Report Tokenization

In this assignment I worked with 4 languages: Cantonese, Upper Sorbian, Buryat, and Udmurt.

Task 1.

1) *Cantonese*: for tokenization of Cantonese, I chose Chinese model, because it is the language from Chinese branch. Despite this, results of tokenization were extremely poor:

Model: chinese-gsd-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	2125	93	0.05%	1.08%	0.09%
Sentences	60	93	30.00%	19.35%	23.53%

Model: chinese-gsdsimp-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1867	93	0.05%	1.08%	0.10%
Sentences	60	93	30.00%	19.35%	23.53%

2) *Upper Sorbian* is a West Slavic language, in the same branch with Czech, Polish, and Slovak. Therefore, I decided to try all models of these 3 languages for tokenization.

- Czech models

Model: czech-pdt-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1671	1490	78.34%	87.85%	82.82%
Sentences	70	92	35.71%	27.17%	30.86%

Model: czech-cac-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1671	1490	78.34%	87.85%	82.82%
Sentences	61	92	27.87%	18.48%	22.22%

Model: czech-cltt-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1669	1490	78.55%	87.99%	**83.00%**
Sentences	85	92	56.47%	52.17%	**54.24%**

Model: czech-fictree-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1671	1490	78.34%	87.85%	82.82%
Sentences	99	92	48.48%	52.17%	50.26%

- Polish models

Model: polish-pdb-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1670	1490	78.44%	87.92%	82.91%
Sentences	58	92	29.31%	18.48%	22.67%

Model: polish-lfg-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1670	1490	78.44%	87.92%	82.91%
Sentences	61	92	27.87%	18.48%	22.22%

- Slovak model

Model: slovak-snk-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1669	1490	78.55%	87.99%	83.00%
Sentences	64	92	28.12%	19.57%	23.08%

The best model based on F1 score is the Czech model czech-cltt-ud-2.5-191206. It demonstrates reliable score for words - 83%, but not so high for sentences – 54.24% (still the best one among all models).

3) *Buryat* language is a language of Mongolic family. However, there is no model for Mongolian language, so I decided to try tokenizing it with Russian models, just because it is spoken in some Russian regions and uses Cyrillic writing system.

Model: russian-gsd-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1801	1420	60.63%	76.90%	**67.81%**
Sentences	98	107	83.67%	76.64%	**80.00%**

Model: russian-syntagrus-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1803	1420	60.90%	77.32%	68.14%
Sentences	93	107	78.49%	68.22%	73.00%

Model: russian-taiga-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	1837	1420	58.36%	75.49%	65.83%
Sentences	104	107	74.04%	71.96%	72.99%

The best model based on F1 score is the russian-gsd-ud-2.5-191206: it has the highest score for sentences – 80%, and a high score for words – 67.81% (not the highest one, but the difference is small).

4) *Udmurt* language is a language of Uralic family, of Finno-Ugric group. Therefore, there are 3 ways to tokenize this language: with Finnish models, with Estonian models, or with Hungarian models.

Model: finnish-ftb-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	2617	2320	77.46%	87.37%	82.11%
Sentences	156	119	21.15%	27.73%	24.00%

Model: finnish-tdt-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	2584	2320	79.64%	88.37%	83.93%
Sentences	78	119	32.05%	21.01%	25.38%

Model: estonian-edt-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	2608	2320	77.99%	87.67%	82.55%
Sentences	75	119	37.33%	23.53%	28.87%

Model: estonian-ewt-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	2609	2320	77.92%	87.63%	**82.49%**
Sentences	89	119	46.07%	34.45%	**39.42%**

Model: hungarian-szeged-ud-2.5-191206.udpipe.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	2609	2320	77.92%	87.63%	82.49%
Sentences	73	119	32.88%	20.17%	25.00%

The best model is the Estonian model estonian-ewt-ud-2.5-191206. All models have almost the same score for words, but this one has the highest score for sentences – 39.42%.

Task 2.

1) Cantonese

I trained the tokenizer using Cantonese data from UDHR dataset, splitting it into training and test part.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	25	25	100.00%	100.00%	100.00%
Sentences	25	25	100.00%	100.00%	100.00%

Resulting accuracy:

- Words: better than in Chinese models
- Sentences: better than in Chinese model

2) Upper Sorbian

I trained the tokenizer using Upper Sorbian data from UDHR dataset, splitting it into training and test part.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	377	377	100.00%	100.00%	100.00%
Sentences	23	20	73.91%	85.00%	79.07%

Resulting accuracy:

- Words: better than in other languages models
- Sentences: better than in other languages model

3) Buryat

To train this tokenizer, I created a dataset using texts from Wikipedia.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	342	342	100.00%	100.00%	100.00%
Sentences	18	28	50.00%	32.14%	39.13%

Resulting accuracy:

- Words: better than in Russian models
- Sentences: worse than in other languages model

4) Udmurt

To train this tokenizer, I created a dataset using texts from Wikipedia.

Accuracy check:

	System	Gold	Precision	Recall	F1
Words	529	529	100.00%	100.00%	100.00%
Sentences	23	33	34.78%	24.24%	28.57%

Resulting accuracy:

- Words: better than in other languages models
- Sentences: worse than in the best (Estonian) model, but almost the same as in other models.