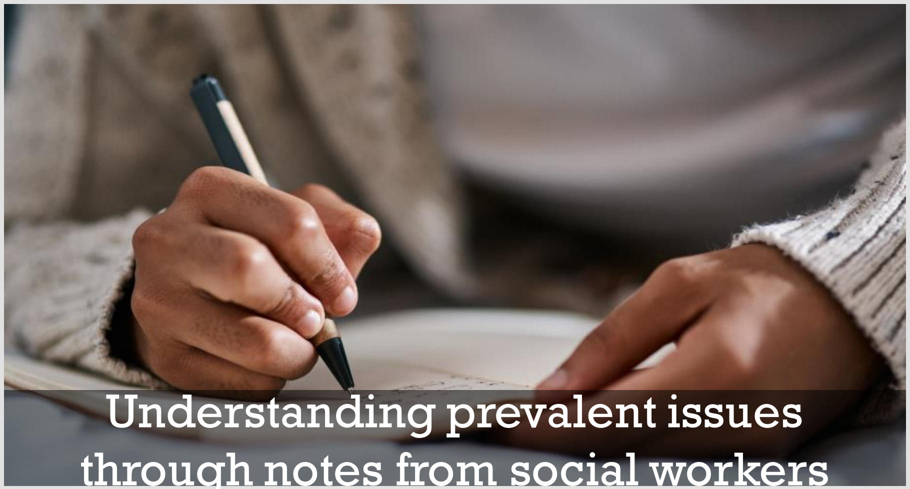

1. Activity
2. Diet
3. Lifestyle

Notes about topic modeling

Requires strong domain knowledge

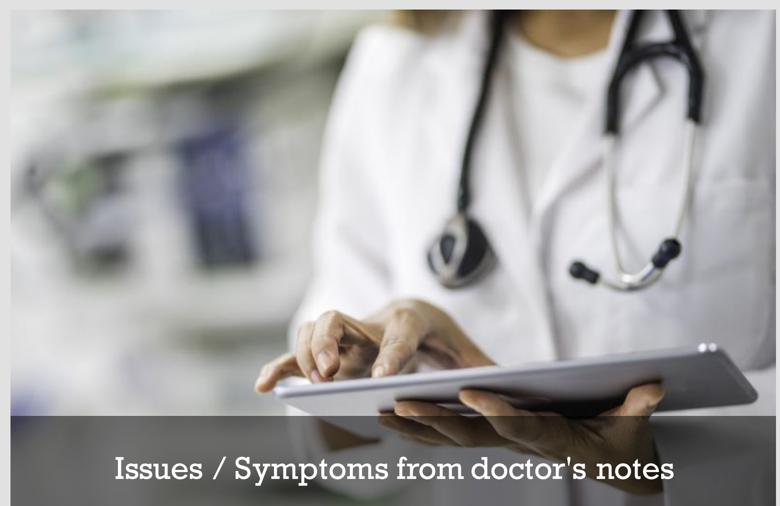
Topic modeling ≠ summary

Examples of topic models in PH / SW

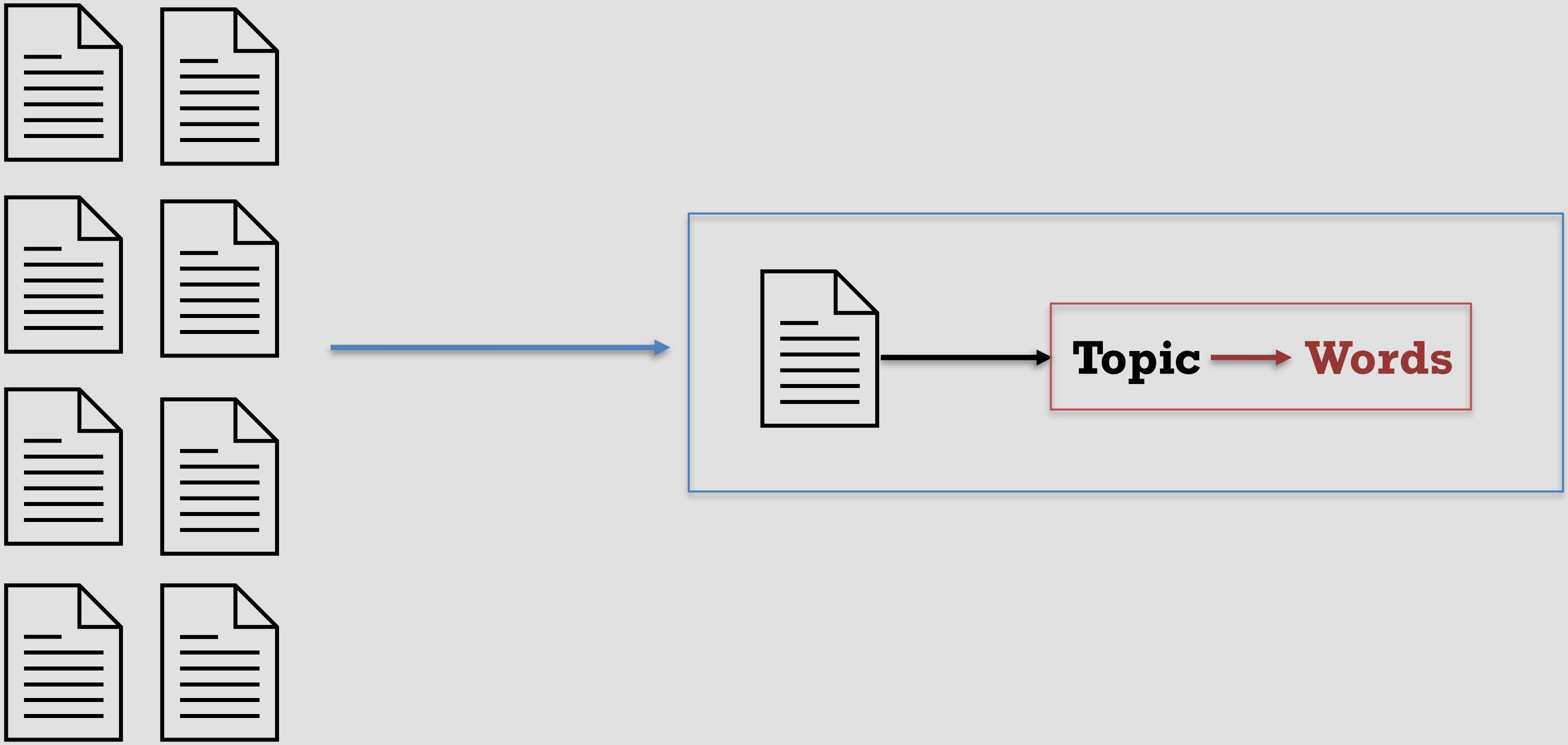

Understanding prevalent issues through notes from social workers

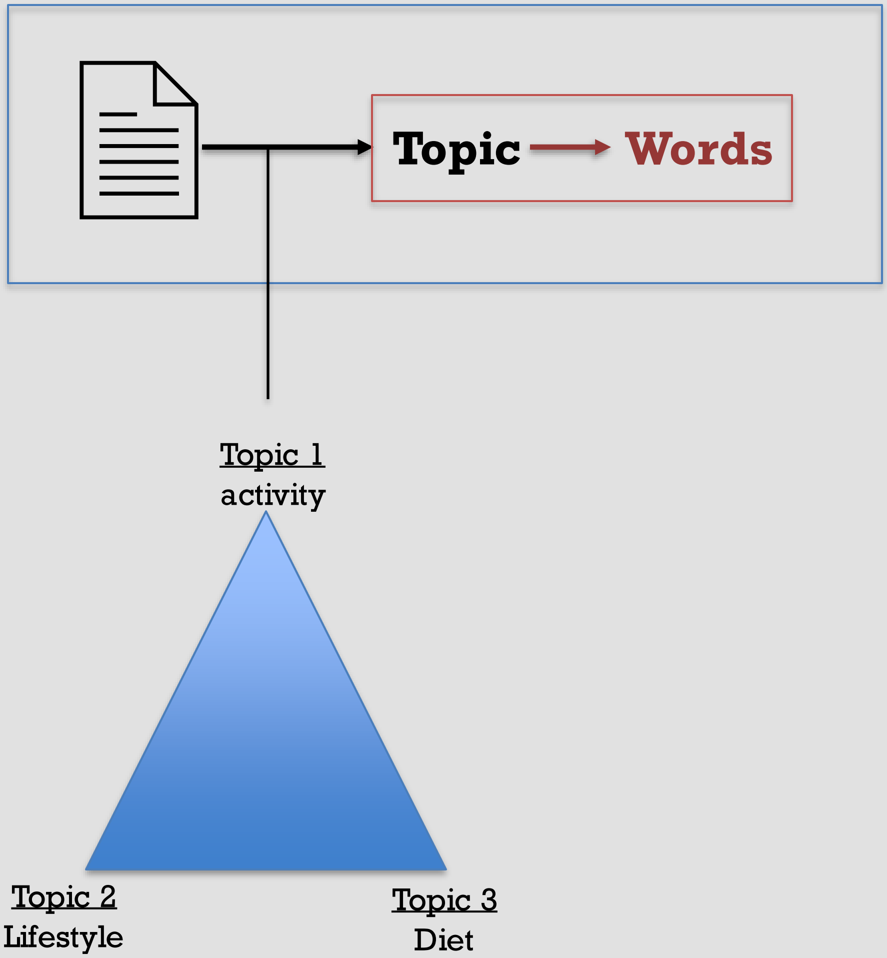

Understanding topics concerning an issue from social media mining posts (eg healthcare in America)


Understanding prevalent issues from news coverage

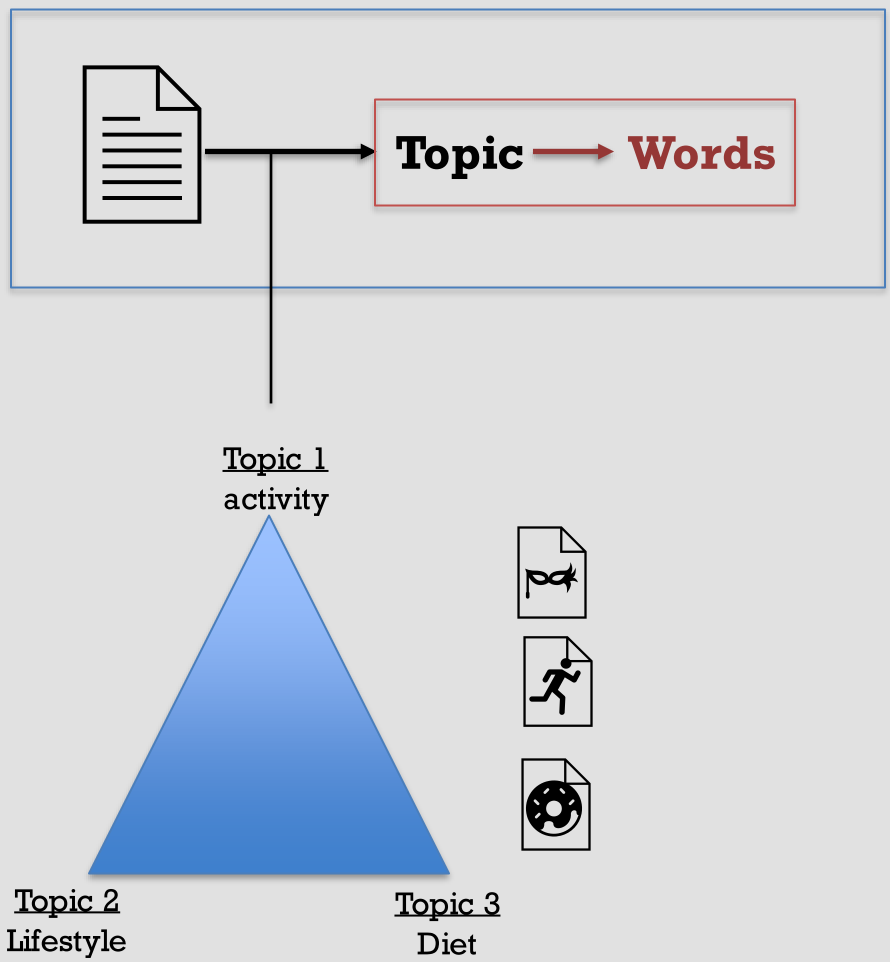

Issues / Symptoms from doctor's notes

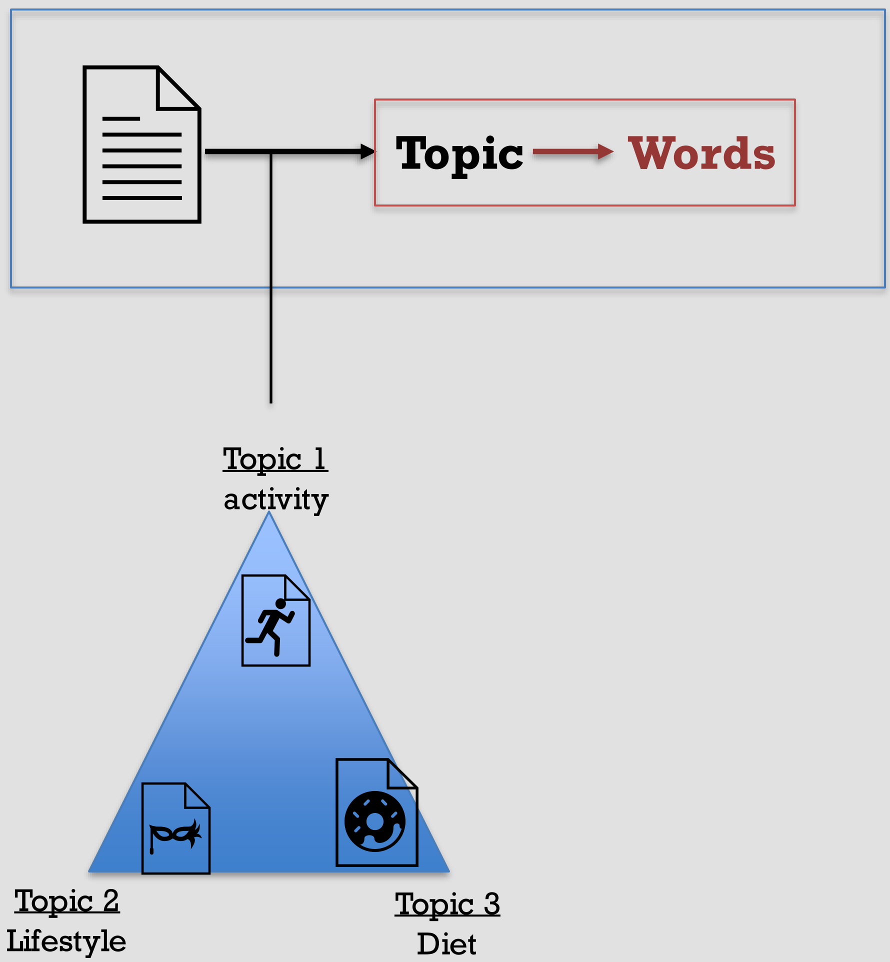
Traditional method – Latent **Dirichlet** allocation (LDA)

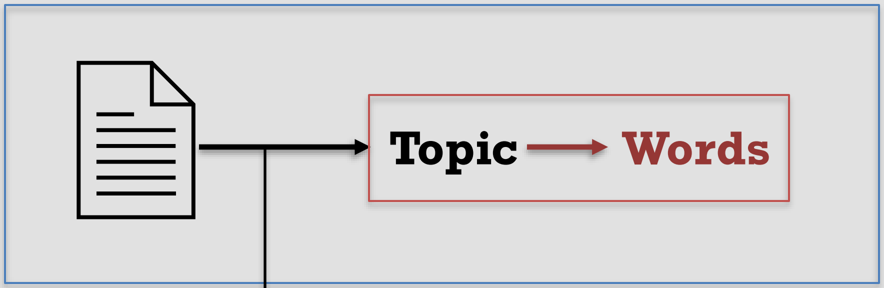


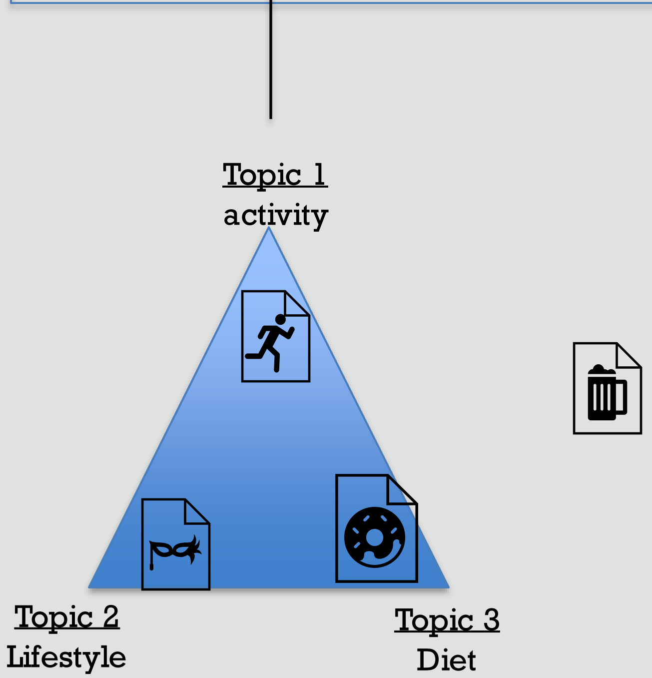


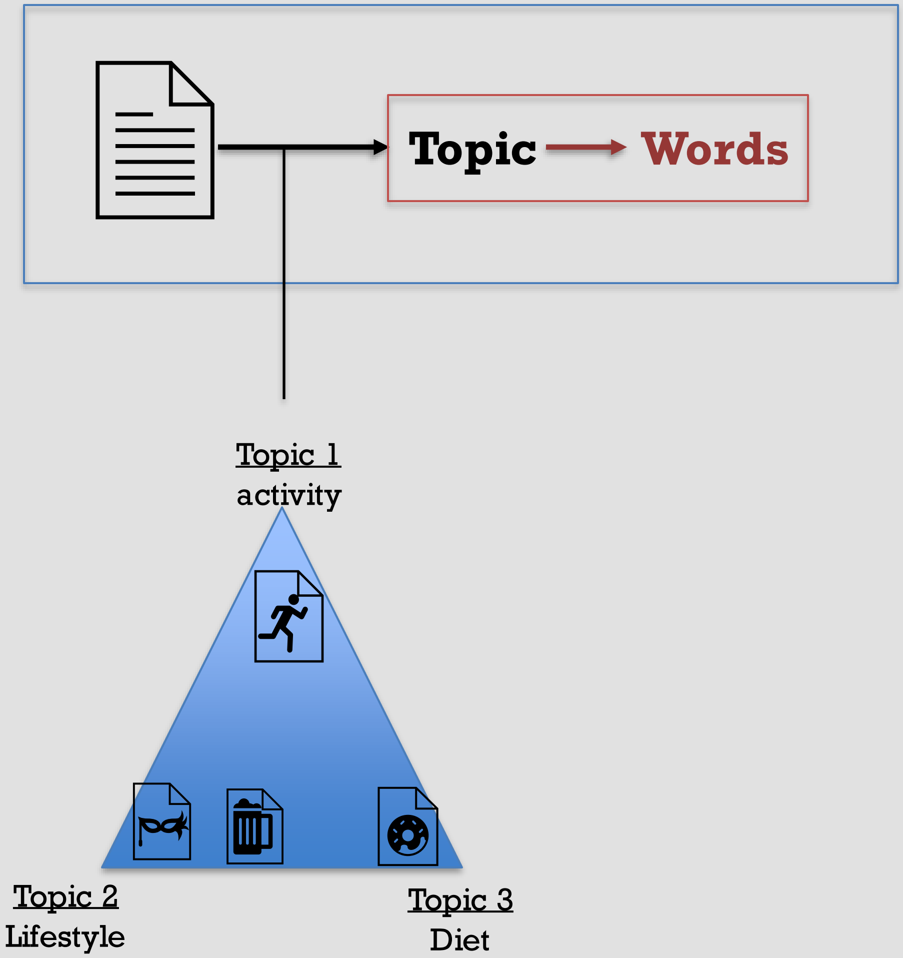
Topic
Words
Topic 1
activity
Topic 2
Lifestyle
Topic 3
Diet

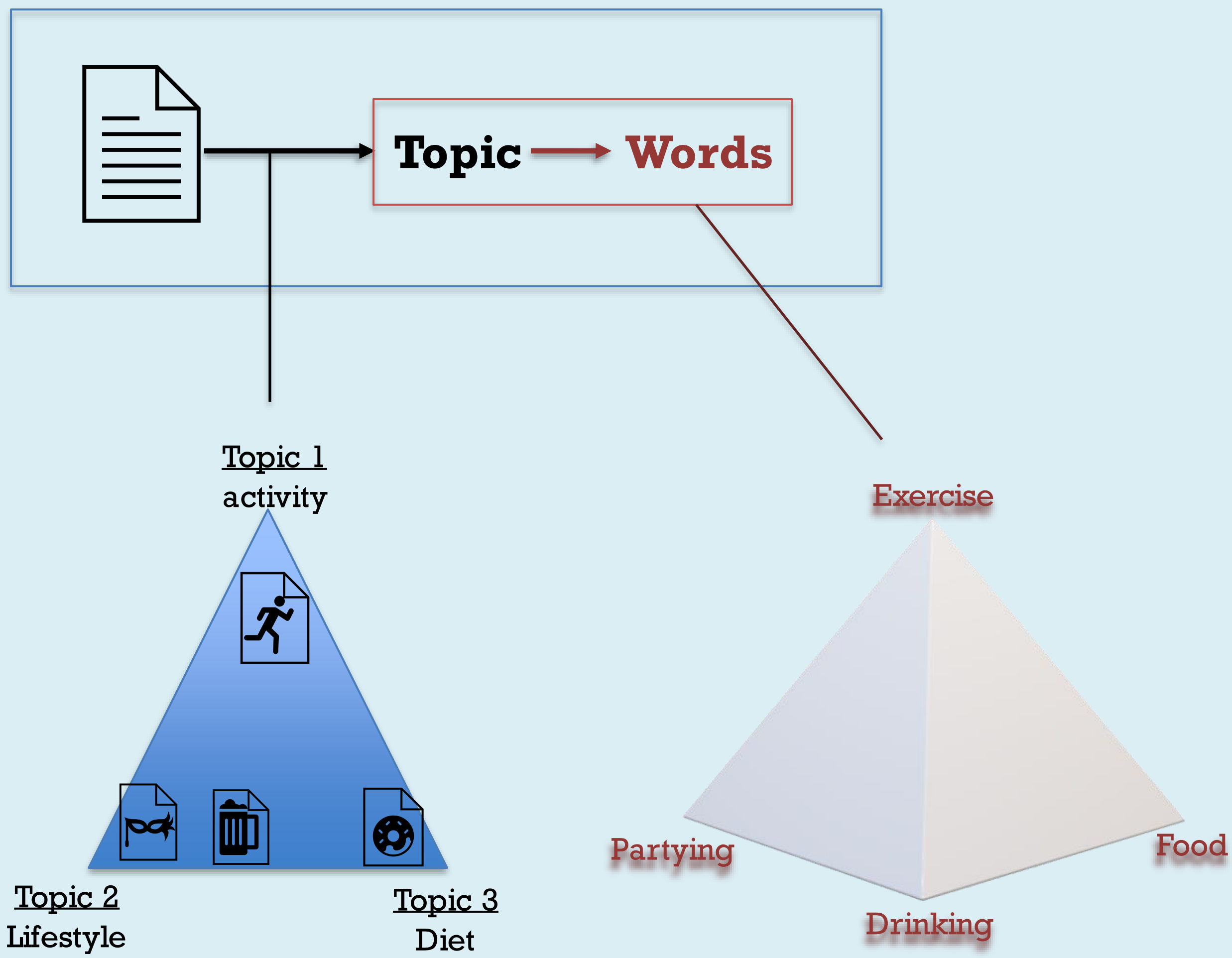

Topic
Words
Topic 1
activity
Topic 2
Lifestyle
Topic 3
Diet

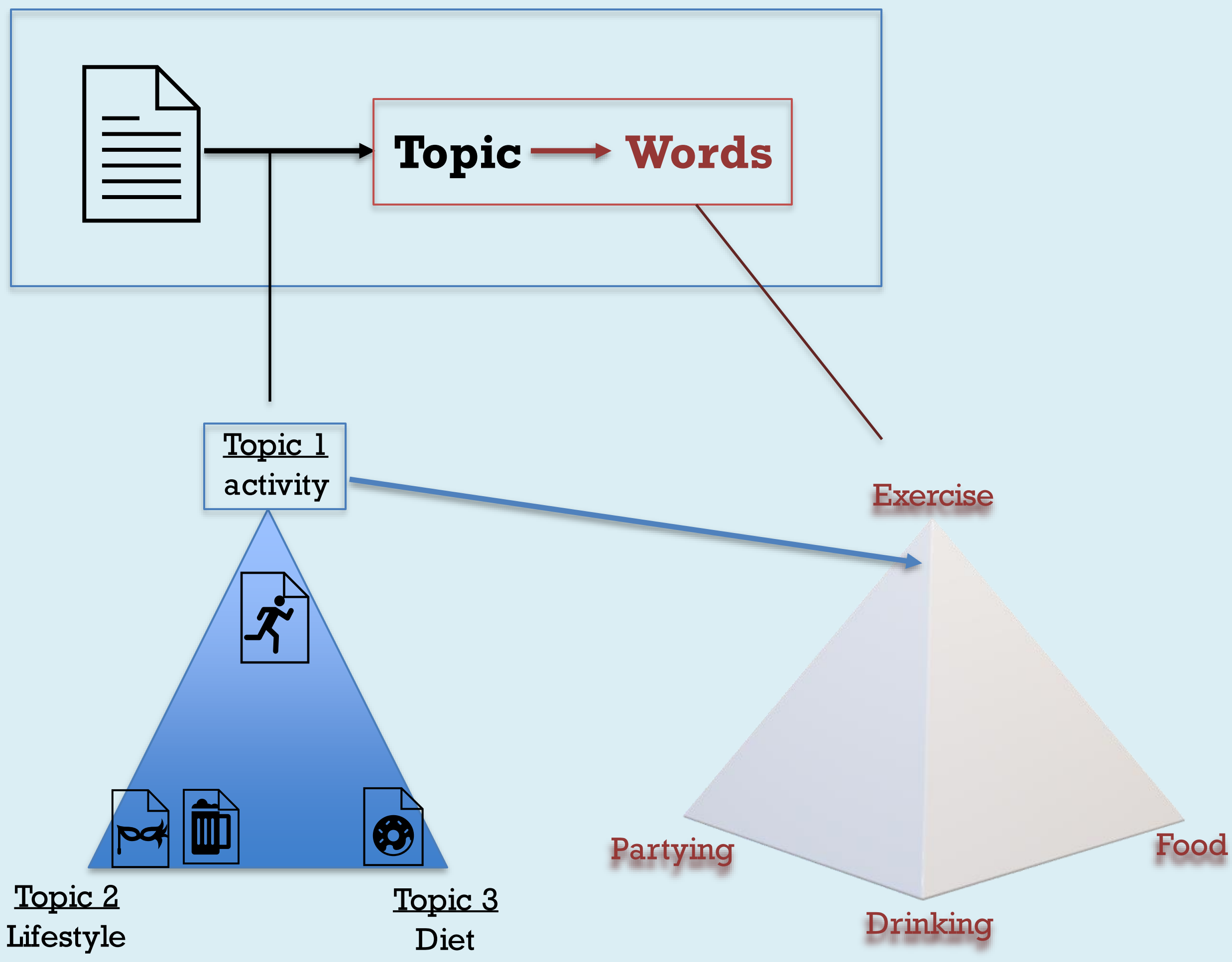

Topic
Words
Topic 1
activity
Topic 2
Lifestyle
Topic 3
Diet


Topic
Words

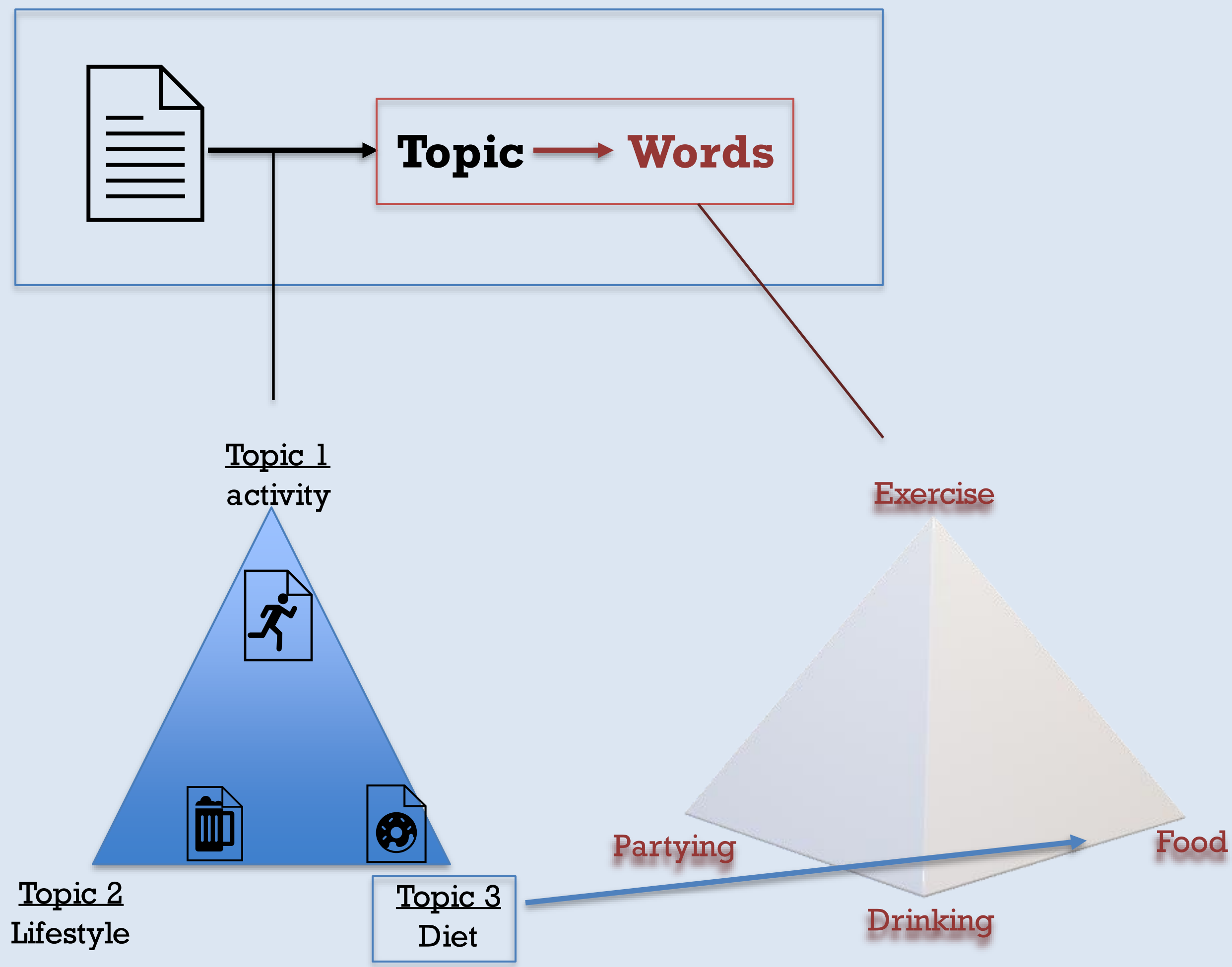

Topic 1
activity
Topic 2
Lifestyle
Topic 3
Diet

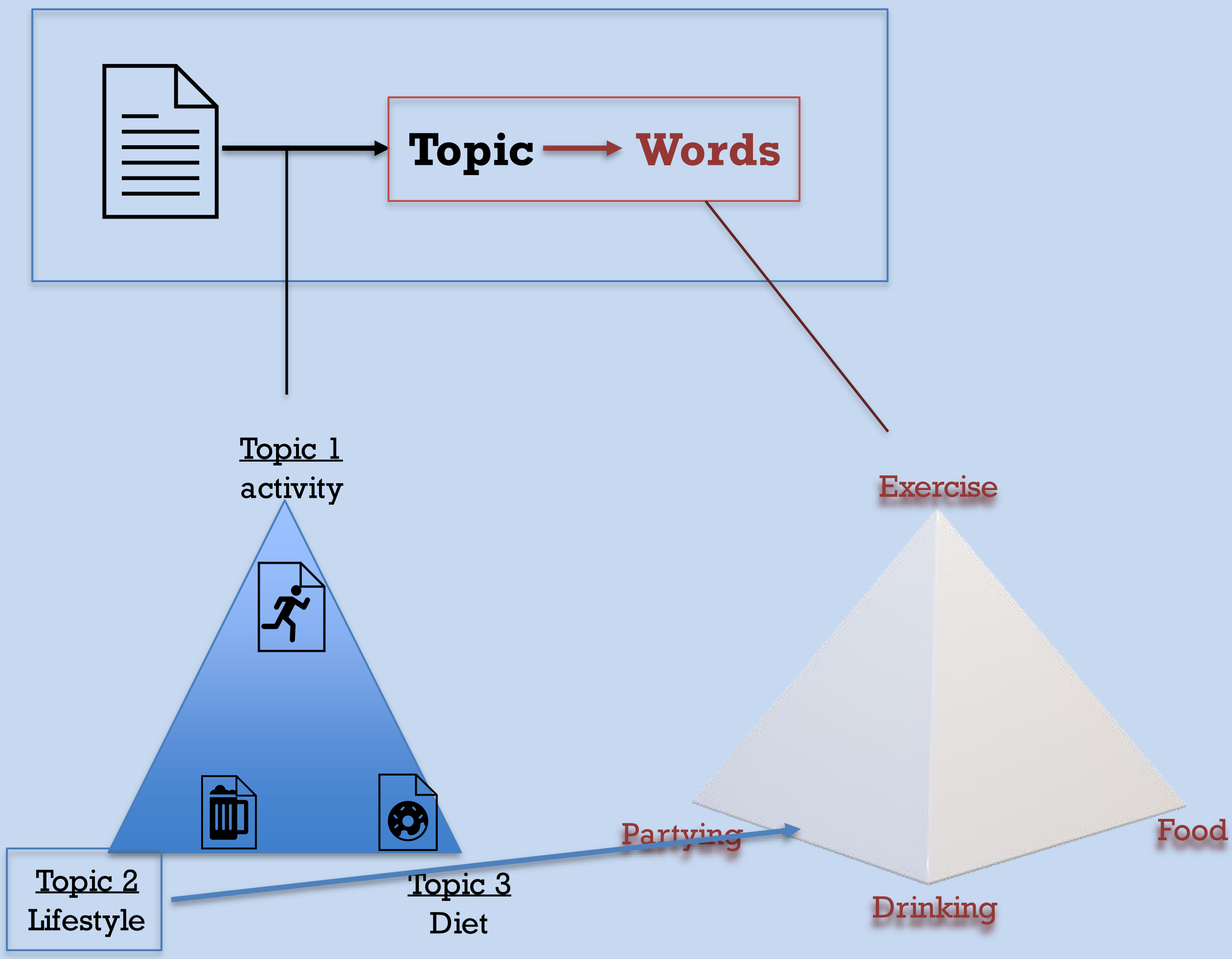

Topic
Words
Topic 1
activity
Topic 2
Lifestyle
Topic 3
Diet


Topic
Words
Topic 1
activity
Topic 2
Lifestyle
Topic 3
Diet
Exercise
Partying
Drinking
Food



Topic
Words
Topic 1
activity
Topic 2
Lifestyle
Topic 3
Diet
Exercise
Partying
Drinking
Food


Topic
Words
Topic 1
activity
Topic 2
Lifestyle
Topic 3
Diet
Exercise
Partying
Food
Drinking



Topic
Words
Topic 1
activity
Topic 2
Lifestyle
Topic 3
Diet
Exercise
Partying
Food
Drinking

FYI

This iterative process is done through **gibbs sampling**

- a Monte-Carlo Markov-chain (MCMC) method

Problems with LDA

Polysemy

Attention & Long range dependencies

Polysemy

Eg: Documents about things to do in LA

"Let's go to a LA **Galaxy** soccer game"

"Let's go learn about the **Galaxy** in the Griffith Observatory"

Polysemy

"Lets go to a LA **Galaxy** soccer game"

"Lets go learn about the **Galaxy** in the Griffith Observatory"

Attention & long-range dependency

<u>Low-income communities</u> are more prevalent to obesity due to their lack of investment in unhealthy food environments.

Attention & long-range dependency

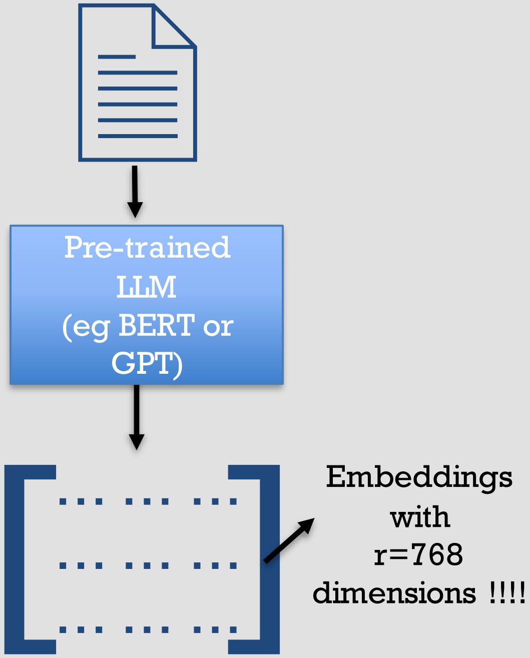
Low-income communities are more prevalent to obesity due to **their lack of investment in unhealthy food environments**.

Solution

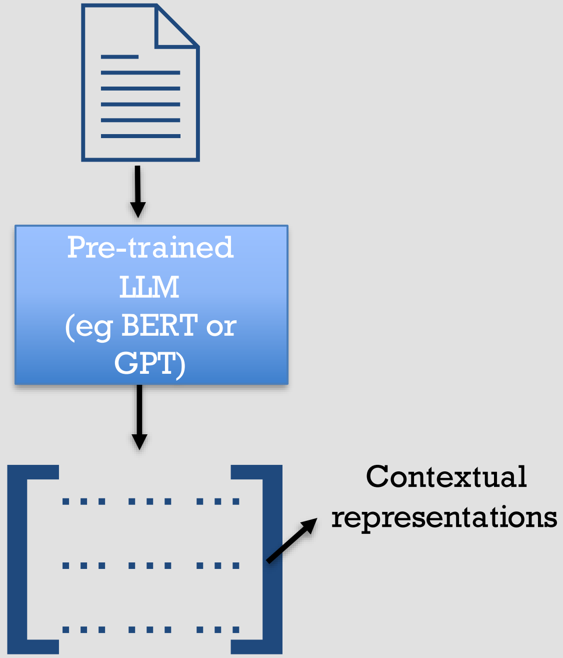
Pre-trained Large Language models (eg GPT or BERT)

Problems in using pre-trained LLMs in Topic Modeling

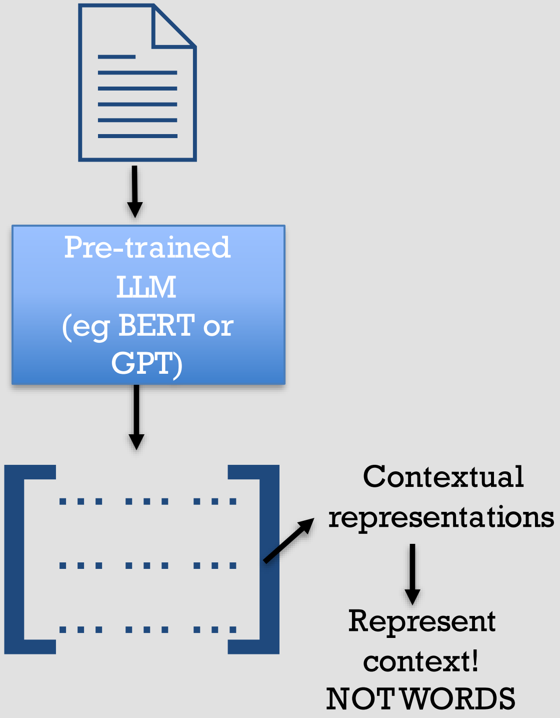
Curse of dimensionality



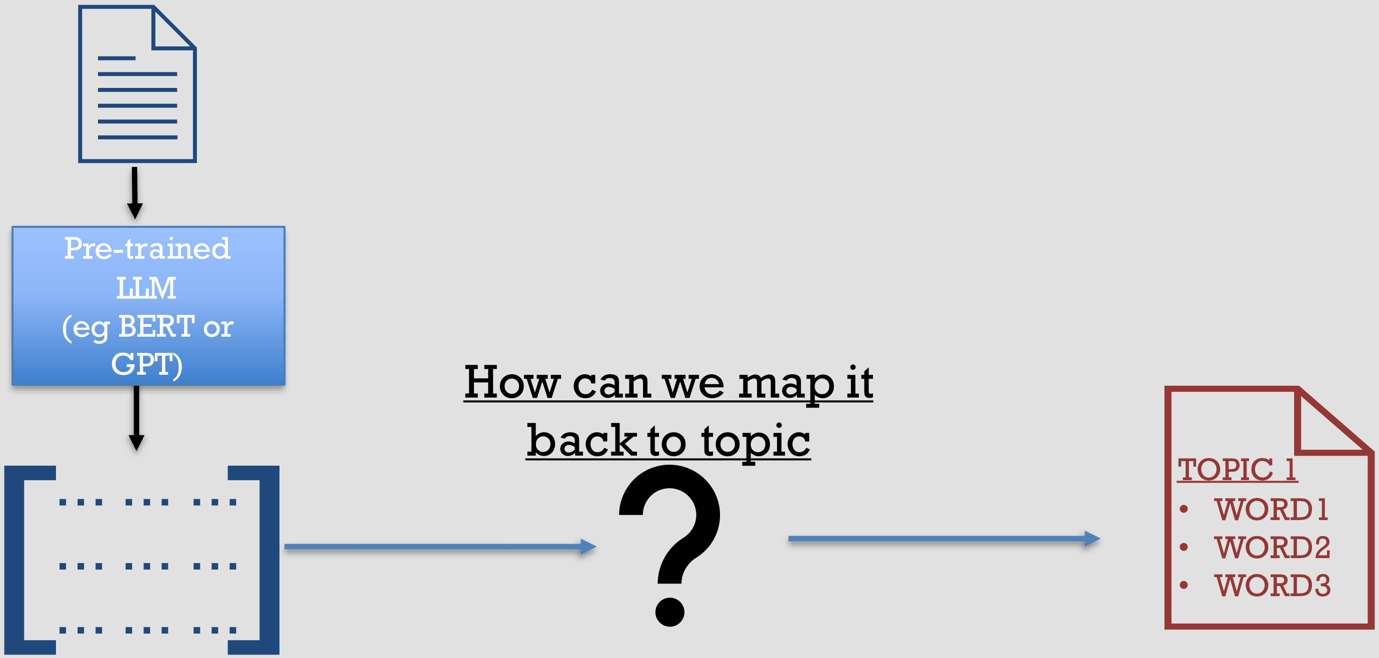
Lack of good document representation



Lack of good document representation



Lack of good document representation



which needs to be represented by words???

Unsuitable of PLMs for clustering

Recap:

1. The main training objectives of BERT-based models are:
    1. Masked Language Modeling (MLM)
    2. Next Sentence Prediction (NSP)

Unsuitable of PLMs for clustering

Recap:

1. The main training objectives of BERT-based models are:
   1. **Masked Language Modeling (MLM)**
   2. Next Sentence Prediction (NSP)

Unsuitable of PLMs for clustering

Masked Language Modeling

- original sentence

- Obesity can be caused by a lack of exercise, poor nutritional diets, and excessive drinking.

- Masked Sentence

- **[MASK]** can be caused by a lack of exercise, poor **[MASK]** diets, and excessive drinking.

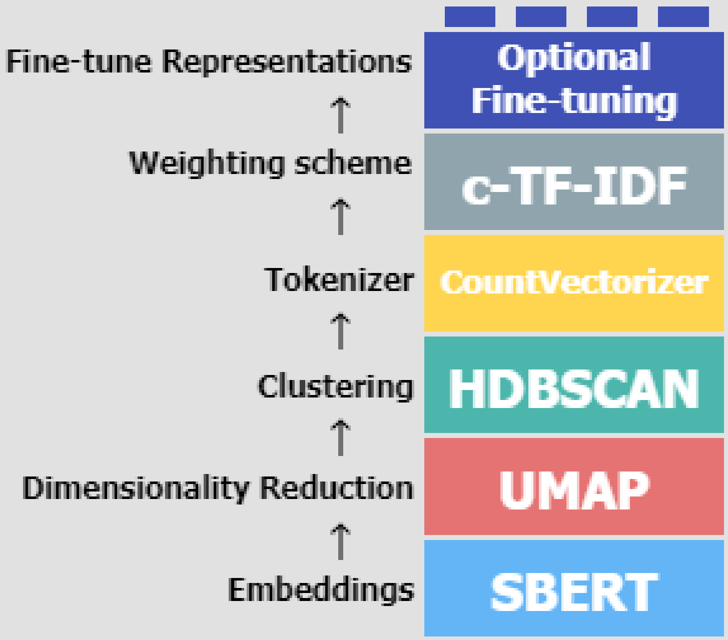
Model is trained to predict the **[MASK]**

Unsuitable of PLMs for clustering

- If that were the case:
  - Assume we have V number of tokens in pre-trained BERT
    - *# of tokens in the model |V|*
  - The 'optimal' number of topics (T):

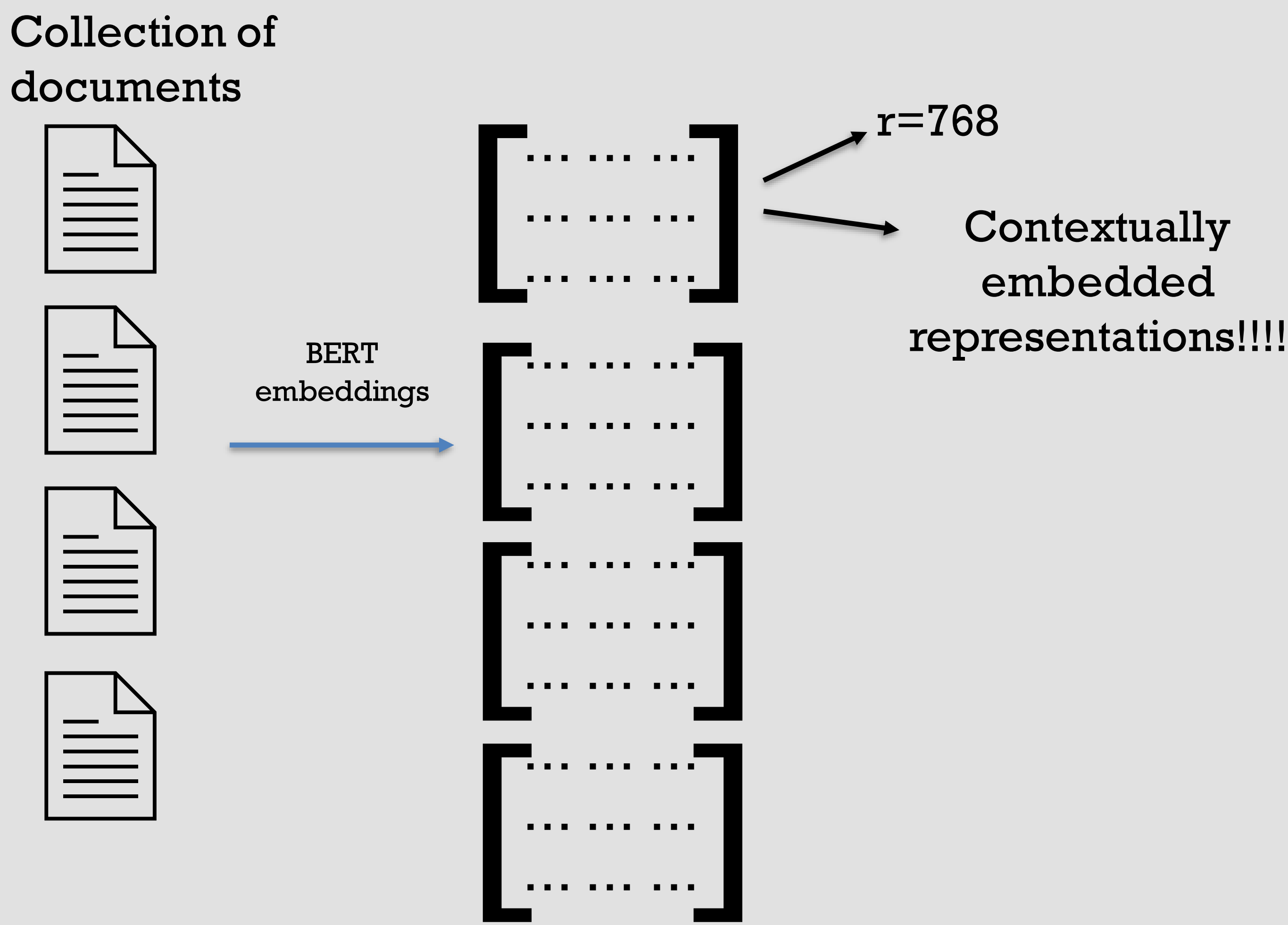
    T=|V|≈30,000

BERTopic



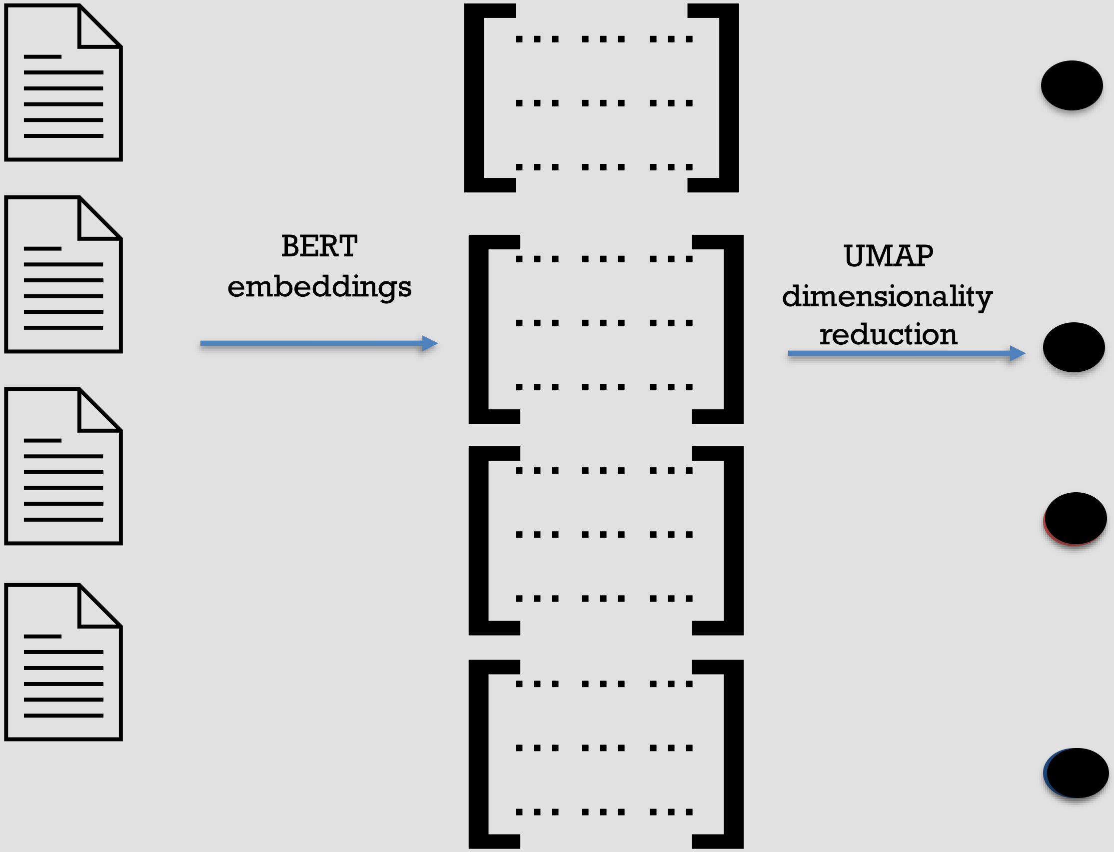
Fine-tune Representations
Optional Fine-tuning
Weighting scheme
c-TF-IDF
Tokenizer
CountVectorizer
Clustering
HDBSCAN
Dimensionality Reduction
UMAP
Embeddings
SBERT

Part I: Grouping documents into topics



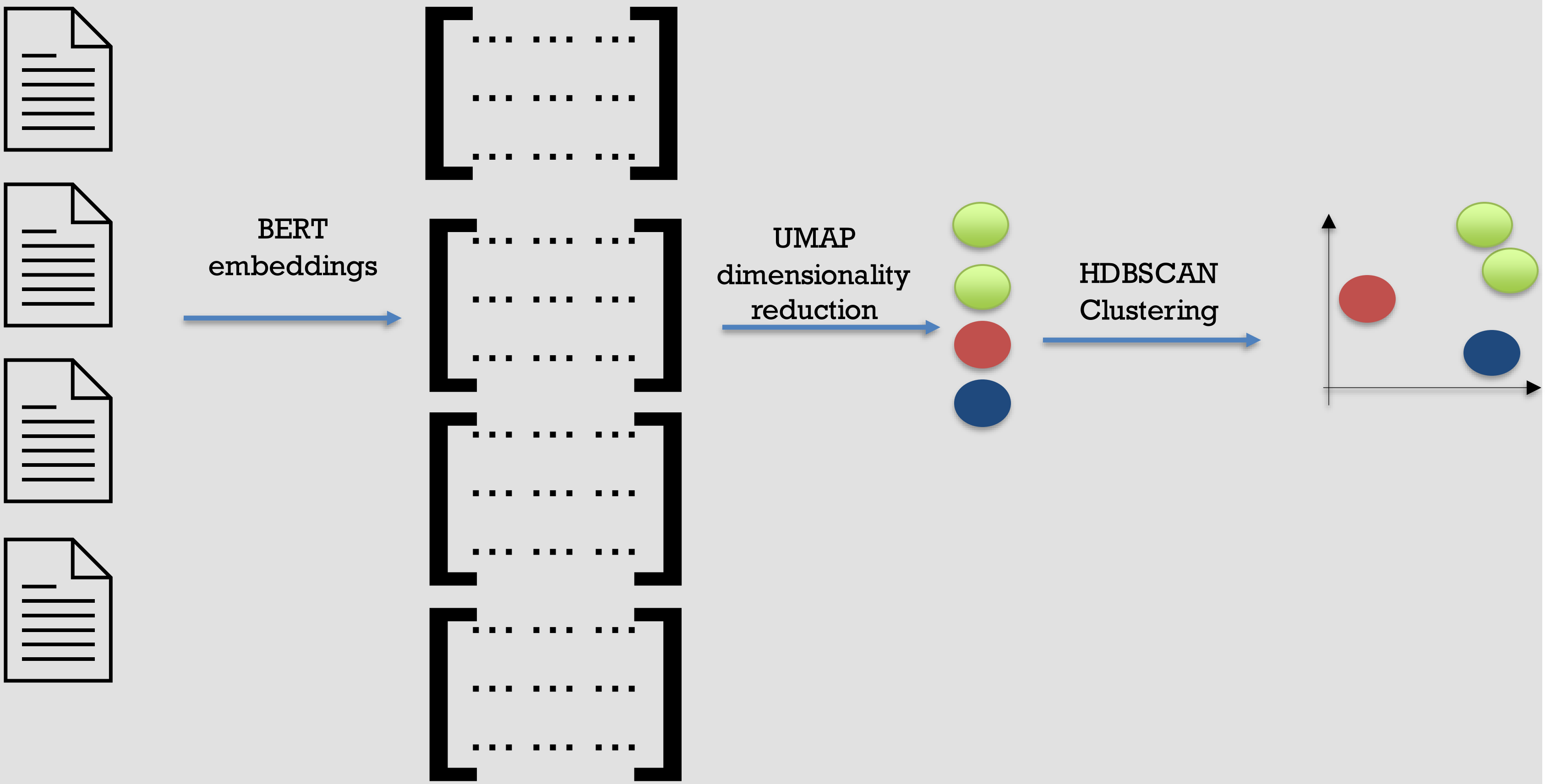
Part I: Grouping documents into topics

Collection of documents

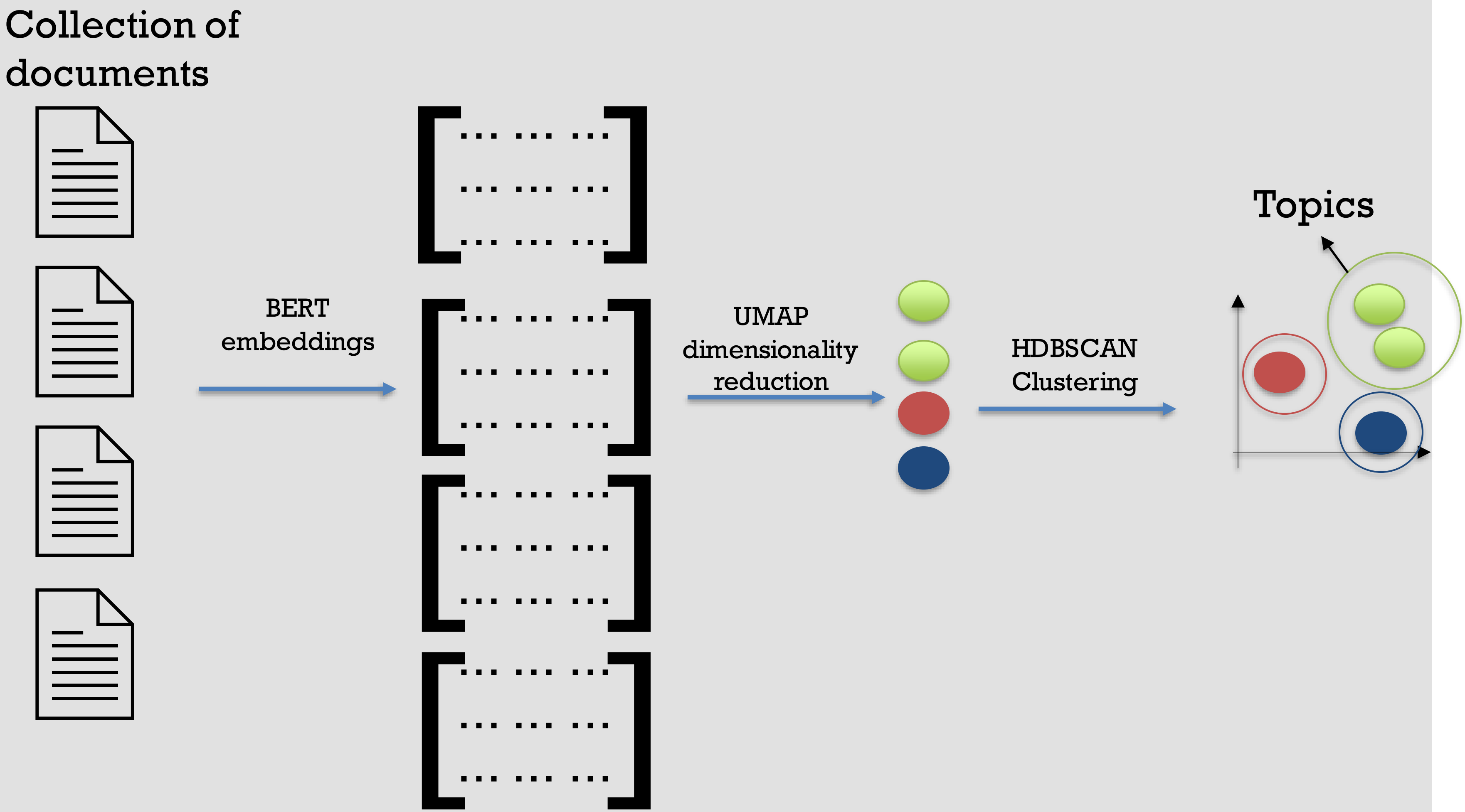


Part I: Grouping documents into topics

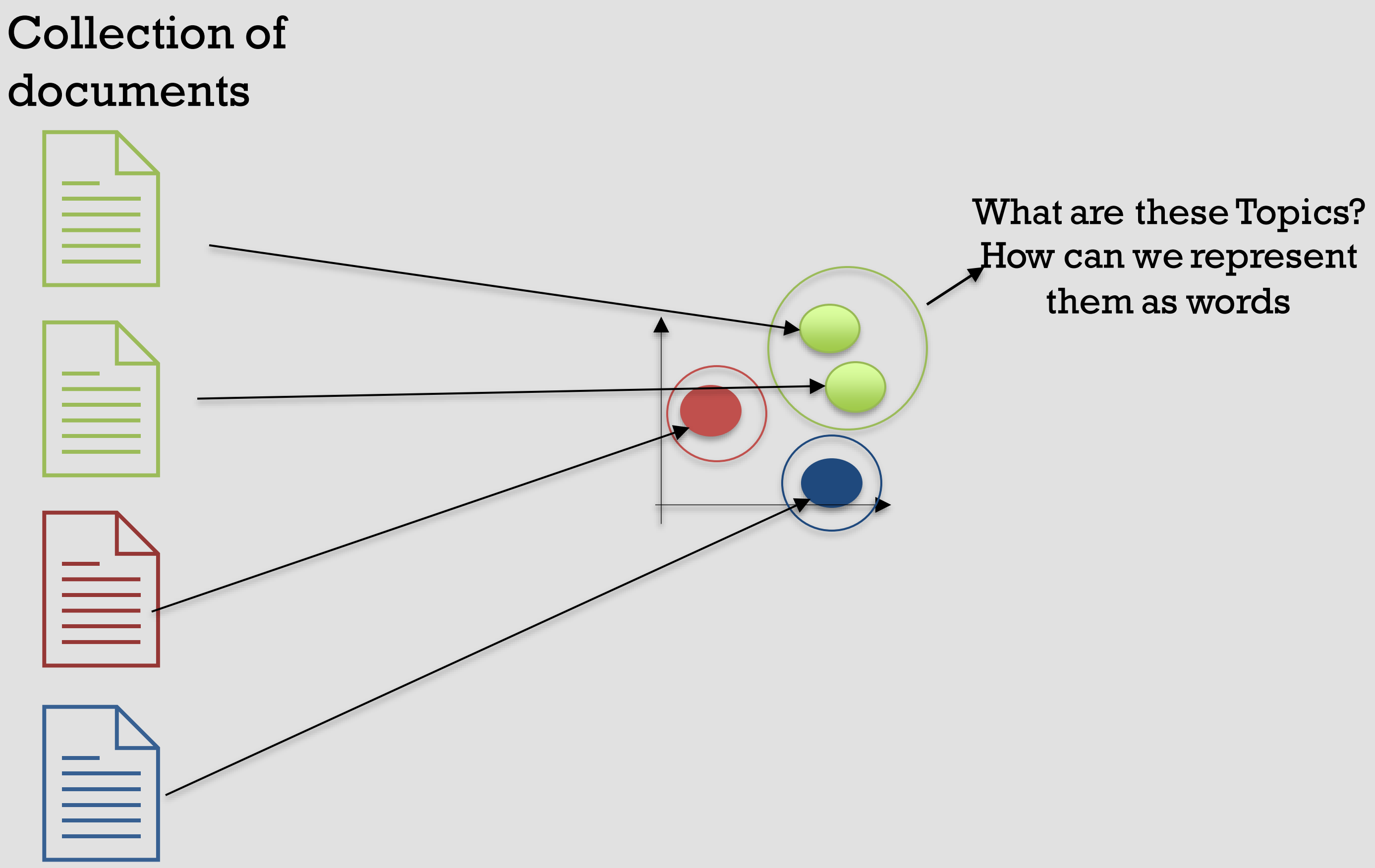
Collection of documents

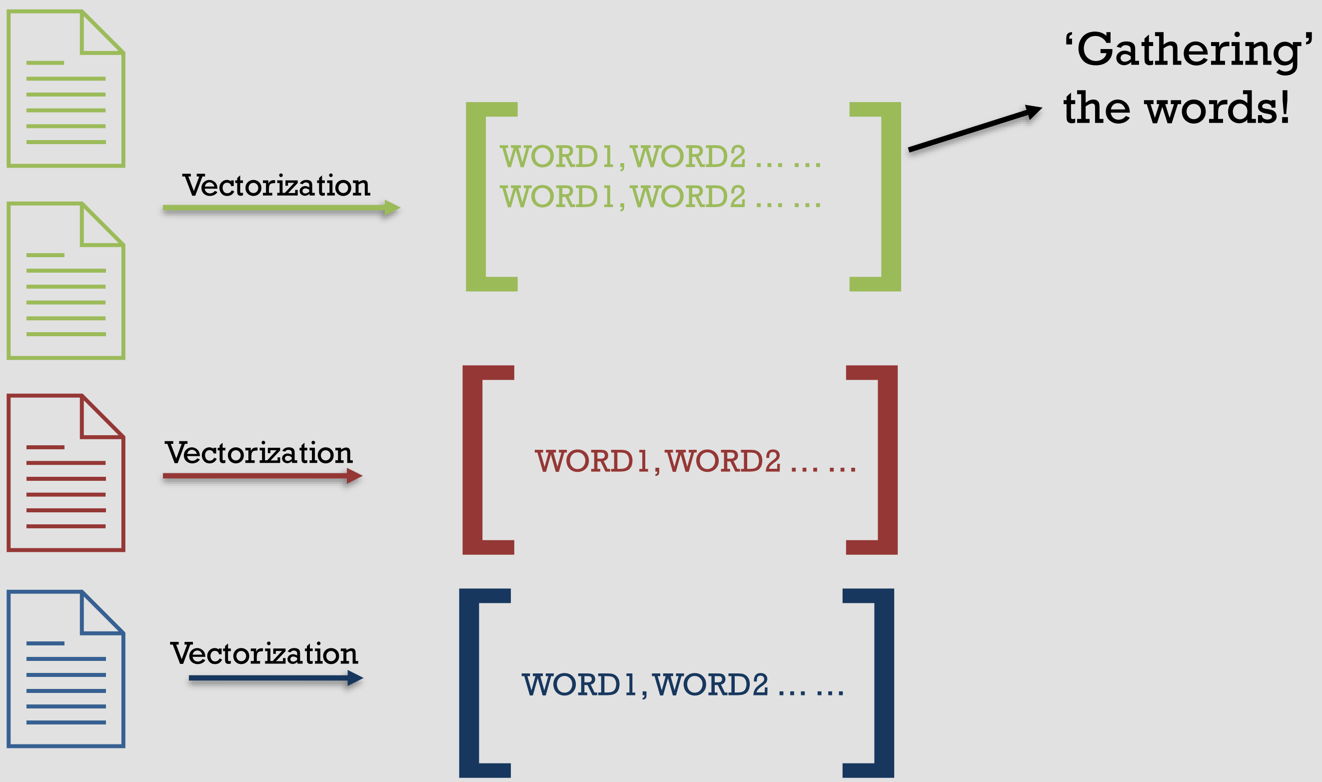


Part I: Grouping documents into topics

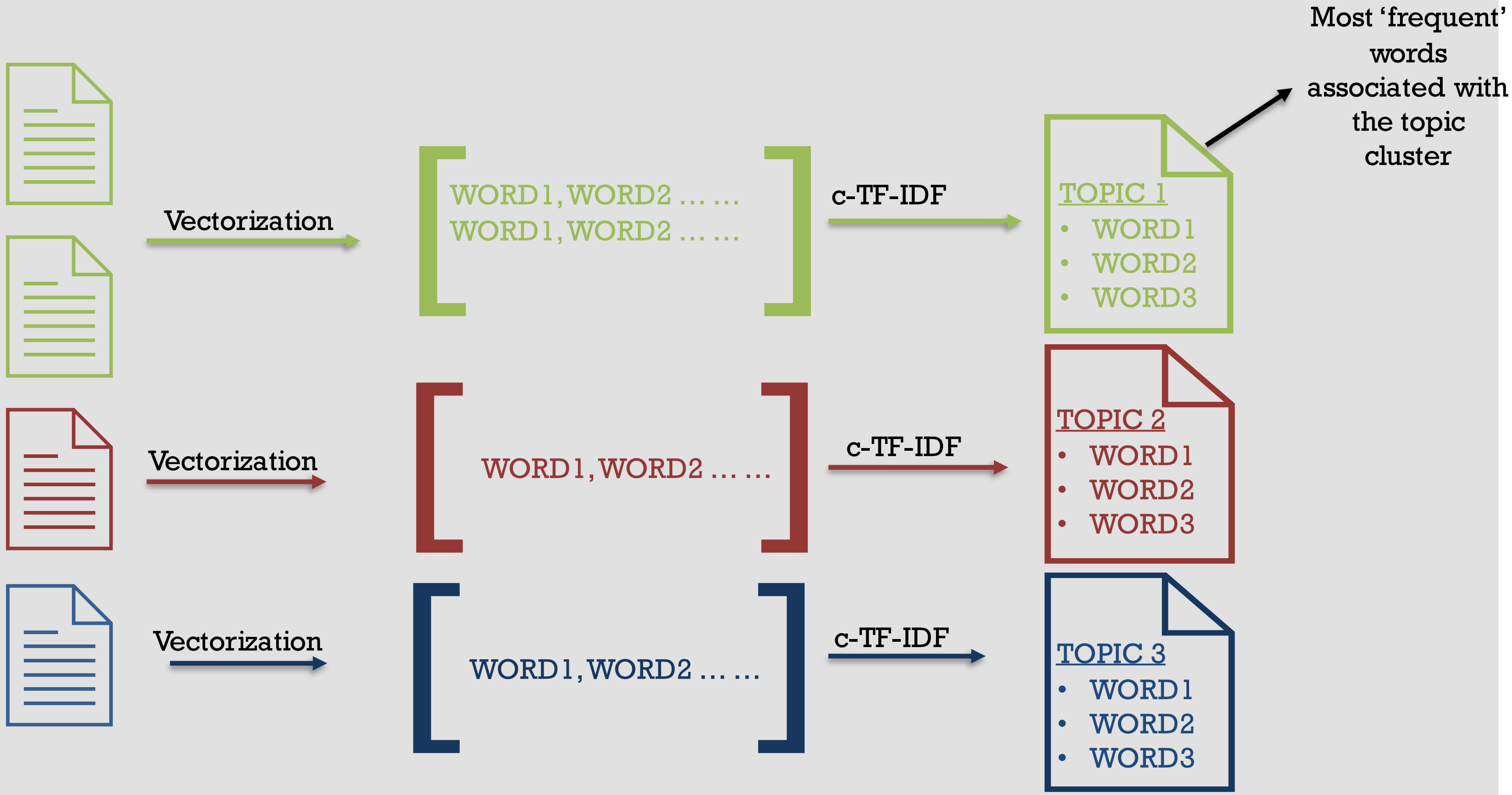


Part II: Obtaining word representations for each topic





‘Gathering’ the words!
Vectorization
WORD1, WORD2 … …
WORD1, WORD2 … …
Vectorization
WORD1, WORD2 … …
Vectorization
WORD1, WORD2 … …



Most 'frequent' words associated with the topic cluster
Vectorization
WORD1, WORD2 … …
WORD1, WORD2 … …
c-TF-IDF
TOPIC 1
WORD1
WORD2
WORD3
Vectorization
WORD1, WORD2 … …
c-TF-IDF
TOPIC 2
WORD1
WORD2
WORD3
Vectorization
WORD1, WORD2 … …
c-TF-IDF
TOPIC 3
WORD1
WORD2
WORD3

What is c-tf-idf

Average number of words in this topic / cluster

$$\boxed{W_{t,c}} = \boxed{tf_{t,c}} \cdot \log\left(1 + \frac{\boxed{A}}{\boxed{tf_t}}\right)$$

How important is this word in this topic / cluster?

How many times does this word appear in the topic / cluster?

How many times does this word appear in ALL DOCUMENTS

c-tf-idf: Idea

Words that appear frequently in this topic / cluster
BUT are **RARE** in other clusters

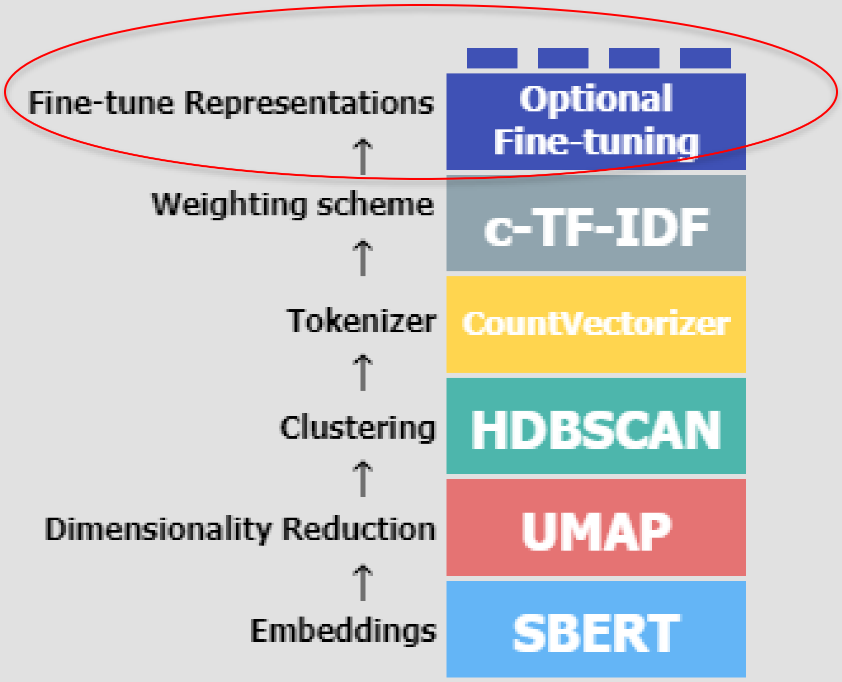
IMPORTANT

Words that appear frequently in this topic / cluster
AND are **also frequent** in other clusters

not so important

DEMO

Finetuning step

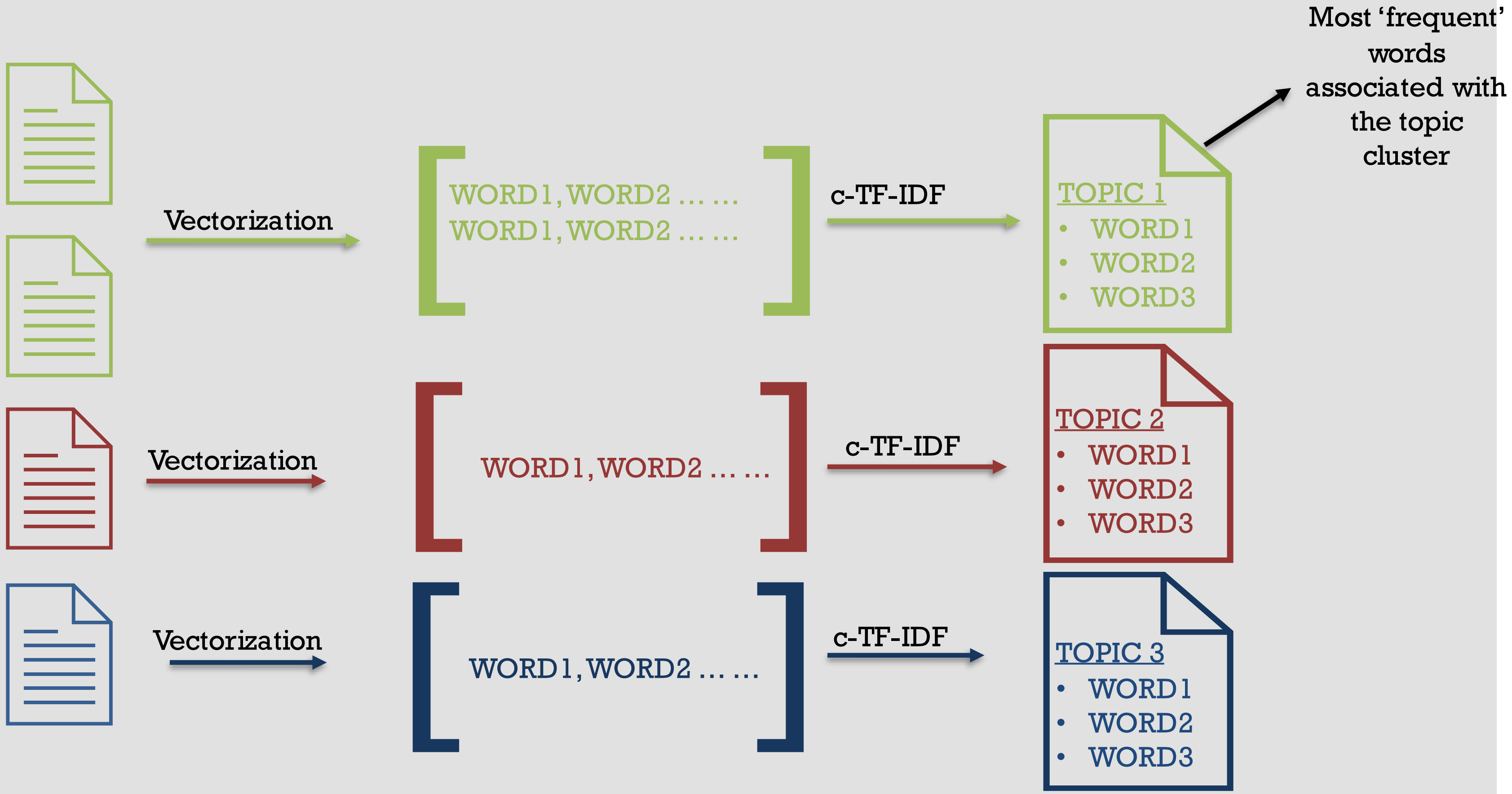


Different methods to fine-tune:

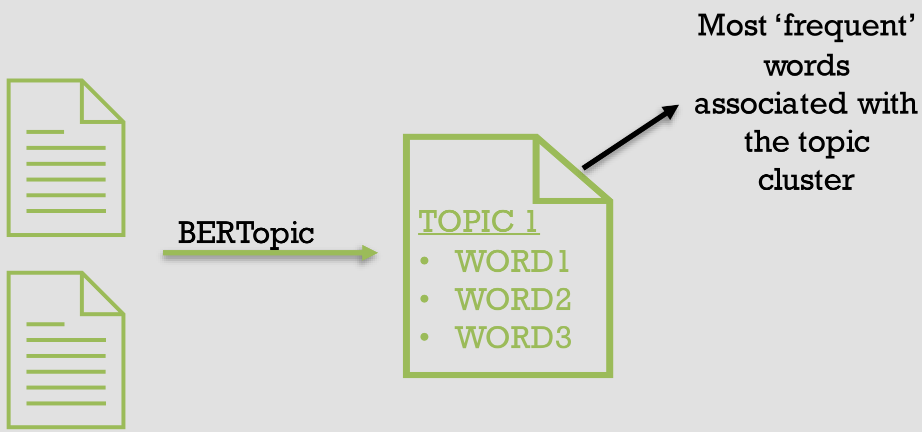
1. Maximal Marginal Relevance
2. KeyBERT
3. Zero-shot classification

All are self-supervised in nature!

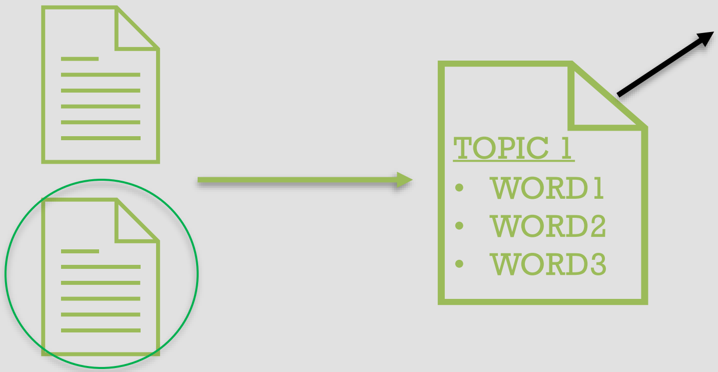
KeyBERT


Most 'frequent' words associated with the topic cluster
Vectorization
WORD1, WORD2 … …
WORD1, WORD2 … …
c-TF-IDF
TOPIC 1
WORD1
WORD2
WORD3
Vectorization
WORD1, WORD2 … …
c-TF-IDF
TOPIC 2
WORD1
WORD2
WORD3
Vectorization
WORD1, WORD2 … …
c-TF-IDF
TOPIC 3
WORD1
WORD2
WORD3

KeyBERT

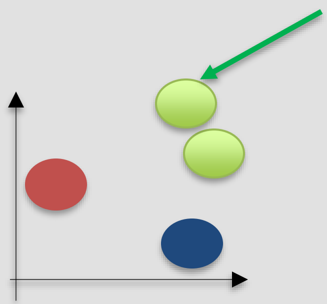

Most 'frequent' words associated with the topic cluster
BERTopic
TOPIC 1
WORD1
WORD2
WORD3

KeyBERT



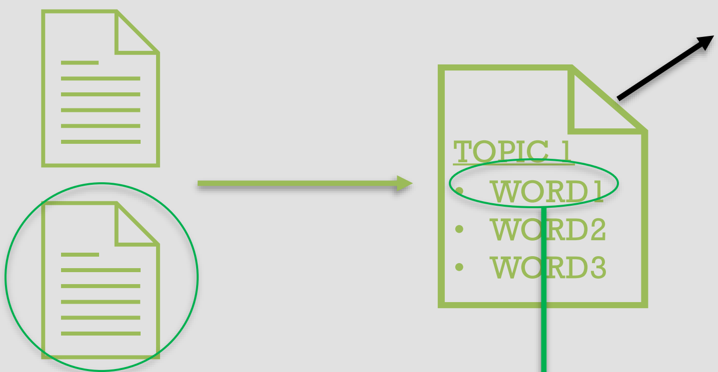
Most 'frequent' words associated with the topic cluster

Extract the document that best represents the identified topic clusters



KeyBERT

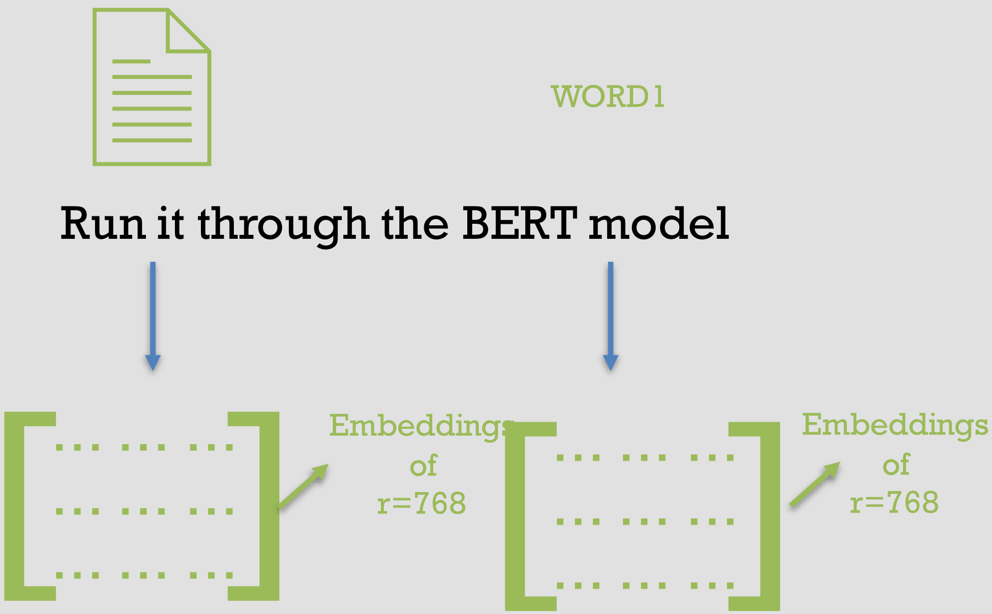
Most ‘frequent’ words associated with the topic cluster



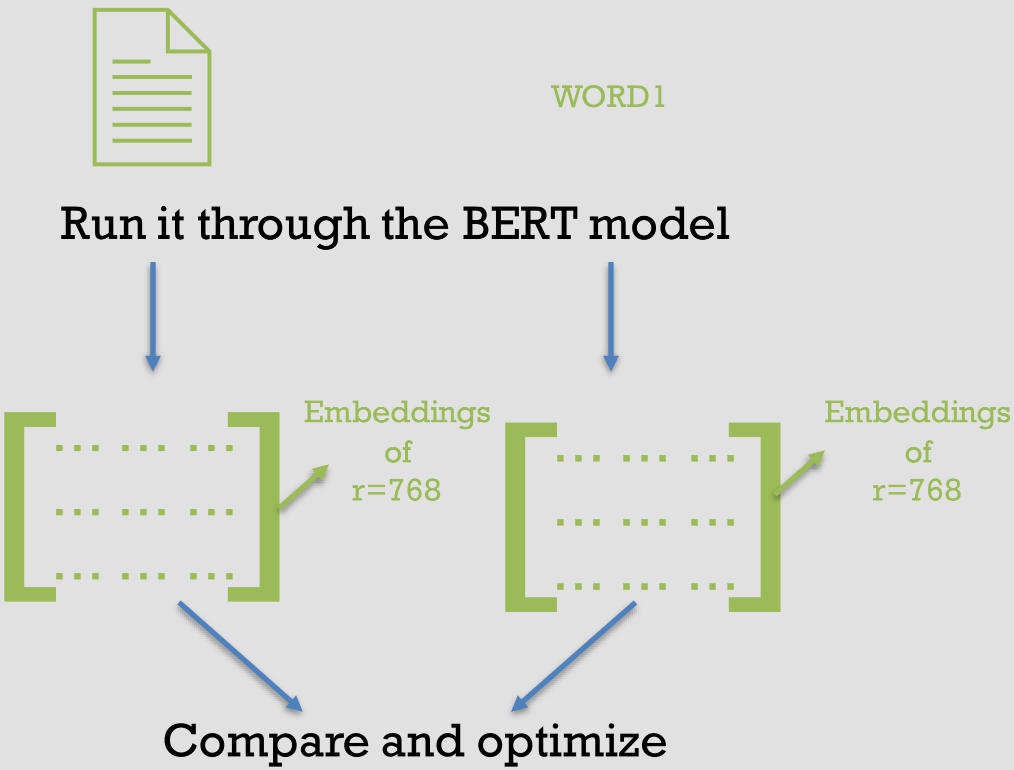
Extract the document(s) that best represents the identified topic clusters

Extract the most frequent word(s)

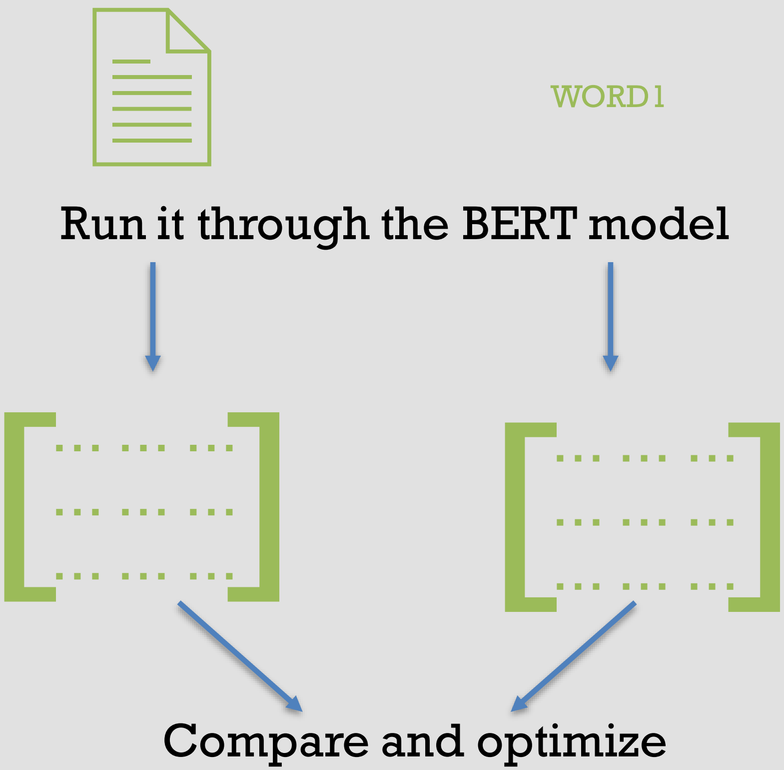
KeyBERT


WORD1
Run it through the BERT model
Embeddings of r=768
Embeddings of r=768

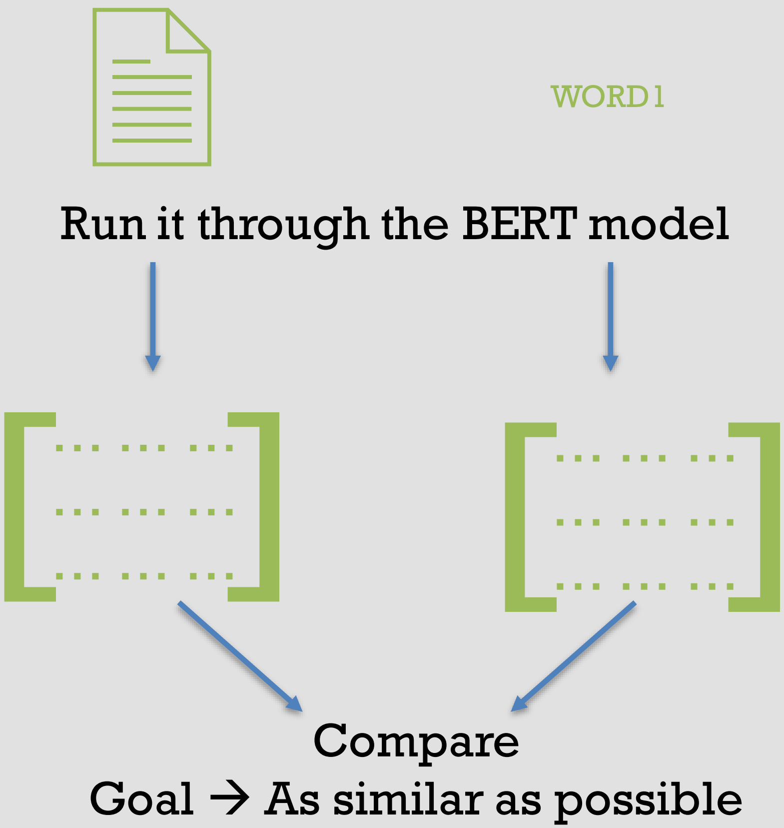
KeyBERT


WORD1
Run it through the BERT model
Embeddings of r=768
Embeddings of r=768
Compare and optimize

KeyBERT

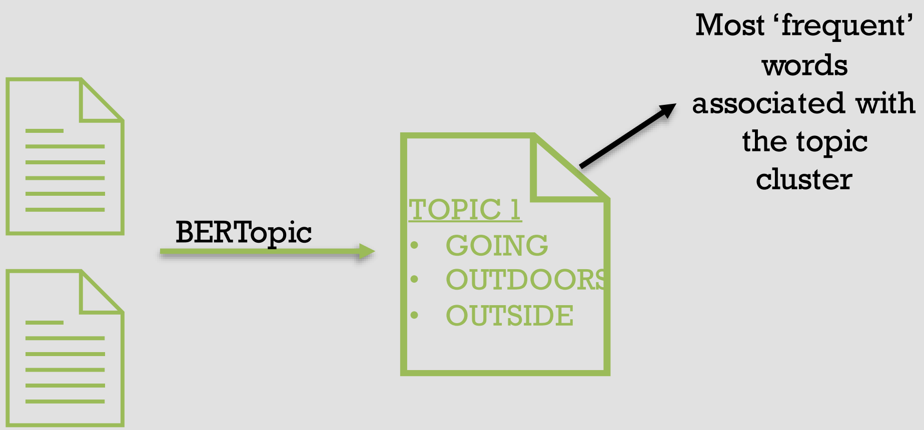


KeyBERT


WORD1
Run it through the BERT model
Compare
Goal → As similar as possible

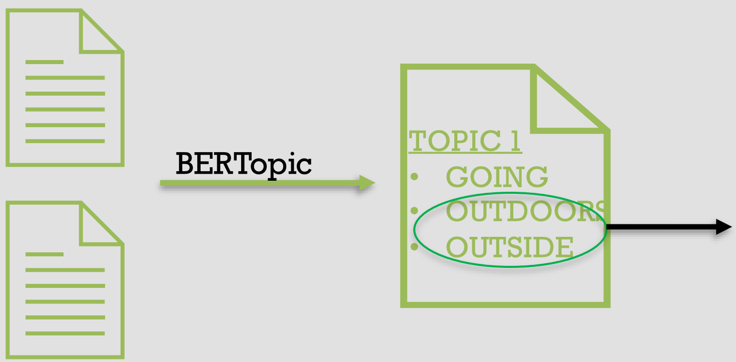
Maximal Marginal Relevance

$$MMR = arg \max_{D_i \in R \setminus S} [\lambda Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j)]$$



Maximal Marginal Relevance

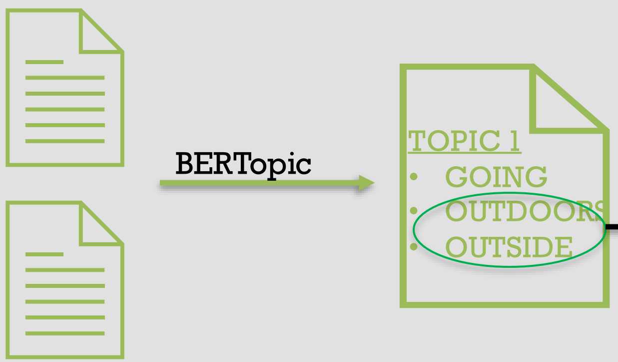
$$MMR = arg \max_{D_i \in R \setminus S} [\lambda Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)]$$



Almost identical w.r.t the topic

Maximal Marginal Relevance

$$MMR = arg \max_{D_i \in R \setminus S} [\lambda Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)]$$

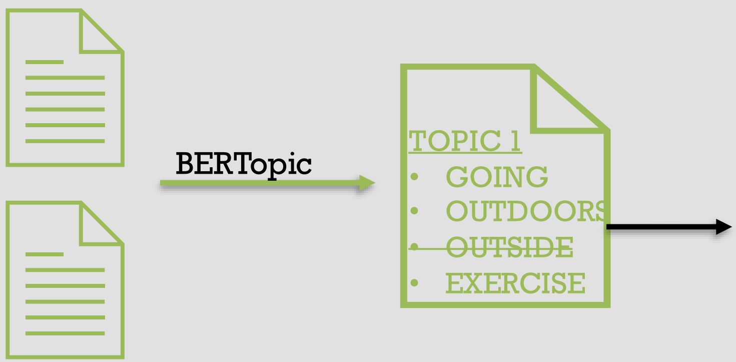


Almost identical w.r.t the topic

Identical words 'clog-up' the outputs

Maximal Marginal Relevance

$$MMR = \arg \max_{D_i \in R \setminus S} \left[\lambda Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right]$$

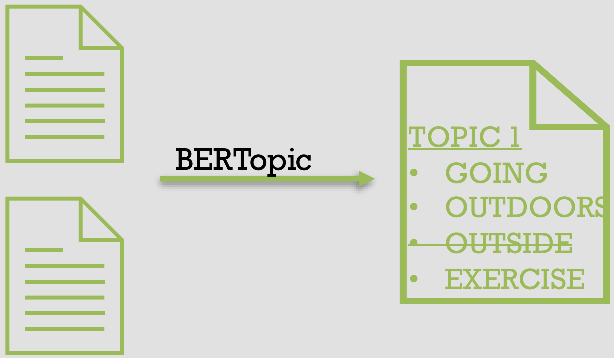


Almost identical w.r.t the topic

Identical words 'clog-up' the outputs

Zero-shot classification

Sometimes, by looking at labels, you have an 'idea' of what the potential topics could be
→ You want to generate more representative keywords

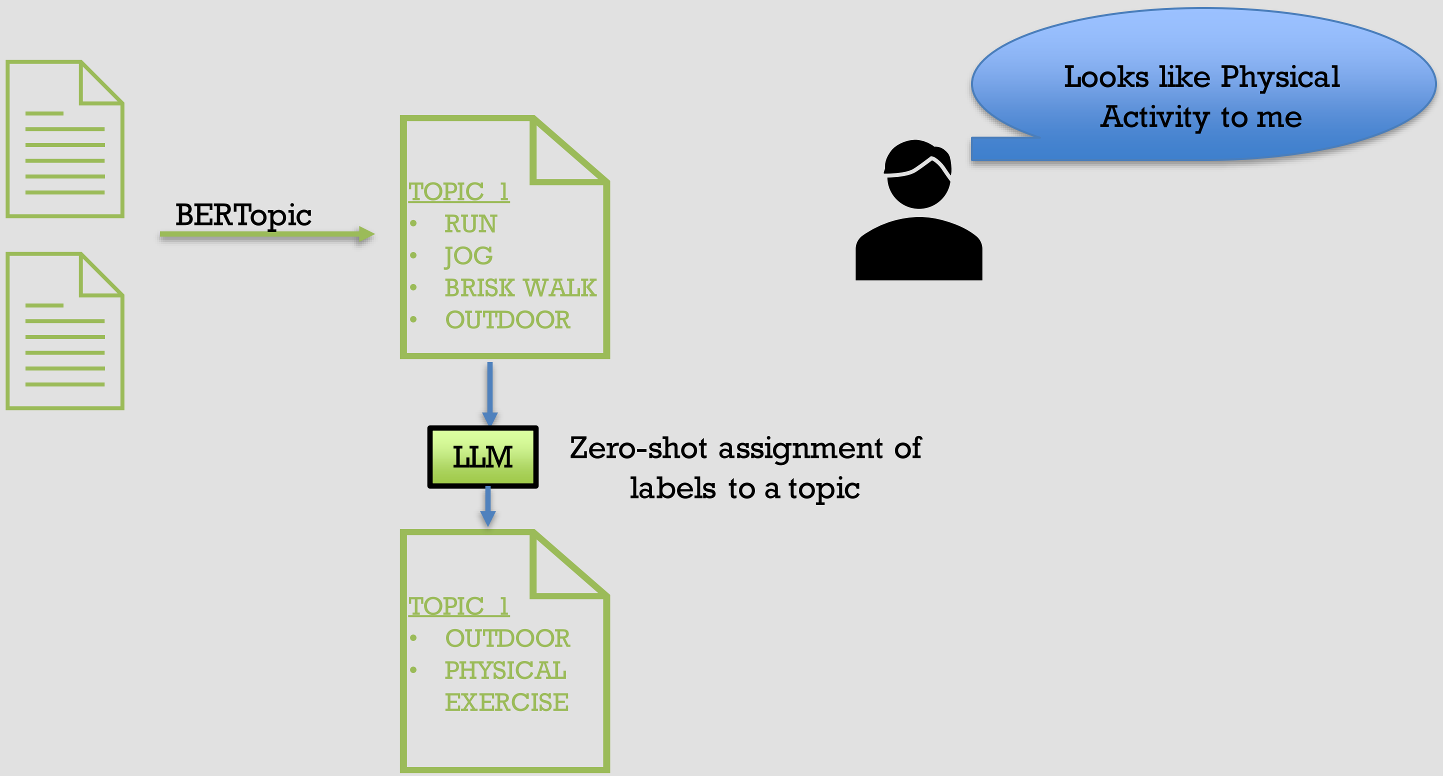


Zero-shot classification

Sometimes, by looking at labels, you have an ‘idea’ of what the potential topics could be
→ You want to generate more representative keywords

Zero-shot classification

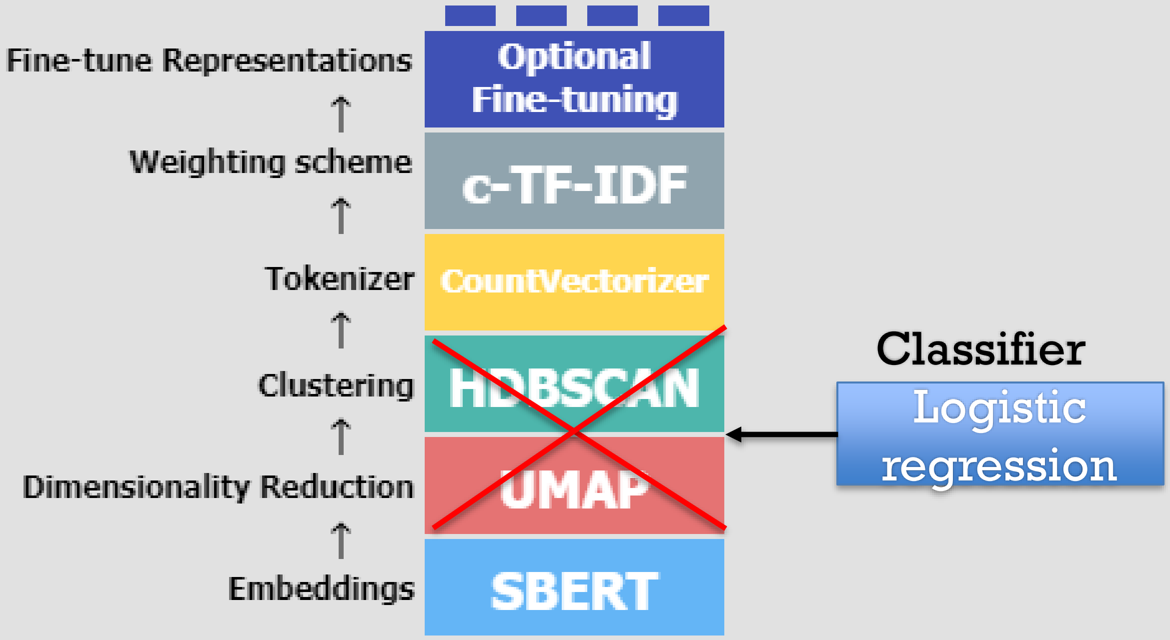
Sometimes, by looking at labels, you have an 'idea' of what the potential topics could be
→ You want to generate more representative keywords



Demo

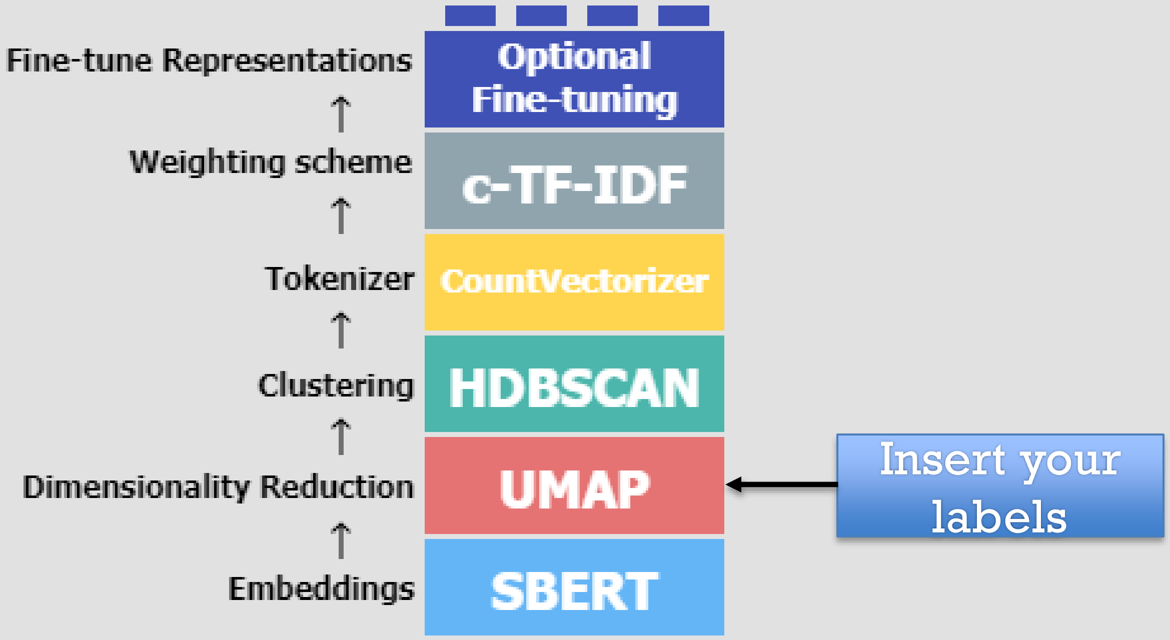
Supervised

What if my data is already labelled with topics?
I just want to CREATE A PREDICTIVE MODEL THAT CAN CLASSIFY TOPICS from the keywords



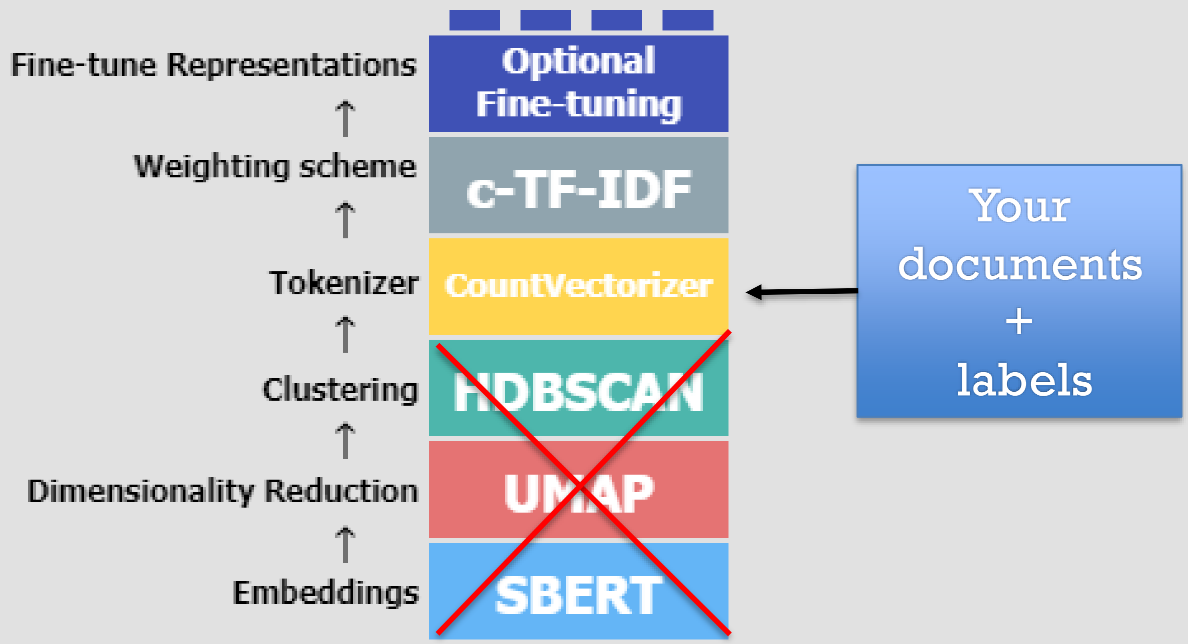
Semi-supervised

What if my data is partially labelled?
Some topics have already been identified?
But I don't know the others?



Manual

What if my data is already labelled with topics?
I want a better understanding of each topic
I just want to understand which keywords are associated with the topics.



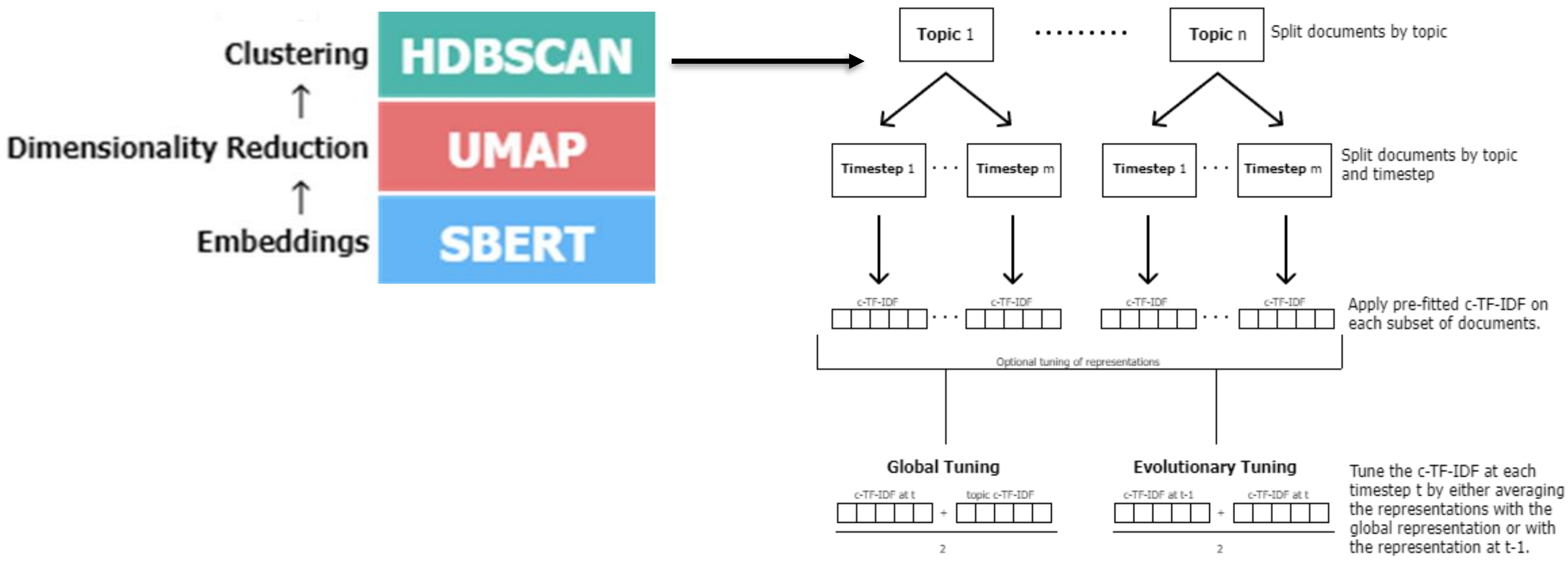
Dynamic topic models

Sometimes documents could span across different timespans

- Eg different dates, years,

You want to observe the relationship across the timespan

Dynamic topic models



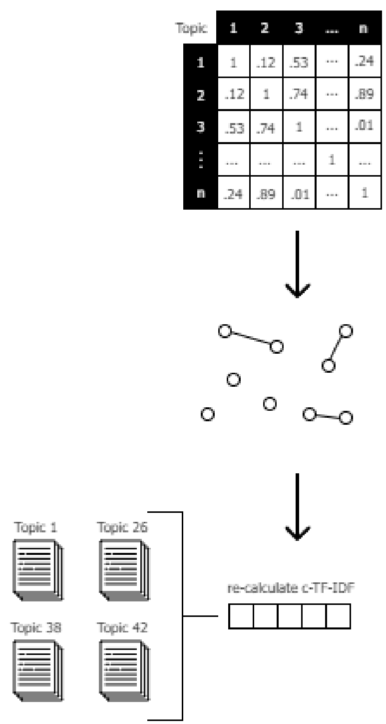
Hierarchical Topic Modeling

Sometimes topics produced may be a subset of a 'bigger topic' in a hierarchical manner

- You want to group these topics into a bigger topic
- Understand the hierarchical relationship

Maybe useful if you produce hundreds of topic, making it difficult to analyze each of them individually

Hierarchical Topic Modeling



Create a distance matrix by calculating the cosine similarity between c-TF-IDF representations of each topic.

Apply a linkage function of choice on the distance matrix to model the hierarchical structure of topics.

Update the c-TF-IDF representation based on the collection of documents across the merged topics.

Limitations to BERTopic

- ✅ Curse of dimensionality
- ✅ Lack of good document representation
- ❌ Unsuitable of PLMs for clustering