SEZNAM.CZ

Náš život s jazykovými modely: ‚Sumarizace dokumentu‘

Martin Bachura

Vektorová reprezentace dokumentu

Martin Bachura

Text / Document summarization

Text extraction vs Text generation

Vectors

Research vs Engineering

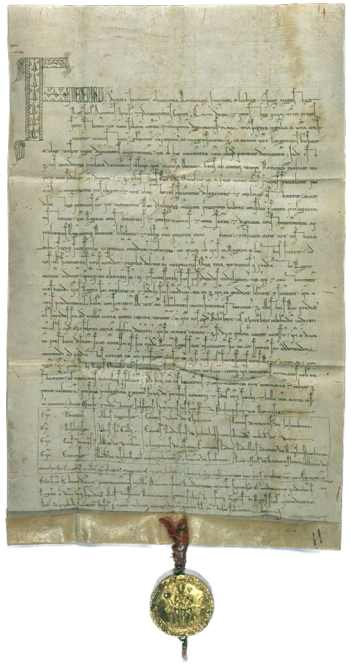
Model designs vs industrial applications

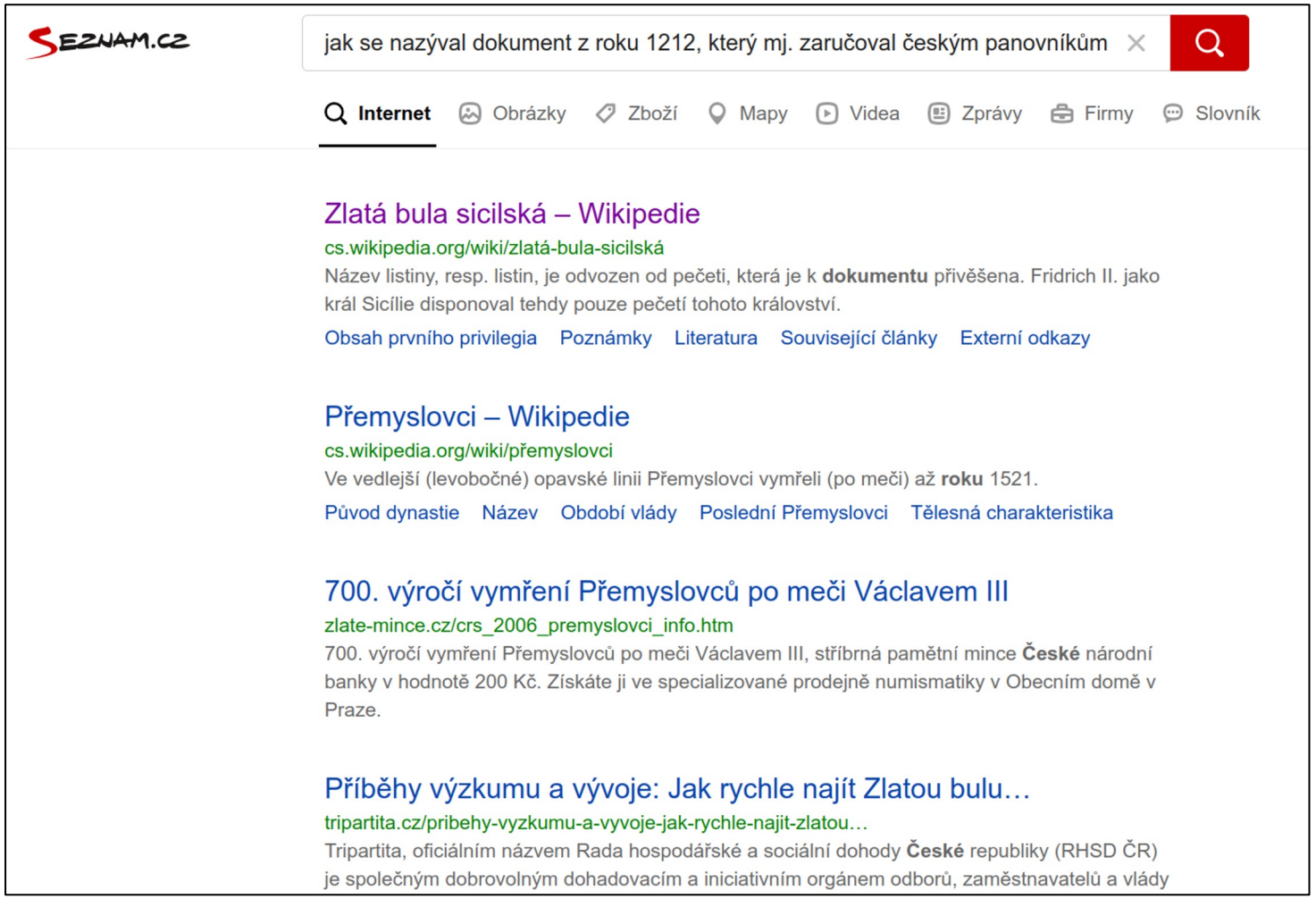
Evaluation

GPT vs Smaller LM models

Our problem

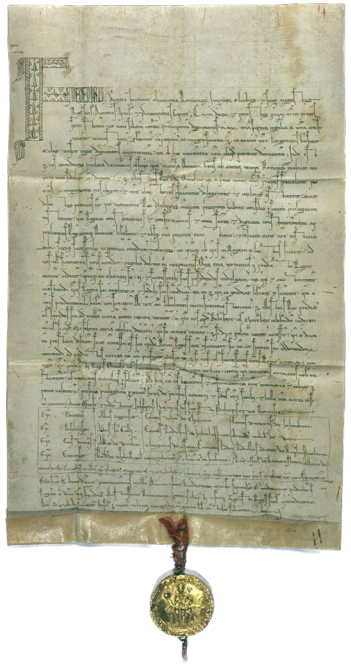
Retrieve appropriate documents / websites from index for given query

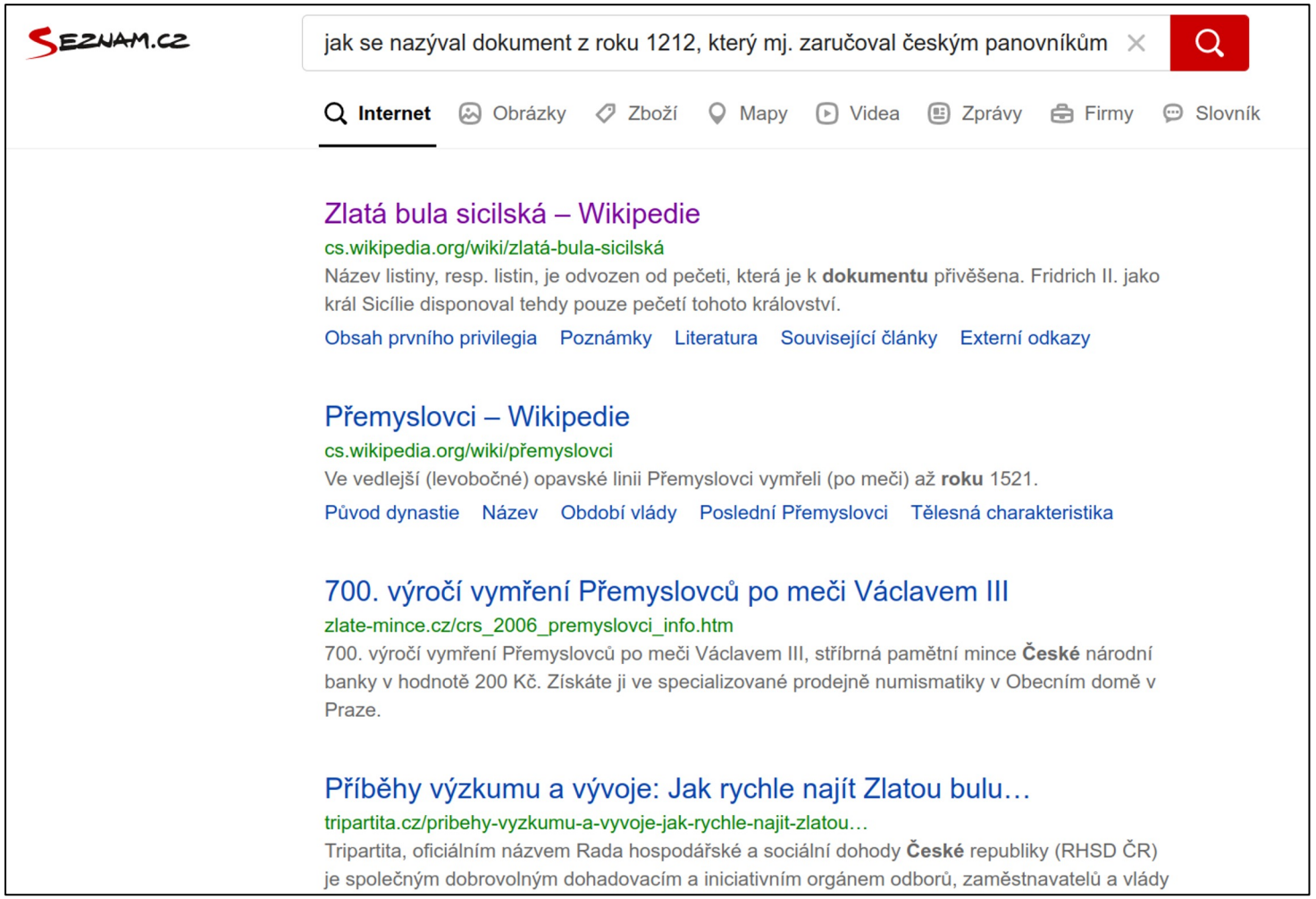




Our problem

Retrieve appropriate documents / websites from **vector** index for given query



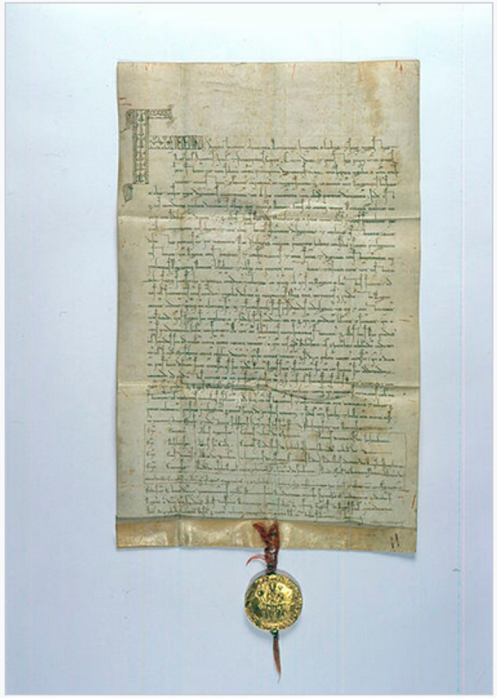


Our problem

Jak se nazýval dokument z roku 1212 který mj. zajišťoval českým panovníkům dědičný královský titul

Zlatá bula sicilská (lat. *Bulla Aurea Siciliæ*) je označení pro jednu hlavní nebo soubor tří souvisejících listin, které vydal 26. září 1212 v Basileji budoucí římský král Fridrich II. českému Přemyslu Otakarovi I. Listina především obsahuje privilegium, které potvrzuje dědičnost titulu českého krále a stanovuje jeho volbu na domácí půdě a to, že římský panovník bude volbu jen stvrzovat.

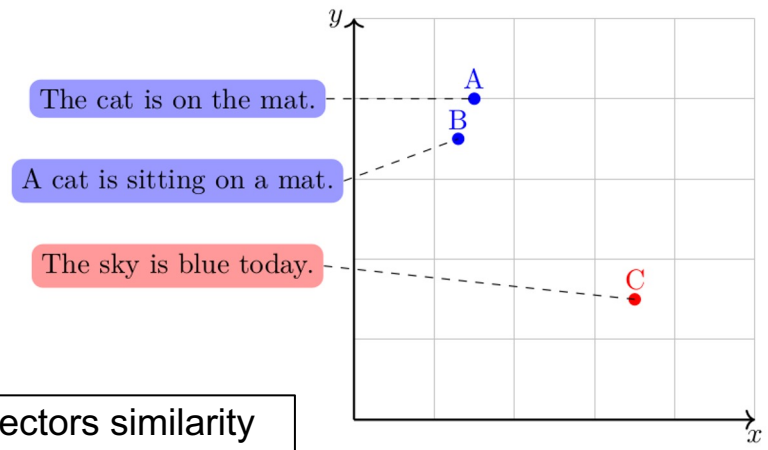
Přemysl Otakar ji získal jako odměnu za podporu, již Fridrichovi poskytl pro volbu římským králem. Novodobý název listiny, resp. listin, je odvozen od zlaté buly (kovové pečeti) sicilského krále, která je k dokumentu přivěšena. Fridrich v té době ještě neměl takto slavnostní pečeť římského krále, použil svou pečeť krále jižní Itálie.[1]



Zlatá bula sicilská

Our problem

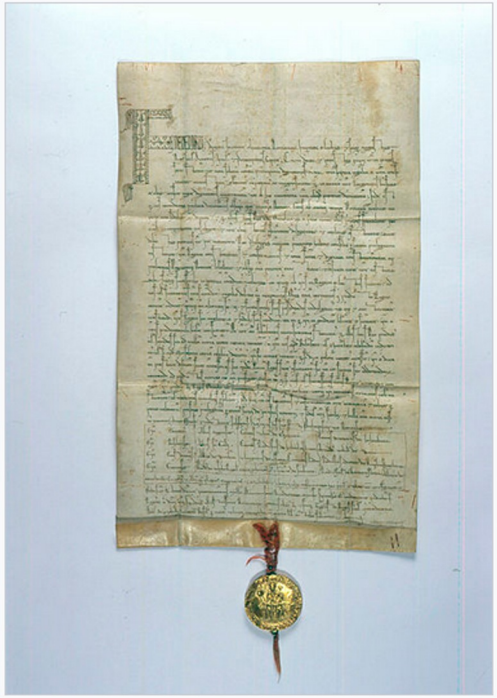
Jak se nazýval dokument z roku 1212 který mj. zajišťoval českým panovníkům dědičný královský titul



Vectors similarity measure

Zlatá bula sicilská (lat. *Bulla Aurea Siciliæ*) je označení pro jednu hlavní nebo soubor tří souvisejících listin, které vydal 26. září 1212 v Basileji budoucí římský král Fridrich II. českému Přemyslu Otakarovi I. Listina především obsahuje privilegium, které potvrzuje dědičnost titulu českého krále a stanovuje jeho volbu na domácí půdě a to, že římský panovník bude volbu jen stvrzovat.

Přemysl Otakar ji získal jako odměnu za podporu, již Fridrichovi poskytl pro volbu římským králem. Novodobý název listiny, resp. listin, je odvozen od zlaté buly (kovové pečeti) sicilského krále, která je k dokumentu přivěšena. Fridrich v té době ještě neměl takto slavnostní pečeť římského krále, použil svou pečeť krále jižní Itálie.[1]

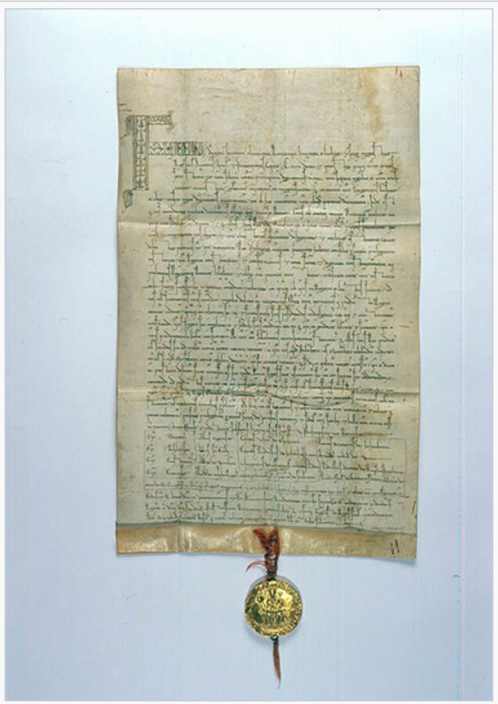

Zlatá bula sicilská

Our problem

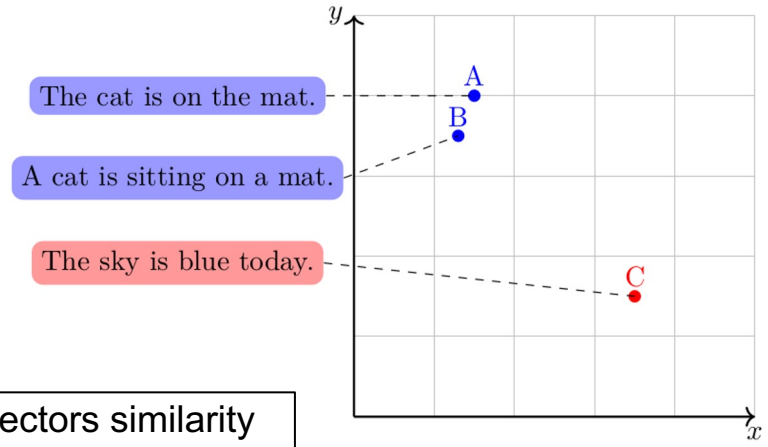
Jak se nazýval dokument z roku 1212 který mj. zajišťoval českým panovníkům dědičný královský titul

Zlatá bula sicilská (lat. *Bulla Aurea Siciliæ*) je označení pro jednu hlavní nebo soubor tří souvisejících listin, které vydal 26. září 1212 v Basileji budoucí římský král Fridrich II. českému Přemyslu Otakarovi I. Listina především obsahuje privilegium, které potvrzuje dědičnost titulu českého krále a stanovuje jeho volbu na domácí půdě a to, že římský panovník bude volbu jen stvrzovat.

Přemysl Otakar ji získal jako odměnu za podporu, již Fridrichovi poskytl pro volbu římským králem. Novodobý název listiny, resp. listin, je odvozen od zlaté buly (kovové pečeti) sicilského krále, která je k dokumentu přivěšena. Fridrich v té době ještě neměl takto slavnostní pečeť římského krále, použil svou pečeť krále jižní Itálie.[1]



Zlatá bula sicilská



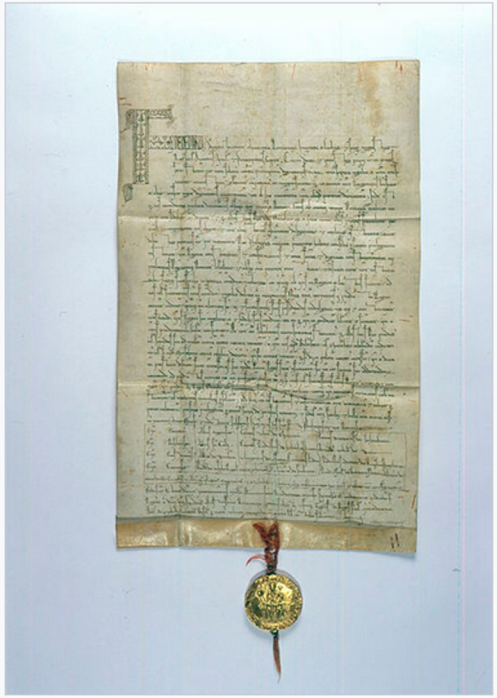
Vectors similarity measure

Our problem

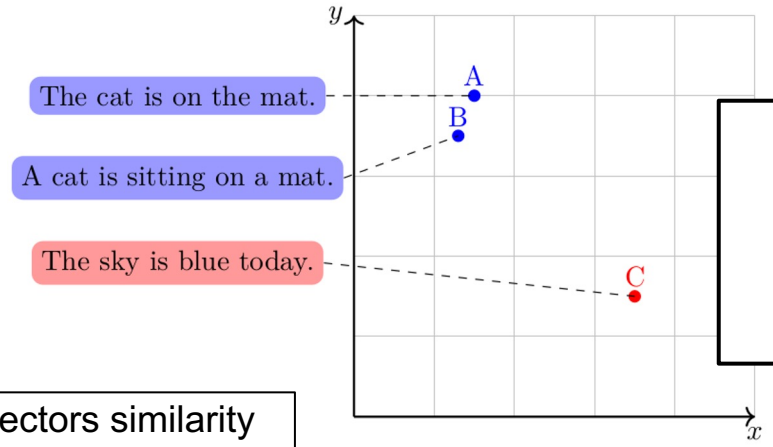
Jak se nazýval dokument z roku 1212 který mj. zajišťoval českým panovníkům dědičný královský titul

Zlatá bula sicilská (lat. *Bulla Aurea Siciliæ*) je označení pro jednu hlavní nebo soubor tří souvisejících listin, které vydal 26. září 1212 v Basileji budoucí římský král Fridrich II. českému Přemyslu Otakarovi I. Listina především obsahuje privilegium, které potvrzuje dědičnost titulu českého krále a stanovuje jeho volbu na domácí půdě a to, že římský panovník bude volbu jen stvrzovat.

Přemysl Otakar ji získal jako odměnu za podporu, již Fridrichovi poskytl pro volbu římským králem. Novodobý název listiny, resp. listin, je odvozen od zlaté buly (kovové pečeti) sicilského krále, která je k dokumentu přivěšena. Fridrich v té době ještě neměl takto slavnostní pečeť římského krále, použil svou



Zlatá bula sicilská

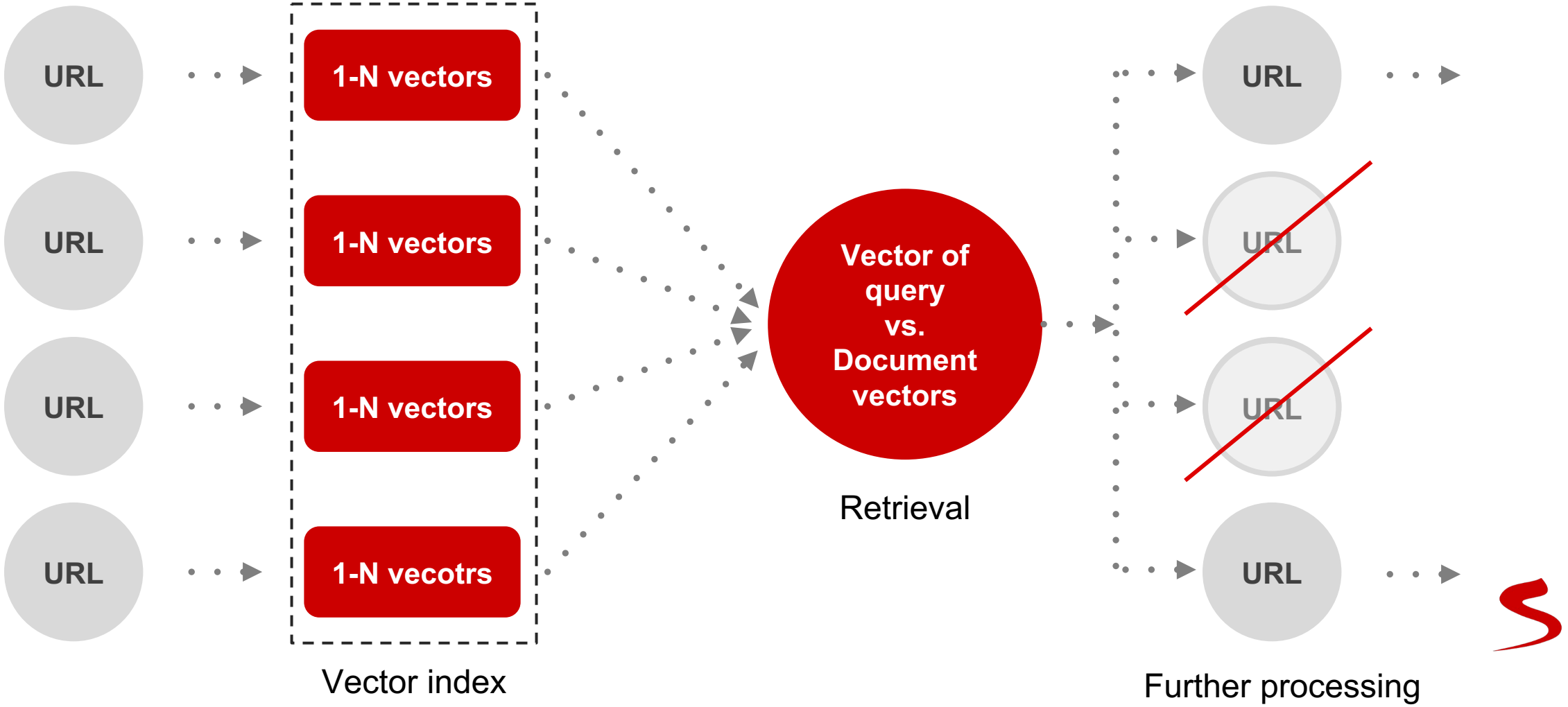


Vectors similarity measure

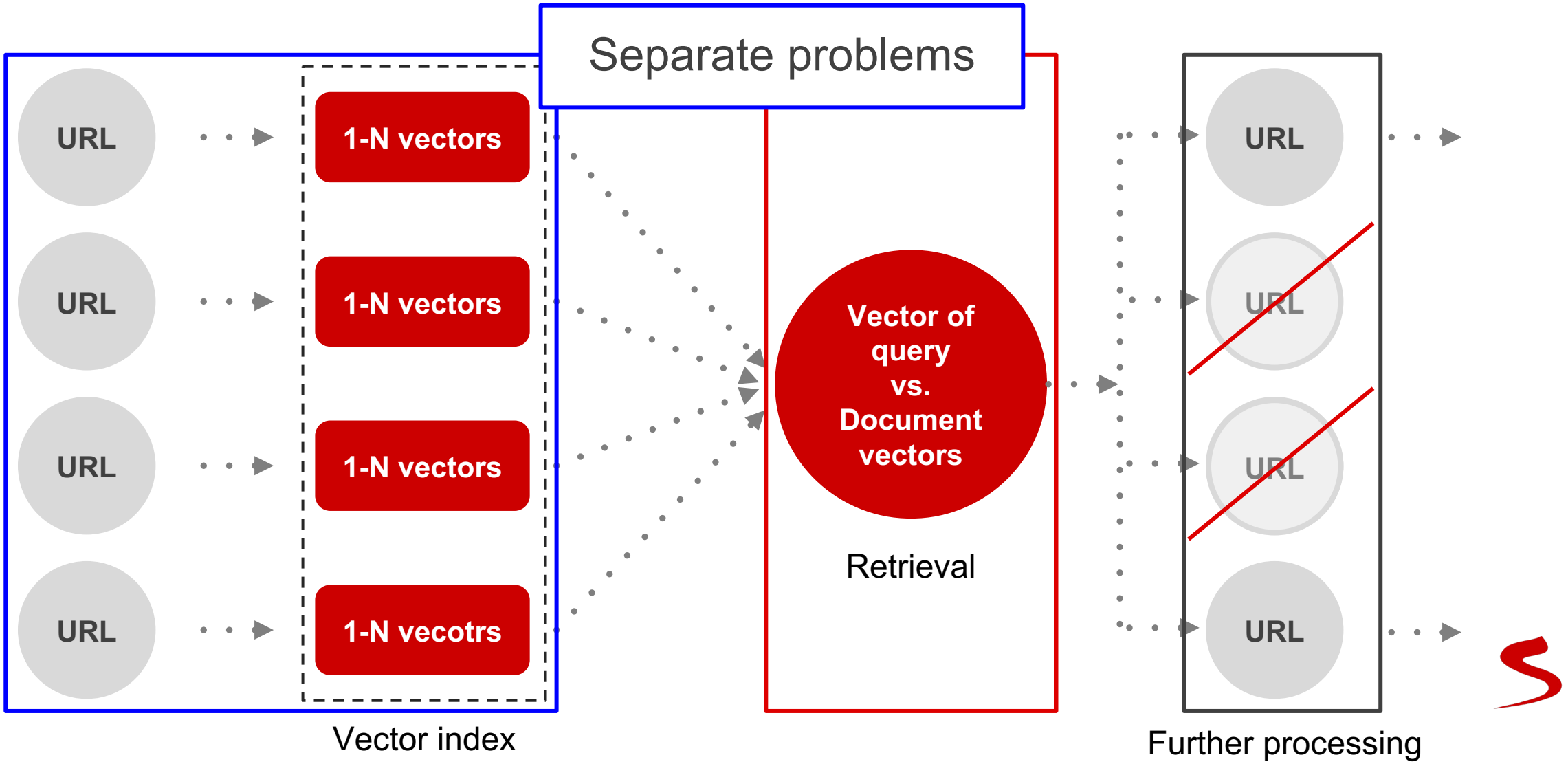
Vectorization of proper text segments: selection of proper segments

Vector branch of Seznam search

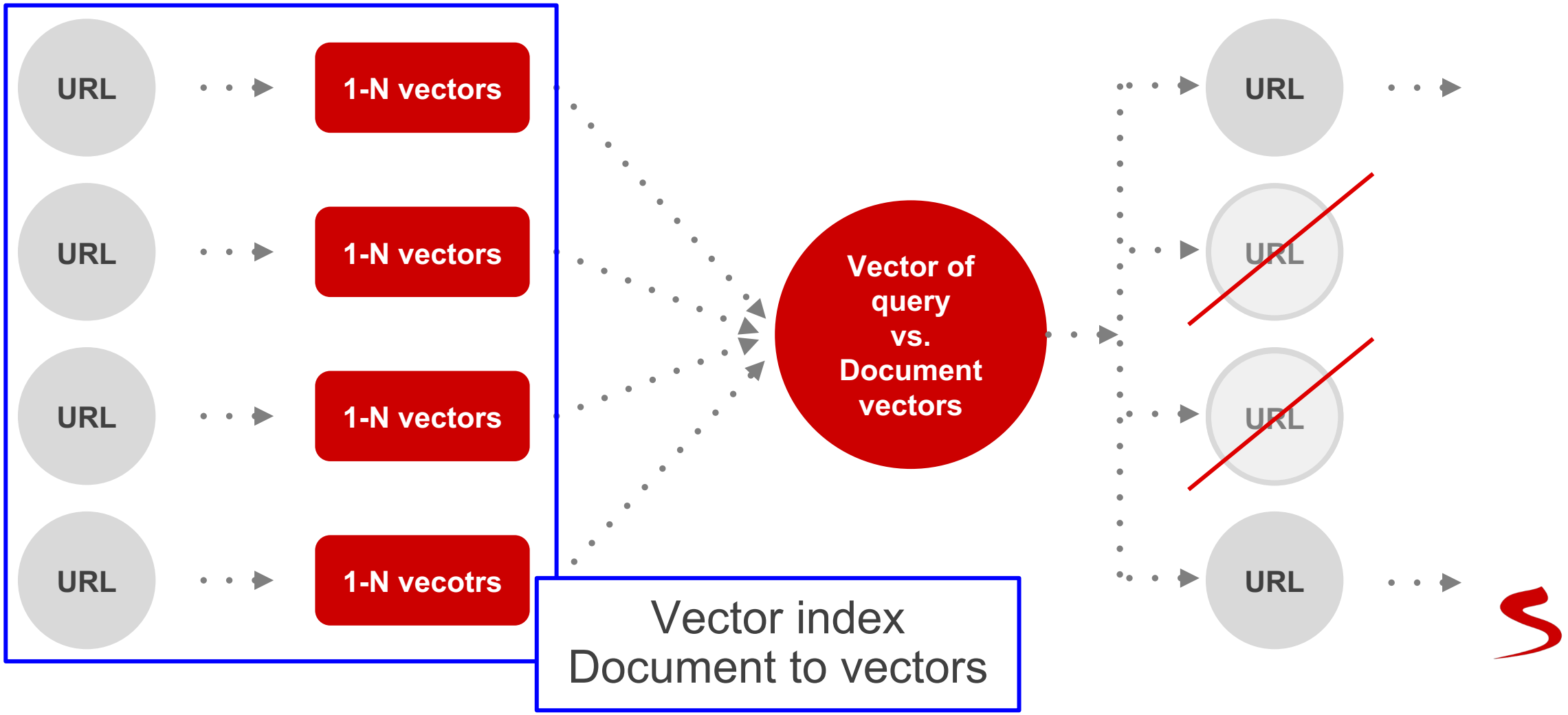
Vector branch of our search engine



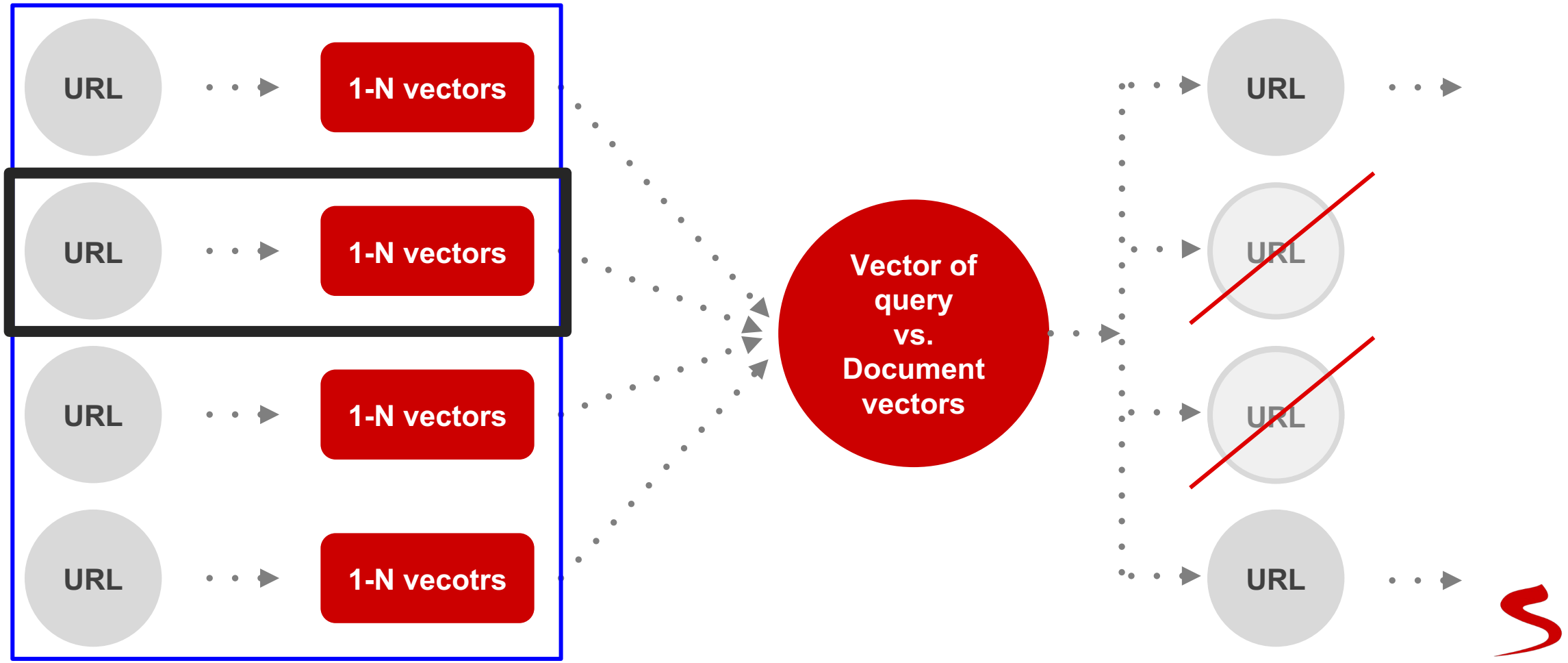
Vector branch of our search engine



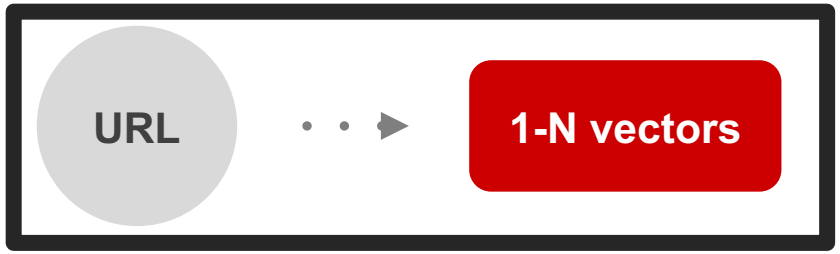
Vector branch of our search engine - our problem



Vector branch of our search engine - our problem

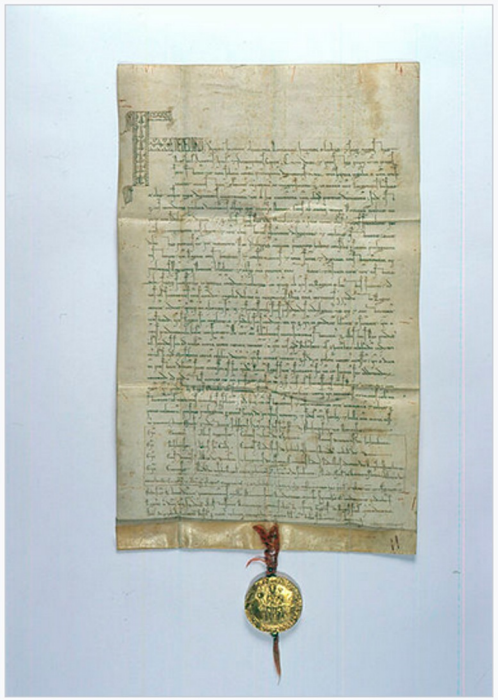


Our problem



Zlatá bula sicilská (lat. *Bulla Aurea Siciliæ*) je označení pro jednu hlavní nebo soubor tří souvisejících listin, které vydal 26. září 1212 v Basileji budoucí římský král Fridrich II. českému Přemyslu Otakarovi I. Listina především obsahuje privilegium, které potvrzuje dědičnost titulu českého krále a stanovuje jeho volbu na domácí půdě a to, že římský panovník bude volbu jen stvrzovat.

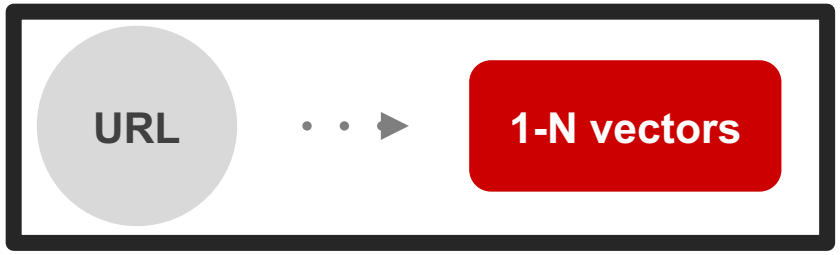
Přemysl Otakar ji získal jako odměnu za podporu, již Fridrichovi poskytl pro volbu římským králem. Novodobý název listiny, resp. listin, je odvozen od zlaté buly (kovové pečeti) sicilského krále, která je k dokumentu přivěšena. Fridrich v té době ještě neměl takto slavnostní pečeť římského krále, použil svou pečeť krále jižní Itálie.[1]


Zlatá bula sicilská

Vectorization of proper text segments: selection of proper segments

Vectors

Vectors - properties



- Storage limitations - how many vectors?
- >> 1 document, whole Czech internet, refreshed
- Limited computational power and $$
- Documents are large
- Retrieval task - one vector one topic, reasonable vector space properties
- Vectors from text, texts - interpretability

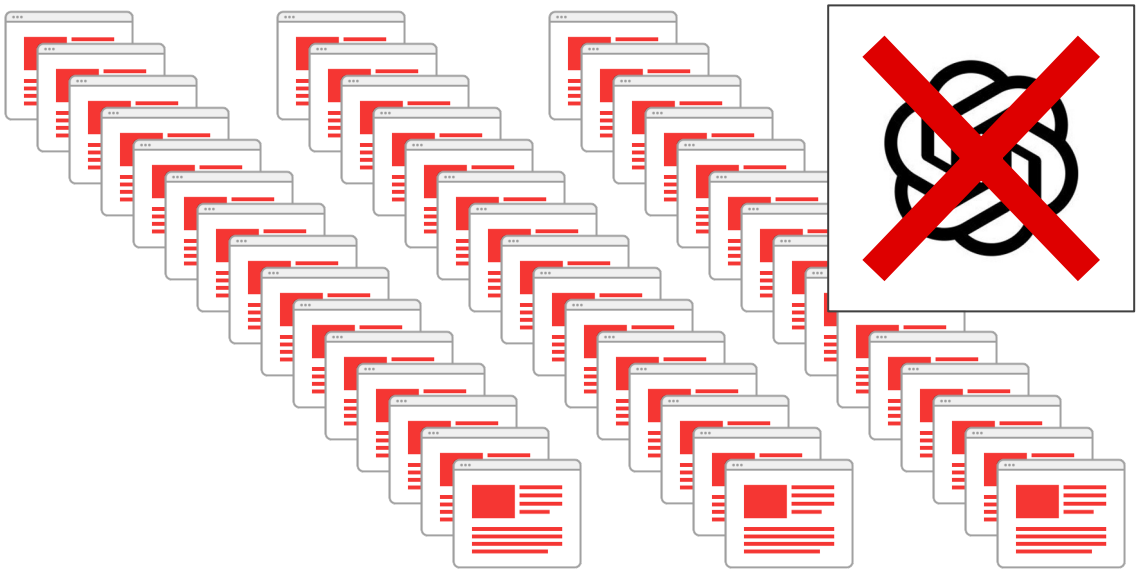
Vectors - properties

- Storage limitations - how many vectors?
- >> 1 document, whole Czech internet, refreshed
- Limited computational power and $$
- Documents are large
- Retrieval task - one vector one topic, reasonable vector space properties
- Vectors from text, texts - interpretability

Vectors - properties

- Unlimited storage - up to N vectros per document
- >> 1 document, whole Czech internet, refreshed
- Limited computational power and $$
- Documents are large
- Retrieval task - one vector one topic, reasonable vector space properties
- Vectors from text, texts - interpretability

Vectors - properties

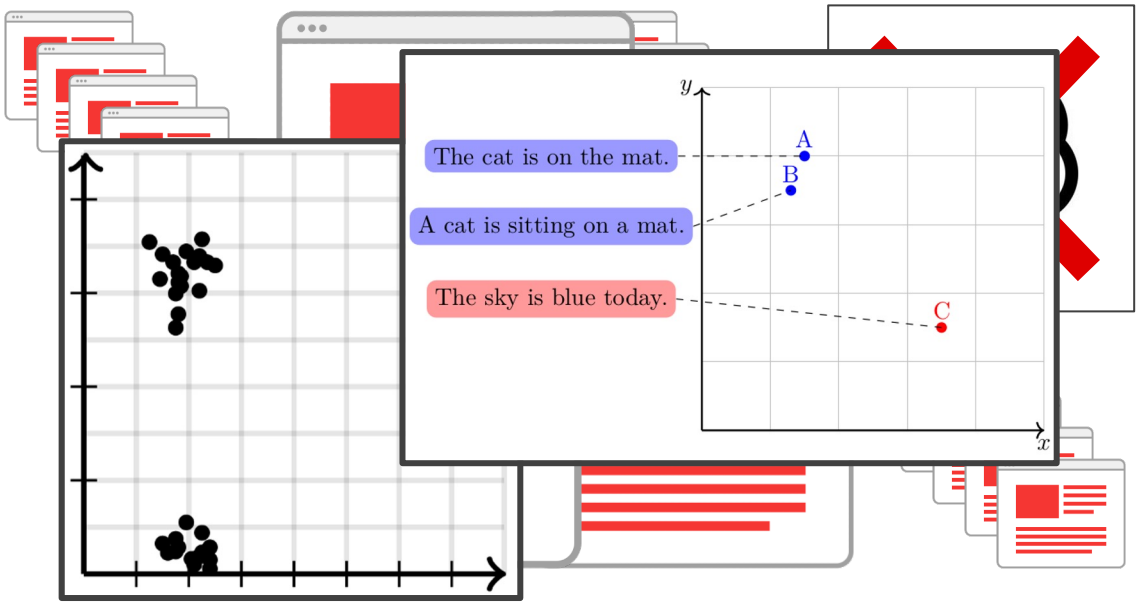


- Unlimited storage - up to N vectros per document
- >> 1 document, whole Czech internet, refreshed
- Limited computational power and $$
- Documents are large
- Retrieval task - one vector one topic, reasonable vector space properties
- Vectors from text, texts - interpretability

Vectors - properties

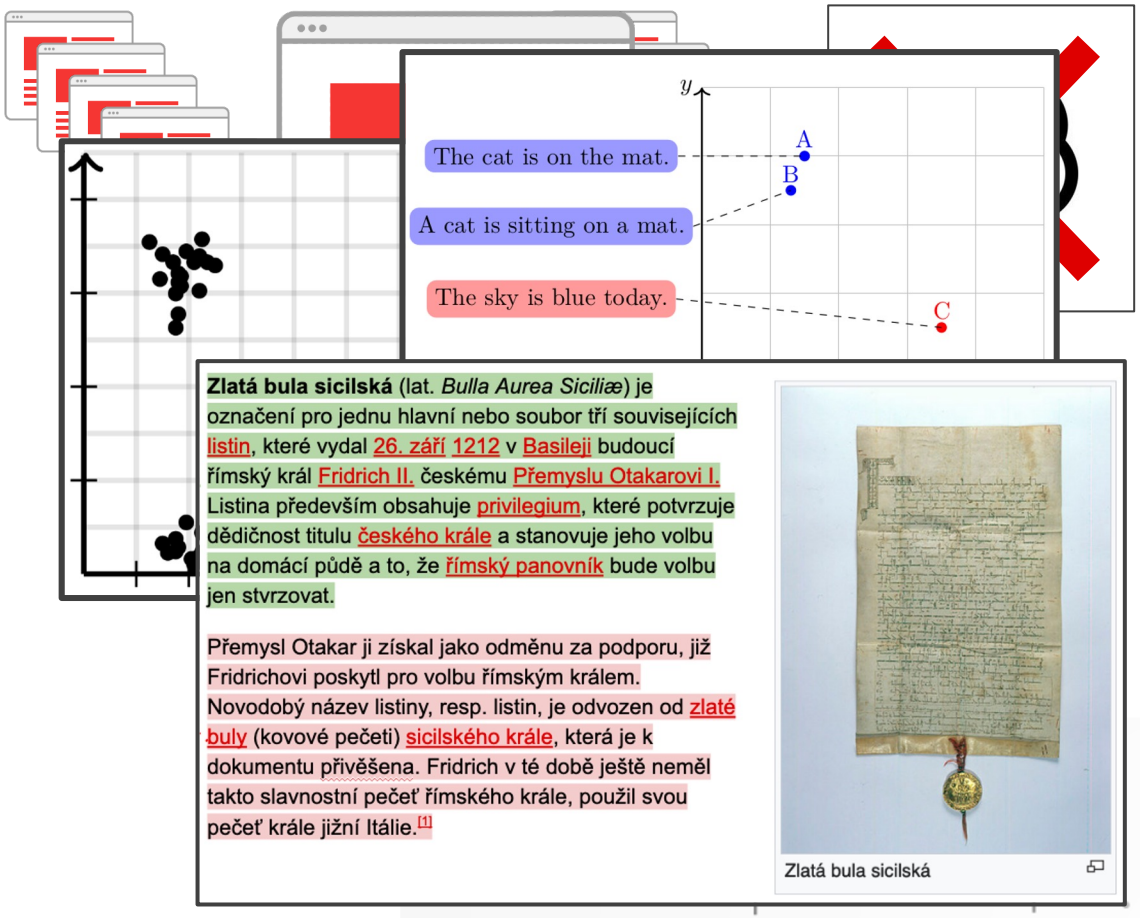
- Unlimited storage - up to N vectros per document
- >> 1 document, whole Czech internet, refreshed
- Limited computational power and $$
- Documents are large
- Retrieval task - one vector one topic, reasonable vector space properties
- Vectors from text, texts - interpretability

Vectors - properties



- Unlimited storage - up to N vectros per document
- >> 1 document, whole Czech internet, refreshed
- Limited computational power and $$
- Documents are large
- Retrieval task - one vector one topic, reasonable vector space properties
- Vectors from text, texts - interpretability

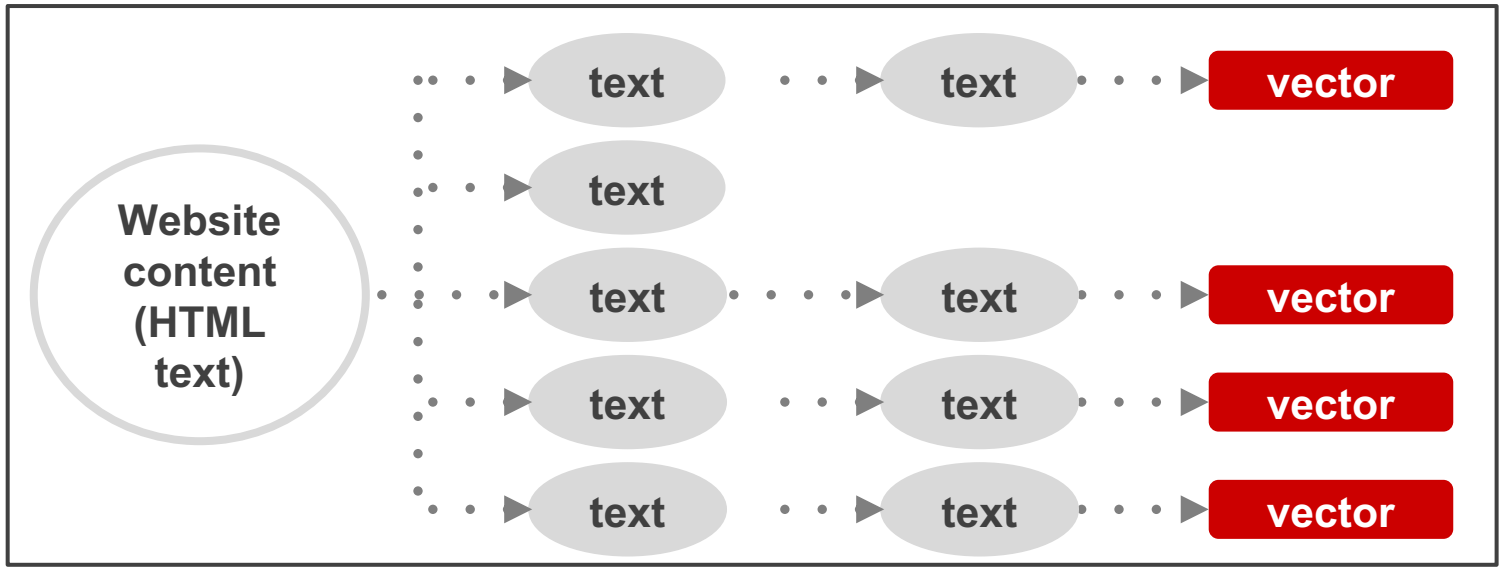
Vectors - properties



- Unlimited storage - up to N vectros per document
- >> 1 document, whole Czech internet, refreshed
- Limited computational power and $$
- Documents are large
- Retrieval task - one vector one topic, reasonable vector space properties
- Vectors from text, texts - interpretability

Our research

Our solution



Works in context of vector search

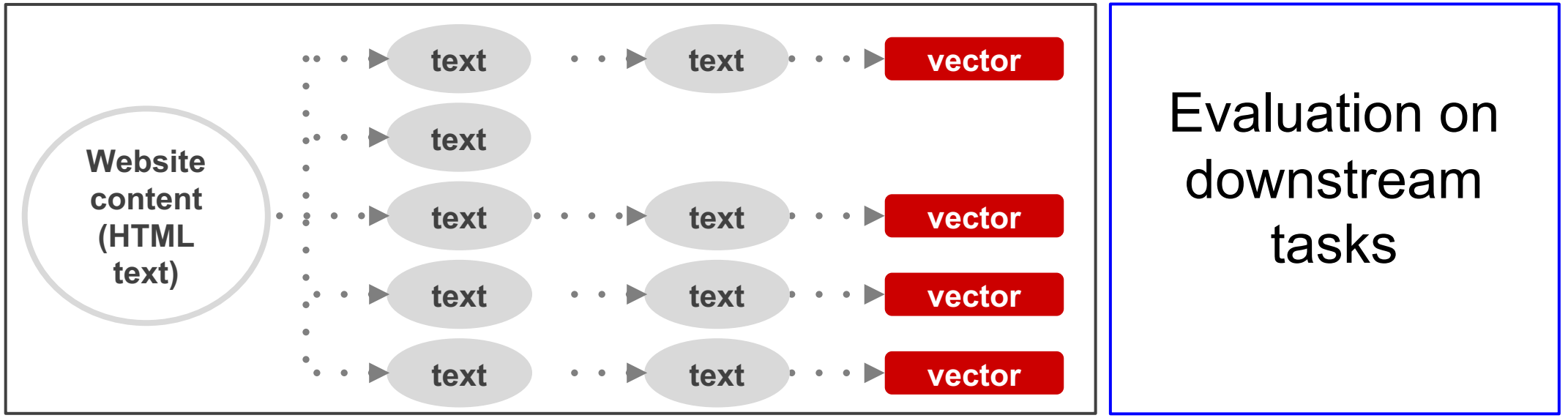
Applicable to production

‘Single topic’ vectors

Rather smaller number of vectors

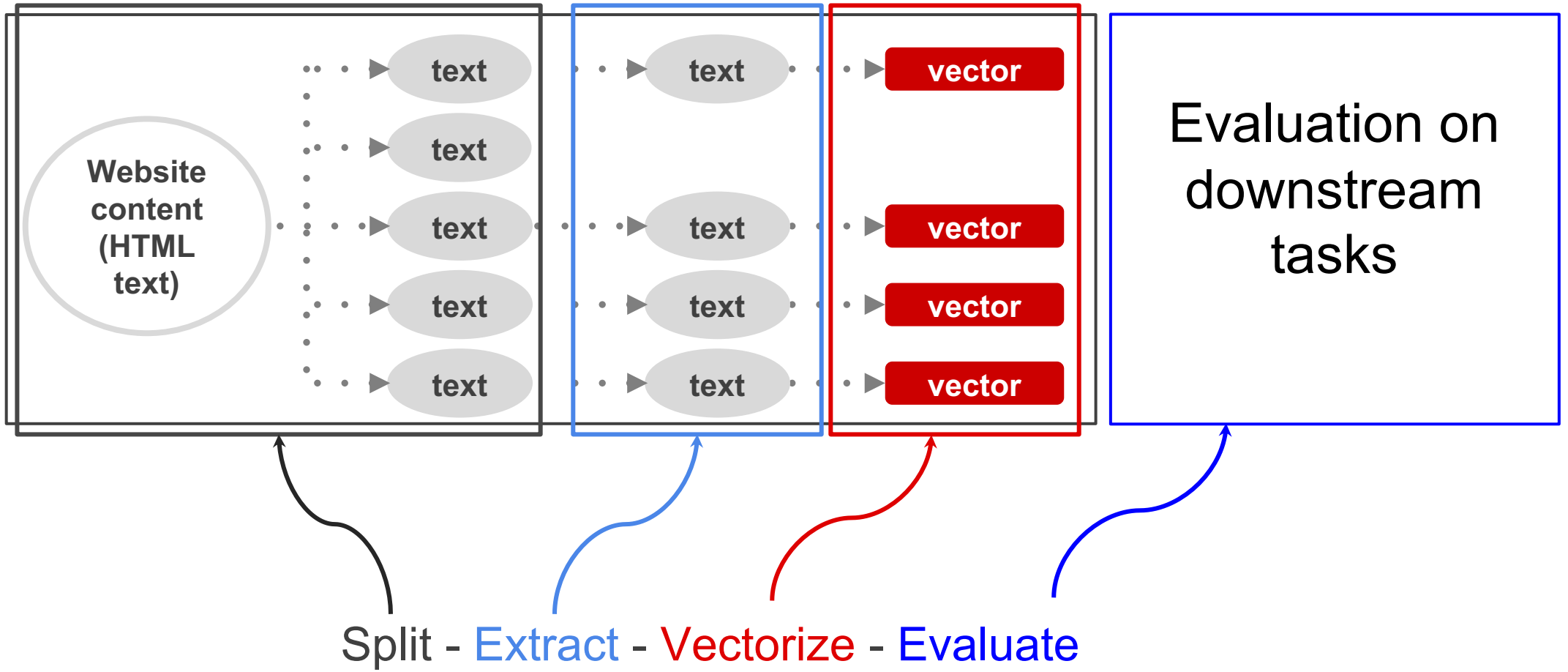
Split - Extract - Vectorize

Our research

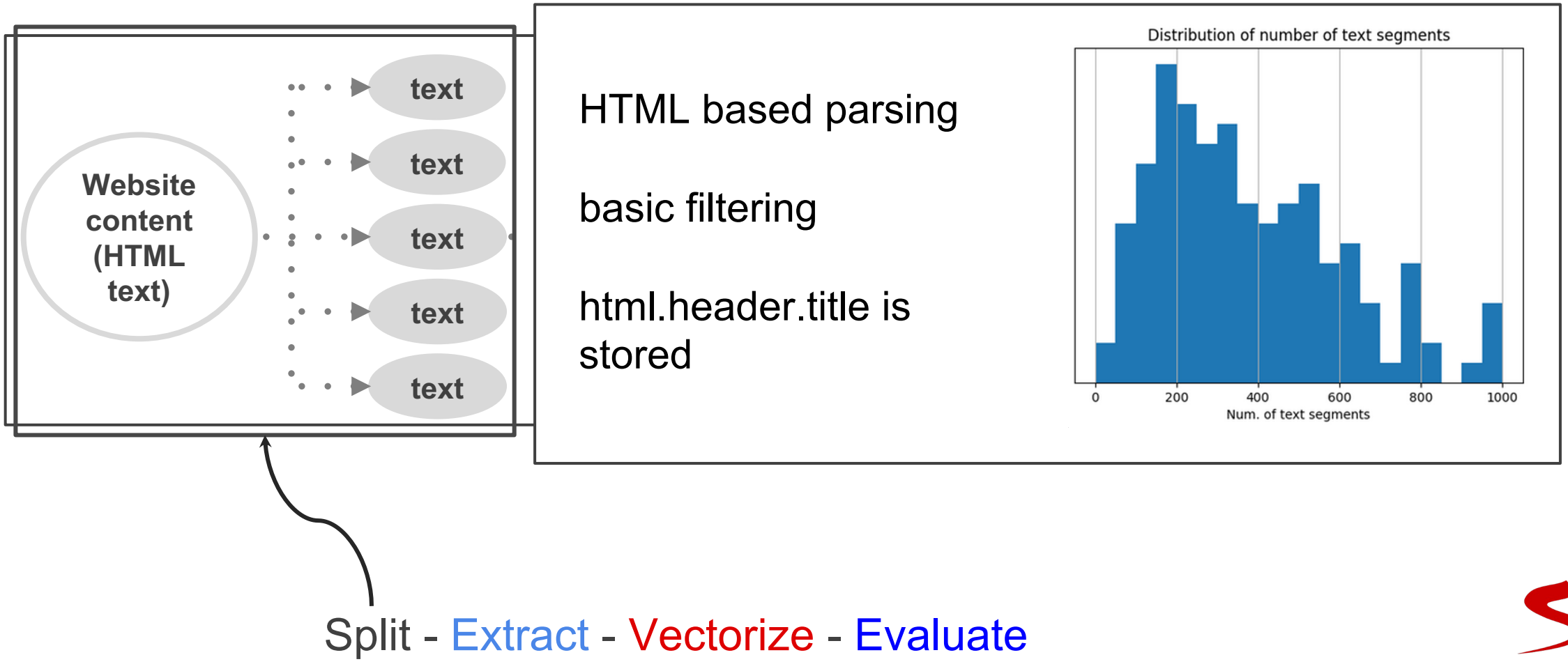


Split - Extract - Vectorize - Evaluate

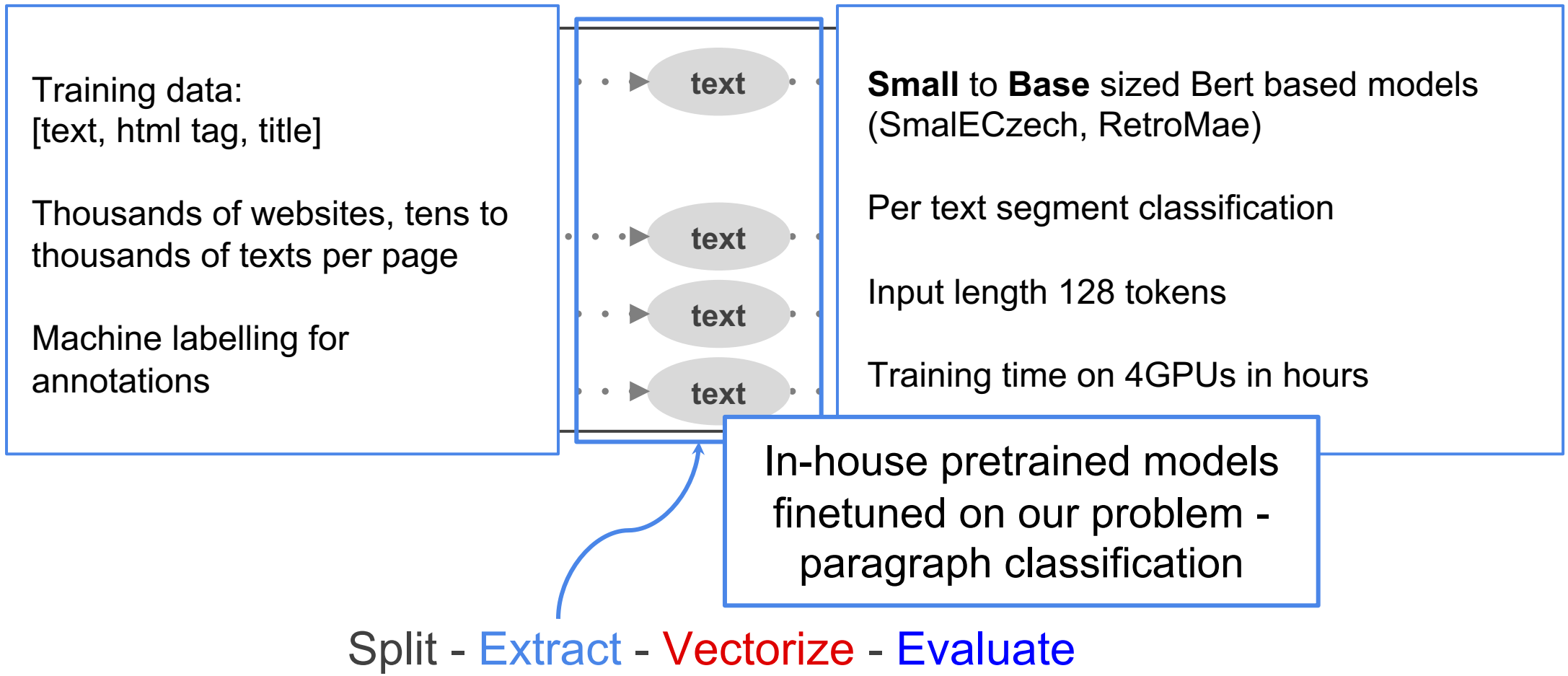
Our research



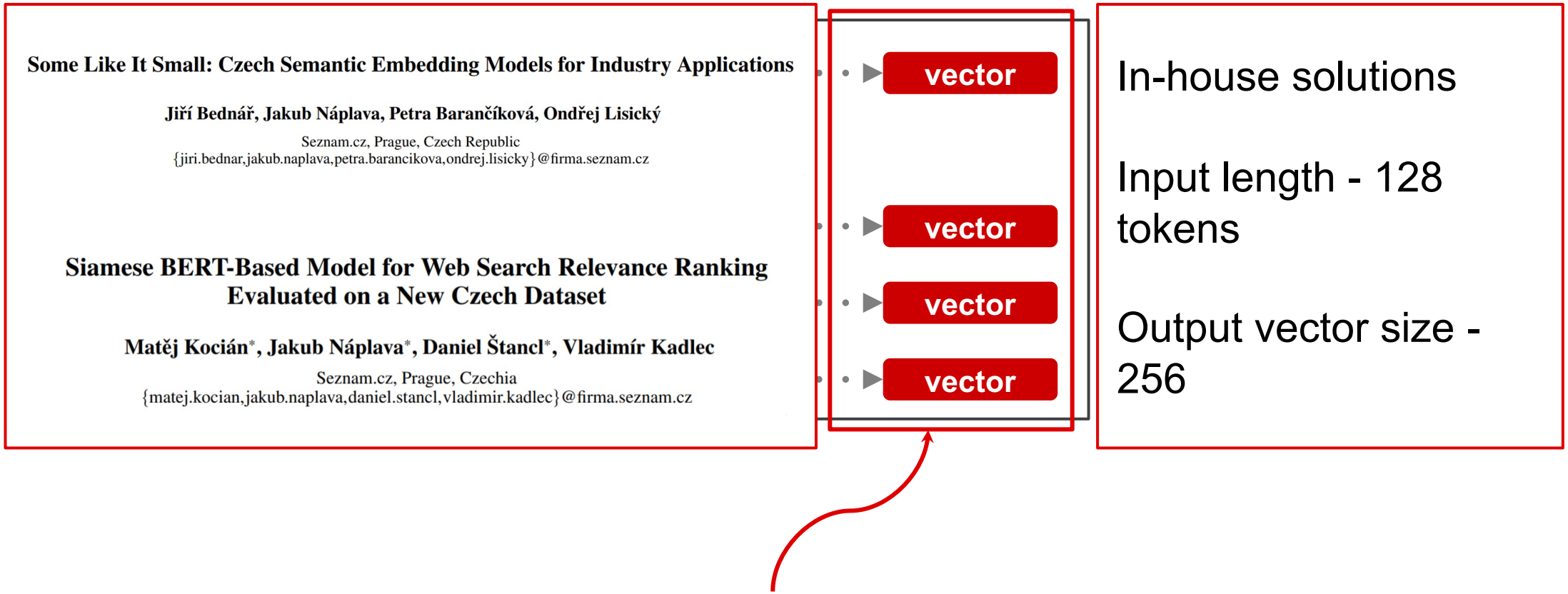
Split


Website content (HTML text)
text
text
text
text
text
HTML based parsing
basic filtering
html.header.title is stored
Distribution of number of text segments
0
200
400
600
800
1000
Num. of text segments
Split - Extract - Vectorize - Evaluate

Extract


Training data:
[text, html tag, title]
Thousands of websites, tens to thousands of texts per page
Machine labelling for annotations
text
text
text
text
Small to Base sized Bert based models (SmalECzech, RetroMae)
Per text segment classification
Input length 128 tokens
Training time on 4GPUs in hours
In-house pretrained models finetuned on our problem - paragraph classification
Split - Extract - Vectorize - Evaluate

Vectorize



Split - Extract - Vectorize - Evaluate

Evaluate

Retrieval task:

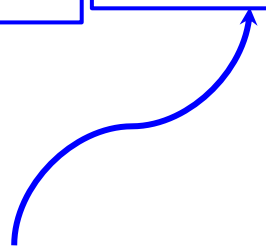
Retrieve *k* documents
How many of good documents did we retrieved?
rec@k
1 0 1 0 1 1| 0 0 1 0 0

Relevance task:

Use our text selection in relevance sorting.
prec@k

1 2 4 3 5 7 | 9 8 10

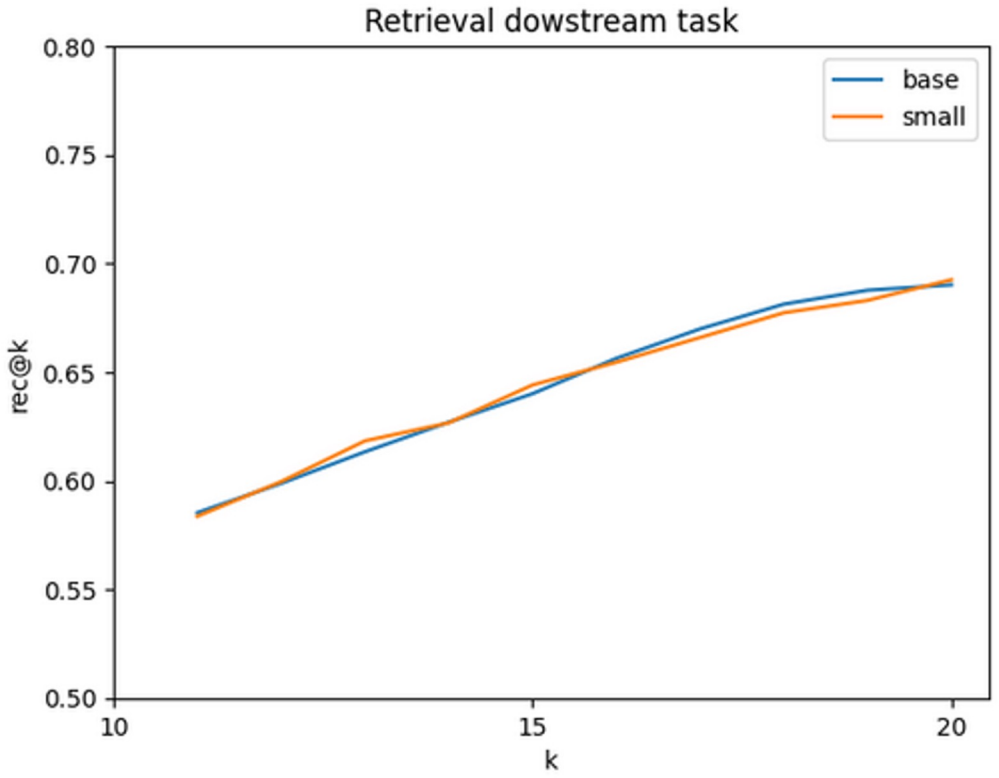
Evaluation on downstream tasks



Split - Extract - Vectorize - Evaluate

Results

Evaluate Results (very preliminary)



Relevance sorting task		
	small	base
prec@k (mod.)	0.692	0.7

https://zvukoveizolacnimaterialy.blogspot.com/2016/12/lamino-desky-na-miru.html

Prodej molitanu na míru Praha Molitanové desky patří k nejprodávanějším druhům pěn na trhu s čalounickým a dalším materiálem. Jedná se nejen o kvalitní desky z molitanové pur pěny, …

Deskový materiál na dřevěné bázi (překližky, lamino desky , dřevotřísky) a jejich následná úprava, to je Dřevobis, s. Firmám, ale i běžným spotřebitelům jsme schopni. Laminované dřevotřískové desky …

Dekory lamina a hrany se dají volně kombinovat. Kombinace barev si můžete vyzkoušet v aplikaci dodavatele lamin Egger. Vyberte si z přírodních materiálů jako je borovice, dub, buk, olše, třešeň, u dýhy a lamina je v nabídce …

….

Small vs Base sized models

Potential to improve

Resume

Self Question-Answering session

Why paragraph extraction approach over text summarization?

Easier, cheaper, in-house (at the moment), the approach do not hallucinate, the approach is `readable`.

Self Question-Answering session

Why paragraph extraction approach over text summarization?

Easier, cheaper, in-house (at the moment), the approach do not hallucinate, the approach is `readable`.

How can we be sure that model can work with separate paragraphs only?

Document is represented by its html.header.title, which goes as an input to training and to inference as well.

Self Question-Answering session

Why paragraph extraction approach over text summarization?

Easier, cheaper, in-house (at the moment), the approach do not hallucinate, the approach is `readable`

How can we be sure that model can work with separate paragraphs only?

Document is represented by its html.header.title, which goes as an input to training and to inference as well.

How precise does it have to be?

Errors propagate, BUT errors are also suppressed in later search steps: term search, relevance filtering, sorting…

Self Question-Answering session

Why paragraph extraction approach over text summarization?

Easier, cheaper, in-house (at the moment), the approach do not hallucinate, the approach is `readable`

How can we be sure that model can work with separate paragraphs only?

Document is represented by its html.header.title, which goes as an

How precise does it have to be?

Errors propagate, BUT errors are als suppressed in later search steps: te search, relevance filtering, sorting…

Then, why larger, or even generative models?

Better text selection (smarter than html segment), more applications (direct answers, related questions…), good for wide range of document types.

Future Plans

Future work - research and application

- Having fun with data: more data, better data, various url types…
- Having fun with models: small, base, distilled small, longer context, HTML language models
- Having fun with methodology: sequences instead of paragraphs
- Having fun with LLM: apply in-house LLM solution (e.g. for better web types coverage)

Martin Bachura

ML Research Team Lead

martin.bachura@firma.seznam.cz

Prostor pro dotazy