Authorship Verification with Skip N-grams and Termsets

Chao Chen, Master of Science in Computer Science

University of Dublin, Trinity College, 2020

Supervisor: Carl Vogel, Erwan Moreau

The aim of this work is to go through a complete process to solve Authorship Verification problems. A set of problems are given in the Authorship Verification task, in each problems, several known texts and one unknown texts are present. The aim is to judge whether these texts are written by the same author.

The experiments focus the difference between two kinds of features: skip n-grams and termsets. A skip n-gram is a special n-gram that can take whatever types of token into account in some of its positions. Weighted termsets are terms without spacial relations, there weighting depends on the occurrence across the problem texts, by whether they appear simultaneously or alone. The termset weighting method is adapted to Authorship Verification base on small amount of texts.

This work follows the rules and uses the datasets from PAN at CLEF shared tasks in 2013, 2014 and 2015, where classifications are implemented with corpora from different genres with different topics. Classifications with SVMs, random decision trees are performed. The coefficients of linear kernel SVMs are examined to identify the influences of features during classification.

A modified Impostor method and an adaption of Universum Inference are also selected as option strategies. The modified Impostor method details the scoring for each repetition and adds a filter for impostors with better similarities before scoring. The Universum Inference is simplified to solve the problems with small sets of texts.

The genre may have impact on the skip n-gram and termset performance. Skip n-gram features seem to be more effective than the termset features in a corpus close to oral form.