

# Language Recognizer

## Povinné parametry

Parametr	Popis
<b>file</b>	Soubor s texty. Je možné zadat více souborů.
<b>model</b>	Model, který se má použít při rozpoznávání jazyka. Je potřeba zadat minimálně 2 různé modely. Modely musí být ve složce "Models" v hlavní složce aplikace.

## Volitelné parametry

Parametr	Popis	Možné hodnoty	Defaultní hodnota
<b>models all</b>	Použije všechny modely ze složky "Models" v hlavní složce aplikace.		
<b>order</b>	Stupeň n-gramů, který se použije pro rozpoznávání jazyka. Pokud se zvolí vyšší stupeň než mají modely, použije se nejvyšší stupeň modelů.	Celé kladné číslo	5
<b>difference</b>	Určuje, o kolik musí mít rozpoznávaný model vyšší log(p) proti modelu s druhým nejvyšším log(p) pro danou větu. Pokud je rozdíl menší než tento parametr, věta se zařadí mezi nejisté výsledky, jinak se zařadí mezi jisté výsledky.	Celé kladné číslo	0
<b>write_uncertain_results</b>	Určuje, zda se na výstup mají zapisovat nejisté výsledky.	true/false	true
<b>folder</b>	Jako vstupní textové soubory použije všechny soubory ze zadaného adresáře a všech jeho podadresářů.	Textový řetězec	
<b>encoding</b>	Kódování vstupního textového souboru. Stejně kódování se použije pro výstupní soubory.	Textový řetězec	UTF-8
<b>separators</b>	Oddělovače vět v odstavci. Oddělování odstavců je automatické. Pokud se nezadají žádné oddělovače, bude text rozdělen jen podle odstavců.	Textový řetězec	
<b>min_length</b>	Minimální délka vět včetně oddělovače. [znaky]	Celé kladné číslo	1
<b>ignore_next_separator</b>	Pokud je délka věty kratší než parametr min_length, pak ignoruje následující oddělovač, tak aby získal delší větu.	true/false	false
<b>split_text</b>	Určuje, zda se má výstup rozdělit podle zadaných oddělovačů (každá věta na samostatný řádek).	true/false	false
<b>lower_case</b>	Převede všechny znaky rozpoznávaného textu na malá písmena. Při použití tohoto parametru by i modely měly být natrénovány na textech s malými písmeny.	true/false	false
<b>remove_names</b>	Určuje, zda se z rozpoznávaných vět mají odstranit „jména“ (slova ve větě začínající velkým písmenem, kromě prvního slova). Tento parametr má vliv pouze na texty, podle kterých se rozpoznává jazyk, výstupní texty budou beze změny (se „jmény“).	true/false	false

## Výstup

Výstupní soubory jsou ve stejné složce jako vstupní textový soubor. Kódování těchto souborů je podle parametru "encoding", defaultně UTF-8. Výsledky jsou rozděleny podle parametru "difference" na jisté a nejisté. Značení názvů souborů je následující:

Jisté výsledky:	[název souboru s texty]-[název modelu]
Nejisté výsledky:	[název souboru s texty]-[název modelu]-uncertain

## Příklad

LanguageRecognizer file data.txt folder C:\Texty encoding windows-1250 model Czech model Slovak order 6 separators .?! ignore\_next\_separator true difference 30