Acquisition of a Large Domain-specific Corpus for Fine-tuning of Language Models

Student: B.Sc. Tim Flegelskamp

Supervisor: Anastasia Zhukova

Date: 02.02.2022

Agenda

- Introduction
- Objective
- Wikidata structure
- Extracting information from wikidata
- Query expansion
- Conclusion

Introduction

- Design, implement, and evaluate a system that collects text data from a large open source database, given keywords of a specific domain

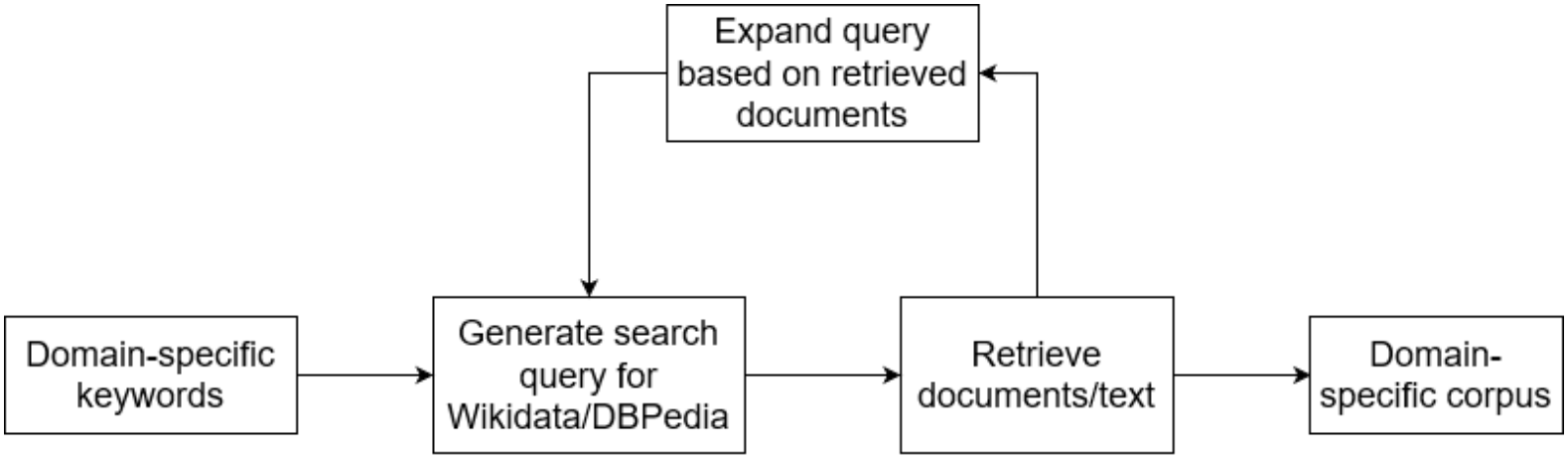
WIKIDATA

[1]

- Fine-tuning of one state-of-the-art language model is not part of the project anymore

Objective

- Design and implement an approach to collect the domain-specific data from Wikidata



Wikidata structure

- Wikidata organizes the data from different wiki projects:
  - Wikipedia
  - Wikibooks
  - Wikiquote

[2]

[3]

[4]

Wikidata structure

- Each entity in wikidata consists of:
  - Item label
  - Item identifier
  - Short description
  - Aliases

machine learning (Q2539)

scientific study of algorithms and statistical models that computer systems use to perform tasks without explicit instructions

edit

ML | statistical learning

In more languages

Configure

Language	Label	Description	Also known as
English	machine learning	scientific study of algorithms and statistical models that computer systems use to perform tasks without explicit instructions	ML statistical learning
German	maschinelles Lernen	Algorithmen zur Erkennung nützlicher Muster in Datenmengen	ML
French	apprentissage automatique	champ d'étude de l'intelligence artificielle	machine learning apprentissage statistique apprentissage machine
Bavarian	No label defined	No description defined	

All entered languages

Wikidata structure

Wikipedia (69 entries) edit

ar	تعلم الآلة
ary	تعلام ماكيني
as	যন্ত্র শিক্ষণ
az	Maşın öyrənməsi
bg	Машинно самообучение
bh	जंत्र शिक्षण
bn	যন্ত্রীয় শিখন
ca	Aprenentatge automàtic
ckb	فێربوونی مەکینە
cs	Strojové učení
cy	Dysgu peirianyddol
da	Maskinlæring
de	Maschinelles Lernen
el	Μηχανική μάθηση
en	Machine learning
es	Aprendizaje automático
et	Masinõppimine
eu	Ikasketa automatiko
fa	یادگیری ماشینی
fiu_vro	Massinoppus
fi	Koneoppiminen
fr	Apprentissage automatique
gl	Aprendizaxe automática

Wikibooks (1 entry) edit

it	Applicazioni pratiche di machine learning

Wikinews (0 entries) edit

Wikiquote (1 entry) edit

en	Machine learning

Wikisource (0 entries) edit

Wikiversity (1 entry) edit

en	Machine learning

Wikivoyage (0 entries) edit

Wiktionary (0 entries) edit

Multilingual sites (0 entries) edit

Wikidata structure

topic's main category	Category:Machine learning	edit
	0 references	
		+ add reference
		+ add value

Wikipedia (46 entries) edit

af	Kategorie:Masjienleer
ar	تصنيف:تعلم الآلة
az	Kateqoriya:Maşın öyrənməsi
be_x_old	Катэгорыя:Машыннае навучаньне
be	Катэгорыя:Машыннае навучанне
bn	বিষয়শ্রেণী:যান্ত্রিক শিখন
ca	Categoria:Aprenentatge automàtic
ckb	پۆل:فێربوونی مەکینە
cs	Kategorie:Strojové učení
de	Kategorie:Maschinelles Lernen
el	Κατηγορία:Μηχανική μάθηση
en	Category:Machine learning
es	Categoría:Aprendizaje automático
et	Kategooria:Masinõpe
eu	Kategoria:Ikasketa automatikoa
fa	رده:یادگیری ماشینی
fi	Luokka:Koneoppiminen
fr	Catégorie:Apprentissage automatique
he	קטגוריה:למידת מכונה
hy	Կատեգորիա:Մեքենայական ուսուցում
id	Kategori:Pemelajaran mesin

Wikibooks (0 entries) edit

Wikinews (0 entries) edit

Wikiquote (0 entries) edit

Wikisource (0 entries) edit

Wikiversity (0 entries) edit

Wikivoyage (0 entries) edit

Wiktionary (0 entries) edit

Multilingual sites (1 entry) edit

commons	Category:Machine learning

Extracting information from wikidata

- Support for different languages
- For extracting the links from the wiki projects two libraries are used:
  - Requests [5]
  - BeatifulSoup4 [6]
- Requests is used for HTTP-GET requests:
  - Generation of a random user-agent string [7]
    - Mozilla/5.0 (X11; Linux i686; rv:21.0) Gecko/20100101 Firefox/21.0
- Content (html) of the HTTP-GET request is used as input for BeautifulSoup4
  - Lxml parser parses the html of the website to the format for the library

Extracting information from wikidata

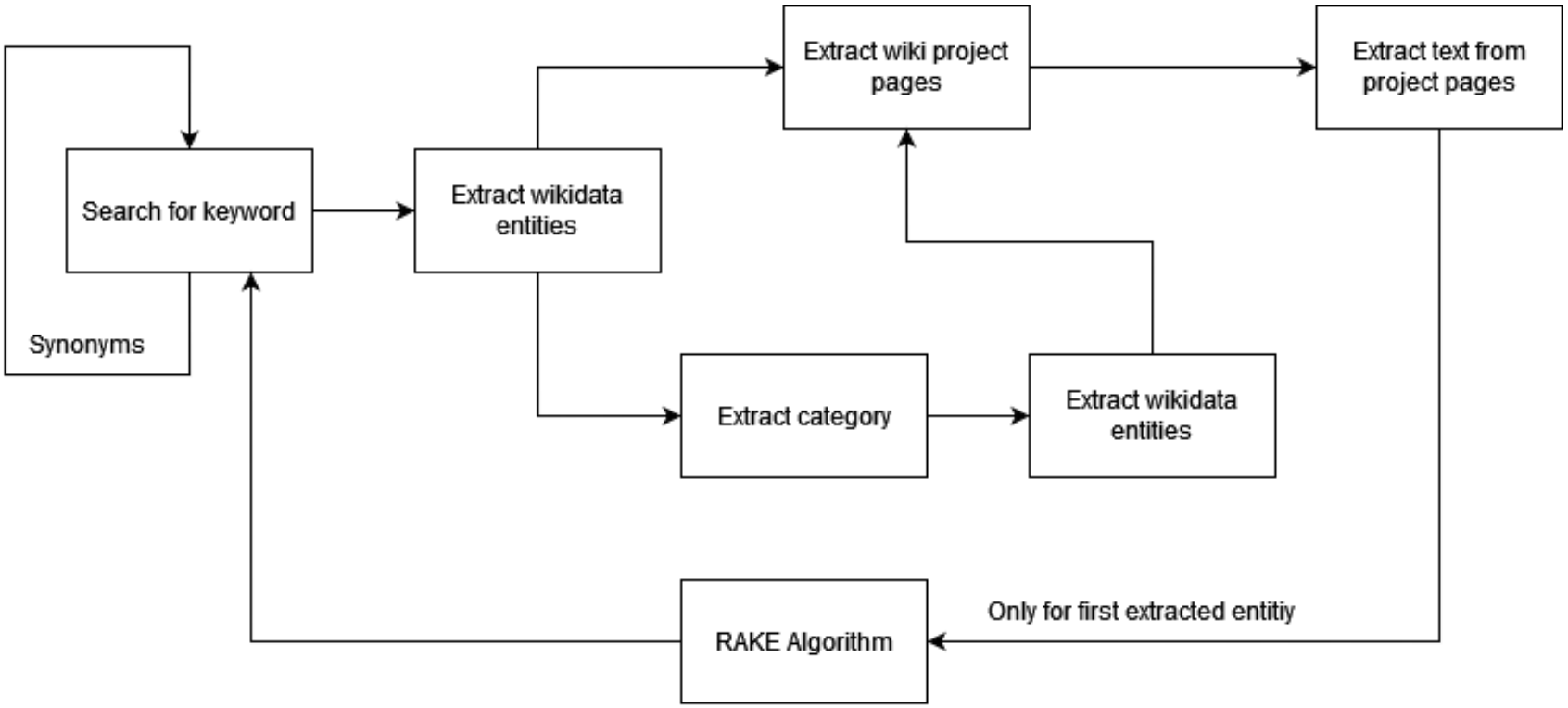
1. Start a search query with a keyword
2. Extract the different entities from wikidata
3. Extract the different wiki project pages and the main category from all entities if possible
   - Multithreading for speed up
4. Extract categories from the wiki pages
5. Extract all text from the pages from the categories
   - Multithreading for speed up
   - Preprocess the raw text and clean it
   - Save it to a text file
   - Speed: 3 Million characters in two minutes

Query expansion

- Pseudo relevance feedback:
  - Consider the first extracted entity on wikidata as the most important
  - Extract the text and use the RAKE (Rapid Automatic Keyword Extraction Algorithm) to extract important keywords from the text
  - Use the new keywords to start a new search query

- Thesaurus based:
  - Use a source to get most similar words/synonyms for a keyword
  - Start a new search query with the synonyms/most similar words
  - FastText multilingual model is used with gensim to retrieve those words

Query expansion



Conclusion

- Extraction process works well
- Multithreading is necessary
- A lot of entities do not have a category assigned
  - Use of Wikipedia categories is necessary

References

[1] https://de.wikipedia.org/wiki/Wikidata#/media/Datei:Wikidata-logo-en.svg

[3] https://upload.wikimedia.org/wikipedia/commons/thumb/9/9e/Wikipedia-logo-v2-de.svg/250px-Wikipedia-logo-v2-de.svg.png

[4] https://de.m.wikipedia.org/wiki/Datei:Wikibooks-logo-de.svg

[5] https://de.wikipedia.org/wiki/Wikiquote#/media/Datei:Wikiquote-logo-en.svg

[6] https://docs.python-requests.org/en/latest/

[7] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[8] https://github.com/hellysmile/fake-useragent

Questions