ВЪВЕДЕНИЕ В ИЗВЛИЧАНЕТО НА ИНФОРМАЦИЯ

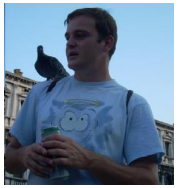
или как работят търсачките

СУ, ФМИ
Катедра “Компютърна информатика”
Мария Матева

Съдържание

- 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 поддържаме 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 “30Х” - *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

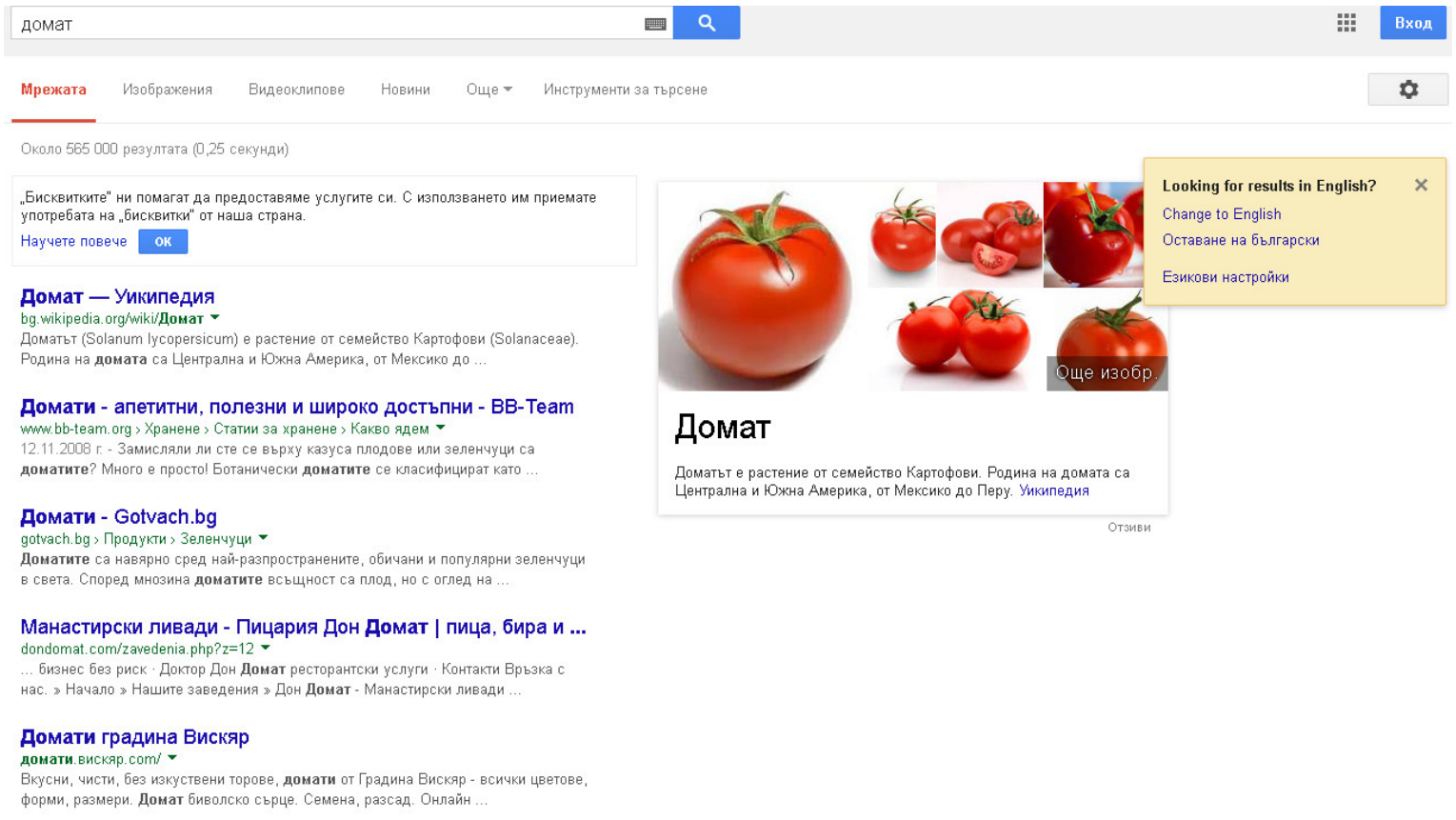
Обратно инженерство (на английски: Reverse engineering) е процесът по преоткриване и възстановяване (реконструиране) на технологичните принципи и механизми, по които е създаден определен обект, машина или програма.

Как работи търсачката на Google?

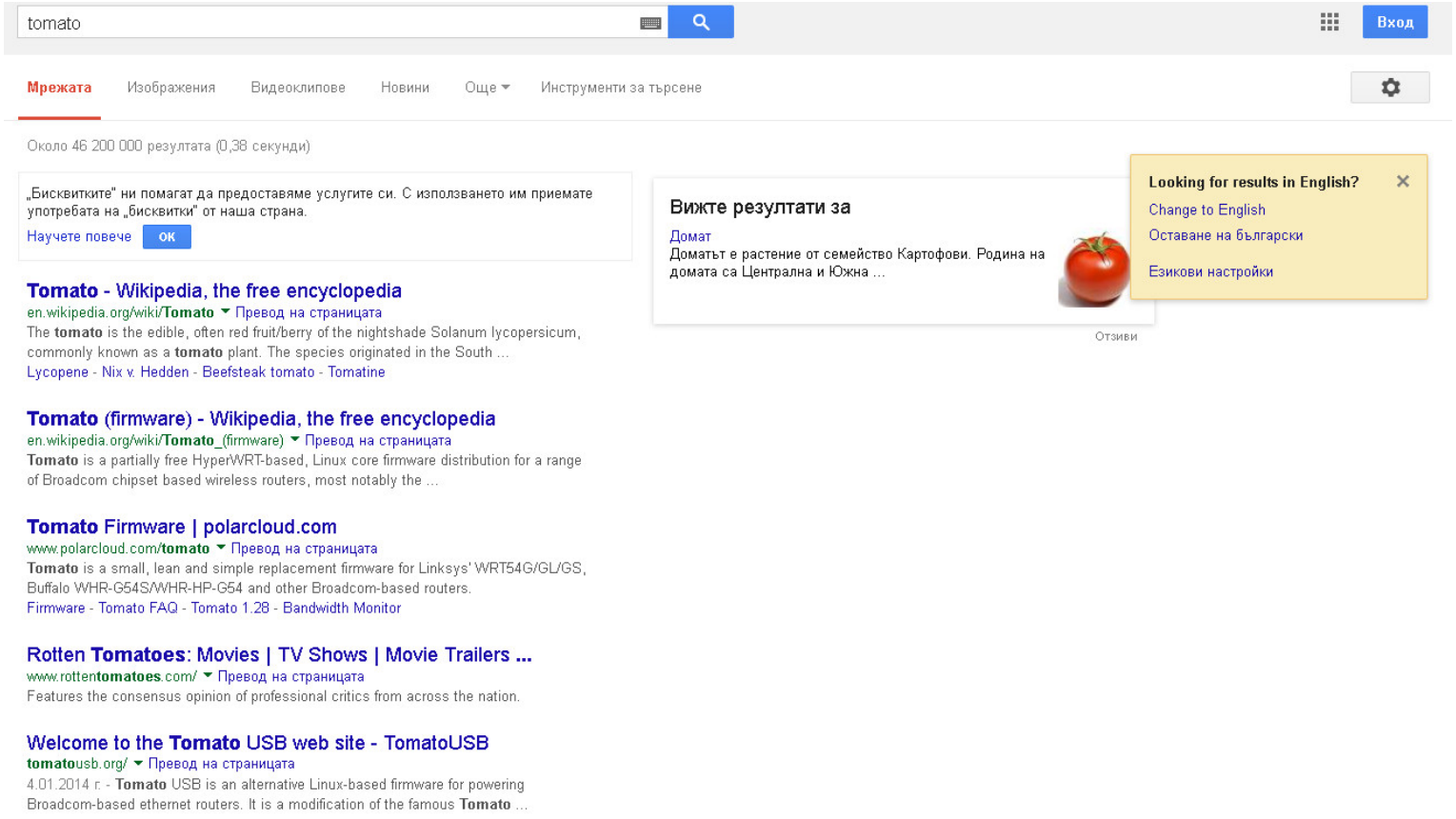
- Задачата е **да се намерят интересни/неочаквани поведения в търсачката, които доказват конкретни характеристики на работата ѝ.**
- Добрите предложения ще получат бонус :)

- Ще повторим предизвикателството накрая на семестъра.

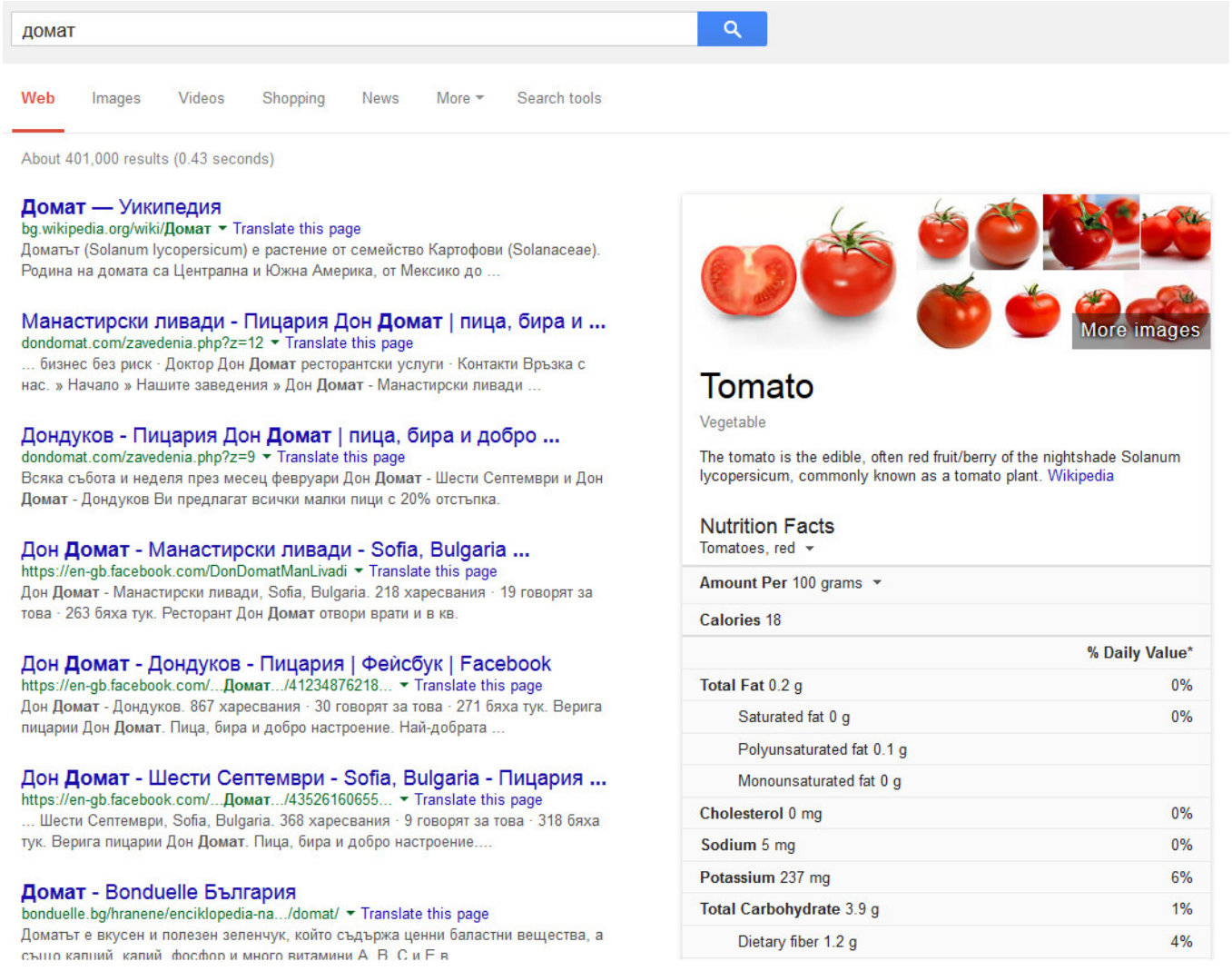
Предизвикателството - пример: домат, БГ



Предизвикателството - пример: tomato, БГ



Предизвикателството - пример: домат, UK



Предизвикателството - пример: tomato, UK

tomato

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 даденна информационна нужда, избрани измежду колекция от такива ресурси. Търсенето се основава на съдържанието на ресурсите или на техните метаданни.

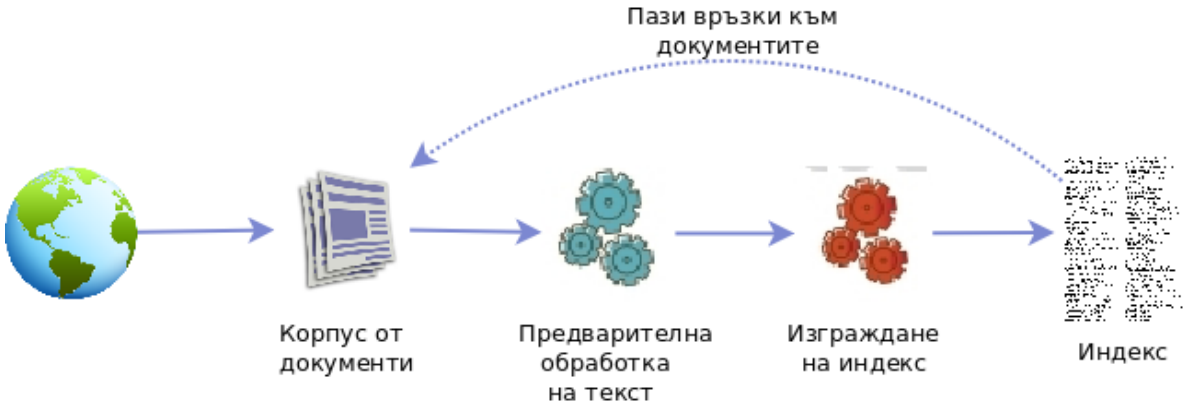
- Изкуствен интелект(Aritificial Intelligence)
- Машинно самообучение(Machine Learning)
- Обработка на естествен език(Natural Language Processing)
- Анализ на текст(Text analysis; Text Mining)
- Извличане на закономерности от данни(Data mining)
- Анализ на данни(Data Analysis)
- Структурирано извличане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

Важни понятия

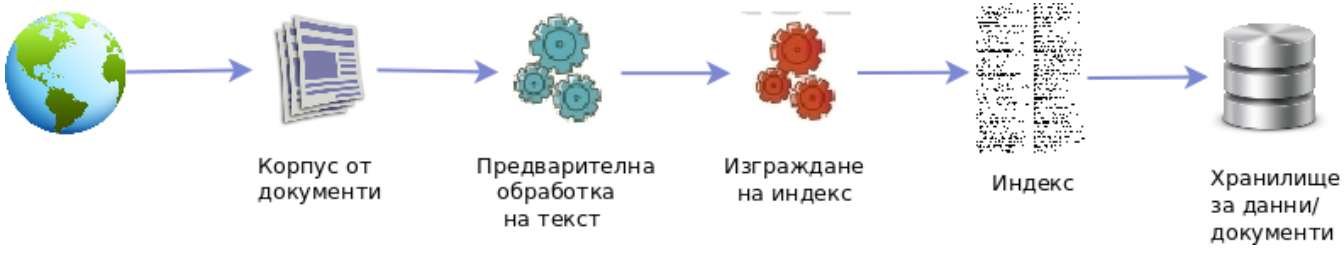
Важни (уводни) понятия в извличането на информация

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

Индексиране



или



Извличане


Пази връзки към документите
Потребители с информационна нужда
Заявка
Предварителна обработка на заявката
Търсене в индекс
Индекс
Резултатни документи
Хранилище за данни/ документи

Булев модел

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

- d1 = { Иван затвори вратата. }
- d2 = { Петър погледна навън. }
- d3 = { 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 докуемент
- 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, пъддържаме списъка със съответстващите документи

Задачка

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

- d1 = { Иван затвори вратата. }
- d2 = { Иван затвори една врата. }
- d3 = { Иван излезе навън. }
- d4 = { Петър погледна навън. }
- d5 = { 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! :)