Проект по ПОЕЕ
на Мария Георгиева Терзиева
специалност Изкуствен интелект
ф.н. 25026

Тема на проекта: Разпознаване на еднакви артикули от различни сайтове и предлагане на подобни артикули на даден артикул по заглавие, вид на обувките (мъжки, женски) и цвят.

Част от проекта, свързана с курса по ПОЕЕ: Разпознаване на еднакви артикули от различни сайтове и предлагане на подобни артикули на даден артикул по заглавие, вид на обувките (мъжки, женски) и цвят. (Разработването на търсачка и извличането на данни са за курса по Извличане на информация).

Подход: Използвана е библиотеката elasticsearch. Elasticsearch има функционалност, която позволява да се извлекат данните, чието поле (напр. title) съвпада на 33%, 66%, 75% или 100% с определена заявка.

Първоначалната ми идея беше да използвам полето title на един артикул като текст, по който да търся артикули, чието поле title съвпада на 75% с него.

В заглавията на данните обаче се оказа, че има много думи (напр. Обувки, мъжки, спортни, ботуши...), които се срещат много често в базата и ако много от тях се срещат в текста, по който търсим, ще получим голямо съвпадение при съвсем неподобни обувки. Затова реших да филтрирам текста за търсене, като премахна от него незначещите думи ('спортни', 'обувки', 'ботуши', 'туристически', 'джапанки', 'апрески', 'футболни', 'платформа', 'сандали', 'дамски', 'кецове', 'маратонки', 'зимни', 'чехли', 'мъжки', 'балеринки', 'black', 'white', 'red', 'blue', 'grey').

Премахвайки от текста незначещите думи, той придоби вид, в който на първо място стои марка и оттам нататък име на модел + всеки сайт може да добави към края някакви кодове, състоящи се от цифри или буква, последвана от цифри, които не са ни полезни при търсенето на еднакви/подобни обувки, а тъкмо обратното – могат да предизвикат голямо съвпадение между заглавията на обувки, които всъщност не са подобни. От тази гледна точка филтрирах и кодовете.

Реших да филтрирам и самата марка обувки, защото моделът осигурява достатъчна уникалност и марките също биха могли да доведат до подвеждащи резултати, що се отнася до подобността/еднаквостта на обувки.

Частта от текста за търсене, която остана, в някои случаи се състои от 1 или 2 думи. В тези случаи е ясно, че става въпрос за марката на обувките. Тогава искаме да ни се върнат резултати, в които се среща този текст на 100%. В останалите случаи искаме съвпадението между текста, по който търсим, и резултатите да е 75%.

Отделно, от получените резултати се избират тези, които са подобни на оригиналния артикул, като се гледа да са от същия вид(мъжки, женски) и да имат същия цвят. За тази цел са използвани more_like_this заявки върху филтрираните от match заявката артикули.