

ВЪВЕДЕНИЕ В ИЗВЛИЧАНЕТО НА ИНФОРМАЦИЯ

или как работят търсачките



СУ, ФМИ
Катедра “Компютърна информатика”
Мария Матева

Съдържание

- 1 Вашите водещи(на упражнения)
- 2 Организационни
- 3 Предизвикателство #1: Обратно инженерство на търсачката на Google
- 4 Понятия
- 5 Обща архитектура на търсачките
- 6 Булев модел
- 7 Предварителна обработка на текст
- 8 Обърнат индекс
- 9 Въпроси и отговори

- Мария Матева



- Ясен Кипров



Завършили са “Компютърни науки” и “Изкуствен интелект” във
ФМИ, СУ.

Имат опит в изграждането на специализирани търсачки, както
и в задачи от машинното самообучение.

Мария в момента работи върху софтуер за аналитични
решения.

Ясен е докторант и има интереси в областта на обработката на
естествен език. (Той ще разкаже повече.)

Moodle

- Регистрирайте се в курса в moodle
<http://moodle.openfmi.net/>
- Курсове » Магистри, зимен семестър 2014/2015 » ИИОЗ
» Извличане на информация 14/15
- IR_2014!
- там ще намерите тази презентация! ;)
- както и всички други материали
- във форума ще поддържахме актуална организационна информация
- можете да задавате всякакви въпроси там

Провеждане

- Упражненията биват **теоретични и практически**
- Ще бъдат в **понеделник, 19:30 - 21:00, бл.2 на СУ**
- Теоретичните упражнения ще бъдат в зала "С"
- Практическите упражнения ще бъдат в зали "30X" -
очаквайте подробности
- Ще ви осведомяваме във форума своевременно
- За практическите упражнения можете да си носите собствени лаптопи
- За теоретичните упражнения е добре да си носите нещо за писане

Софтуер, с който ще работим

На практическите упражнения ще се запознаем с:

- GATE 8, <https://gate.ac.uk>
- Apache Lucene, <http://lucene.apache.org/core/>
- Apache Solr, <http://lucene.apache.org/solr/>
- Apache Nutch, <http://nutch.apache.org> и други кролери
- други

Оценяване

Крайното оценяване се базира на

- проект
- два теста
- домашни(по желание)
- участие в упражнения(бонуси)
- предизвикателства(бонуси)

Предизвикателство #1: Обратно инженерство на търсачката на Google

Обратно инженерство (на английски: Reverse engineering) е процесът по преоткриване и възстановяване (реконструиране) на технологичните принципи и механизми, по които е създаден определен обект, машина или програма.

Как работи търсачката на Google?

- Задачата е да се намерят интересни/неочаквани поведения в търсачката, които доказват конкретни характеристики на работата ѝ.
- Добрите предложения ще получат бонус :)



- Ще повторим предизвикателството накрая на семестъра.

Предизвикателството - пример: домат, БГ

домат

Искания

Изображения

Видеоклипове

Новини

Още

Инструменти за търсене

Около 565 000 резултата (0,25 секунди)

„Бисквити“ ни помагат да предоставяме услуги си. С използването им приемате употребата на „Бисквити“ от наша страна.

Научете повече

OK

Домат — Уикипедия

bg.wikipedia.org/wiki/Домат

Доматът (Solanum lycopersicum) е растение от семейство Картофови (Solanaceae). Родина на **домата** са Централна и Южна Америка, от Мексико до ...

Домати - апетитни, полезни и широко достъпни - BB-Team

www.bb-team.org > Хранене > Статии за хранене > Какво идем

12.11.2008 г. : Замисляли ли сте се върху кауза за плодове или зеленчуци са **доматите**? Много е просто! Ботанически **доматите** се класифицират като ...

Домати - Golvach.bg

golvach.bg > Продукти > Зеленчуци

Доматите са навърно сред най-разпространените, обичани и популярни зеленчуци в света. Според мнозина **доматите** всъщност са плод, но с оглед на ...

Манастирски ливади - Пицария Дон Домат | пица, бира и ...

dondomat.com/?zaveneni.php?z=12

... Бизнес без риск - Доктор Дон **Домат** ресторанти услуги - Контакти Връзка с нас > Начало > Нашите заведения > Дон **Домат** - Манастирски ливади ...

Домати градина Вискяр

domati.viskyar.com/

Вкусни, чисти, без изкуствени торове, **домати** от Градина Вискяр - всички цветове, форми, размери. **Домат** биологично съвсем. Семена, разсад. Онлайн ...

Looking for results in English?

Change to English

Оставане на български

Езикови настройки



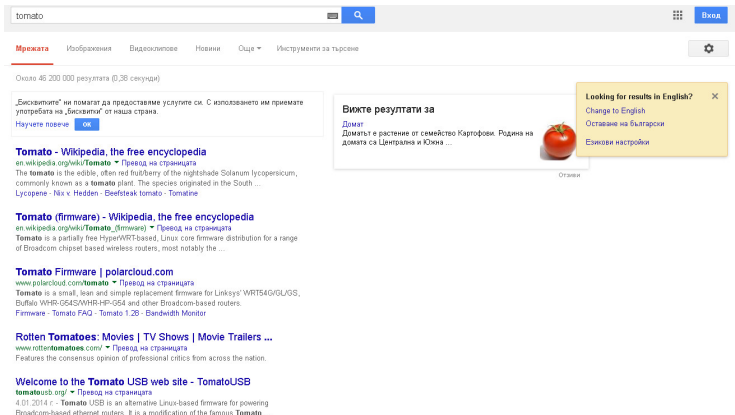
Още избор

Домат

Доматът е растение от семейство Картофови. Родина на домата са Централна и Южна Америка, от Мексико до Перу. [Уикипедия](#)

Отвори

Предизвикателството - пример: tomato, БГ



tomato

Мрежата Изображения Видеоклипове Новини Още Инструменти за търсене

Около 46 200 000 резултата (0,36 секунди)

„Бисквитките“ ни помагат да предоставим услугите си. С използването им приемате употребата на „Бисквитки“ от наша страна.
[Научете повече](#) [ОК](#)

Вижте резултати за

[Домат](#)
Доматът е растение от семейство Картофки. Родина на домата са Централна и Южна ...

Looking for results in English? [X](#)
[Change to English](#)
[Остава на български](#)
[Езикови настройки](#)

Tomato - Wikipedia, the free encyclopedia
<en.wikipedia.org/wiki/Tomato> [Превод на страницата](#)
The **tomato** is the edible, often red fruit/berry of the nightshade *Solanum lycopersicum*, commonly known as a **tomato** plant. The species originated in the South ...
[Lycopene](#) - [Nix v. Hedden](#) - [Beefsteak tomato](#) - [Tomatine](#)

Tomato (firmware) - Wikipedia, the free encyclopedia
[en.wikipedia.org/wiki/Tomato_\(firmware\)](en.wikipedia.org/wiki/Tomato_(firmware)) [Превод на страницата](#)
Tomato is a partially free HyperWRT-based, Linux core firmware distribution for a range of Broadcom chipset based wireless routers, most notably the ...

Tomato Firmware | polarcloud.com
www.polarcloud.com/tomato [Превод на страницата](#)
Tomato is a small, lean and simple replacement firmware for Linksys' WRT54G/GU/GS, Buffalo WHR-G54S/WHR-HP-G54 and other Broadcom-based routers.
[Firmware](#) - [Tomato FAQ](#) - [Tomato 1.2B](#) - [Bandwidth Monitor](#)

Rotten Tomatoes: Movies | TV Shows | Movie Trailers ...
www.rottentomatoes.com/ [Превод на страницата](#)
Features the consensus opinion of professional critics from across the nation.

Welcome to the Tomato USB web site - TomatoUSB
tomatousb.org/ [Превод на страницата](#)
4.01.2014 г. - **Tomato USB** is an alternative Linux-based firmware for powering Broadcom-based ethernet routers. It is a modification of the famous **Tomato** ...

Предизвикателството - пример: домати, UK

[Web](#)
[Images](#)
[Videos](#)
[Shopping](#)
[News](#)
[More ▾](#)
[Search tools](#)

About 401,000 results (0.43 seconds)

Домат — Уикипедия
bg.wikipedia.org/wiki/Домат ▾ [Translate this page](#)
 Доматът (Solanum lycopersicum) е растение от семейство Картофови (Solanaceae).
 Родина на домата са Централна и Южна Америка, от Мексико до ...

Манастирски ливади - Пицария Дон Домат | пица, бира и ...
dondomat.com/zavedenia.php?z=12 ▾ [Translate this page](#)
 ... бизнес без риск - Доктор Дон Домат ресторантски услуги - Контакти Връзка с нас. » Начало » Нашите заведения » Дон Домат - Манастирски ливади ...

Дондуков - Пицария Дон Домат | пица, бира и добро ...
dondomat.com/zavedenia.php?z=9 ▾ [Translate this page](#)
 Всяка събота и неделя през месец февруари Дон Домат - Шести Септември и Дон Домат - Дондуков Ви предлагат всички малки пици с 20% отстъпка.

Дон Домат - Манастирски ливади - Sofia, Bulgaria ...
<https://en-gb.facebook.com/DonDomatManLvadi> ▾ [Translate this page](#)
 Дон Домат - Манастирски ливади, Sofia, Bulgaria. 218 харесвания · 19 говорят за това · 263 бяха тук. Ресторант Дон Домат отвори врати в кв.

Дон Домат - Дондуков - Пицария | Фейсбук | Facebook
https://en-gb.facebook.com/.domat_/41234876218... ▾ [Translate this page](#)
 Дон Домат - Дондуков. 867 харесвания · 30 говорят за това · 271 бяха тук. Верига пицарии Дон Домат. Пица, бира и добро настроение. Най-добрата ...

Дон Домат - Шести Септември - Sofia, Bulgaria - Пицария ...
https://en-gb.facebook.com/.domat_/43526160655... ▾ [Translate this page](#)
 ... Шести Септември, Sofia, Bulgaria. 368 харесвания · 9 говорят за това · 318 бяха тук. Верига пицарии Дон Домат. Пица, бира и добро настроение....

Домат - Bonduelle България
bonduelle.bg/hranene/enciklopedia-na.../domat/ ▾ [Translate this page](#)
 Доматът е вкусен и полезен зеленчук, който съдържа ценни балансни вещества, а също калий, калий, фосфор и много витамини А, В, С и Е.



Tomato
 Vegetable

The tomato is the edible, often red fruit/berry of the nightshade *Solanum lycopersicum*, commonly known as a tomato plant. [Wikipedia](#)

Nutrition Facts
 Tomatoes, red ▾

Amount Per 100 grams ▾

Calories 18

	% Daily Value*
Total Fat 0.2 g	0%
Saturated fat 0 g	0%
Polyunsaturated fat 0.1 g	
Monounsaturated fat 0 g	
Cholesterol 0 mg	0%
Sodium 5 mg	0%
Potassium 237 mg	6%
Total Carbohydrate 3.9 g	1%
Dietary fiber 1.2 g	4%

Предизвикателството - пример: tomato, UK

[Web](#)
[Images](#)
[Shopping](#)
[Videos](#)
[News](#)
[More ▾](#)
[Search tools](#)

About 152,000,000 results (0.41 seconds)

Tomato - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Tomato ▾
 The tomato is the edible, often red fruit/berry of the nightshade *Solanum lycopersicum*, commonly known as a **tomato plant**. The species originated in the South ...
[Lycopene](#) - [Nix v. Hedden](#) - [Beefsteak tomato](#) - [Tomatine](#)

tomato
www.tomato.co.uk/ ▾
 tomato is a collective of artists, designers, musicians and writers engaged in publishing, exhibitions, live performances, advertising, architecture, fashion, public ...

Tomatoes - The World's Healthiest Foods
www.whfoods.com/genpage.php?tname=foodspice&dbid=44 ▾
 What's New and Beneficial About Tomatoes. Did you know that tomatoes do not have to be a deep red color to be an outstanding source of lycopene? Lycopene ...

Tomato | BBC Good Food
www.bbcgoodfood.com/glossary/tomato ▾
 The British tomato season runs from June to October. In winter, you could use more canned tomatoes to save on food miles (the environmental cost of food ...

News for tomato



Rotten Tomatoes: Farmers Pay the Price for a False Food ...
[Businessweek](#) - 7 hours ago
 Graves Williams, a farmer in Quincy, Fla., was just a few days into a six-week tomato harvest in June 2008, when the U.S. Food and Drug ...

[Heinz unveils new Tomato Ketchup TV ad](#)
[Talking Retail](#) - 2 hours ago

[Recipe: Ricotta and Tomato Jam Jar \(9.29.14\)](#)
[fox13now.com](#) - 17 minutes ago



Tomato
 Vegetable
 The tomato is the edible, often red fruit/berry of the nightshade *Solanum lycopersicum*, commonly known as a tomato plant. [Wikipedia](#)

Nutrition Facts
 Tomatoes, red ▾

Amount Per 100 grams ▾	
Calories 18	
	% Daily Value*
Total Fat 0.2 g	0%
Saturated fat 0 g	0%
Polyunsaturated fat 0.1 g	
Monounsaturated fat 0 g	
Cholesterol 0 mg	0%
Sodium 5 mg	0%
Potassium 237 mg	6%
Total Carbohydrate 3.9 g	1%
Dietary fiber 1.2 g	4%

Мрежата от близки/свързани понятия

Извличане на информация(Information retrieval)

представлява придобиването на (информационни) ресурси, които отговарят на дадена информационна нужда, избрани измежду колекция от такива ресурси. Търсенето се основава на съдържанието на ресурсите или на техните метаданни.

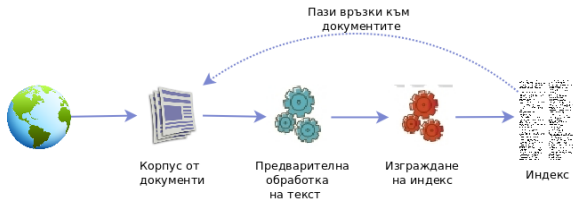
- Изкуствен интелект(Artificial Intelligence)
- Машинно самообучение(Machine Learning)
- Обработка на естествен език(Natural Language Processing)
- Анализ на текст(Text analysis; Text Mining)
- Извличане на закономерности от данни(Data mining)
- Анализ на данни(Data Analysis)
- Структурирано извличане на информация(Information Extraction)
- Извличане на значещи обекти(Named-entity Recognition, Semantic Annotation)
- Препоръчващи системи(Recommendation Systems), и още,

Важни понятия

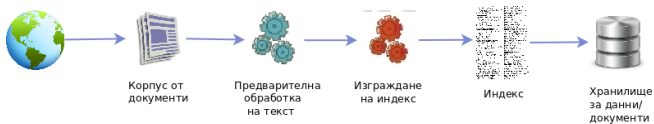
Важни (уводни) понятия в извличането на информация

- документ
- корпус
- информационна нужда
- заявка
- индекс
- резултат
- метрики за оценка на точност: precision, recall, accuracy, F1

Индексиране



ИЛИ



Извличане



Булев модел

Направете матрица на инцидентност на следните документи и думите в тях:

- $d1 = \{ \text{Иван затвори вратата.} \}$
- $d2 = \{ \text{Петър погледна навън.} \}$
- $d3 = \{ \text{Иван и Петър излязоха навън през вратата.} \}$

Изпълнете заявките: “Иван”, “навън”, “Петър” OR “вратата”,
“Петър” AND “вратата”

Основни стъпки в обработката на текст:

- определяне какво е документ
- кодировка на текста(encoding)
- определяне на езика на текста
- определяне на азбуката на писане(кирилица или шльокавица?)
- лексикален анализ - извличане на наредено множество от части на речта
- токенизация - отделяне за думи (и фрази), които по-нататък считаме за единици при анализа
- премахване на стоп-думи
- определяне на части на речта
- стеминг - определяне на корен на думата чрез прости евристични правила
- лематизация - определяне на корен на думата чрез по-сложни, морфо-синтактични правила
- синтактичен анализ - намиране на връзки на подчинение между отделните части на речта в рамките на изречението

Обърнат индекс

След първоначална обработка на текста:

- за всяка дума правим списък от документите, в които се среща
- първо обработваме всеки документ - създаваме двойки дума - ID на документ
- сортираме по думи(азбучен ред)
- сливаме повторенията, формирайки списък с ID-та на документите, съдържащи дадена дума
- докато сливаме повторенията, пддържаме списъка със съответстващите документи

Задача

Прилагайки стеминг, Съставете обърнат индекс на следните документи и думите в тях:

- $d1 = \{ \text{Иван затвори вратата.} \}$
- $d2 = \{ \text{Иван затвори една врата.} \}$
- $d3 = \{ \text{Иван излезе навън.} \}$
- $d4 = \{ \text{Петър погледна навън.} \}$
- $d5 = \{ \text{Иван и Петър излязоха навън през вратата.} \}$

Операции върху множества: обединение, разлика, сечение
Пропускащи списъци(skip-lists) и операции при голяма разлика
в дължината на списъците

Въпроси?

Следващия път...

- **Практическо упражнение**
- Ще говорим за предварителната обработка на текст
- И GATE. Моля, изтеглете си GATE оттук:
<https://gate.ac.uk/download/>

Благодаря за вниманието! :)