Lab 1: Crawling, indexation and webpage research

During this web mining laboratory, we explored the website crawling with Crawler4j and its indexation with Solr via jsoup, which is HTML parser. We finally used the indexation to perform searches.

We used docker-solr as docker image for our Solr.

1 Crawler

1.1 Core

We started by creating a core for Solr, named *core_one* using the following commands:

```
$ docker exec -u 0 -it docker-solr bash to attach to the docker bash
```

then

```
$ bin/solr create -c core_one to create a default core.properties
```

As we set *update.autoCreateFields* to **false** in our core. We had to create two custom fields: *doc_title_en* and *doc_body_en*, which will be used to store the title and the body of each retained pages by the crawler.

1.2 Crawler4j

We configured our crawler, **MyCrawler.java** to work with our Solr core *core_one*. Concerning our first crawler configuration:

- starting page: wikipedia.org at "Veganism" page
- domain limitation: yes
- maximum pages to fetch: 70
- maximum deepness: 3
- politeness delay: 500
- https: yes
- FILTERS: custom binary files

1.2.1 shouldVisit function

We are applying the **FILTERS** pattern matching on the URL and verify that the domain.

1.2.2 visit function

The crawler retrieves the HTML page and parses them with jSoup. It creates a Solr document with the id field (hashcode of the page), the *doc_title_en* (title from the page), and *doc_body_en* (body content from the page) and finally adds the document into the current Solr instance.
To avoid Solr overloads, we set a loader for the commits. Indeed, the program will stock 50 documents before committing them to the Solr as a batch.
Each visited page has its content indexed by Solr.

1.3 Tries and fails

- We first indexed **Vegan.com**, but the pages were not meaningful from a feature point of view, indeed it had no categories.
- We then indexed **arxiv.org**, but their *"robots.txt"* policy was not allowing us to go anywhere.

2 Specific Indexation

As a continuation of our work on the crawler, and a starting point for a more advanced specification, we duplicated our class **MyCrawler** into **MyCrawler2**. Indeed, the purpose here is to upgrade our crawler to gather more meaningful information.

In a will to make our indexation more performant, we increased the page limit to 2000 and removed the deepness limit.
We are parsing the following elements:

- en_doc_title: The title of the page (Fig. 1)
- en_doc_body: The body of the page, its content
- en_doc_categories: The categories of the page (Fig. 2)
- en_doc_topics: The topics of the page, the <h3> (Fig. 3)
- en_doc_infobox: The infobox of the page (Fig. 4)
- en_doc_language: The language of the page
- en_doc_navigations: The navigation of the page (Fig. 5)
- en_doc_url: The URL of the page

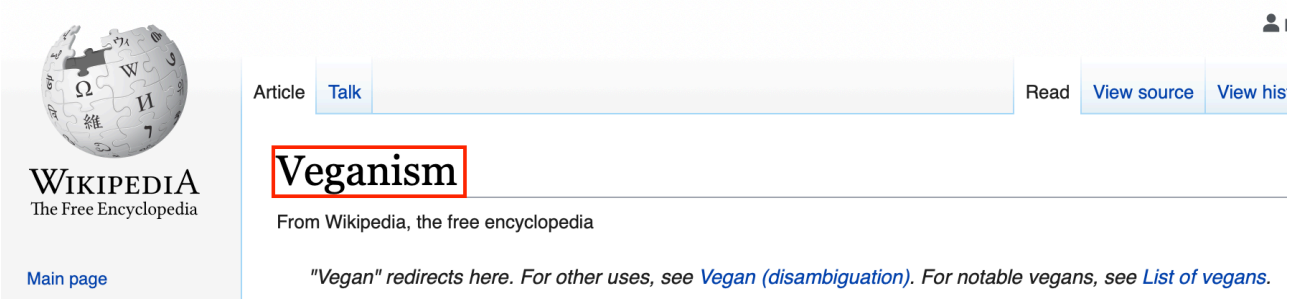

Figure 1: Wikipedia Title

Figure 2: Wikipedia Categories

Origins

Further information: History of vegetarianism

Vegetarian etymology

The term "vegetarian" has been in use since around 18
on scientific linguistic principles explain its origin as an
humanitarian).[38] The earliest-known written use is attri
plantation in 1838–1839.[i]

History

The practice can be traced to Indus Valley Civilization ir
Pakistan.[44] Early vegetarians included Indian philosop
Chandragupta Maurya and Ashoka; Greek philosopher
playwright Seneca the Younger.[45][46] The Greek sage
is disputed whether he ever advocated any form of veg
woolen garments.[49] Eudoxus of Cnidus, a student of A
killers that he not only abstained from animal foods, but
al-Maʿarri (c. 973 – c. 1057).[a][50] Their arguments were
Abstinentia ab Esu Animalium ("On Abstinence from Ar

Vegetarianism established itself as a significant movem
entirely.[52] In 1813, the poet Percy Bysshe Shelley pub
1815, William Lambe, a London physician, claimed that
a "habitual irritation", and argued that "milk eating and f
Graham's meatless Graham diet—mostly fruit, vegetab
1830s in the United States.[55] Several vegan communi
Louisa May Alcott, opened the Temple School in 1834 a
community at Alcott House on Ham Common, in 1838.[

Vegetarian Society

Further information: Vegetarian Society

Figure 3: Wikipedia Topics

Veganism

Clockwise from top-left:
Seitan pizza; roasted sprouts, tofu, and pasta; cocoa–avocado brownies; leek-and-bean cassoulet with dumplings.

Pronunciation	/ˈviːgənɪzəm/ *VEE-gə-niz-əm* Vegan /ˈviːgən/ *VEE-gən*
Description	Elimination of the use of animal products, particularly in diet
Earliest proponents	Al-Maʿarri (c. 973 – c. 1057)[a] Roger Crab (1621–1680)[2] Johann Conrad Beissel (1691–1768)[3] James Pierrepont Greaves (1777–1842)[4] Amos Bronson Alcott (1799–1888)[5] Sarah Bernhardt (1844–1923)[6] Donald Watson (1910–2005)[7]
Term coined by	Donald Watson, November 1944[8]
Notable vegans	List of vegans

Figure 4: Wikipedia Infobox

v·t·e	Veganism and vegetarianism	[hide]
Perspectives	Veganism	Animal-free agriculture · Fruitarianism · History · Juice fasting · Low-carbon diet · Raw veganism · Nutrition · Vegan organic gardening · Vegan studies
	Vegetarianism	Economic vegetarianism · Environmental vegetarianism · History · Lacto vegetarianism · Ovo vegetarianism · Ovo-lacto vegetarianism · Cuisine · Vegetarian Diet Pyramid · Ecofeminism · Nutrition · By country
	Lists	Vegans · Vegetarians · Vegetarian festivals · Vegetarian organizations · Vegetarian restaurants
Ethics	Secular	Animal rights · Animal welfare · Carnism · Deep ecology · Environmental vegetarianism · Ethics of eating meat · Meat paradox · Nonviolence · Sentientism · Speciesism · Tirukkural
	Religious	Buddhism · Christianity · Hinduism (Sattvic · Ahimsa) · Jainism · Judaism · Pythagoreanism · Rastafari · Sikhism
Food, drink		Agar · Agave nectar · Meat analogue (List of meat substitutes) · Miso · Mochi · Mock duck · Nutritional yeast · Plant cream · Plant milk · Quinoa · Quorn · Seitan · Soy yogurt · Tempeh · Tofu · Tofurkey · Cheese · Vegepet · Hot dog · Vegetarian mark · Sausage · Beer · Wine · Veggie burger
Groups and events	Vegan	American Vegan Society · Beauty Without Cruelty · Food Empowerment Project · *Go Vegan* · Movement for Compassionate Living · Physicians Committee for Responsible Medicine · Plamil Foods · Vegan Awareness Foundation · Vegan flag · Vegan Ireland · Vegan Outreach · Vegan Prisoners Support Group · The Vegan Society · Veganuary · Veganz · World Vegan Day
	Vegetarian	American Vegetarian Party · Boston Vegetarian Society · Christian Vegetarian Association · European Vegetarian Union · Happidog · Hare Krishna Food for Life · International Vegetarian Union · Jewish Veg · Meat-free days (Meatless Monday · Friday Fast) · Swissveg · Toronto Vegetarian Association · Vegetarian Society · Vegetarian Society (Singapore) · Veggie Pride · Viva! Health · World Esperantist Vegetarian Association · World Vegetarian Day
Companies		List of vegetarian and vegan companies
Books, reports		*Thirty-nine Reasons Why I Am a Vegetarian* (1903) · *The Benefits of Vegetarianism* (1927) · *Ten Talents* (1968) · *Diet for a Small Planet* (1971) · *Moosewood Cookbook* (1977) · *Fit for Life* (1985) · *Diet for a New America* (1987) · *The Sexual Politics of Meat* (1990) · *Vegetarian Cooking for Everyone* (1997) · *The China Study* (2004) · *Skinny Bitch* (2005) · *Livestock's Long Shadow* (2006) · *Eating Animals* (2009) · *Why We Love Dogs, Eat Pigs, and Wear Cows* (2009) · *Meat Atlas* (annual)
Films		*Meet Your Meat* (2002) · *Peaceable Kingdom* (2004) · *Earthlings* (2005) · *A Sacred Duty* (2007) · *Fat, Sick and Nearly Dead* (2010) · *Planeat* (2010) · *Forks Over Knives* (2011) · *Vegucated* (2011) · *Live and Let Live* (2013) · *Cowspiracy* (2014) · *What the Health* (2017) · *Carnage* (2017)
Magazines		*Naked Food* · *Vegetarian Times* · *VegNews*
Physicians, academics		Carol J. Adams · Neal D. Barnard · Rynn Berry · T. Colin Campbell · Caldwell Esselstyn · Gary L. Francione · Joel Fuhrman · Greta Gaard · Michael Greger · Melanie Joy · Michael Klaper · John A. McDougall · Reed Mangels · Jack Norris · Dean Ornish · Richard H. Schwartz · Laura Wright
Related		Semi-vegetarianism (Macrobiotic diet · Pescetarianism) · Vegetarian and vegan dog diet · Vegetarian and vegan symbolism
		Veganism portal · Vegetarianism portal

v·t·e	Animal rights	[show]

Figure 5: Wikipedia Navigation

3 Research

Based on our MyCrawler2, we have asked to index up to 2000 webpages. Using the Solr web interface, we found out that 1995 documents were indeed indexed.
Continuing with Solr web interface, we can also search with a specific query such as Lactose into a specific field we extracted via our crawler. For example: en_doc_body:Lactose.
We implemented a searcher, Searcher.java, which also returns the score for each returned document. Moreover, we are using a weighting system on each extracted features. Indeed, the meaningfulness of each feature is not equivalent. After some tweaking, we came out with the following formula, where <query> is the user's query. : q: $(\text{en\_doc\_title:<query>})^6(en\_doc\_topics :< query >)^5(en\_doc\_infobox :< query >)^4(en\_doc\_categories :< query >)^3(en\_doc\_navigations :< query >)^2(en\_doc\_body :< query >)^1$
We give a weight of 6 for the title, a weight of 5 for the topics and so one. The lower is the weight; the lower is the importance is.
The following are query examples:

4 Theorical questions

Please explain what strategy should be adopted for indexing pages in several languages (each page is composed of only one language, but the corpus includes pages in several languages). What should you watch out for? Please explain the process you propose.

We imagine different solutions to handle multiple languages with Solr:

- The first solution is to defined different schema fields for every language like "`title_fr`" or "`title_en`" and applying filters to each language. The downside of this solution is the high memory consumption and complexity when many languages are present.
- The second solution is to use a collection of same fields for all languages, add a field to store the langue and then apply a filter on it (e.g. `fq=language:english`). The downside of this solution is that we cannot use language specific features like lemmatisation and stemming.
- The third solution is to create a Solr core for each language and route the queries to the right core.

If we have few languages to querying we proposed to use the first solution but if we have a lot of languages we propose to use the third solution.

Solr allows by default to do a fuzzy search. Please explain what it is and how Solr implements it. Some first names may have a lot of spelling variations (eg Caitlin : Caitilin, Caitlen, Caitlinn, Caitlyn, Caitlyne, Caitlynn, Cateline, Catelinn, Catelyn, Catelynn, Catlain, Catlin, Catline, Catlyn, Catlynn, Kaitlin, Kaitlinn, Kaitlyn, Kaitlynn, Katelin, Katelyn, Katelynn, etc). Is it possible to use, while keeping a good performance, the fuzzy research made available by Solr to do research taking into account such variations? If not what alternative(s) do you see, please justify your answer.

The fuzzy search matches a term when the edit distance between the query and the document's term is under an arbitrary threshold. According to the JavaDoc, this threshold is by default at 2.

Yes, it is possible to use the fuzzy search and keeping good performance. Solr is based on Lucene and from its version 4 this library uses a Levenshtein Automaton, a deterministic automaton (DFA) that accepts only the terms within edit distance N. It is possible to compute this automaton of degree N for an input word W time linear in the length of W.

For more informations: `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.16.652`