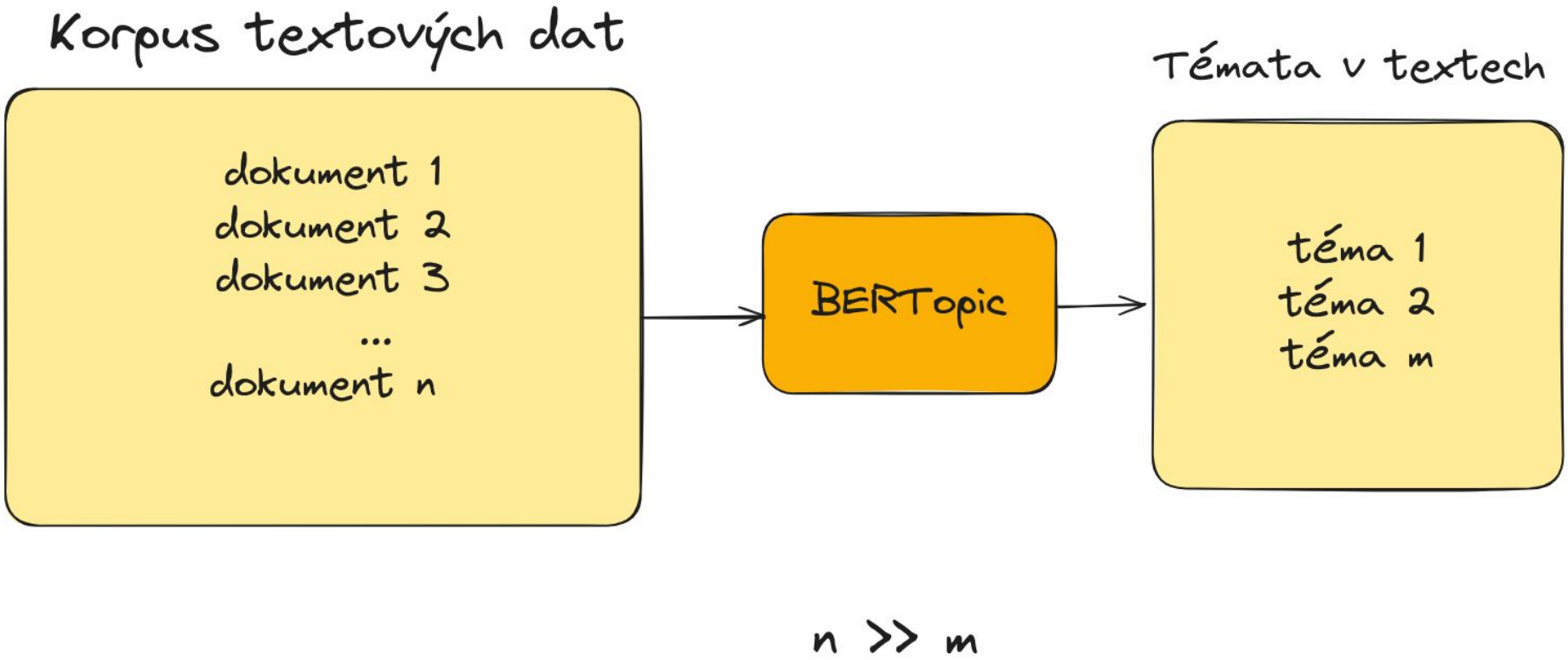
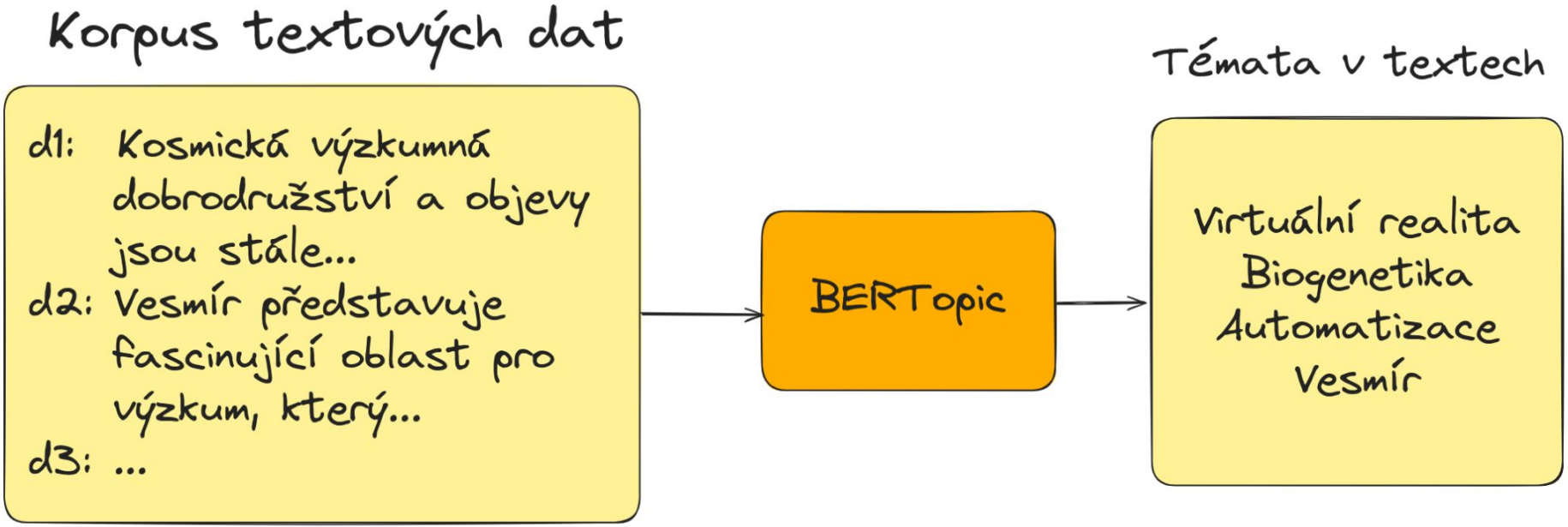
BERTopic

Vojtěch Eichler, Ondřej Vlček, Antonín Jarolím


Korpus textových dat
dokument 1
dokument 2
dokument 3
...
dokument n
BERTopic
Témata v textech
téma 1
téma 2
téma m
n >> m


Korpus textových dat
d1: Kosmická výzkumná dobrodružství a objevy jsou stále...
d2: Vesmír představuje fascinující oblast pro výzkum, který...
d3: ...
BERTopic
Témata v textech
Virtuální realita
Biogenetika
Automatizace
Vesmír

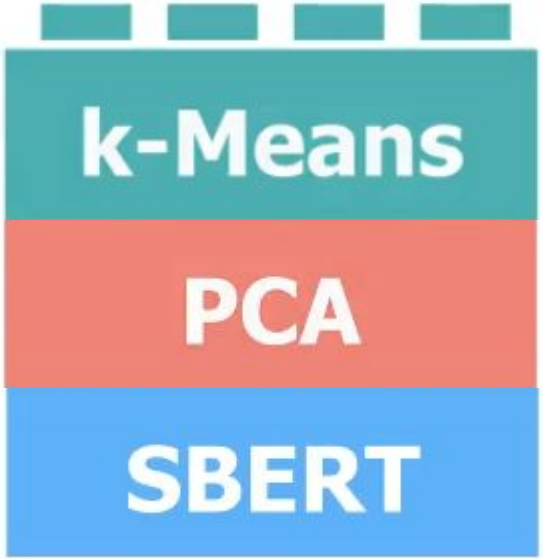
Embedding dokumentů

- Sentence-Transformers
- Předpoklad je, že texty se stejným tématem jsou sémanticky podobné
- Je možné použít prakticky jakoukoliv techniku pro získávání embeddingů, model však musí být natrénovaný na sémantickou podobnost

Redukce dimenzionality

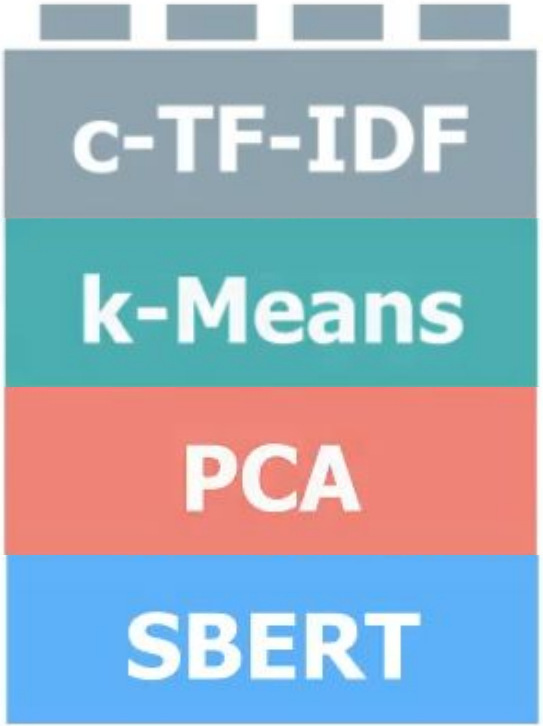
- *Curse of dimensionality*
- PCA, UMAP, …
- UMAP:
  - Velmi rychlý
  - Lepší globální struktura
  - Stochastický

Clustering



- Sémantická vzdálenost
- k-Means, HDBSCAN, …
- HDBSCAN:
  - Soft-clustering
  - Více témat na dokument
  - Odolnost vůči šumu

Reprezentace témat - Class-based TF-IDF



- Rozšířená verze TF-IDF
- Jak zjistit která slova reprezentují témata?
  - Chtěli bychom, aby se v daném clusteru nacházela s vysokou frekvencí
  - Zároveň by bylo vhodné, aby se v ostatních clusterech moc nevyskytovala

$$W_{t,c} = f_{t,c} \cdot log\left(1 + \frac{A}{f_t}\right)$$

Build Your Topic Model