



IPSSC +
CMBO

STUDENTS' CONFERENCE

Jožef Stefan International Postgraduate School
and Young Researchers' Day CMBO

19 and 20 April

POS tagger and lemmatizer evaluation: The TextFlows open science approach

Matej Martinc,
Jožef Stefan International Postgraduate School

Poster number: 35

What is my research?

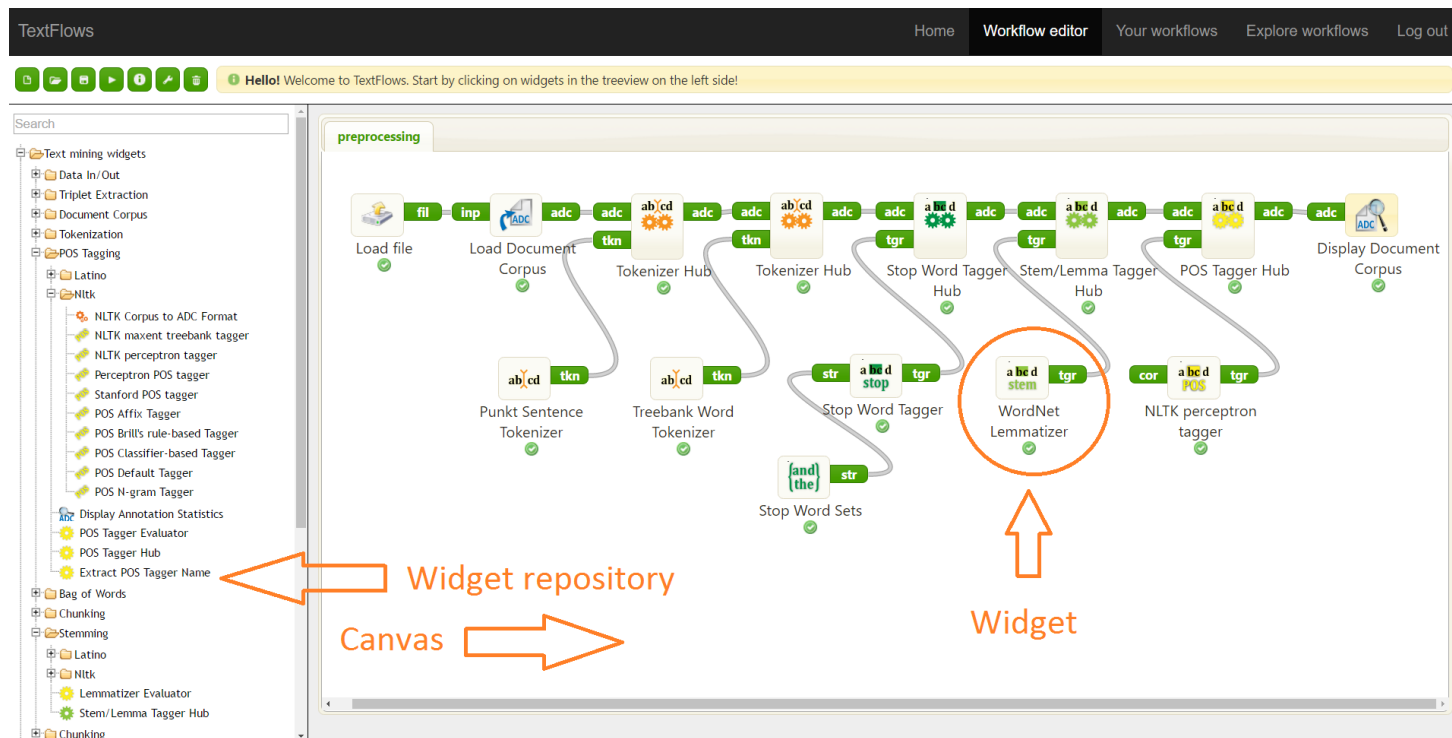
- Text mining
 - Subfield of artificial intelligence
 - Automatic extraction of information from language
- Tools that help computer „understand“ language
 - POS tagging (a:determiner dyslexic:adjective man:noun walks:verb into:preposition a:determiner bra:noun)
 - Lemmatization (blue:blue is:to be greener:green than:than purple:purple)

What is the problem?

- Nobody properly evaluated these tools,
- So we did it!
 - Number of different experiments were conducted
 - We identified different factors that influence the performance of these tools
 - We compared these tools to humans (gold standard corpora)
 - We tested how these tools perform in real life text mining tasks (gender classification)

How did we do it?

- TextFlows (<http://www.TextFlows.org>)



- Everything is published online, anyone can reproduce the study!

What did we achieve?

- We gave researchers and students a better understanding on what tools to use in a variety of different scenarios
- We help them reduce the time and cost of their research
- We help them improve the quality of their research