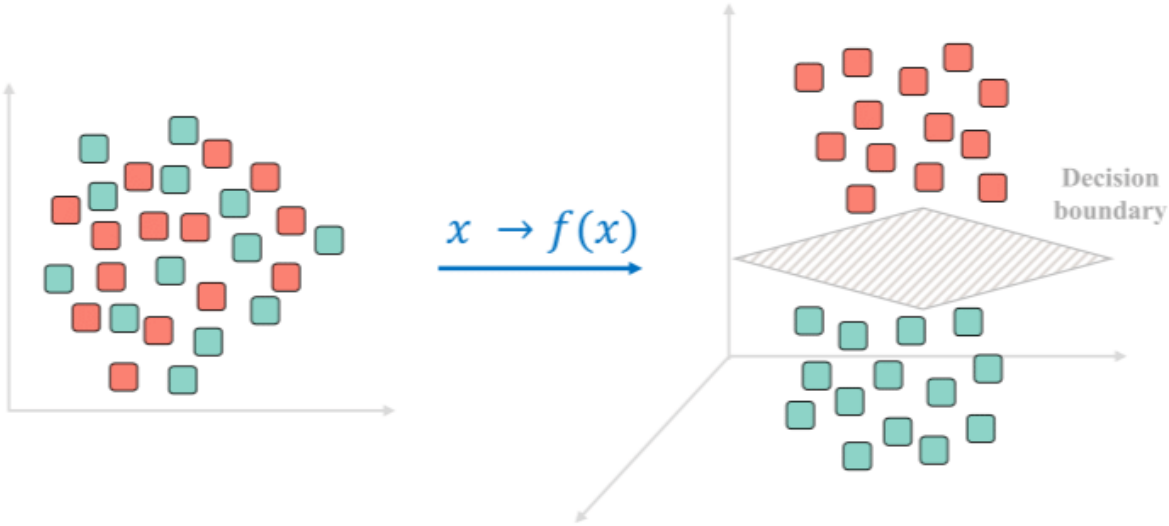
ИЗВЕШТАЈ ЗАДАТКА 3

Решавање класификационог проблема употребом SVM класификатора



Душан Бркић, Филип Живанац

19. цветња лета Господњег 2022.

Софтверско инжењерство и информационе технологије

Факултет техничких наука

Универзитет у Новом Саду

ЗАДАТАК

Класификовати наслове онлајн медијских чланака на енглеском језику (text) у две класе (clickbait): 0 - наслов није кликбејт, 1 - наслов јесте кликбејт. Задатак је успешно урађен уколико се на комплетном тестном скупу података добије микро f1 мера (micro f1 score) већа од 0.90. Задатак се решава употребом SVM класификатора.

ПРИСТУП ПРОБЛЕМУ

За решење проблема смо на почетку из свих наслова уклонили знакове интерпункције како би се, након векторизације, речи које се уз њих налазе биле препознате и повезане једне с другима. Затим смо искористили векторајзер CountVectorizer и TF-IDF (term frequency-inverse document frequency) трансформер из python библиотеке scikit-sklearn. Поред уклањања интерпункције, искористили смо још неке feature extraction методе: Направили смо унираме и биграме, уклањали речи које се превише често појављују, уклонили речи које се појављују у само једном документу и одрадили сублинеарно tf скалирање. Задатак смо онда решили коришћењем SVM класификатора и сигмоидног кернела, трансформацијом и тренирањем датих вектора.

ИСПРОБАНЕ МЕТОДЕ

ПРЕДПРОЦЕСИРАЊЕ РЕЧИ

Испобали смо методе предпроцесирања као што су уклњање интерпунције, претварање у мала слова и уклањање зауставних речи, затим уклањање високо фреквентних (зауставних) речи, нискофреквентних речи, убацивање униграма и биграма. Од ових метода једина која је имала негативан ефекат јесте претварање свих слова у мала, јер је један од главних одлика кликбејт сајтова велико почетно слово сваке речи.

TF-IDF (term frequency-inverse document frequency)

Статистичка мера за рачунање колико је дата реч релевантна у документу. Ова метрика се добија множењем двеју метрика, односно метрике која показује колико се пута реч појављује у документз и метрике инверзне фреквенције речи у свим документима. Такође смо користили и сублинеарно TF скалирање.

ОДАБИР КЕРНЕЛА

За кернел смо искористили сигмоидни кернел gamma вредности ‘1’. Приметили смо да тип

кернела не утиче драстично на решење, сем ако није изабран полиномијални, он даје ужасан резултат.

РЕЗУЛТАТИ

SVM класификатор са сигмоидним кернелом

Игнорисање великих слова	TF-IDF	Кернел	Гама	micro f1 скор
ДА	ДА	Полинимионалан	1	1.0
НЕ	ДА	Полиномионалан	1	0.55
НЕ	ДА	Полиномионалан	1	1.0
НЕ	**ДА**	Сигмоидан	1	**1.0**
НЕ	ДА	Сигмоидан	0.01	0.55

ОДАБРАНО РЕШЕЊЕ

За решење проблема смо изабрали:

- Уклањање интерпункције
- Уклањање високофреквентних речи
- Уклањање нискофреквентних речи
- Креирање биграма
- Векторизовање речи
- Трансформација вектора помоћу TF-IDF (term frequency-inverse document frequency) трансформера из python библиотеке scikit-sklearn, уз сублинеарно tf скалирање.
- SVM класификатор и сигмоидни кернел