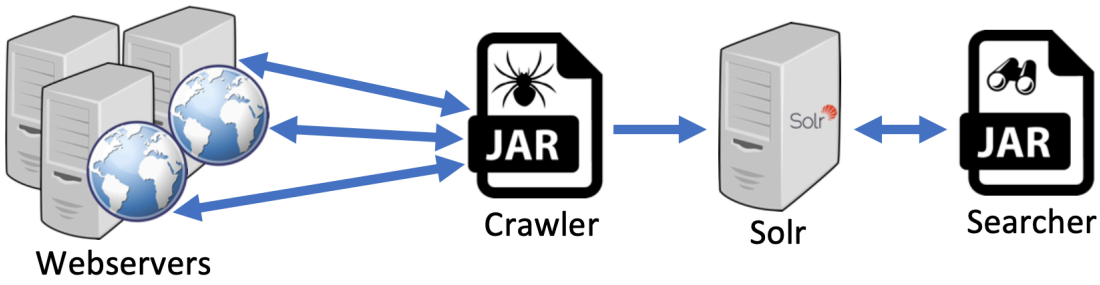
Lab 1: Crawling, indexation and webpage research

In this web mining laboratory, we will implement a crawler that visits a website and push the data to a Solr server. We will also create some fields for indexing interesting data. Then we will create a searcher in Java that query data from the Solr server. The architecture is represented below:



To implement the crawler, we used the library called Crawler4j and to parse the HTML we used Jsoup. These dependencies are automatically downloaded from Maven Central using Gradle. The following command runs our application:

```
$ gradle run
```

To simplify the installation, 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

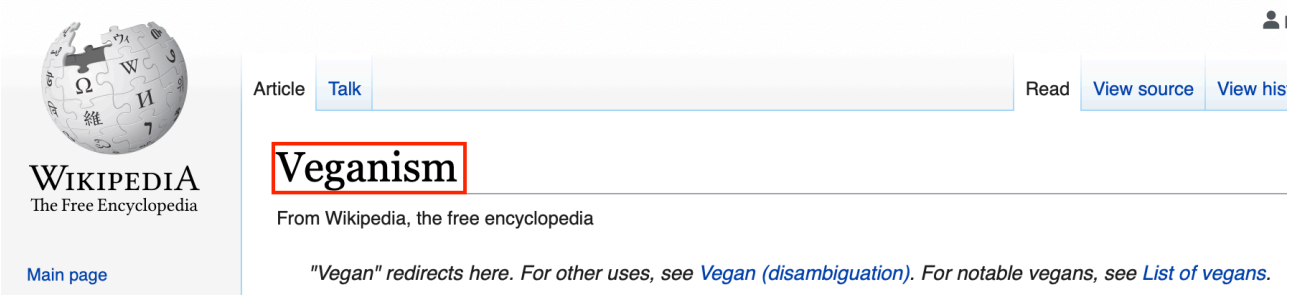


Figure 1: Wikipedia Title

Figure 2: Wikipedia Categories

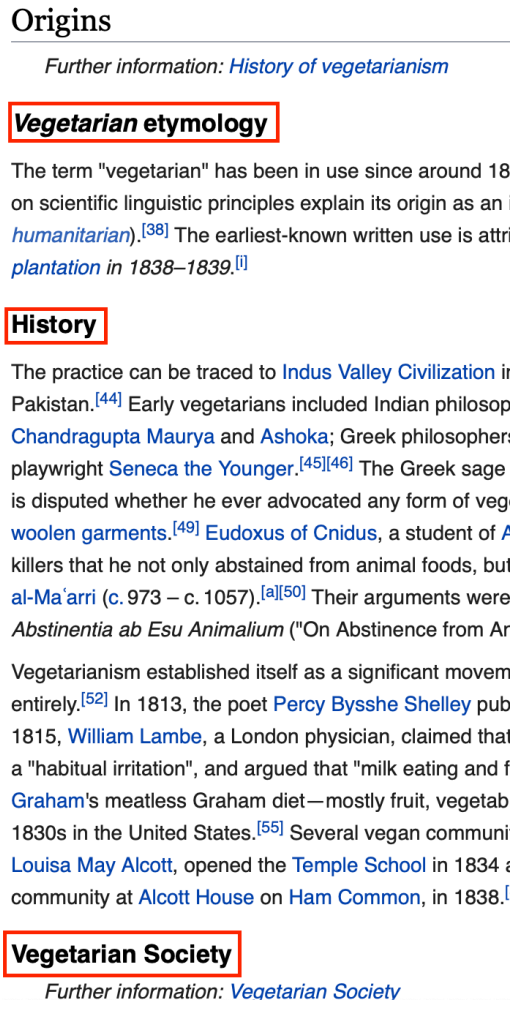


Figure 3: Wikipedia Topics

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: (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 gave 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:

```
Enter a query:
Meat
==================
QUERY: Meat
DOCUMENTS FOUND: 734
TOTAL INDEX: 1995
==================
title: [Meat Atlas]
score: 114.9512
url: [https://en.wikipedia.org/wiki/Meat_Atlas]

title: [Psychology of eating meat]
score: 83.08943
url: [https://en.wikipedia.org/wiki/Meat_paradox]

title: [Carol J. Adams]
score: 69.00029
url: [https://en.wikipedia.org/wiki/Carol_J._Adams]

title: [Carnism]
score: 68.9441
url: [https://en.wikipedia.org/wiki/Carnism]

title: [Ground meat]
score: 62.052734
url: [https://en.wikipedia.org/wiki/Ground_meat]

title: [Meat industry]
score: 59.226562
url: [https://en.wikipedia.org/wiki/Meat_industry]

title: [Meat analogue]
score: 54.775394
url: [https://en.wikipedia.org/wiki/Meat_analogue]

title: [Meat-free days]
score: 52.16121
url: [https://en.wikipedia.org/wiki/Meat-free_days]

title: [List of meat substitutes]
score: 46.795784
url: [https://en.wikipedia.org/wiki/List_of_meat_substitutes]

title: [Pork]
score: 44.41754
url: [https://en.wikipedia.org/wiki/Pork]
```

Figure 6: Search Query: meat

```
Enter a query:
Lactose
==================
QUERY: Lactose
DOCUMENTS FOUND: 58
TOTAL INDEX: 1995
==================
title: [Vegetarianism and beer]
score: 41.519524
url: [https://en.wikipedia.org/wiki/Vegetarianism_and_beer]

title: [Milk substitute]
score: 41.04914
url: [https://en.wikipedia.org/wiki/Milk_substitute]

title: [FODMAP]
score: 38.218586
url: [https://en.wikipedia.org/wiki/FODMAP]

title: [Sucrose]
score: 29.845587
url: [https://en.wikipedia.org/wiki/Sucrose]

title: [Probiotic]
score: 25.252743
url: [https://en.wikipedia.org/wiki/Probiotic]

title: [Date sugar]
score: 17.966122
url: [https://en.wikipedia.org/wiki/Date_sugar]

title: [Pine honey]
score: 17.824799
url: [https://en.wikipedia.org/wiki/Pine_honey]

title: [Maple sugar]
score: 17.481277
url: [https://en.wikipedia.org/wiki/Maple_sugar]

title: [Palm sugar]
score: 17.402197
url: [https://en.wikipedia.org/wiki/Palm_sugar]

title: [Soy formula]
score: 17.03336
url: [https://en.wikipedia.org/wiki/Soy-based_infant_formula]
```

Figure 7: Search Query: lactose

```
Enter a query:
dog food
==================
QUERY: dog food
DOCUMENTS FOUND: 12
TOTAL INDEX: 1995
==================
title: [Dog food]
score: 76.46058
url: [https://en.wikipedia.org/wiki/Dog_food]

title: [Vegepet]
score: 29.962826
url: [https://en.wikipedia.org/wiki/Vegepet]

title: [Happidog]
score: 28.393778
url: [https://en.wikipedia.org/wiki/Happidog]

title: [Vegetarian and vegan dog diet]
score: 3.5341976
url: [https://en.wikipedia.org/wiki/Vegetarian_and_vegan_dog_diet]

title: [Tallow]
score: 2.9409788
url: [https://en.wikipedia.org/wiki/Tallow]

title: [Animal product]
score: 2.3677258
url: [https://en.wikipedia.org/wiki/Animal_product]

title: [Methionine]
score: 2.3360817
url: [https://en.wikipedia.org/wiki/Methionine]

title: [Cat food]
score: 2.1195178
url: [https://en.wikipedia.org/wiki/Cat_food]

title: [Kelp]
score: 2.0879176
url: [https://en.wikipedia.org/wiki/Kelp]

title: [Cat health]
score: 1.8037103
url: [https://en.wikipedia.org/wiki/Cat_health]
```

Figure 8: Search Query: dog food

```
Enter a query:
School
==================
QUERY: School
DOCUMENTS FOUND: 697
TOTAL INDEX: 1995
==================
title: [School meal programs in the United States]
score: 75.741776
url: [https://en.wikipedia.org/wiki/School_meal_programs_in_the_United_States]

title: [National Film and Television School]
score: 46.44328
url: [https://en.wikipedia.org/wiki/National_Film_and_Television_School]

title: [National School Lunch Act]
score: 46.020695
url: [https://en.wikipedia.org/wiki/National_School_Lunch_Program]

title: [Temple School (Massachusetts)]
score: 35.03086
url: [https://en.wikipedia.org/wiki/Temple_School_(Massachusetts)]

title: [Alasdair Cochrane]
score: 34.837685
url: [https://en.wikipedia.org/wiki/Alasdair_Cochrane]

title: [Gary L. Francione]
score: 31.54172
url: [https://en.wikipedia.org/wiki/Gary_Francione]

title: [Gary L. Francione]
score: 31.54172
url: [https://en.wikipedia.org/wiki/Gary_L._Francione]

title: [Joel Fuhrman]
score: 30.111942
url: [https://en.wikipedia.org/wiki/Joel_Fuhrman]

title: [Alice Crary]
score: 29.522297
url: [https://en.wikipedia.org/wiki/Alice_Crary]

title: [Tom Beauchamp]
score: 29.506216
url: [https://en.wikipedia.org/wiki/Tom_Beauchamp]
```

Figure 9: Search Query: School

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 could 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 lemmatization and stemming.
- The third solution is to create a Solr core for each language and route the queries to the right core.
- A fourth solution would be to index specific subdomains with specific tags or cores, such as en.wikipedia.org, fr.wikipedia.org, de.wikipedia.org, etc.
- In a fifth solution, we could try to modify Crawler's header by asking a specific language.

If we have a few languages to querying, we recommend to use the first solution, but if we have many languages, we recommend the third or fourth 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 version 4, its 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.

Concerning the Caitlin example, by using the operator, we can specify a distance, for exemple aitlin 2. If it's still not enough, we can combine the fuzzy search with the OR operator: aitlin 2 OR Kaitlynn 2.
However, in most of the cases, stemming gives similar results to the fuzzy search, and we could also mention the lemmatization.

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