Language Recognizer

Aplikace slouží pro třídění textů podle jazyka. Aplikaci v parametru předáte soubor s texty, který chcete roztřídit, a řeknete, které modely má použít při identifikaci jazyka. Dále aplikaci předáte seznam oddělovačů, podle kterých má rozdělovat text na části, na kterých se bude jazyk identifikovat. Implicitně se oddělují pouze odstavce. Aplikaci lze předat ještě mnohem více parametrů, viz níže. Výsledkem jsou roztříděné texty, umístěné v jednotlivých textových souborech a označené identifikovaným modelem.

Povinné parametry

Parametr	Popis
file	Soubor s texty. Je možné zadat více souborů.
model	Model, který se použije při identifikaci jazyka. Je potřeba zadat minimálně 2 různé modely. Modely musí být ve složce „Models“, která je v hlavní složce aplikace.

Volitelné parametry

Parametr	Popis	Možné hodnoty	Defaultní hodnota
models all	Použije všechny modely ze složky „Models“, která je v hlavní složce aplikace.		
order	Stupeň n-gramů, který se použije pro identifikaci jazyka. Pokud se zvolí vyšší stupeň než mají modely, použije se nejvyšší stupeň modelů.	Celé kladné číslo	5
difference	Určuje, o kolik musí mít identifikovaný 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
write_incertain_results	Určuje, zda se na výstup mají zapisovat nejisté výsledky.	true/false	true
folder	Jako vstupní textové soubory použije všechny soubory ze zadaného adresáře a všech jeho podadresářů.	Textový řetězec	
encoding	Kódování vstupního textového souboru. Stejné kódování se použije pro výstupní soubory.	Textový řetězec	UTF-8
separators	Oddělovače vět v odstavci. Oddělování odstavců je automatické. Pokud se nezadají žádné oddělovače, bude text rozdělen jen podle odstavců.	Textový řetězec	
min_length	Minimální délka vět včetně oddělovače. [znaky]	Celé kladné číslo	1
ignore_next_separator	Pokud je délka věty kratší než parametr „min_lenght“, ignoruje následující oddělovač, tak aby získal delší větu.	true/false	false
split_text	Určuje, zda se má výstup rozdělit podle zadaných oddělovačů (každá věta na samostatný řádek).	true/false	false
lower_case	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
remove_names	Určuje, zda se z rozpoznávaných vět odstraní „jména“ (slova ve větě začínající velkým písmenem, kromě prvního slova). Tento parametr má vliv pouze na texty, podle kterých se identifikuje 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