КУРСОВ ПРОЕКТ

Свързване на обекти с Магелан

|  |  |
| --- | --- |
| **Факултет по Математика и Информатика** **Студент:** Борислав Стоянов Марков **Факултетен номер:** **0MI3400048** **Учебен план:** **Изкуствен Интелект (редовно, магистър)** Курс: **Курс 1**; Група: **Група 1** **Активен период**: 2021/2022 летен, магистри **Дисциплина**: Семантичен Уеб | **Факултет по Математика и Информатика** **Студент:** Кирил Димов Георгиев **Факултетен номер:** **1MI3400098** **Учебен план:** **Изкуствен Интелект (редовно, магистър)** Курс: **Курс 1**; Група: **Група 1** **Активен период**: 2021/2022 летен, магистри **Дисциплина**: Семантичен Уеб |

# 1. Съдържание

[1. Съдържание 2](#_Toc103542121)

[2. Увод 3](#_Toc103542122)

[3. Реализация 3](#_Toc103542123)

[3.1 Алгоритъм 3](#_Toc103542124)

[3.2 Библиотека 3](#_Toc103542125)

[3.3 Корпус с данни 3](#_Toc103542126)

[3.4 Редукция на входните множества 3](#_Toc103542127)

[3.4.1 Допълнително редуциране с нови полета 3](#_Toc103542128)

[3.5 Свързване на обектите 3](#_Toc103542129)

[3.5.1 Подобряване на сравнението с нови полета 3](#_Toc103542130)

[3.6 Използване на приложението 3](#_Toc103542131)

[3.6 Оценка на резултатите 3](#_Toc103542132)

[4. Недостатъци и подобрения 3](#_Toc103542133)

[5. Източници и използвана литература 3](#_Toc103542134)

[Приложения 3](#_Toc103542135)

[1. Сорс код (Source code) 3](#_Toc103542136)

# 2. Увод

# 3. Реализация

## 3.1 Алгоритъм

След като са обработени първоначално двете множества, новите атрибути са добавени и са използвани различни блокинг техники за създаване на кандидатите, трябва обектите от двете множества да бъдат свързани. Свързването на обектите се осъществява посредством алгоритъм от областта на машинното самообучение, който трябва да предскаже с някаква вероятност, дали дадени два обекта са едни и същи. За да получим максимален брой точни свързвания, трябва да бъде избран подходящ алгоритъм, който да бъде максимално добре обучен върху тренировъчните данни и да бъде изпробван върху тестовите данните. Ще бъде използван подходът ,,К-Fold Cross Validation”, за да изберем най-добрият алгоритъм, като метриката за измерване на алгоритмите е точност (precision). Алгоритмите, измежду които ще бъде избирано, са Дърво на решенията (Decision Tree), support vector machine, случайна гора (random forest), логистична и линейна регресия (logistic and linear regression). След това бива превръщан изчислителния сет във фийчър вектори за тренировъчните и тестовите данни. Сега биват използвани фийчър векторите от тренировъчния сет, за да бъде трениран алгоритъма, и го изпробваме върху тестовото множество.

## 3.2 Библиотека Магелан

Библиотеката Магелан предоставя и набор от алгоритми за машинно самообучение, които могат да бъдат използвани за свързване на обектите. Te са следните:

* DecisionTree (Дърво на решенията)
* Random Forest (Случайна гора)
* Support Vector Machine
* LinReg (линеарна регресия)
* LogReg (логистична регресия)

Също така този фреймуорк предоставя функцията ,,select\_matcher”, която чрез ,,К-Fold Cross Validation” метода избира най-добрия алгоритъм, който да бъде използван. За да може да бъде трениран алгоритъма, данните от „суров вид“ трябва да бъдат преобразувани във фийчър вектори. Това става посредством фукцията „extract\_feature\_vecs“.

## 3.3 Корпус с данни

## 3.4 Редукция на входните множества

### 3.4.1 Допълнително редуциране с нови полета

## 3.5 Свързване на обектите

### Първо, за да свържем обектите, трябва да заредим двете нови таблици с новите колони, както и лейбълнатите данни.След това използваме функциите ,,clean\_price(s)“ и ,, guess\_screen\_size(s)“, чрез които добавяме новите колони за цената на продукта и за размера на екрана само за електронните продукти. Ще бъде използван подходът ,,К-Fold Cross Validation”, за да изберем най-добрият алгоритъм, като к = 6 и метриката за измерване на алгоритмите е точност (precision). Алгоритмите, измежду които ще бъде избирано, са Дърво на решенията (Decision Tree), support vector machine, случайна гора (random forest), логистична и линейна регресия (logistic and linear regression). След изпълнението на ,,К-Fold Cross Validation” получените резултати показват, че алгоритъмът Дърво на решенията се справя най-добре. След това бива превръщан изчислителния сет във фийчър вектори за тренировъчните и тестовите данни. Сега използваме фийчър векторите от тренировъчния сет, за да тренираме алгоритъма, и го изпробваме върху тестовото множество. Получаваме следните резултати: точност, пълнота и F1 оценка имат стойност около 50%.

### 3.5.1 Подобряване на свързването с нови полета

За да бъдат подобрени резултатите, ще бъдат добавени новите атрибут при тренирането и оценяването на кандидате, получени от предишните стъпки. Това е атрибутът цена (Price) и атрибутът, които описва инчовете на електронните устройства. За тази цел са създадени две функции: ,,clean\_price“ и ,,guess\_screen\_size“. Първата функция взема текстовият низ, който репрезентира цената на продукта, и я превръща в число. Втората функция съединява текстовите полета- името, описанието и характеристиките на продукта и търси определени числа (пример 17.3) и в случай, че намери някое от изброените числа, то продуктът получава стойност за инчовете. Продукти, които не са електронни, нямат стойност. След възпроизвеждане на предишната стъпка за трениране на модела, то моделът получава по-добри резултати, а те имено са: Точност, пълнота и F1 оценка имат стойност 72%.

## 3.6 Използване на приложението

## 3.6 Оценка на резултатите

# 4. Недостатъци и подобрения

# 5. Източници и използвана литература

[1] <TITLE>, AUTHORS, YEAR, ISBN, 978-XXX  
link

# Приложения

## 1. Сорс код (Source code)

Кодът е публичен и качен в платформата Гитхъб.

<https://github.com/borkox/uni-sofia-entity-linking-magellan/>