Софийски университет „Св. Кл. Охридски”

Факултет по математика и информатика

Курсов Проект

на тема:

“Система на български с информация за гъби”

Студенти:

Цветелина Михайлова Пашовска, Ф.Н. 7MI3400669**,**

Вероника Марк Мюние, Ф.Н. 2MI3400578

Курс: „Извличане на информация“,

Учебна година: 2024/2025

Преподаватели: **проф. Иван Койчев, Димитър Димитров**

Декларация за липса плагиатство:

- Плагиатство е да използваш, идеи, мнение или работа на друг, като претендираш, че са твои. Това е форма на преписване.
- Тази курсова работа е моя, като всички изречения, илюстрации и програми от други хора са изрично цитирани.
- Тази курсова работа или нейна версия не са представени в друг университет или друга учебна институция.
- Разбирам, че ако се установи плагиатство в работата ми ще получа оценка “Слаб”.

13.02.25 г. Подпис на студента: ЦП, ВМ

Съдържание

Кодът на проекта е достъпен на адрес: https://github.com/ccvetanska/MushroomBot

1 Увод

Съществуват много видове гъби, които често трудно се разпознават без експертни знания. Това прави автоматизираната помощ полезна за любители миколози. Нашата цел е да създадем система, която разпознава гъба на база потребителско описание на нейни характеристики. Системата трябва също така да дава и кратко резюме на информацията за гъбата, която е отчела като най-близка спрямо отговорите на потребителя. Такъв инструмент би могъл да бъде много полезен за български потребители, тъй като е ориентиран конкретно към видовете гъби в страната ни.

2 Преглед на областта

Към момента, голяма част от любителите и професионалните миколози, използват приложения за разпознаване на гъби по снимка. Такива са например *iNaturalist* и *Picture Mushroom*. Някои от техните основни недостатъци са това, че не разглеждат едновременно всички важни белези за идентификация (цвят, форма, размер, ламели, пори, пънче и хабитат), тъй като е трудно всички те да бъдат обединени в една снимка. Според миколозите и гъбарите с повече опит, друг недостатък на приложенията за разпознаване по снимка е, че те често не са образователни за ползвателите. Смятаме, че система, базирана на текстово описание, е подходяща за образователна цел на начинаещите в сферата и стимулира потребителя да разгледа гъбата внимателно и да запомни белезите ѝ.

3 Проектиране

Разделихме задачата на няколко основни части. Първата от тях беше да извлечем данните, с които ще работим. Свързахме се със създателите на сайта *www.manatarka.org* и с тяхно позволение избрахме да го използваме като източник на информация. Следващата задача се състоеше в това да имплементираме чатбот, който задава въпроси на потребителя и дава предложение за гъба, която е най-близка до отговорите му. Избрахме да приложим и друг подход към тази задача, за да сравним резултатите. Трябваше да зададем идентични въпроси на потребителя, както в чатбота, но в този случай да използваме отговорите му като query в търсачка над същата база данни и отново да предложим най-близката гъба.

4 Реализация

4.1 Използвани технологии, платформи и библиотеки

За реализация на системата използвахме *Python* и библиотеките *Beautiful Soup* [5] (*Web scraping*), *pandas* (за работа със .csv файлове), *nltk* (предварителна обработка), *Stanza* [4]

(предварителна обработка на български език), *scikit-learn* (векторизация), *Elasticsearch* [6] (конфигуриране на търсачка) и *networkx* (резюмиране).

4.2 Web scraping (Цветелина)

За да извлечем данните от сайта www.manatarka.org, първо проучихме структурата му. Идентифицирахме страница, съдържаща индекс с хипервръзки към индивидуалните страници на всяка гъба. Индексната страница съдържаше списък на гъбите, организиран по техните български наименования. Тъй като някои гъби имат повече от едно наименование, беше необходимо предварително нормализиране и групиране на записите, за да се елиминират дублиращите се записи и да се получи набор от уникални екземпляри. Запазихме резултата в *.csv* файл. Използвайки този индекс като изходна точка, извлякохме и запазихме в отделен файл *HTML* съдържанието на съответните страници. Тази междинна стъпка ни даде възможност да експериментираме с по-нататъшната обработка без да се налага да правим нови заявки към сайта. Обработихме всяка отделна страница и, възползвайки се от приблизително сходната им структура, разделихме информацията по отделни характеристики. Запазихме я като масив от *JSON* обекти от този вид:


```
{
        "latinTitle": "Tricholoma aurantium",
        "bgName": "Оранжева есенна гъба",
        "bgAlias": "Оранжева есенна гъба.",
        "worldAlias": "Orange knight (английски), Tricholome orangé (френски), Orangeroter Ritterling (немски), Рядовка золотистая (руски).",
        "edibility": "Неядлива гъба.",
        "cap": "Отначало изпъкнала, по-късно плоско-изпъкнала, често с широка, ниска гърбица в центъра. Ръждивооранжева до оранжево-кафява, понякога с присъстващи маслиненозелени нюанси. Повърхността е гладка до финолюспеста, суха или мазна в зависимост от времето. Ръбът отначало е силно подвит и остава подвит почти до края на развитието. Диаметър до 12 cm.",
        "body": "Няма информация",
        "stem": "Цилиндрично или стесняващо се в основата. Повърхността е покрита със зебровидна шарка с цвета на шапката в долната част на пънчето. В по-късата горна част е бяло. Двете зони са рязко разграничени с линия. Височина до 12 cm.",
        "flesh": "Бяло, дебело и твърдо. С брашняна миризма и слабогорчив вкус.",
        "gills": "Отначало бели, но скоро с ръждиви петна. Гъсти и прираснали със зъбче.",
        "tubes": "Няма информация",
        "pores": "Няма информация",
        "underside": "Ламели",
        "ring": "Няма информация",
        "spores": "Споровият прашец е бял. Спорите са с размери 4.5-6 x 3.5-4 µm.",
        "habitat": "Иглолистни гори, като образува микориза с борове (Pinus), смърч (Picea) или ела (Abies). Предпочита варовита почва. Плододава поединично или на групи. Сезон – от средата на лятото до есента. Рядък вид у нас.",
```

```
        "toxins": "Няма информация",
        "similarSpecies": "Tricholoma aurantium е лесен за разпознаване вид
благодарение на уникалната зебровидна шарка по пънчето. Срещат се други видове
Tricholoma с кафяви или ръждивокафяви шапки, но се открояват с различни шарки по
техните пънчета.",
        "images": [
            "http://manatarka.org/files/2018/08/Tricholomaaurantium1-300x225.jpg",
            "http://manatarka.org/files/2018/08/Tricholomaaurantium2-300x225.jpg",
            "http://manatarka.org/files/2018/08/Tricholomaaurantium3-300x225.jpg",
            "http://manatarka.org/files/2018/08/Tricholomaaurantium4-300x225.jpg"
        ],
       "fullDescription": "Tricholoma aurantium (Schaeff.)...”
}
```

Използвахме получения списък от 489 екземпляра за корпус от данни.

4.3 Чатбот (Вероника)

Тъй като информацията за всяка характеристика на гъбите, извлечени от сайта manatarka.org, са под формата на свободен текст, направихме предварителна обработка на текста, за да повишим точността на познаване. Нормализацията се състои основно от токенизация, лематизация и отделяне на носещите информация думи чрез *POS-tagging*.

За идентифицирането на гъба решихме да използваме *TF-IDF* векторизация. За целта, за всяка значима характеристика създадохме *TfidfVectorizer*, който обучихме върху всички налични данни в корпуса за съответната характеристика.

Работата на чатбота се състои в задаването на въпроси на потребителя относно различните такива характеристики. Събраната информация преминава през същия процес на предварителна обработка, чийто резултат е аналогичен нормализиран *JSON*. След това той се векторизира чрез вече обучените *TfidfVectorizer*-и за всяка характеристика.

Накрая, описанието на всяка гъба в корпуса се сравнява с това на потребителската гъба. Сравнението се състои в пресмятането на косинус-близостта между съответните вектори за всяка характеристика и взимане на средното аритметично. Разпознатата гъба е тази с най-голяма близост до входната.

4.4 Търсачка (Цветелина)

Стартирахме локален *ElasticSearch* сървър. Инициализирахме индекс със следната схема на полетата:

```
       "mappings": {
            "properties": {
                "habitat": {"type": "text"},
                "cap": {"type": "text"},
                "body": {"type": "text"},
```

```
            "stem": {"type": "text"},
            "flesh": {"type": "text"},
            "gills": {"type": "text"},
            "tubes": {"type": "text"},
            "pores": {"type": "text"},
            "underside": {"type": "keyword"},
            "ring": {"type": "text"},
            "spores": {"type": "text"},
        }
    }
```

Полето *"underside"* е от тип *"keyword"*, защото ще се използва за търсене на точно съвпадение. Във функцията за търсене, избрахме да добавим *"fuzziness":"AUTO"*, за да бъде търсачката толерантна към грешки.

Зададохме на потребителя въпроси за характеристиките на гъбата, която иска да разпознае, и използвахме отговорите като заявка към търсачката.

5 Провеждане на експерименти

За да сравним резултата от двата подхода, проведохме ръчно десет теста над всеки от тях, като използвахме едни и същи отговори на въпросите. Анализирахме броя на съвпадения между целевата гъба и разпознатата гъба при всеки от методите. Подходът с чатбот се представи добре на 6 от 10-те теста, а търсачката позна вярно на 8 от тестовете. На тази база, можем да кажем, че първият подход има 60% успеваемост, а вторият: 80%. За да получим по-надежда и точна оценка, е добре да бъдат проведени много повече тестове. Точната формулировка на тестовете е показана на ***Фигура 1*** и ***Фигура 2.***

6 Заключение

Смятаме, че успяхме да извлечем и структурираме една добра начална база данни за гъби в България. Двата подхода към задачата за разпознаване на гъби се справиха сравнително добре предвид естеството на задачата и недостатъчно големия корпус от данни за решаването ѝ.

За подобряване на точността на познаване на чатбота, може да се използва подход, който взима предвид семантичното значение на описанията, както в корпуса, така и в потребителското описание. За целта може да се разгледат подходи за *synonym expansion* и/или *sentence embeddings*.

За подобряване на точността на търсачката, можем да я разширим с речник на синоними.

Системата в момента работи само с описания на гъби на български език. Възможно е да се разшири корпусът от данни, така че да се поддържат повече видове гъби, както и описания на различни езици.

Към системата може да се добавят още - функционалност за разпознаване на гъба по снимка, както и възможност за идентифициране на гъбно отравяне и препоръки за конкретни действия в такава ситуация.

	въпроси	Тест №1	Тест №2	Тест №3	Тест №4
1	Какви са формата и размерът на шапката? Изпъкнала или вдлъбната е в средата? Набразден ли е ръбът ѝ?	изпъкнала и голяма	кръгла, кълбовидна, изпъкнала в средата	Кръгла. Ръбът ѝ е подвит навътре	кръгла, с издутина по средата
2	Какъв цвят е шапката отгоре?	бяла с кафяви петна	сив	светлокафяв, бежов	светлокафява
3	Какви са формата, размерът и цветът на плодното тяло(цялата гъба)?		голямо		малко
4	Как изглежда пънчето: цвят, височина, форма. Има ли в основата си (частта, която излиза от земята) нещо?	бяло с кафява шарка	жълто и червено, с мрежа	Бяло с липави бразди	тъничко и кухо
5	Какво има под шапката: ламели (ресни), пори или тръбички(дълбоки пори)?	ламели	пори	ламели	ламели
6	Как изглеждат ламелите: цвят, гъстота, сраснати ли са с пънчето или са отделени(свободни)? / Как изглеждат порите? Какъв цвят са и той променя ли се при нараняване или натиск? / Как изглеждат тръбичките? Какъв цвят са и той променя ли се при нараняване или натиск?	бели, свободни от пънчето	червени, посиняват при натиск	гъсти, свободни от пънчето, бледокафяви	бели, рески и свободни от пънчето
7	Има ли гъбата пръстен? Ако да, как изглежда: цвят, големина, вид? Отделя ли се лесно от пънчето?	пръстенът е голям и се отделя лесно от пънчето	не	не	не
8	Какъв е цветът на споровият прашец?	бял			
9	Какъв е цветът и консистенцията на месото? Има ли миризма? Променя ли цвета си при нараняване/отчупване?	бяло	бяло, посинява при отчупване	бяло	Белезникаво, мирише приятно
10	В каква среда намерихте тази гъба: гора или поляна? Какви дървета имаше наблизо?	акациева гора	широколистна гора	на поляна, в парка до блока ми	на полянка, в тревата
11	През кой сезон я намерихте?	есен	есен	зима, през декември	късна пролет

Целева гъба:	**Сърнела (Macrolepiota porcera)**	**Дяволска гъба (Rubroboletus satanas)**	**Ливадна виолетка (Lepista personata)**	**Обикновена челядинка (Marasmius oreades)**
Разпозната от Класическия MushroomBot гъба:	**Жълта мухоморка**	**Дяволска гъба (Rubroboletus satanas)**	**Ливадна виолетка (Lepista personata)**	**Сурогрижка (Amanita vaginata)**
Разпозната от Search MushroomBot гъба:	**Сърнела (Macrolepiota porcera)**	**Дяволска гъба (Rubroboletus satanas)**	**Ливадна виолетка (Lepista personata)**	**Обикновена челядинка (Marasmius oreades)**

Фигура 1: *Първите четири от проведените тестове, състоящи се от отговори на въпросите в колона "Въпроси", зададени към двата чат бота. На последните три реда се виждат съответно целевата гъба, разпознатата от класическия чат бот и разпознатата чрез query към търсачката. В зелено са маркирани правилно разпознатите гъби, а в червено - сгрешените*

Тест №5	Тест №6	Тест №7	Тест №8	Тест №9	Тест №10
с несиметрична форма	обла, изпъкнала	полукълбовидна	изпъкнала	кълбовидна	кръгла
сива	тъмнокафява, почти черна	бяла	ярко червена, с бели **издатинки** по нея	червен на бели **точки**	червена с бели **петна**
средноголямо				средно голяма, около 20см	
Пънчето е много късо, почти не се вижда	Много разширено в основата и тясно горе	гладко, кухо и удебелено в основата	бяло	Цилиндрично, бяло	бяло, високо, има ципа в основата си
ламели	тръбички	ламели	ламели	ламели	ламели
Бели, редки и много сраснати с пънчето	бели	розови, гъсти и свободни от пънчето	не са сраснати, бели са на цвят	бели и свободни от пънчето	бели
не	не	да, широк	Да, бял	да, голям, бял и парцалив	да, висящ и не се отделя от пънчето лесно
бяло и меко	бяло	бялезникаво, мирише на анасон	не видях	бяло	бяло
Беше в голяма група с още гъби, върху стеблото на мъртво дърво	беше до дъб	в овощна градина	в гора	в борова гора	в иглолистна гора, до смърч
зима	есен	есен	лято	есен	лято

Обикновена кладница (Pleurotus ostreatus)	**Бронзова манатарка (Boletus aereus)**	**Ливадна печурка (Agaricus arvensis)**	**Червена мухоморка (Amanita muscaria)**	**Червена мухоморка (Amanita muscaria)**	**Червена мухоморка (Amanita muscaria)**
Обикновена кладница (Pleurotus ostreatus)	**Бронзова манатарка (Boletus aereus)**	**Ливадна печурка (Agaricus arvensis)**	**Сиво-розова зърненка (Cystoderma carcharias)**	**Червена мухоморка (Amanita muscaria)**	**Яйцевидна мухоморка (Amanita ovoidea)**
Обикновена кладница (Pleurotus ostreatus)	**Бронзова манатарка (Boletus aereus)**	**Ливадна печурка (Agaricus arvensis)**	**Красива гълъбка, (Russula rosea)**	**Червена мухоморка (Amanita muscaria)**	**Сърнела, (Macrolepiota procera)**

***Фигура 2**: Останалите шест теста. Тест №8, Тест №9 и Тест №10 са с една и съща целева гъба: Червена мухоморка, и демонстрират как в някои случаи използването на различни думи (синоними) за описание на гъбите, влошава успеха и на двата метода. Наличието на няколко различни описания на една и съща гъба в корпуса от данни би помогнало срещу този проблем. Речник на синоними също би бил ефективен.*

7 Използвана литература

[1] “Извличане на информация”, Записки към курса по „Извличане на информация“ четен от Иван Койчев

[2] Sam Campbell, Natural Language Processing (NLP) for Beginners: Comprehensive Guide to Understanding language with Machines

[3] Манатарка.org - сайт за гъбите в България, www.manatarka.org

[4] Stanford NLP Group, “Stanza: A Python NLP Library for Many Human Languages.”, https://stanfordnlp.github.io/stanza/

[5] Beautiful Soup: a library that makes it easy to scrape information from web pages, https://pypi.org/project/beautifulsoup4/

[6] Elasticsearch: the leading distributed, RESTful, open source search and analytics engine https://www.elastic.co/docs