

# Taalmodellen voor de Automatische Categorisatie van Documenttitels

Martijn Spitters  
Textkernel BV

TEXTKERNEL RAPPORT IN HET KADER VAN HET RNA PROJECT

## Abstract

Textkernel heeft een aantal experimenten uitgevoerd waarin de haalbaarheid wordt onderzocht van de automatische categorisatie van documenttitels. De experimenten zijn uitgevoerd op een door de Koninklijke Bibliotheek samengestelde dataset binnen het domein *onderwijs*. Dit document geeft een beknopt overzicht van de beste categorisatiemethode en de daarmee behaalde resultaten.

## 1 Inleiding

De technologie voor het automatisch categoriseren (labelen) van tekstuele informatie heeft het laatste decennium een enorme ontwikkeling doorgemaakt. Voor de meeste teksttypen en -onderwerpen zijn met de huidige state-of-the-art technieken zeer accurate categorisatieresultaten te behalen.

Toch blijven bepaalde categorisatietaken moeilijk oplosbaar. Een voorbeeld van zo'n taak is categorisatie van documenten met zeer weinig informatie, zoals de documenttitels in deze studie. Bij automatische tekstcategorisatie worden typisch woorden, woordfragmenten, of frasen gebruikt als *features*, d.w.z. de kenmerken van de tekst waaraan het categorisatiealgoritme de correcte categorie 'herkent'. Vrijwel alle classificatiealgoritmen gebruiken reeds gecategoriseerde voorbeelddocumenten waaruit deze kenmerken worden geleerd. Bij zeer korte documenten is het gevaar van *data sparseness* echter erg groot. Door de variëteit en beknoptheid van dergelijke data zijn er vaak weinig kenmerken te vinden die door een substantieel deel van de documenten van een bepaalde categorie worden gedeeld. De kans is groot dat ongeziene, te categoriseren documenten volledig steunen op unieke woorden die nog niet aanwezig zijn in de voorbeelddata, tenzij de classifier wordt getraind met aanzienlijke hoeveelheden voorbeelddocumenten voor iedere categorie. Dit verschijnsel heeft over het algemeen een minder grote impact op specifieke categorieën dan op 'brede' categorieën, vanwege het beperkter vocabulair. Dynamische categorieën zijn uiteraard ook gevoelig voor data sparseness, vanwege het voorturend veranderend vocabulair (bijv. langlopende onderwerpen in het nieuws).

Om te ontdekken welke categorisatiemethode het meest geschikt is voor de data in deze studie, hebben we een aantal verschillende state-of-the-art classificatietechnieken toegepast. We hebben machine learning-gebaseerde classifiers (*k-nearest neighbour*, *support vector machines*, en *naive bayes*) getraind voor alle 225 categorieën gezamenlijk (*n*-way classifiers) en voor iedere individuele categorie (binair classifiers). Deze laatste aanpak resulteert in 225 classifiers die aan een testdocument voor ‘hun’ categorie 1 of 0 toekennen. Hiernaast hebben we classifiers gebouwd op basis van statistische taalmodellen (*language models*) van de documenten. Mede door de ongebalanceerde training data leverden de machine learning classifiers relatief inconsistente resultaten. Ongebalanceerde training data (ongelijk verdeelde aantallen voorbeelddocumenten voor de verschillende categorieën) is vooral een probleem bij binaire classifiers, waarbij het aantal negatieve voorbeelden vele malen groter is dan het aantal positieve. Het trainen van een classifier die beide categorieën als gelijkwaardig beschouwt, vereist dan selectietechnieken waarmee de trainingdata in balans wordt gebracht. Hiermee gaat echter veel mogelijk nuttige negatieve informatie verloren. De aanpak met de statistische taalmodellen leverde de hoogste accuraatheid, maar ook met deze techniek varieerde die enorm van categorie tot categorie.

Dit document geeft beknopte achtergrondinformatie over classificatie met statistische taalmodellen en tevens over de gebruikte evaluatiemethoden. Tot slot presenteren we de tot dusver best behaalde categorisatieresultaten.

## 2 Taalmodellen voor Tekstcategorisatie

Statistische taalmodellen worden vooral gebruikt voor *information retrieval* (IR) (zie bijvoorbeeld [Croft and Lafferty, 2003] en [Hiemstra, 2001]) en in een complexere vorm voor o.a. spraakherkenning. Statistische taalmodellen zijn ook succesvol ingezet voor verschillende tekstclassificatietaken (zie bijvoorbeeld [Spitters and Kraaij, 2001] en [Peng et al., 2003]). Bij een information retrieval taak willen we documenten rangschikken op basis van hun relevantie voor een bepaalde query (zoekvraag), welke typisch weinig informatie bevat (enkele termen, zoals de documenttitels in deze studie). In een tekstclassificatietask is het doel niet zozeer om documenten te rangschikken, maar om voor ieder (test) document te beslissen of een bepaalde categorie wel of niet moet worden toegekend. Als we de vergelijking maken met language modeling voor IR kan het testdocument worden opgevat als de zoekvraag en de representaties van de categorieën als de (potentieel relevante) gezochte documenten.

In onze aanpak wordt eerst voor iedere categorie een unigram taalmodel gegenereerd op basis van de voorbeelddocumenten van de betreffende categorie. Zo’n taalmodel kan worden gezien als een eenvoudige kansdistributie van de woorden die voorkomen in de documenten van de categorie. De classificatiescore van een categorie  $C$  voor een bepaald testdocument is gebaseerd op de kans dat de woorden waaruit die is opgebouwd ‘gegenereerd’ worden uit het taalmodel van  $C$ . Deze kans wordt afgezwakt (*smoothing*) door middel van een achtergrond-taalmodel (een generiekere kansdistributie met een groot vocabulair), zodat bijvoorbeeld wordt gecompenseerd voor veel voorkomende woorden met weinig informatiewaarde. De categorie wordt toegekend aan het testdocument indien de score hoger is dan een bepaalde, voor die categorie vastgestelde drempel.

### 3 Evaluatiemethoden

De resultaten in dit document zijn beschreven in termen van:

- **Precision:** de precision voor categorie  $C$  is het percentage testinstanties dat correct geclassificeerd is als  $C$ , van het totaal aantal testinstanties dat geclassificeerd is als  $C$
- **Recall:** de recall voor categorie  $C$  is het percentage testinstanties dat correct geclassificeerd is als  $C$ , van het totaal aantal testinstanties dat werkelijk tot  $C$  behoort (d.w.z. handmatig van dat label is voorzien)
- **F-measure:** de F-measure is het harmonisch gemiddelde van precision en recall: ( $F = 2PR/(P + R)$ ).

### 4 Resultaten

Tabellen 1-3 (bijgevoegd aan het eind van dit document) geven voor alle categorieën de precision, recall en f-measure, alsmede de threshold en het aantal voorbeelddocumenten waaruit het taalmodel voor de betreffende categorie is opgebouwd. De gegeven threshold is geoptimaliseerd op basis van de f-measure, d.w.z. bij toekening van de betreffende categorie aan alle testdocumenten met een score  $>$  threshold is de f-measure voor die classifier het hoogst. De scores in deze tabellen zijn gebaseerd op de door de KB gemaakte train/test verdeling van de data.

Om het probleem van *data sparseness* te illustreren, hebben we extra classificierens gedraaid, waarbij steeds werd getraind op basis van *alle* beschikbare documenten, *behalve* het testdocument dat op dat moment wordt geclassificeerd: Tabellen 4-6). Dit betekent dus dat een taalmodel voor een categorie is opgebouwd uit alle trainingdocumenten voor die categorie, *plus* de testdocumenten voor die categorie die op dat moment niet worden geclassificeerd. De resultaten van deze classifiers laten in feite zien wat de resultaten van de classifiers uit Tabellen 1-3 *zouden* zijn, na bijtraining met enkele nieuwe documenten. De sterk verbeterde resultaten suggereren dat dergelijke classifiers zichzelf snel sterk kunnen verbeteren, indien ze worden ingezet als incrementeel suggestiesysteem. In een dergelijke situatie worden bijvoorbeeld de vijf meest waarschijnlijke categorieën gesuggereerd voor een nieuw document, waarna een redacteur de correcte categorieën goed- en de onjuiste afkeurt. Tevens kan de redacteur categorieën toekennen die door het systeem niet zijn gevonden. Het systeem kan dan eenvoudig van deze kennis bijleren door zijn taalmodellen (direct) aan te passen.

### 5 Suggestieclassifiers

Tabellen 1-6 geven de resultaten zoals ze zouden zijn in een classificatieworkflow *zonder* menselijke interventie. Voor enkele categorieën lijkt menselijke controle niet nodig (bijvoorbeeld *kinderboek* en *lerarenopleidingen*), maar voor de meeste categorieën is controle, zeker in de beginfase, totdat de classifier tot de gewenste accuraatheid is bijgetraind, echter vereist. Een *suggestieclassifier* kan in deze situatie een nuttig en tijdsbesparend hulpmiddel zijn. Bij een suggestieclassifier worden alle door het systeem relevant geachte categorieën (alle categorieën waarvan de score voor het testdocument hoger is dan hun threshold) op overzichtelijke wijze voorgesteld aan de gebruiker. De

gebruiker heeft dan de mogelijkheid, bijvoorbeeld door middel van vinkboxjes, om snel de juiste categorieën aan het testdocument toe te kennen. Ook biedt een dergelijke interface de mogelijkheid om categorieën toe te kennen die door het systeem over het hoofd worden gezien, bijvoorbeeld door deze direct in te vullen of door snel te zoeken in de gekoppelde thesaurus. Textkernel ontwikkelt momenteel zo'n volledig geïntegreerd classificatiesysteem met overzichtelijke gebruikersinterface.

Bij een suggestieclassificatie wordt de classificatieaccuraatheid bepaald door een flexibelere maatstaf, namelijk of de correcte categorie(ën) voor het testdocument in de lijst van gesuggereerde categorieën voorkomt/-komen. Beschouw ter illustratie onderstaand classificatievoorbeeld.

Op opvoeding aangewezen.  
Een kritiek op de wijze van omgaan met kinderen in onze cultuur.

Suggesties: [1] kinderpsychologie  
              [2] opvoeding  
              [3] kunstonderwijs

De onjuiste toekenning van categorie *kunstonderwijs* draagt bij aan diens relatief lage werkelijke precision (zie Tabel 1). Echter, dit is in deze situatie niet relevant omdat de redacteur deze categorie eenvoudigweg kan negeren (of 'uitzetten'). In een situatie zonder menselijke controle zou deze categorie echter onterecht worden toegekend aan het testdocument. De correcte categorie (*opvoeding*) is echter wel aanwezig in de suggestielijst. De niet handmatig toegekende term *kinderpsychologie* is mogelijk ook relevant.

## 5.1 Consistente Classificatie

Het volgende voorbeeld laat zien dat een suggestieclassificatie kan helpen bij een consistentere classificatie. Dit testdocument is niet gelabeld met de categorie *leerplicht*, ondanks diens overduidelijke relevantie. De redacteur heeft alleen de term *onderwijspsychologie* toegekend. Ons insziens is het beter en consistentier om dit testdocument te labelen met beide categorieën, zodat het document bij een zoekopdracht via beide thesaurustermen kan worden teruggevonden.

Verstand op nul: leerplicht en intelligentieverloop.

Suggesties: [1] onderwijspsychologie  
              [2] leerplicht  
              [3] onderwijsstimulering

## 5.2 Evaluatie

Een andere reden waarom automatische classificatie van documenttitels problematisch kan zijn is het gebrek aan context. In tegenstelling tot de redacteur heeft de classifier in deze experimenten alleen de beschikking over de titel van het boek of document. In vele gevallen is de informatie in de titel zo beperkt dat ook een mens moeite zou hebben de juiste termen toe te kennen. In sommige gevallen lijkt het zelfs onmogelijk, zoals bij de volgende titels:

- *Raamplan.*
- *Er komt een muisje aangelopen.*
- *Rapportage, meer dan een rapport.*

In dergelijke gevallen zal de documentalist het origineel erop moeten naslaan (bijvoorbeeld de flaptekst lezen) om de correcte termen te kunnen toekennen. Omdat dergelijke contextuele informatie in deze experimenten niet beschikbaar is voor de geautomatiseerde classifier, worden dergelijke gevallen niet of onjuist geclassificeerd. We zijn van mening dat dit de resultaten in Tabellen 1-6 voor sommige categorieën enigszins vertekent.

Om een ‘eerlijker’ oordeel te kunnen vellen over de classificatieaccuraatheid van een suggestieclassifier, hebben we de door de suggestieclassifier toegekende labels voor een willekeurige gekozen subset van de test-titels (ongeveer 10%) handmatig gecontroleerd. In deze test had de controleur dus alleen de beschikking over de titels en niet de originele documenten. Indien een titel te weinig informatie bevatte om te kunnen worden geclassificeerd, werd de classificatie voor die betreffende titel niet meegerekend. Van de voor deze evaluatie 121 random gekozen test-titels was dit bij 18 gevallen (15%) van toepassing. Ook titels in een andere taal (in deze evaluatieset in totaal 7 (6%)) werden genegeerd<sup>1</sup>. Gemiddeld kende de suggestieclassifier 2.5 termen toe. De accuracy (het percentage van de titels waarbij de correcte categorie bij de gesuggereerde categorieën zat) is 78%. Hierbij moet worden opgemerkt dat we verwachten dat de accuracy van de suggestieclassifier nog aanzienlijk kan worden verbeterd door toepassing van eenvoudige taalkundige bewerkingen, zoals *compound splitting* en *lemmatisering*, en door verdere experimenten met de normalisatie van de confidence-drempels voor de verschillende categorieën.

## References

- Bruce Croft and John Lafferty, editors. *Language Models for Information Retrieval*. Kluwer Academic Publishers, 2003.
- Djoerd Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, Enschede, 2001.
- Fuchung Peng, Dale Schuurmans, and Shaojun Wang. Language and task independent text categorization with simple language models. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 189–196, 2003.
- Martijn Spitters and Wessel Kraaij. Using language models for tracking events of interest over time. In *Proceedings of the Workshop on Language Models for Information Retrieval (LMIR)*, 2001.

---

<sup>1</sup>Overigens werd door de suggestieclassifier aan 4 van de 7 buitenlandse titels geen term toegekend, en aan de helft van de ‘onmogelijke’ titels. Dit zijn dus in zekere zin ‘correcte’ classificaties, omdat het systeem voor deze gevallen onmogelijk de juiste termen kan toekennen.

Category	Precision	Recall	F-measure	Threshold	N train
1: vaderschap	1.00	0.80	0.89	0.53	20
2: kinderboek	1.00	0.80	0.89	0.58	23
3: montessori-onderwijs	1.00	0.80	0.89	0.45	20
4: gymnasia	1.00	0.80	0.89	0.30	20
5: middenschool	0.71	1.00	0.83	0.35	20
6: adoptie	0.80	0.80	0.80	0.40	20
7: scripties	0.67	1.00	0.80	0.44	21
8: rekenonderwijs	0.82	0.75	0.78	0.47	37
9: speltherapie	1.00	0.62	0.77	0.77	23
10: verkeerseducatie	0.83	0.71	0.77	0.56	22
11: kinderen	1.00	0.60	0.75	0.70	21
12: universiteiten	0.80	0.67	0.73	0.24	24
13: peuterspeelgroepen	0.67	0.80	0.73	0.38	22
14: middelbaar dienstverlenings- en gezondheidszorgonderwijs	0.80	0.67	0.73	0.54	28
15: huiswerk	0.80	0.67	0.73	0.45	21
16: stages	0.71	0.71	0.71	0.43	24
17: schoolbesturen	0.62	0.83	0.71	0.36	28
18: orthodidactiek	1.00	0.56	0.71	0.77	26
19: autisme	0.56	0.83	0.67	0.34	20
20: kinderspelen	0.56	0.83	0.67	0.34	25
21: biologie-onderwijs	0.80	0.57	0.67	0.99	26
22: medezeggenschap in het onderwijs	1.00	0.50	0.67	0.56	24
23: buitenschoolse activiteiten	1.00	0.50	0.67	0.45	23
24: pesten	0.75	0.60	0.67	0.32	23
25: agogie	0.75	0.60	0.67	0.67	26
26: godsdienstonderwijs	0.58	0.70	0.64	0.23	28
27: milieueducatie	0.55	0.75	0.63	0.17	46
28: management ; onderwijs	0.56	0.71	0.63	0.36	31
29: lichamelijke opvoeding	0.62	0.62	0.62	0.39	22
30: brede scholen	0.45	1.00	0.62	0.24	20
31: leesstoornissen	0.50	0.80	0.62	0.65	25
32: studiehuis	0.50	0.80	0.62	0.35	22
33: begaafdheid	0.57	0.67	0.62	0.36	22
34: spelling	0.57	0.67	0.62	0.25	24
35: schoolverlaters	1.00	0.44	0.62	0.52	26
36: dyslexie	0.67	0.57	0.62	0.27	20
37: technisch onderwijs	1.00	0.43	0.60	0.81	23
38: recht ; opleiding	0.60	0.60	0.60	0.49	20
39: leermiddelen ; omgangskunde	0.60	0.60	0.60	0.63	20
40: voorbereidend beroepsonderwijs	0.75	0.50	0.60	0.49	24
41: hoger onderwijs	0.62	0.57	0.59	0.26	68
42: bijzonder onderwijs	0.67	0.50	0.57	0.30	24
43: medisch onderwijs	1.00	0.40	0.57	0.50	29
44: modulair onderwijs	1.00	0.40	0.57	0.48	23
45: kunstonderwijs	0.50	0.67	0.57	0.33	20
46: seksuele voorlichting	0.44	0.80	0.57	0.33	21
47: informaticaonderwijs	0.44	0.80	0.57	0.24	33
48: basiseducatie	0.43	0.86	0.57	0.37	35
49: tekenonderwijs	0.44	0.80	0.57	0.38	21
50: kinderopvang	0.60	0.55	0.57	0.28	37
51: basisvorming	0.50	0.64	0.56	0.35	45
52: landbouwonderwijs	0.60	0.50	0.55	0.30	25
53: middelbaar algemeen vormend onderwijs	0.60	0.50	0.55	0.49	24
54: maatschappijleer	0.55	0.55	0.55	0.34	29
55: babyverzorging	0.60	0.50	0.55	0.63	28
56: schoolloopbaan	0.75	0.43	0.55	0.33	25
57: literatuuronderwijs	0.50	0.60	0.55	0.22	26
58: leermethoden	1.00	0.38	0.55	1.21	27
59: kinderdagverblijven	0.40	0.80	0.53	0.27	26
60: jeugdwerk	0.50	0.57	0.53	0.26	23
61: vormingswerk	0.57	0.50	0.53	0.31	24
62: didactiek	1.00	0.36	0.53	0.77	50
63: lerarenopleidingen	0.35	1.00	0.52	0.14	29
64: schrijfonderwijs	0.38	0.71	0.50	0.24	23
65: jeugdhulpverlening	0.67	0.40	0.50	0.37	38
66: scholengemeenschappen	1.00	0.33	0.50	0.50	21
67: leerplicht	0.50	0.50	0.50	0.24	22
68: jeugdgezondheidszorg	0.50	0.50	0.50	0.35	22
69: lichamelijk gehandicapten	0.44	0.57	0.50	0.34	22
70: probleemgestuurd onderwijs	1.00	0.33	0.50	0.56	22
71: beeldende vorming	0.67	0.40	0.50	0.96	23
72: vakkenpakketkeuze	1.00	0.33	0.50	3.00	23
73: schoolgebouwen	1.00	0.33	0.50	0.78	21
74: beroepsvoorlichting	1.00	0.33	0.50	0.55	36
75: kindermishandeling	0.50	0.50	0.50	0.29	20

Table 1: KB education test data. Classification based on language modeling.

Category	Precision	Recall	F-measure	Threshold	N train
76: christelijk onderwijs	0.43	0.60	0.50	0.28	22
77: openbaar onderwijs	0.60	0.43	0.50	0.44	24
78: schoolverzuim	0.44	0.57	0.50	0.19	25
79: projectonderwijs	0.50	0.50	0.50	0.26	21
80: scheikunde-onderwijs	0.67	0.40	0.50	0.75	26
81: kleuteronderwijs	0.40	0.67	0.50	0.21	27
82: geloofsopvoeding	0.67	0.40	0.50	0.52	23
83: gezinsbegeleiding	0.60	0.43	0.50	0.50	22
84: speciaal onderwijs	0.57	0.42	0.48	0.22	65
85: Engelse taalkunde	0.33	0.86	0.48	0.19	24
86: Nederlands voor anderstaligen	0.40	0.60	0.48	0.28	25
87: leerlingwezen	0.67	0.36	0.47	0.50	27
88: computers in het onderwijs	0.50	0.43	0.47	0.17	63
89: leesonderwijs	0.46	0.46	0.46	0.28	36
90: leesbevordering	0.50	0.43	0.46	0.38	25
91: onderwijswetgeving	0.38	0.60	0.46	0.26	38
92: onderwijspsychologie	1.00	0.29	0.44	0.93	24
93: sportonderwijs	0.40	0.50	0.44	0.27	21
94: taalververving	0.40	0.50	0.44	0.31	27
95: ouderparticipatie	0.50	0.40	0.44	0.32	26
96: school-maatschappelijk werk	0.50	0.40	0.44	0.36	22
97: alfabetiseringsprojecten	0.67	0.33	0.44	0.58	23
98: geografie-onderwijs	0.67	0.33	0.44	1.21	24
99: katholiek onderwijs	0.67	0.33	0.44	0.33	22
100: dramatische vorming	0.67	0.33	0.44	0.56	27
101: voorbereidend middelbaar beroepsonderwijs	0.33	0.67	0.44	0.25	27
102: bedrijfsopleidingen	0.67	0.33	0.44	0.32	30
103: schoolkeuze	0.67	0.33	0.44	0.29	23
104: Nederlandse taal ; onderwijs	0.50	0.38	0.43	0.26	32
105: onderwijsstimulering	0.45	0.42	0.43	0.20	33
106: educatieve software	0.33	0.60	0.43	0.37	23
107: natuuronderwijs	0.38	0.50	0.43	0.34	23
108: peuters	0.50	0.36	0.42	0.28	24
109: verstandelijk gehandicapten	0.38	0.45	0.42	0.28	28
110: middelbaar beroepsonderwijs	0.29	0.70	0.41	0.20	39
111: baby's	0.40	0.40	0.40	0.45	26
112: puberteit	0.40	0.40	0.40	0.54	21
113: logopedie	0.40	0.40	0.40	0.33	22
114: afstandsonderwijs	0.40	0.40	0.40	0.38	24
115: onderwijsrecht	0.33	0.50	0.40	0.36	21
116: gezondheidszorg	0.50	0.33	0.40	0.43	21
117: hulpverlening	0.67	0.29	0.40	0.62	23
118: remedial teaching	1.00	0.25	0.40	1.12	23
119: verpleging ; opleiding	0.50	0.33	0.40	0.51	22
120: natuurbeschermingseducatie	0.30	0.60	0.40	0.39	26
121: bewegingsonderwijs	0.50	0.33	0.40	0.35	35
122: natuurkunde-onderwijs	0.38	0.43	0.40	0.53	27
123: wetenschappelijk onderzoek	0.38	0.43	0.40	0.33	30
124: volwasseneneducatie	0.30	0.58	0.39	0.13	51
125: opvoeding	0.53	0.31	0.39	0.19	120
126: onderwijs aan anderstaligen	0.33	0.44	0.38	0.21	36
127: kleuters	0.30	0.50	0.37	0.28	27
128: filosofieonderwijs	0.30	0.50	0.37	0.27	21
129: pedagogiek	0.60	0.27	0.37	0.48	56
130: gezondheidsopvoeding	0.50	0.29	0.36	0.40	34
131: leermoeilijkheden	0.40	0.33	0.36	0.32	23
132: taalonderwijs	0.80	0.24	0.36	0.46	64
133: wereldoriëntatie	0.40	0.33	0.36	0.29	24
134: PABO	0.25	0.60	0.35	0.28	23
135: studietoetsen ; algemeen	0.27	0.50	0.35	0.29	24
136: wiskunde-onderwijs	0.27	0.50	0.35	0.31	32
137: onderwijsvernieuwing	1.00	0.21	0.35	0.33	48
138: examens	0.26	0.50	0.34	0.23	23
139: gezondheidsvoorlichting	0.23	0.71	0.34	0.13	24
140: leerlingbegeleiding	0.26	0.50	0.34	0.15	57
141: beroepskeuze	0.23	0.64	0.33	0.16	37
142: omgangsvormen	0.27	0.43	0.33	0.28	21
143: intercultureel onderwijs	0.30	0.33	0.32	0.29	41
144: sociale pedagogiek	0.33	0.30	0.32	0.37	28
145: jeugdbeleid	0.19	1.00	0.31	0.14	23
146: leesvaardigheid	0.25	0.40	0.31	0.28	24
147: leerplannen	0.32	0.30	0.31	0.20	53
148: tweetaligheid	0.29	0.33	0.31	0.31	24
149: orthopedagogiek	0.25	0.38	0.30	0.18	58
150: hogescholen	0.21	0.50	0.30	0.23	20

Table 2: **KB education test data.** Classification based on language modeling.

Category	Precision	Recall	F-measure	Threshold	N train
151: lager beroepsonderwijs	0.18	0.80	0.30	0.21	28
152: voortgezet onderwijs	0.26	0.34	0.30	0.13	124
153: basisonderwijs	0.20	0.51	0.29	0.03	180
154: geneeskunde ; opleiding	0.50	0.20	0.29	0.44	20
155: onderwijsstatistiek	0.50	0.20	0.29	0.55	20
156: wetenschapsbeleid	0.50	0.20	0.29	0.39	23
157: ervaringsleren	0.50	0.20	0.29	0.55	21
158: gesprekstechniek	0.50	0.20	0.29	1.00	21
159: onderwijssociologie	0.25	0.33	0.29	0.16	35
160: middelbaar technisch onderwijs	0.50	0.20	0.29	0.51	25
161: sociale psychologie	0.50	0.20	0.29	0.59	21
162: muziekonderwijs	0.50	0.20	0.29	0.47	22
163: antroposofisch onderwijs	0.50	0.20	0.29	0.39	20
164: permanente educatie	0.50	0.20	0.29	0.64	20
165: studiefinanciering	0.50	0.20	0.29	0.33	20
166: hoger beroepsonderwijs	0.22	0.40	0.29	0.28	31
167: gehandicapte kinderen	0.29	0.29	0.29	0.36	29
168: schoolexamens ; vmbo/mavo/havo/vwo	0.17	0.83	0.28	0.14	23
169: studievoorlichting	0.29	0.25	0.27	0.47	30
170: jeugdboek	0.22	0.33	0.27	0.32	28
171: onderwijs	0.31	0.24	0.27	0.18	46
172: gehandicaptenzorg	0.21	0.33	0.26	0.27	21
173: leraren	0.17	0.60	0.26	0.03	30
174: voortgezet onderwijs ; tweede fase	0.33	0.20	0.25	0.52	21
175: studenten	0.50	0.17	0.25	0.45	24
176: taalstoornissen	0.33	0.20	0.25	0.43	22
177: lezen	0.17	0.40	0.24	0.21	26
178: economieonderwijs	0.17	0.40	0.24	0.46	26
179: hoger technisch onderwijs	0.17	0.40	0.24	0.21	21
180: ouderschap	0.15	0.60	0.24	0.12	28
181: lesgeven	0.16	0.40	0.23	0.09	22
182: gedragsstoornissen	0.14	0.55	0.23	0.07	25
183: gehandicapten ; onderwijs	0.25	0.20	0.22	0.27	23
184: kinderpsychologie	0.18	0.30	0.22	0.19	23
185: leerdoelen	0.50	0.14	0.22	0.74	22
186: leerstoornissen	0.50	0.14	0.22	0.86	27
187: onderwijsresearch	0.29	0.18	0.22	0.19	50
188: onderwijsbeleid	0.16	0.33	0.21	0.10	65
189: beroepsopleidingen	0.18	0.25	0.21	0.29	29
190: studiebegeleiding	0.17	0.29	0.21	0.42	26
191: lager onderwijs	0.20	0.20	0.20	0.28	20
192: schrijven	0.12	0.50	0.20	0.20	26
193: schoolwerkplannen	0.13	0.40	0.20	0.27	26
194: onderwijsconomie	0.33	0.14	0.20	0.45	25
195: taalbeheersing	0.17	0.25	0.20	0.27	28
196: wetenschappelijk onderwijs	0.17	0.22	0.20	0.11	63
197: studeren	0.14	0.33	0.19	0.19	35
198: spelen	0.13	0.33	0.19	0.23	24
199: kunstzinnige vorming	0.20	0.17	0.18	0.39	31
200: beroepsonderwijs	0.15	0.19	0.17	0.16	44
201: kinderverzorging	0.50	0.10	0.17	0.70	24
202: gezinssociologie	0.17	0.17	0.17	0.36	22
203: pedagogische academies	0.11	0.33	0.16	0.11	21
204: voorbereidend wetenschappelijk onderwijs	0.14	0.18	0.16	0.26	37
205: onderwijsbegeleiding	0.20	0.12	0.15	0.36	31
206: geschiedenisonderwijs	0.10	0.33	0.15	0.15	28
207: studielessen	0.08	0.80	0.15	0.08	20
208: jeugd	0.09	0.38	0.15	0.12	23
209: leermiddelen	0.14	0.14	0.14	0.52	28
210: onderwijsstechnologie	0.07	0.40	0.12	0.08	24
211: onderwijs en emancipatie	0.06	0.44	0.11	0.03	25
212: hoger algemeen vormend onderwijs	0.08	0.11	0.10	0.32	28
213: speelleervormen	0.06	0.20	0.09	0.18	24
214: wetenschap en samenleving	0.05	0.67	0.09	-0.02	20
215: leerpsychologie	0.05	0.17	0.08	0.13	22
216: leerprocessen	0.04	0.43	0.08	0.02	27
217: schooltoets	0.03	0.20	0.05	0.11	20
218: ontwikkelingspsychologie	0.03	0.33	0.05	0.02	24
219: onderwijskunde	0.01	0.22	0.03	-0.01	35
220: leermiddelen ; lezen	0.01	1.00	0.01	-0.08	33
221: sociaal-pedagogisch onderwijs	0.00	0.00	0.00	0.00	27

Table 3: **KB education test data.** Classification based on language modeling.

Category	Precision	Recall	F-measure	Threshold	N train
1: kinderboek	1.00	1.00	1.00	0.43	23
2: lerarenopleidingen	0.89	1.00	0.94	0.14	29
3: babyverzorging	0.86	1.00	0.92	0.56	28
4: stages	1.00	0.86	0.92	0.38	24
5: autisme	1.00	0.83	0.91	0.24	20
6: universiteiten	1.00	0.83	0.91	0.09	24
7: kinderspelen	1.00	0.83	0.91	0.57	25
8: brede scholen	0.83	1.00	0.91	0.27	20
9: vaderschap	1.00	0.80	0.89	0.60	20
10: peuterspeelgroepen	1.00	0.80	0.89	0.29	22
11: montessori-onderwijs	1.00	0.80	0.89	0.89	20
12: gymnasia	1.00	0.80	0.89	0.28	20
13: wiskunde-onderwijs	1.00	0.75	0.86	0.10	32
14: intercultureel onderwijs	0.80	0.89	0.84	0.20	41
15: scripties	0.83	0.83	0.83	0.64	21
16: middelbaar dienstverlenings- en gezondheidszorgonderwijs	0.83	0.83	0.83	0.56	28
17: schoolexamens ; vmbo/mavo/havo/vwo	0.83	0.83	0.83	0.58	23
18: verkeerseducatie	1.00	0.71	0.83	0.33	22
19: orthodidactiek	0.88	0.78	0.82	0.20	26
20: literatuuronderwijs	1.00	0.70	0.82	0.25	26
21: adoptie	0.80	0.80	0.80	0.31	20
22: schoolbesturen	0.67	1.00	0.80	0.26	28
23: opvoeding	0.77	0.83	0.80	0.02	120
24: speltherapie	1.00	0.62	0.77	0.65	23
25: middenschool	0.62	1.00	0.77	0.25	20
26: jeugdhulpverlening	0.73	0.80	0.76	0.09	38
27: onderwijswetgeving	0.73	0.80	0.76	0.20	38
28: volwasseneneducatie	0.71	0.79	0.75	0.03	51
29: leerlingwezen	0.69	0.82	0.75	0.11	27
30: onderwijsvernieuwing	0.71	0.79	0.75	0.05	48
31: kinderen	1.00	0.60	0.75	0.37	21
32: medisch onderwijs	1.00	0.60	0.75	0.46	29
33: pesten	1.00	0.60	0.75	0.28	23
34: logopedie	1.00	0.60	0.75	0.38	22
35: openbaar onderwijs	1.00	0.57	0.73	0.31	24
36: schoolverzuim	1.00	0.57	0.73	0.29	25
37: begaafdheid	0.80	0.67	0.73	0.24	22
38: studiehuis	0.67	0.80	0.73	0.30	22
39: informaticaonderwijs	0.67	0.80	0.73	0.31	33
40: huiswerk	0.80	0.67	0.73	0.34	21
41: leraren	0.67	0.80	0.73	0.04	30
42: speciaal onderwijs	0.70	0.74	0.72	0.16	65
43: leerplannen	0.74	0.70	0.72	0.22	53
44: kinderdagverblijven	0.56	1.00	0.71	0.21	26
45: bijzonder onderwijs	0.83	0.62	0.71	0.27	24
46: omgangsvormen	0.71	0.71	0.71	0.21	21
47: management ; onderwijs	0.71	0.71	0.71	0.35	31
48: verpleging ; opleiding	0.62	0.83	0.71	0.13	22
49: voorbereidend beroepsonderwijs	0.62	0.83	0.71	0.29	24
50: medezeggenschap in het onderwijs	0.83	0.62	0.71	0.40	24
51: voortgezet onderwijs	0.63	0.82	0.71	0.02	124
52: sociale pedagogiek	0.86	0.60	0.71	0.16	28
53: computers in het onderwijs	0.93	0.57	0.70	0.14	63
54: kinderopvang	0.78	0.64	0.70	0.10	37
55: Nederlands voor anderstaligen	0.70	0.70	0.70	0.10	25
56: voorbereidend wetenschappelijk onderwijs	0.67	0.73	0.70	0.31	37
57: hoger onderwijs	0.89	0.57	0.70	0.30	68
58: leerlingbegeleiding	0.64	0.75	0.69	0.08	57
59: basisvorming	0.67	0.71	0.69	0.13	45
60: taalonderwijs	0.73	0.65	0.69	0.04	64
61: basisonderwijs	0.60	0.80	0.69	-0.02	180
62: milieueducatie	0.65	0.69	0.67	0.10	46
63: orthopedagogiek	0.73	0.62	0.67	0.08	58
64: schrijfonderwijs	0.62	0.71	0.67	0.19	23
65: landbouwonderwijs	1.00	0.50	0.67	0.21	25
66: leesstoornissen	0.57	0.80	0.67	0.61	25
67: sportonderwijs	0.71	0.62	0.67	0.45	21
68: onderwijsstimulering	0.78	0.58	0.67	0.17	33
69: biologie-onderwijs	0.80	0.57	0.67	1.80	26
70: leermiddelen ; lezen	0.50	1.00	0.67	0.18	33
71: onderwijs aan anderstaligen	0.67	0.67	0.67	0.24	36
72: schoolverlaters	0.67	0.67	0.67	0.15	26
73: pedagogiek	0.70	0.64	0.67	0.06	56
74: voorbereidend middelbaar beroepsonderwijs	0.67	0.67	0.67	0.22	27
75: dyslexie	0.62	0.71	0.67	0.22	20

Table 4: KB education test data. Classification based on language modeling.

Category	Precision	Recall	F-measure	Threshold	N train
76: scheikunde-onderwijs	1.00	0.50	0.67	0.29	26
77: buitenschoolse activiteiten	1.00	0.50	0.67	0.26	23
78: gehandicapten ; onderwijs	0.75	0.60	0.67	0.22	23
79: onderwijsstatistiek	0.75	0.60	0.67	0.54	20
80: recht ; opleiding	0.75	0.60	0.67	0.51	20
81: gesprekstechniek	0.75	0.60	0.67	0.85	21
82: studielessen	0.75	0.60	0.67	0.79	20
83: agogie	0.75	0.60	0.67	0.72	26
84: leesonderwijs	0.67	0.62	0.64	0.22	36
85: hoger algemeen vormend onderwijs	0.54	0.78	0.64	0.26	28
86: onderwijsbegeleiding	0.50	0.88	0.64	0.07	31
87: gedragsstoornissen	0.64	0.64	0.64	0.17	25
88: verstandelijk gehandicapten	0.75	0.55	0.63	0.18	28
89: godsdienstonderwijs	0.67	0.60	0.63	0.20	28
90: middelbaar algemeen vormend onderwijs	0.50	0.83	0.62	0.24	24
91: onderwijs	0.67	0.59	0.62	0.02	46
92: beroepsonderwijs	0.58	0.67	0.62	0.03	44
93: seksuele voorlichting	0.50	0.80	0.62	0.31	21
94: beeldende vorming	0.50	0.80	0.62	0.62	23
95: maatschappijleer	0.53	0.73	0.62	0.13	29
96: lichamelijk gehandicapten	0.67	0.57	0.62	0.32	22
97: leesbevordering	0.67	0.57	0.62	0.26	25
98: rekenonderwijs	0.64	0.58	0.61	0.45	37
99: leermethoden	1.00	0.44	0.61	0.31	27
100: onderwijsrecht	0.75	0.50	0.60	0.29	21
101: katholiek onderwijs	0.75	0.50	0.60	0.28	22
102: schoolloopbaan	1.00	0.43	0.60	0.39	25
103: jeugdwerk	1.00	0.43	0.60	0.33	23
104: leermiddelen ; omgangskunde	0.60	0.60	0.60	0.55	20
105: vakkenpakketkeuze	0.75	0.50	0.60	0.30	23
106: middelbaar beroepsonderwijs	0.45	0.90	0.60	0.11	39
107: geelofsopvoeding	0.60	0.60	0.60	0.39	23
108: didactiek	0.62	0.57	0.59	0.05	50
109: lesgeven	0.71	0.50	0.59	0.02	22
110: bewegingsonderwijs	0.58	0.58	0.58	0.09	35
111: lichamelijke opvoeding	0.67	0.50	0.57	0.32	22
112: kinderverzorging	1.00	0.40	0.57	0.69	24
113: basiseducatie	0.57	0.57	0.57	0.31	35
114: vormingswerk	0.41	0.88	0.56	0.06	24
115: Engelse taalkunde	0.45	0.71	0.56	0.17	24
116: jeugdbeleid	0.38	1.00	0.56	0.15	23
117: Nederlandse taal ; onderwijs	0.50	0.62	0.55	0.30	32
118: leerplicht	0.60	0.50	0.55	0.23	22
119: onderwijstechnologie	0.50	0.60	0.55	0.14	24
120: geografie-onderwijs	0.60	0.50	0.55	0.92	24
121: beroepskeuze	0.55	0.55	0.55	0.16	37
122: gezondheidsvoorlichting	0.75	0.43	0.55	0.37	24
123: peuters	1.00	0.36	0.53	0.21	24
124: kleuteronderwijs	0.44	0.67	0.53	0.17	27
125: probleemgestuurd onderwijs	0.38	0.83	0.53	0.26	22
126: studietoetsen ; algemeen	0.38	0.83	0.53	0.25	24
127: taalbeheersing	0.45	0.62	0.53	-0.00	28
128: ouderschap	0.56	0.50	0.53	0.12	28
129: gezondheidszorg	1.00	0.33	0.50	0.40	21
130: technisch onderwijs	0.44	0.57	0.50	0.07	23
131: jeugdgezondheidszorg	0.75	0.38	0.50	0.33	22
132: puberteit	0.67	0.40	0.50	0.61	21
133: pedagogische academies	0.50	0.50	0.50	0.17	21
134: taalverwerving	0.75	0.38	0.50	0.25	27
135: educatieve software	0.43	0.60	0.50	0.45	23
136: kinderpsychologie	0.67	0.40	0.50	0.22	23
137: schoolgebouwen	1.00	0.33	0.50	0.47	21
138: natuuronderwijs	0.50	0.50	0.50	0.29	23
139: afstandsonderwijs	0.67	0.40	0.50	0.40	24
140: christelijk onderwijs	0.67	0.40	0.50	0.15	22
141: hogescholen	0.40	0.67	0.50	0.11	20
142: school-maatschappelijk werk	0.67	0.40	0.50	0.32	22
143: projectonderwijs	0.50	0.50	0.50	0.25	21
144: modulair onderwijs	0.67	0.40	0.50	0.53	23
145: gezinsbegeleiding	0.60	0.43	0.50	0.25	22
146: onderwijsbeleid	0.41	0.60	0.49	0.04	65
147: onderwijs en emancipatie	0.35	0.78	0.48	0.02	25
148: tekenonderwijs	0.33	0.80	0.47	0.30	21
149: hoger beroepsonderwijs	0.38	0.60	0.46	0.16	31
150: spelling	0.43	0.50	0.46	0.15	24

Table 5: **KB education test data.** Classification based on language modeling.

Category	Precision	Recall	F-measure	Threshold	N train
151: natuurkunde-onderwijs	0.50	0.43	0.46	0.38	27
152: onderwijssociologie	0.43	0.50	0.46	0.08	35
153: onderwijsseconomie	0.50	0.43	0.46	0.22	25
154: kindermishandeling	0.43	0.50	0.46	0.25	20
155: wetenschappelijk onderwijs	0.47	0.44	0.46	-0.02	63
156: onderwijsresearch	0.45	0.45	0.45	0.15	50
157: onderwijspsychologie	0.33	0.71	0.45	0.10	24
158: studievoorzichting	0.30	0.88	0.45	0.09	30
159: examens	0.50	0.40	0.44	0.21	23
160: hulpverlening	1.00	0.29	0.44	0.47	23
161: schoolwerkplannen	0.50	0.40	0.44	0.47	26
162: leerstoornissen	1.00	0.29	0.44	0.59	27
163: scholengemeenschappen	0.67	0.33	0.44	0.44	21
164: filosofieonderwijs	0.67	0.33	0.44	0.84	21
165: dramatische vorming	0.33	0.67	0.44	0.21	27
166: tweetaligheid	0.67	0.33	0.44	0.23	24
167: beroepsvoorzichting	0.38	0.50	0.43	0.29	36
168: gezondheidsopvoeding	0.43	0.43	0.43	0.27	34
169: natuurbeschermingseducatie	0.33	0.60	0.43	0.35	26
170: leermoeilijkheden	0.31	0.67	0.42	0.15	23
171: lager beroepsonderwijs	0.29	0.80	0.42	0.23	28
172: baby's	0.40	0.40	0.40	0.44	26
173: ouderparticipatie	0.40	0.40	0.40	0.21	26
174: jeugdboek	0.33	0.50	0.40	0.17	28
175: kleuters	0.33	0.50	0.40	0.26	27
176: studeren	0.50	0.33	0.40	0.19	35
177: remedial teaching	1.00	0.25	0.40	0.99	23
178: bedrijfsopleidingen	0.50	0.33	0.40	0.24	30
179: permanente educatie	0.30	0.60	0.40	0.22	20
180: schoolkeuze	0.33	0.50	0.40	0.10	23
181: wereldoriëntatie	0.33	0.50	0.40	0.22	24
182: kunstzinnige vorming	0.50	0.33	0.40	0.38	31
183: beroepsopleidingen	0.43	0.38	0.40	0.12	29
184: wetenschappelijk onderzoek	0.38	0.43	0.40	0.21	30
185: jeugd	0.31	0.50	0.38	0.10	23
186: gehandicaptenzorg	0.43	0.33	0.38	0.21	21
187: onderwijskunde	1.00	0.22	0.36	0.09	35
188: ontwikkelingspsychologie	0.40	0.33	0.36	0.15	24
189: kunstonderwijs	0.40	0.33	0.36	0.38	20
190: PABO	0.25	0.60	0.35	0.28	23
191: leerprocessen	0.25	0.57	0.35	0.02	27
192: gehandicapte kinderen	0.40	0.29	0.33	0.25	29
193: geschiedenisonderwijs	0.25	0.50	0.33	0.15	28
194: leermiddelen	0.27	0.43	0.33	0.05	28
195: spelen	0.29	0.33	0.31	0.22	24
196: speelleervormen	0.24	0.40	0.30	0.09	24
197: voortgezet onderwijs ; tweede fase	0.50	0.20	0.29	0.47	21
198: economieonderwijs	0.50	0.20	0.29	0.90	26
199: wetenschapsbeleid	0.50	0.20	0.29	0.37	23
200: ervaringsleren	0.50	0.20	0.29	0.59	21
201: middelbaar technisch onderwijs	0.50	0.20	0.29	0.57	25
202: sociale psychologie	0.50	0.20	0.29	0.52	21
203: muziekonderwijs	0.50	0.20	0.29	0.36	22
204: antroposofisch onderwijs	0.50	0.20	0.29	0.40	20
205: studiefinanciering	0.50	0.20	0.29	0.33	20
206: gezinsociologie	0.20	0.50	0.29	0.10	22
207: leesvaardigheid	0.22	0.40	0.29	0.27	24
208: leerdoelen	0.21	0.43	0.29	0.09	22
209: hoger technisch onderwijs	0.22	0.40	0.29	0.19	21
210: taalstoornissen	0.17	0.60	0.26	0.16	22
211: geneeskunde ; opleiding	0.33	0.20	0.25	0.38	20
212: studenten	0.50	0.17	0.25	0.39	24
213: lager onderwijs	0.33	0.20	0.25	0.31	20
214: alfabetiseringsprojecten	0.18	0.33	0.24	0.19	23
215: studiebegeleiding	0.20	0.29	0.24	0.38	26
216: lezen	0.13	0.60	0.21	0.12	26
217: schrijven	0.13	0.50	0.21	0.19	26
218: wetenschap en samenleving	0.11	0.67	0.19	0.06	20
219: leerpsychologie	0.10	0.67	0.17	-0.02	22
220: schooltoets	0.03	0.20	0.05	0.11	20
221: sociaal-pedagogisch onderwijs	0.00	0.00	0.00	0.00	27

Table 6: **KB education test data.** Classification based on language modeling.