Aantekeningen automatisch onderwerpsontsluiting ebooks m.b.v. Annif

Thomas Haighton

April 30, 2020

Contents

1 Inleiding

Dit document bevat mijn aantekeningen m.b.t. het testen van Annif t.b.v. het automatisch toekennen van onderwerpen aan E-books binnen de KB.

De KB hanteert een eigen classificatiesysteem/thesaurus genaamd Brinkman catalogus. Op de KB website is meer informatie te vinden m.b.t. de verschillende trefwoordcatalogi binnen de KB.

Bij aanvang van dit onderzoek was er sprake van een mogelijke samenwerking tussen de KB en het CB. Waar de KB de onderwerpen vastlegt d.m.v. Brinkmantrefwoorden, werkt het CB met Thema.

2 Data

2.1 Overeenkomsten Thema en Brinkman

Het eerste onderzoek was gericht op het vinden van overeenkomsten tussen de brinkmanonderwerpen en de onderwerpen zoals deze stonden vastgelegd in Thema. Bij de vergelijking zijn alle termen eerst geconverteerd naar kleine letters (lowercase), omdat Thema onderwerpen altijd beginnen met een hoofdletter en in Brinkman alleen trefwoorden zoals plaatsnamen of persoonsnamen met een hoofdletter beginnen.

_	Brinkman	Thema
Totaal aantal termen	14737	7362
Unieke termen	13729	6355
Overeenkomende termen	980	_

Zie Jupyter Notebook op GitHub

Constatering
Als we kijken naar een trefwoord wat alleen in Brinkman voorkomt, maar waarvan we verwachten dat deze ook in Thema zou moeten staan, b.v. 'autisme'. In Thema wordt 'autisme' vastgelegd in 'autisme en asperger syndroom'en 'omgaan met autisme/asperger'. Het zou dus mogelijk zijn dat er meer overlap mogelijk is.

N.B.: Het CB heeft laten weten (tijdelijk) af te zien van dit project. Verder onderzoek met Thema is na dit bericht gestaakt en is er alleen gekeken naar Brinkmantrefwoorden.

2.2 GGC dataset

De gebruikte dataset is een query aan het GGC als TSV (Tab Seperated Values) tekst bestand. De query bestaat uit verschillende eisen:

- Is een e-book
- Jaar van uitgave is tussen 2015-2019
- Nederlandstalig
- Er is minimaal één Brinkman-trefwoord toegekend
- Samenvattingsveld (KMC 4207) is niet leeg

De verkregen dataset bevat 12243 regels.

Bekijk de sql query: ***ggc_query.sql***
Bekijk de output: ***vraag_20190620.txt***

Tot nu toe zijn hier 3 verschillende datasets mee gemaakt, elk gesplits in een train (15%), eval (5%) en test set (80%).

- ggc1.zip: bevat alleen samenvattingen/flaptekst van alle E-books.
- ggc2.zip: bevat titel, ondertitel (wanneer deze aanwezig is) en samenvatting/flaptekst van alle E-books.
- ggc3.zip: bevat titel, ondertitel (wanneer deze aanwezig is) en samenvatting/flaptekst van E-books die geen vormtrefwoorden hebben toegekend.

Top 20 meest toegewezen Brinkmantrefwoorden

index	Toegewezen Brinkmantrefwoorden	Aantal x toegewezen
0	romans en novellen ; vertaald	2165
1	romans en novellen ; oorspr. - Nederlands	1960
2	jeugdboeken ; verhalen	1265
3	levensbeschrijvingen	193
4	gedichten ; oorspr. - Nederlands	181
5	autobiografieën	99
6	columns	61
7	levenskunst	55
8	stripverhalen	48
9	jeugdboeken ; verhalen — prentenboeken	46
10	geloofsleven	41
11	jeugdboeken ; verhalen — romans en novellen ; ...	39
12	jeugdboeken ; informatie - biologie	35
13	overdenkingen	34
14	prentenboeken — jeugdboeken ; verhalen	34
15	voetbal	33
16	spiritualiteit	33
17	essays	29
18	leidinggeven	26
19	reisverhalen	25

2.3 Vorm- en Zaaktrefwoorden

TODO:

3 Annif

Annif homepage: `http://annif.org/`

3.1 Train en Evalueer Annif model (in vogelvlucht)

Officiele *Getting Started* documentatie

Ik gebruik /Annif/tests als project map; zoals in de documentatie wordt aangeraden.

Annif commands/options help: `annif --help`

1. (optioneel) Maak eerst een configuratie in *projects.cfg*
2. Start Annif Python virtual environment (annif-venv), in annif-venv/bin: `source activate`
3. Navigeer naar project folder (e.g. Desktop/Annif/Annif/tests) en start Annif: `annif`
4. Check of *projects.cfg* gevonden wordt: `annif list-projects`
5. Laad onderwerpen: `annif loadvoc PROJECT_ID [SUBJECT_FILE]`
6. Train model: `annif train PROJECT_ID [PATHS]`
7. Evalueer model: `annif eval PROJECT_ID [PATHS]`
8. Gebruik de getrainde modellen (zie H. 3.5)

3.2 Configuratie

Voorbeeld Annif configuratie en bijbehorende evaluatie.

TF-IDF backend met snowball analyzer

```
[tfidf-brinkman]                # PROJECT_ID
name=TF-IDF Brinkman            # Uitgebreidde naam
language=nl                     # Taal
backend=tfidf                   # Backend (algoritme)
analyzer=snowball(dutch)        # Analyzer (stemmer)
limit=100                       # Aantal onderwerpen
vocab=brinkmanthesaurus_vocab   # Thesaurus
```

3.3 Evaluatie

```
Precision (doc avg):    0.07161500815660683
Recall (doc avg):       0.6252039151712887
F1 score (doc avg):     0.12716049583912553
Precision (conc avg):   0.0033237931737672725
Recall (conc avg):      0.006049485486600793
F1 score (conc avg):    0.003495638750427632
Precision (microavg):   0.07161500815660685
Recall (microavg):      0.5480649188514357
F1 score (microavg):    0.12667724715048334
NDCG:                   0.4647151156485022
NDCG@5:                 0.44466032224552277
```

```
NDCG@10:                0.4647151156485022
Precision@1:            0.29853181076672103
Precision@3:            0.18651441000543773
Precision@5:            0.12854812398042414
LRAP:                   0.39799930172762893
True positives:         439
False positives:        5691
False negatives:        362
Documents evaluated:    613
```

Alle evaluatie uitkomsten
Ik heb tot nu toe de volgende backends getest: *TF-IDF, Fasttext, Ensemble van de twee, en Omikuji*. Daarnaast ook kort de twee verschillende analyzers geprobeerd: *snowball* en *simple*. Bij de twee verschillende analyzers heeft de snowball analyzer een hogere score, dus voor de volgende experimenten zal ik alleen nog deze gebruiken.

Voor alle uitkomsten van de verschillende configuraties zie: ***annif_uitkomsten.xlsx***

3.4 Experimenten

In onderstaand experiment heb ik gebruik gemaakt van het project `tfidf-brinkman`.

3.4.1 Experiment 1

Test Annif via command-line interface: `cat document.txt | annif suggest tfidf-brinkman`
Annif suggestie voor bijbehorende Brinkman termen voor 420818715.txt:

```
cat ./data/Annif-corpora/fulltext/ggc/dev/420818715.txt | annif suggest tfidf-brinkman

<http://data.bibliotheken.nl/id/thes/p075660849>        levenskunst
    0.4163530829438607
<http://data.bibliotheken.nl/id/thes/p075665689>        zelfkennis
    0.39504628028213046
<http://data.bibliotheken.nl/id/thes/p075606178>        filosofie
    0.3769179576596768
<http://data.bibliotheken.nl/id/thes/p075607050>        geloofsleven
    0.3712595743673854
<http://data.bibliotheken.nl/id/thes/p07561765X>        organisatieontwikkeling
    0.3698668389205731
<http://data.bibliotheken.nl/id/thes/p075663910>        spiritualiteit
    0.36965842850694486
<http://data.bibliotheken.nl/id/thes/p075660822>        leidinggeven
    0.36507826867675164
<http://data.bibliotheken.nl/id/thes/p075617846>        overdenkingen
    0.36310030904972374
<http://data.bibliotheken.nl/id/thes/p075603578>        cultuurfilosofie
    0.3629843846776753
<http://data.bibliotheken.nl/id/thes/p075610744>        jeugdboeken ; verhalen
    0.36280384270411503
```

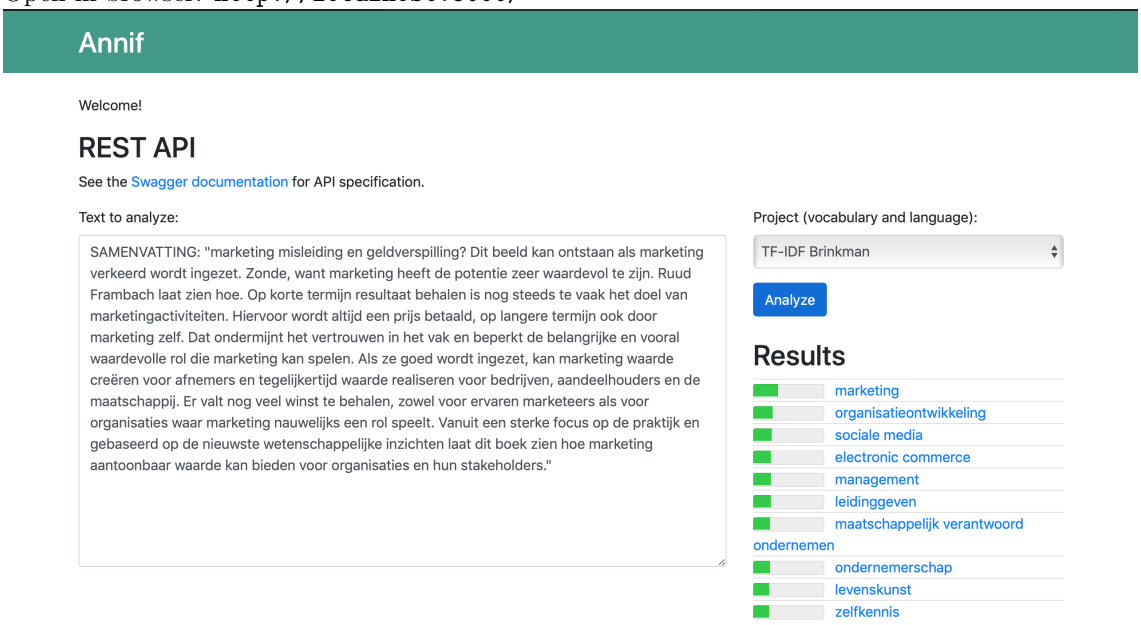
Daadwerkelijk toegekende Brinkman termen - 420818715.tsv:

```
cat ./data/Annif-corpora/fulltext/ggc/dev/420818715.tsv

<http://data.bibliotheken.nl/id/thes/p075600447>        bedrijfsorganisatie
<http://data.bibliotheken.nl/id/thes/p075603012>        citatenverzamelingen
```

Volledige tekst 420818715.txt:

```
cat ./data/Annif-corpora/fulltext/ggc/dev/420818715.txt

Er wordt wat afgeklooid in onze bedrijven en organisaties. Tijd om te ontklooien, dus!
    Hoe graag zouden we luid gillend willen protesteren tegen alle ellende? We blijken
    namelijk slechts 15% van onze tijd bezig te zijn met het creren van waarde. Er zijn
    tienduizenden bullshitjobs. 75% van alle (ict-)projecten halen hun budget, balance,
    alsof werken geen leven is. En wat al niet meer... De trieste waarheid is dat werken
     inderdaad vaak geen leven is. Maar luidop protesteren is een niet zo erg
    carrirebevorderende actie. Durven we onze mond dus wel open te doen? Waarschijnlijk
    niet... Dit boekje snelt u ter hulp. Honderd en een citaten van bekende en minder
    bekende managementexperts, filosofen en wetenschappers die hun flink gepekelde
    vinger in de open wonden leggen. Ze vertellen precies wat u en ik denken. We kunnen
    ermee uitpakken: op ons whiteboard, in een mailtje of anoniem in de kantine. En wij
    blijven buiten schot, want de goeroe heeft het gedaan. Liever op de achtergrond
    blijven? Dan hebt u vast veel gniffelplezier bij het lezen van dit boekje. Bent u
    zelf leidinggevende, manager of directeur? Zoals de Vlaamse dichteres Alice Nahon
    schreef: 't is goed in 't eigen hart te kijken, des avonds voor het slapengaan. May
    the force be with you! Bron: Flaptekst, uitgeversinformatie
```

Test annif als web applicatie: `annif run`
Open in browser: `http://localhost:5000/`



Bijbehorende toegekende Brinkman term:
`<http://data.bibliotheken.nl/id/thes/p075661098> marketing`

3.5 SKOS (Simple Knowledge Organization System)

- SKOS Core Guide
- SKOS Core Vocabulary

SKOS is een toepassing van RDF. SKOS standaard is specifiek te gebruiken voor het vastleggen van een thesaurus, gecontroleerde vocabulair e.d. SKOS maakt o.a. gebruik van synoniemen (`altLabel`) van een term (`Concept`, `prefLabel`) en de hierargische relatie (`broader`, `narrower`) tussen termen in een thesaurus.

Voorbeeld snippet

```
<skos:Concept rdf:about="http://www.yso.fi/onto/yso/p21272">
  <rdf:type rdf:resource="http://www.yso.fi/onto/yso-meta/Concept"/>
  <skos:altLabel xml:lang="en">leaf beetles</skos:altLabel>
  <skos:broader rdf:resource="http://www.yso.fi/onto/yso/p6734"/>
  <skos:closeMatch rdf:resource="http://id.loc.gov/authorities/subjects/sh85025443"/>
  <skos:exactMatch rdf:resource="http://www.yso.fi/onto/allars/Y37803"/>
  <skos:exactMatch rdf:resource="http://www.yso.fi/onto/koko/p57371"/>
  <skos:exactMatch rdf:resource="http://www.yso.fi/onto/ysa/Y158869"/>
  <skos:inScheme rdf:resource="http://www.yso.fi/onto/yso/"/>
  <skos:narrower rdf:resource="http://www.yso.fi/onto/yso/p21619"/>
  <skos:prefLabel xml:lang="en">Chrysomelidae</skos:prefLabel>
  <skos:prefLabel xml:lang="sv">bladbaggar</skos:prefLabel>
  <skos:prefLabel xml:lang="fi">lehtikuoriaiset</skos:prefLabel>
</skos:Concept>
```

3.6 Maui backend

Maui backend Annif Github

Brinkman turtle bestand: ***thes_000001.ttl***
Brinkman download: ***brinkman_dl.txt***

3.6.1 Installatie Maui backend

Maui Server image geinstalleerd op Macbook Pro via Docker 2.2.0.0.
Server kon benaderd worden via browser (localhost).

3.6.2 Brinkman catalogus in SKOS formaat

Maui werkt met thesauri in het SKOS formaat. De Brinkman is niet direct in deze vorm te krijgen, wel in naderende vorm. Ik heb via Rene Voorburg een oude uitdraai van de Brinkman catalogus gekregen in het turtle (.ttl) formaat. En via zijn script `https://github.com/renevoorburg/oai2linerec` een huidige versie proberen te downloaden. De download heb ik na een dag of twee

gestopt.
Brinkman SKOS download m.b.v. script Rene. Start script:

```
sh oai2linerec.sh -p dcx -s GGC-THES -o brinkman_skos_test.txt -b http://services.kb.nl/
    mdo/oai
```

De verkregen download heb ik eerst moeten opschonen, omdat er niet alleen Brinkman trefwoorden in staan; gedaan m.b.v. een code editor en reguliere expressies. Deze heb ik daarna m.b.v. Skosify omgezet naar een voor Annif bruikbaar bestand (default settings).

Het turtle bestand heb ik ook met Skosify bewerkt, omdat de inhoud van het bestand gesorteerd was op parameter.

Beide bestanden gaven niet direct een foutmelding wanneer het `vocab` command gebruik in Annif om het bestand als thesaurus aan te geven. Maar het trainen van een model wilde niet starten. Wanneer er via de browser naar de parameters werd gekeken (json bestand op de achtergrond, te zien in browser) die Maui had kon worden gezien dat er nog geen thesaurus was opgegeven.

Uiteindelijk besloten om eens een test te doen met de bijgevoegde SKOS (*yso-skos-boethius.rdf*) van Annif en deze werd ook niet geaccepteerd. Dit is een indicatie dat het waarschijnlijk niet meteen aan de gemaakte Brinkman SKOS lag (wat mijn eerste gedachte was), maar er ergens anders iets fout zat (bv. installatie docker Maui Server).

3.7 Actie punten

personal Trello board

- ☑ Configuratie aanpassen, i.e. probeer simple analyzer.
- ☐ Probeer verschillende backends - bv ensemble approach.
  - ☑ TF-IDF
  - ☑ Fasttext
  - ☐ Maui
  - ☑ Omikuji
  - ☐ Ensembles
- ☐ Ook titel data gebruiken naast de samenvatting.
  - ☑ Dataset gemaakt ggc2.
  - ☐ Testen, vergelijken met uitkomsten ggc dataset, en uitkomsten documenteren. *nog niet met alle backends getest. De backends die getest zijn blijken net iets beter te werken met titel data.*
  - ☐ Test met weights: 2x titel in data gebruiken.
- ☐ Documentatie! (die lees je nu)
- ☐ SKOS vocab + Maui backend - Brinkman als SKOS.
  [05-03-2020] Met Sara besproken om Maui even te parkeren.

- ☐ Minimale eisen SKOS bestand voor Annif met Maui backend (moeten we onderstaande punten gaan uitzoeken?).
- ☑ Hulp vragen Rene Voorburg om SKOS te genereren via OAI-PMH (GGC_THES) - zie zijn script op `https://github.com/renevoorburg/oai2linerec`.
- ☐ Omvormen verkregen 'SKOS' data naar daadwerkelijk SKOS-XML bestand.
- ☒ Evaluatie model per woord. (gaat Sara oppakken)

☐ Tweak backends en optimaliseer ensembles.

☐ Idee m.b.t evaluatie model
Annif testen tegen 2 collectiespecialisten, ieder probeert een brinkmantrefwoord toe te wijzen, m.b.v. word2vec kijken of de termen dicht bij elkaar liggen (of andere methodes; onderzoeken!)?

3.7.1 Actiepunten voor voortzettten testen Maui backend

☐ Andere manier van installeren via Tomcat (zie `https://github.com/NatLibFi/Annif/wiki/Backend%3A-Maui#setting-up-maui-server-using-tomcat`)

☐ Verse install Annif - gelijk testen met door Annif bijgevoedge SKOS (zou inprincipe direct moeten werken).

☐ Ronald Cornelisen mailen m.b.t. Brinkman SKOS. data.bibliotheken.nl opgezet (extern)