Authorship analysis

Roberto Deresu
Facultatea de Matematica si Informatica

Dataset

- The Federalist Papers Dataset
- PAN12 - Text Authorship Attribution Dataset

Features

- Ranking distance representation of each document, with function words: Mosteller and Wallace or nltk english
- N-grams representation of each document. N-grams range: 3-4, frequency threshold: 100

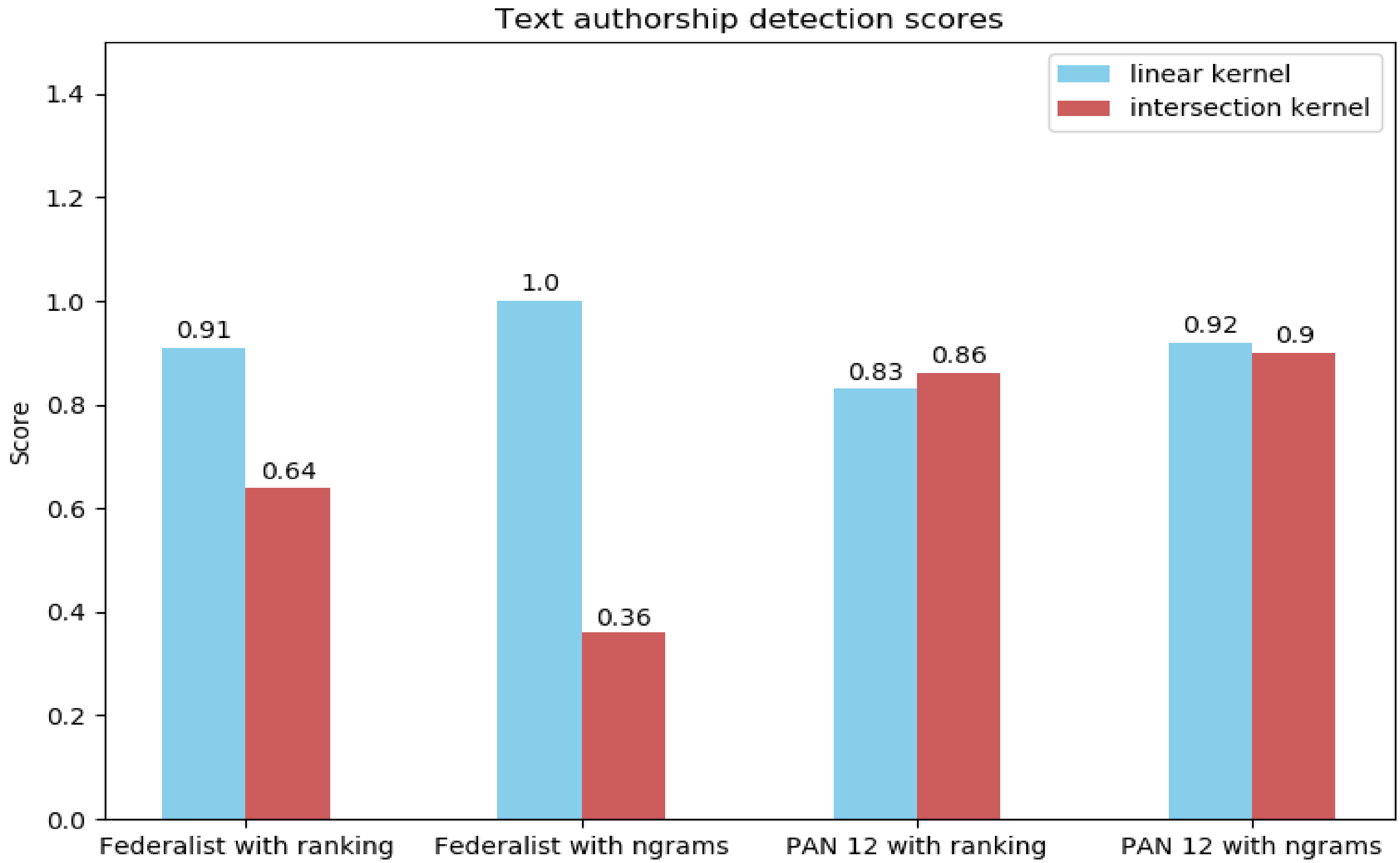
String-kernels

- Linear kernel
- Intersection kernel
- Binary kernel

Training and classification

- SVM with one vs. rest classifier

Results


Text authorship detection scores
linear kernel
intersection kernel
Score
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.91
0.64
1.0
0.36
0.83
0.86
0.92
0.9
Federalist with ranking
Federalist with ngrams
PAN 12 with ranking
PAN 12 with ngrams

THANK YOU!