

Bachelor-Forschungsprojekt Informatik: Relevante OSM-Tags vorschlagen

Marco Hildebrand, XXXX, stXXXX@stud.uni-stuttgart.de
Lukas Baur, 3131138, st141998@stud.uni-stuttgart.de
Felix Bühler, 2973410, st117123@stud.uni-stuttgart.de

October 15, 2018

Abstract

Die vom *Institut für Formale Methoden der Informatik Stuttgart* entwickelte textbasierte Suchmaschine *OSCAR*, die OpenStreetMap-Daten auf Eingabe von OSM-Tags durchsucht, liefert unbefriedigende Ergebnisse auf anderweitige textuelle Eingaben. Im Rahmen unseres Bachelor-Forschungsprojekt Informatik sollte diese Lücke geschlossen werden, indem eine Anfrage an das von uns entwickelte System eine Menge an damit verwandten, relevanten Tags zurückgibt.

1 Einleitendes

1.1 Projektrahmen

Die Arbeit wurde im Rahmen des *Bachelor-Forschungsprojekts Informatik* in der Zeit vom April bis Oktober 2018 angefertigt. Diese Ausarbeitung stellt die inhaltliche Dokumentation des entwickelten Moduls dar.

1.2 Initiale Problemstellung

Grundlage für unsere Arbeit war die Suchmaschine *OSCAR*, die vom *Institut für Formale Methoden der Universität Stuttgart* entwickelt wurde.

OSCAR durchsucht auf Eingabe eines *OpenStreetMap-Tags* die zugehörige Datenbank nach passenden Einträgen und bereitet das Suchresultat grafisch auf. Ein *Tag* ist in OpenStreetMap wie folgt definiert:

$$\textit{key}=\textit{value}$$

Ein *key* wird benutzt, um ein Themenbereich zu charakterisieren, es repräsentiert einen Typ oder beschreibt ein Feature. Außerdem werden Tags vereinzelt als Namespaces verwendet [1].

Der *value*-Teil stellt ein Wert des Features da. Typische Werte sind Eigenschaften oder Zahlen [1]. Beispiele für Tags sind *building=yes*, *building=house* oder *highway=service* [2][3].

Da die Eingabe auf Tags beschränkt ist, benötigt ein User zur Suche einen passenden Tag. Diese Lücke soll mithilfe dieses Projekts geschlossen werden. Das zu entwickelnde System soll auf Eingabe eines natürlichen Wortes der englischen Sprache möglichst eng verwandte, relevante OpenStreetMap-Tags vorschlagen.

1.3 Abgrenzungen

Unsere Arbeit konzentriert sich auf die Suche der relevanten Tags zu einem eingegebenen Wort. Formaler ausgedrückt besteht unsere Eingabe aus genau einem Wort der englischen Sprache, das nicht in der zugrundeliegenden Stop-Word-Liste enthalten ist.

2 Projekt-Durchführung

2.1 Planungsaspekte

Zu Beginn unserer Arbeit grenzten wir unser Projekt thematisch ein und überlegten uns eine grobe Vorstrukturierung. Dazu gliederten wir unser Projekt in **drei** wesentliche Bausteine:

Im zeitlich ersten Arbeitsblock sollten wir uns mit der Darstellung, der Qualität und der Möglichkeit des Zugriffs der Daten vertraut machen. Im Folgenden überlegten wir uns eine aufbereitete brauchbare Daten-Zwischenform, auf deren Grundlage die spätere Suche durchgeführt werden soll. Der dritte Arbeitsbaustein galt der eigentlichen Such-Implementierung.

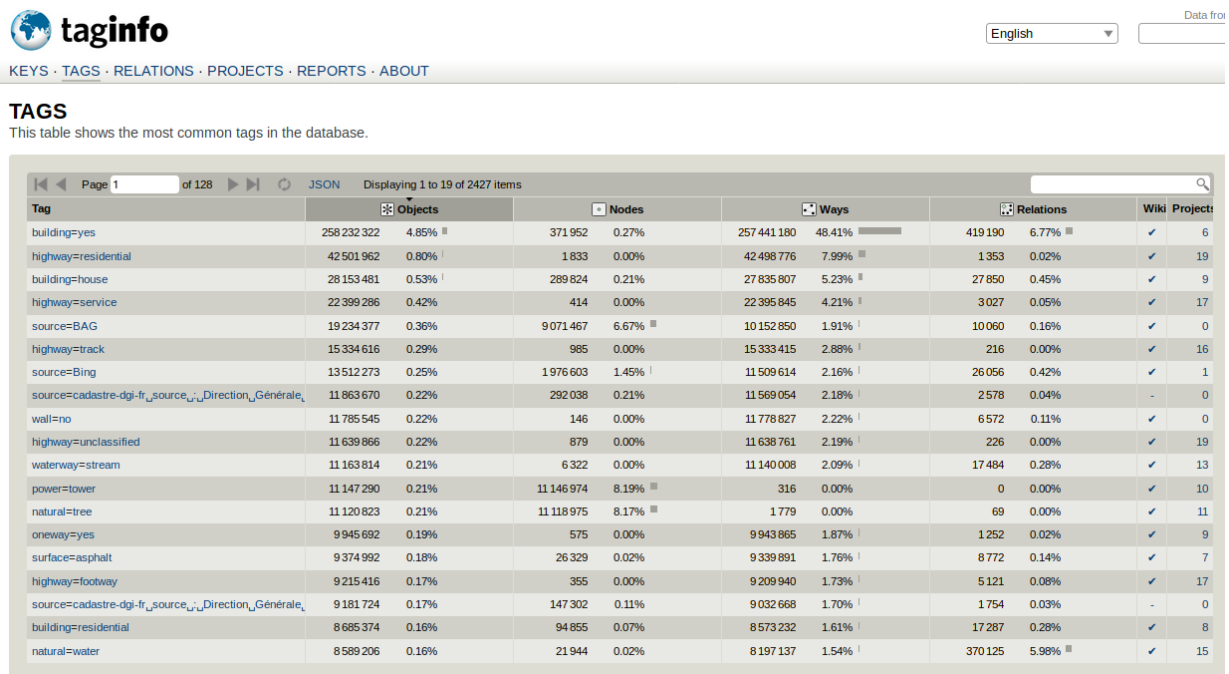
Die bearbeiteten Arbeitspakete werden im folgenden inhaltlich beschrieben. Die Pakete sind intern zeitlich sequentiell beschrieben, überlappen sich allerdings in Ihrer Abarbeitung. Der Grund hierfür sind Abhängigkeiten, wie zum Beispiel, dass die Datenaufbereitung an die Repräsentation des Suchalgorithmus angepasst werden muss, zuvor aber Daten als Grundlage der Suche beschafft sein müssen.

2.2 Datenbeschaffung

Unsere anfängliche Recherche begannen wir mit der Website von OpenStreetMap [4], insbesondere mit dem zugehörigem Wiki [5]. Das OSM-Wiki verfügt über eine ausführliche Dokumentation vieler gängiger OSM-Tags. Unser Ziel war es, auf alle vorhandenen Daten-Tupel, bestehend aus einem gültigen Tag und einer zugehörigen Tag-Beschreibung, lokalen Zugriff zu haben.

Leider besteht keine Möglichkeit das OSM-Wiki herunter zu laden. Wir kontaktierten den Verantwortlichen des Wikis, dieser konnte uns allerdings ebenfalls keine Kopie zukommen lassen. Folglich mussten wir die Webseiten des Wikis automatisiert crawlen.

Zwischenzeitlich versuchten wir alternativ mithilfe der Website *taginfo* [6] an die gesuchten Daten zu gelangen. *taginfo* wurde in Zusammenarbeit von Jochen und Christian Topf entwickelt und sammelt auf Grundlage der OSM-Daten aktuell rund 2.500 Tags inklusive deren statistischen Charakteristika und teilweise Beschreibungen[7]. Zusätzlich besteht die Möglichkeit, die komplette Datenbank herunterzuladen.



The screenshot shows the taginfo website interface. At the top, there's a logo and navigation links: KEYS, TAGS, RELATIONS, PROJECTS, REPORTS, ABOUT. Below the navigation bar, the 'TAGS' section is active, displaying a table of common tags. The table has columns for Tag, Objects, Nodes, Ways, Relations, Wiki, and Project. Each row represents a specific OSM tag with its corresponding counts and percentages for each object type, along with a checkmark indicating if a Wiki page exists and a number for the Project page.

Tag	Objects	Nodes	Ways	Relations	Wiki	Project
building=yes	258 232 322 4.85%	371 952 0.27%	257 441 180 48.41%	419 190 6.77%	✓	6
highway=residential	42 501 962 0.80%	1 833 0.00%	42 498 776 7.99%	1 353 0.02%	✓	19
building=house	28 153 481 0.53%	289 824 0.21%	27 835 807 5.23%	27 850 0.45%	✓	9
highway=service	22 399 286 0.42%	414 0.00%	22 395 845 4.21%	3 027 0.05%	✓	17
source=BAG	19 234 377 0.36%	9 071 467 6.67%	10 152 850 1.91%	10 060 0.16%	✓	0
highway=track	15 334 616 0.29%	985 0.00%	15 333 415 2.88%	216 0.00%	✓	16
source=Bing	13 512 273 0.25%	1 976 603 1.45%	11 509 614 2.16%	26 056 0.42%	✓	1
source=cadastre-dgi-fr;source=Direction_Générale_du_Cadastr	11 863 670 0.22%	292 038 0.21%	11 569 054 2.18%	2 578 0.04%	-	0
wall=no	11 785 545 0.22%	146 0.00%	11 778 827 2.22%	6 572 0.11%	✓	0
highway=unclassified	11 639 866 0.22%	879 0.00%	11 638 761 2.19%	226 0.00%	✓	19
waterway=stream	11 163 814 0.21%	6 322 0.00%	11 140 008 2.09%	17 484 0.28%	✓	13
power=lower	11 147 290 0.21%	11 146 974 8.19%	316 0.00%	0 0.00%	✓	10
natural=tree	11 120 823 0.21%	11 118 975 8.17%	1 779 0.00%	69 0.00%	✓	11
oneway=yes	9 945 692 0.19%	575 0.00%	9 943 865 1.87%	1 252 0.02%	✓	9
surface=asphalt	9 374 992 0.18%	26 329 0.02%	9 339 891 1.76%	8 772 0.14%	✓	7
highway=footway	9 215 416 0.17%	355 0.00%	9 209 940 1.73%	5 121 0.08%	✓	17
source=cadastre-dgi-fr;source=Direction_Générale_du_Cadastr	9 181 724 0.17%	147 302 0.11%	9 032 668 1.70%	1 754 0.03%	-	0
building=residential	8 685 374 0.16%	94 855 0.07%	8 573 232 1.61%	17 287 0.28%	✓	8
natural=water	8 589 206 0.16%	21 944 0.02%	8 197 137 1.54%	370 125 5.98%	✓	15

Figure 1: Beispielhafte Datenbankeinträge der Datenbank von taginfo

Leider stellten wir fest, dass die Beschreibungs-Einträge der Datenbank zu lückenhaft und damit für unsere Zwecke nicht geeignet sind. Schließlich kombinierten wir unsere bisherigen Ansätze, indem wir die Tag-Einträge der heruntergeladenen taginfo-Datenbank als Grundlage für ein Crawlen der OSM-Wiki-Seite verwendeten. Dies war möglich, da neben der lückenhaften Beschreibung zu jedem Tag zusätzlich der Link zur Wiki-Seite abgespeichert war. Dieser generische Link bestand aus den Teilen

wiki.openstreetmap.org/wiki/Tag%3Key%3Dvalue

wobei die Variablen *key* und *value* gemäß obiger Erklärung zu füllen sind.

2.3 Datenaufbereitung

2.4 Suchanfrage beantworten

3 notizen

Phase 1: Planung - Tags und dazugehörige semantische Beschreibung holen - in Struktur bringen
- Suchanfrage an Daten - gensim (Python von Mendel zu Beginn vorgeschlagen) - vorhanden/nicht vorhanden -> bewerbung fehlt - tf-idf -> gut, aber Problem: Mehrere Links auf dieselbe Seite -> Duplikate entfernen -> hohe Gewichtung für kleine Seiten -> Multiplizieren mit log/oder Wurzel 2
- Suchraum expandieren - mit Google Modell Anfrage semantisch auffüllen, Suche durchführen, am meisten Relevanten herausnehmen.

4 Einleitung

4.1 Projektbeschreibung

5 Vorgehensweise

Anschauen von wiki xml dump

unbrauchbare daten, da viel untereinander verlinkt ist.

herunterladen der tags: <https://taginfo.openstreetmap.org/>

6 Gettings started

6.1 languages

einfach eine liste aller sprachen bekommen mithilfe `taginfo-wiki.db`.

Die kann man von <https://taginfo.openstreetmap.org/download> herunterladen.

6.2 export-links

herunterladen der osm-wiki sitemap <https://wiki.openstreetmap.org/sitemap-index-wiki.xml>

davon interessiert uns nur `sitemap-wiki-NS_0-0.xml` der rest enthält daten zu den nutzern, diskussionen und historie

7 crawl

alle gesammelten link in die `links.txt` legen

```
scrapy crawl osmWiki -t json -o keys.json
```

7.1 pretty json

```
python -m json.tool keys.json > keys-pretty.json
```

8 Anhang

References

- [1] wikibooks. Tag. <https://wiki.openstreetmap.org/wiki/Tags>. Accessed: 2018-10-08.
- [2] taginfo. building. <https://taginfo.openstreetmap.org/keys/buildingvalues>. Accessed: 2018-10-08.
- [3] taginfo. building. <https://taginfo.openstreetmap.org/tags/highway=service>. Accessed: 2018-10-08.
- [4] OpenStreetMap Foundation (OSMF). Openstreetmap. <https://www.openstreetmap.org/>. Accessed: 2018-10-15.
- [5] The OpenStreetMap Foundation. De:hauptseite. <https://wiki.openstreetmap.org/wiki/DE:Hauptseite>. Accessed: 2018-10-15.
- [6] The OpenStreetMap Foundation. <https://taginfo.openstreetmap.org/>. Accessed: 2018-10-15.
- [7] The OpenStreetMap Foundation. About. <https://taginfo.openstreetmap.org/about>. Accessed: 2018-10-15.