OSCARS
Open Science Clusters' Action for Research & Society

SlamaTrain – Representative Training Dataset for Slavonic Large Language Models

Marek Medveď, Radoslav Sabol, Aleš Horák

NLP Centre, Faculty of Informatics, Masaryk University

December 6, 2024

Funded by the European Union

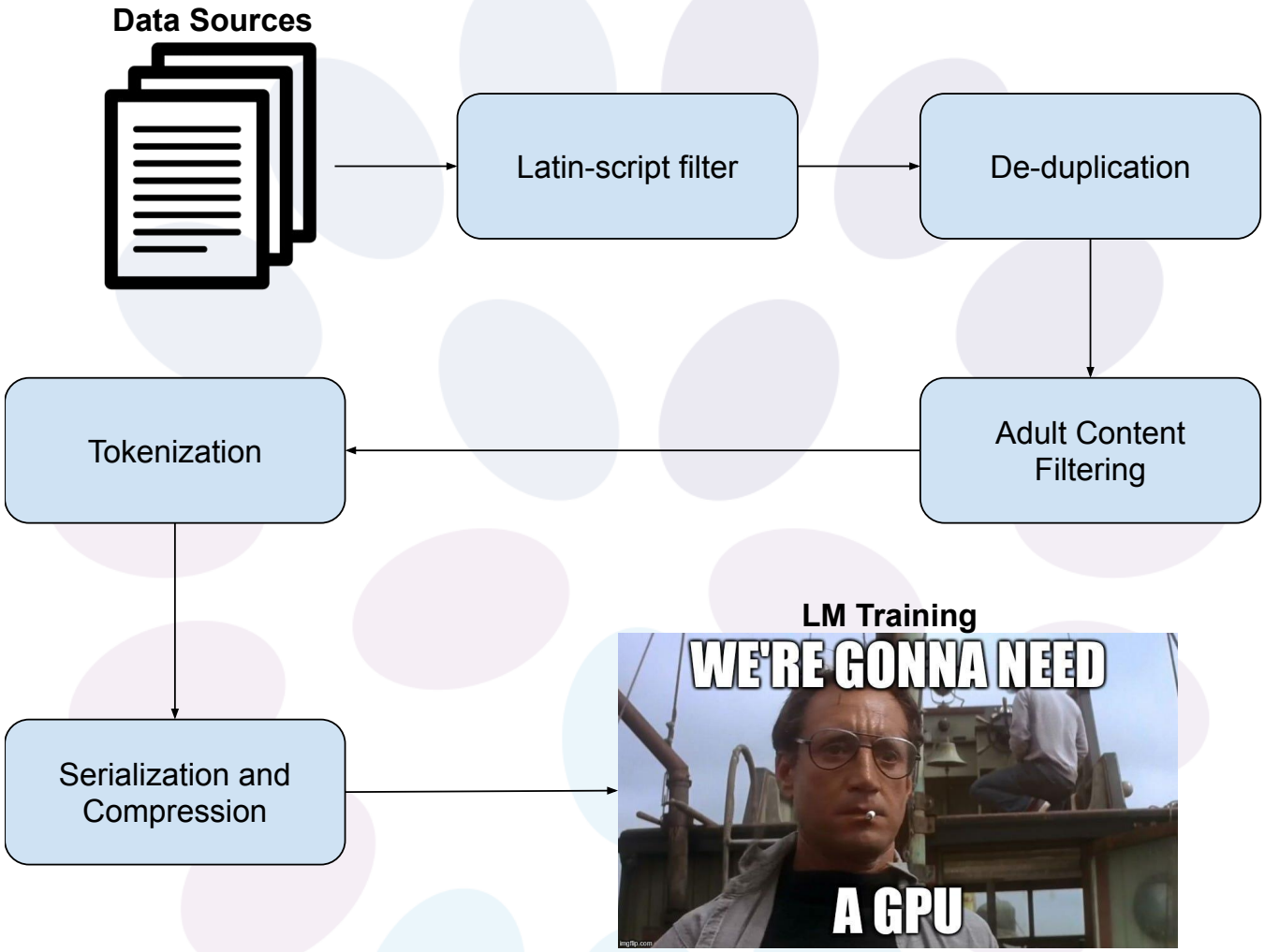
Outline

OSCARS

1. Computational Resources
2. Data Preparation
3. LLM Training
4. Evaluation (BenCzechMark)

Outline – Data Processing

OSCARS



Target Languages

OSCARS

- **Czech**
- **Slovak**
- **Polish**
- **Slovene**
- **Croatian**
- English
- French
- Italian
- German
- Spanish

Data Sources – Base

OSCARS

- **CulturaX**
  - 6.3T tokens
  - 167 languages
- **High Performance Language Technologies (HLPT)**
  - 11TB disk usage
  - 75 languages
- de-duplicated datasets
- base for all target languages

Data Sources – Monolingual

OSCARS

- **TenTen Corpora**
  - CS, SK, > 1B for each corpus
- **Aranea Corpora**
  - Slovak-centric family of corpora
- **czes corpus**
  - Czech newspaper/magazine corpus
- **MaCoCu**
  - family of corpora focused on low-resourced languages
  - Slovenian only

Data Sources – Initial Token Counts

OSCARS

Corpus	# tokens (M)
Araneum Maius (CS)	1,224
cstenten23	5,732
cstenten_all_mj2	14,278
czes2	455
CulturaX	35,617
HLPT	22,713
SUM	80,037

Corpus	# tokens (M)
CulturaX	10,323
skTenTen21	1,198
HLPT	5,913
Araneum Maius (SK)	1,244
skTenTen2	866
SUM	19,545

Latin-script filtering

OSCARS

- sentence-level removal

Slovami jedného z diskutujúcich pod článkom "Клин клином вышибают, но на Украине это поняли слишком буквально - дыры дырами латают..."

Figure: Example of the Latin-script filter.

De-duplication

OSCARS

- paragraph-level
- **Onion**
  - 5-grams
  - Duplicate threshold: 0.9
  - Maximal length of a ”stub”: 10

Post de-duplication statistics (CS)

OSCARS

Corpus	Source	Latin	in %	Deduplication	in %
Araneum Maius (CS)	1 224	1 223	99.9	1 192	97.3
cstenten23	5 747	5 732	99.7	3 248	56.5
cstenten_all_mj2	14 278	14 249	99.8	13 450	94.2
czes2	455	455	100.0	286	62.8
CulturaX	35 617	35 480	99.6	16 835	47.3
HPLT	22 713	22 470	98.9	3 106	13.7
SUM	80 037	79 611	99.5	38 120	47.6

Post de-duplication statistics (SK)

OSCARS

Corpus	Source	Latin	in %	Deduplication	in %
CulturaX	10 323	10 286	99.6	5 911	57.3
skTenTen21	1198	1196	99.8	1194	99.7
HPLT	5 913	5 854	99.0	2 332	39.5
Araneum Maius (SK)	1244	1243	99.9	521	41.9
skTenTen2	866	865	99.9	458	52.9
SUM	19 545	19 446	99.5	10 418	53.3

Borderline Example

OSCARS

Privatportal.sk si zamiluje každý pán, ktorý nie je spokojný so svojim sexuálnym životom. Vyskúšajte niečo nové, netradičné a vzrušujúce vďaka sex ponukám Martin. Zažite strhujúce erotické zážitky s príťažlivými slečnami, ktoré ponúkajú svoje erotické služby v Martine. Uprednostňujete štíhle slečny, ktoré sa s vami nežne pohrajú alebo sú vám bližšie ženy plnších tvarov, ktoré vedia s chlapom zatočiť? Z každého rožka troška, nájdete medzi sex inzerátmi Martin. Sex Martin ponúka inzeráty na erotické služby. Dokonalý prehľad kde v meste sa nachádzajú sex priváty Martin poskytujúce rôzne sex praktiky a sex ponuky za peniaze. Inzercia tiež zahŕňa erotické priváty ponúkajúce cez amatérky alebo profesionálky Zafo služby erotické masáže Martin.

Classifier

OSCARS

- **Support Vector Machines (SVM)**
- **Document Representation:** TF-IDF
  - 10,000 word 1–3grams
  - Stripped accents
  - Minimal document frequency: 2
  - Maximal document frequency (relative): 95%
- **Score for document:** distance from the separation hyperplane

Dataset – CS

OSCARS

- **Initial dataset:** labeled part of BUT-LCC
  - OK: 2,504
  - Adult: 203
- **Issue:** label imbalance
- train a SVM on the initial dataset to create new positive instances
- **Augmented Dataset**
  - OK: 2,504
  - Adult: 2,264

Dataset – SK

- **Initial Dataset:** machine-translated Czech dataset
  - OK: 2,504
  - Adult: 2,264
- **Issue:** bias towards online video sharing platforms
- **Improved Dataset**
  - OK: 3,507
  - Adult: 2,467

Adult Content Filtering

OSCARS

- both CS and SK documents **ranked** with their respective models
- all documents above **pre-determined threshold** are removed
- all domains where **>50% of documents are adult are removed**

Adult Filter Statistics (CS)

OSCARS

Corpus	Deduplication	in %	Adult filter	in %
Araneum Maius (CS)	1 192	97.3	1 168	95.4
cstenten23	3 248	56.5	3 085	53.7
cstenten_all_mj2	13 450	94.2	12 778	89.5
czes2	286	62.8	280	61.5
CulturaX	16 835	47.3	15 825	44.4
HPLT	3 106	13.7	2 889	12.7
SUM	38 120	47.6	36 027	45.0

Adult Filter Statistics (SK)

OSCARS

Corpus	Deduplication	in %	Adult filter	in %
CulturaX	5 911	57.3	5 408	52.4
skTenTen21	1 194	99.7	1 166	97.4
HPLT	2 332	39.5	1 859	31.4
Araneum Maius (SK)	521	41.9	509	40.9
skTenTen2	458	52.9	430	49.7
SUM	10 418	53.3	9 375	48.0

Tokenizers – Experiment Setup

OSCARS

- **Comparison of several tokenizers**
  - GPT-4o
  - BPE (GPT-2 style, Western Slama and Slama)
  - HFT (Western Slama (4 languages), Slama (10 languages)
- train set composed of 10M documents for each language
- BPE/Unitok ratio for a disjoint test set
- Vocabulary sizes: 32k, 50k, 100k, 200k

Tokenizer Evaluation

OSCARS

Tokenizer	cs	sk	pl	en	sl	hr
GPT-4o (Tiktoken)	1.98	1.95	2.26	1.11	1.87	1.92
Slama 32k (HF)	1.95	1.77	2.06	1.41	1.76	2.12
Slama 52k (HF)	1.85	1.68	1.91	1.33	1.66	2.04
Slama 100k (HF)	1.72	1.55	1.74	1.24	1.51	1.93
Slama 200k (HF)	1.58	1.43	1.58	1.17	1.39	1.82
Slama 52k (hftoks)	1.75	1.71	1.74	1.33	1.64	2.04
Slama 100k (hftoks)	1.60	1.55	1.53	1.22	1.46	1.93
Slama 200k (hftoks)	**1.45**	**1.42**	**1.36**	**1.14**	**1.32**	**1.83**
Western Slama 100k	1.59	1.47	1.55	1.19	1.94	2.14

Compression and Conversion

OSCARS

- MosaicML streaming dataset format
- ZSTD-compressed 64MB shards (reduced to 17MB via compression)

Dataset	Disk Usage	# data shards	# tokens (B)
Slama	688 GB	44 914	358 B
Western Slama	301 GB	19 494	166 B

Final Token Counts

OSCARS

Language	Size in B words	Size in B tokens
Czech	36.2	57.28
Slovak	9.37	13.97
Polish	35	53.55
Slovene	10.65	14.69
Croatian	1.22	2.22
Spanish	35	42.12
English	35	41.65
French	35	42.03
Italian	35	42.35
German	35	48.65
SUM	267.28	358.38

Conclusion

OSCARS

- a large new dataset focused on Slavonic languages
- 358 B tokens, 71 B in Czech and Slovak
- one of the largest cleaned collections for Czech
- already used for pre-training of Slama models


OSCARS
Thank you