Identifying Literary Texts with Bigrams

Andreas.van.Cranenburgh@Huygens.knaw.nl
Corina Koolen — C.W.Koolen@UvA.nl

Institute for Logic, Language and Computation
University of Amsterdam

Survey

As part of the project *The Riddle of Literary Quality*, a large (13k respondents) survey was held of the Dutch reading public (cf. www.hetnationalelezersonderzoek.nl).
Participants rated books they had read on two scales:

Literariness 1–7 scale
Bad/good 1–7 scale

Method

- linear Support Vector Machine (SVM)
- trained on bigram counts and mean of ratings.
- 10-fold cross-validation
- Tasks:

  Classification binary classification after applying threshold to scale (predict high vs. low rating)
  Regression predict numeric rating

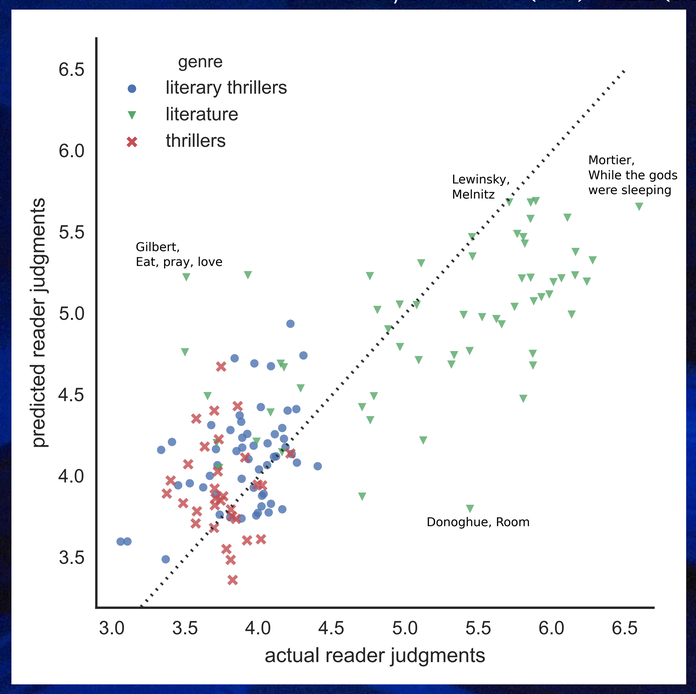
Results

Classification results (accuracy; percentage correct):

Features	Literary	Bad/good
Content	90.4	63.7
Style	89.0	63.0

Evaluation of the regression models; R^2 scores (percentage of variation explained), root mean squared error in parentheses (1–7).

Features	Literary	Bad/Good
Content	61.3 (0.65)	33.5 (0.49)
Style	57.0 (0.67)	22.2 (0.52)



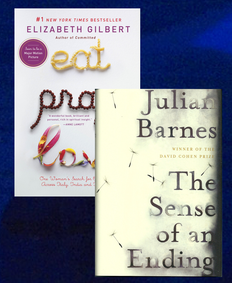
Regression results for literariness using content bigrams.

Texts

A subset (cf. Corpus box) of the Riddle corpus of 400 recent, successful novels, containing thrillers, literary thrillers, and literary novels.

Features: Bag-of-Words model based on bigrams that appear in 60–90 % of texts.

Content bigrams all words.
Style bigrams function words plus content words replaced with POS tags.



Interpretation

Weights of the top 20 most important content features (above) and top 5 most important style features (below), indicative of literary (left), and non-literary texts (right).

weight	literary features, content		weight	non-literary features, content	
12.1	*de oorlog*	the war	-6.1	*de moeder*	the mother
8.1	*het bos*	the forest	-5.1	*keek op*	looked up
8.1	*de winter*	the winter	-4.9	*mijn hoofd*	my head
6.6	*de dokter*	the doctor	-4.9	*haar moeder*	her mother
5.8	*zo veel*	so much	-4.7	*mijn ogen*	my eyes
4.8	*nog altijd*	yet still	-4.7	*ze keek*	she looked
4.5	*de meisjes*	the girls	-4.5	*mobiele telefoon*	mobile telephone
4.3	*zijn vader*	his father	-4.2	*de moord*	the murder
4.0	*mijn dochter*	my daughter	-4.0	*even later*	a while later
3.9	*het boek*	the book	-3.8	*nu toe*	(until) now
3.8	*de trein*	the train	-3.5	*zag ze*	she saw
3.7	*hij hem*	he him	-3.4	*ik voel*	I feel
3.7	*naar mij*	at me	-3.3	*mijn man*	my husband
3.5	*zegt dat*	says that	-3.2	*tot haar*	to her
3.5	*het land*	the land	-3.2	*het gebouw*	the building
3.5	*een sigaret*	a cigarette	-3.2	*liep naar*	walked to
3.4	*haar vader*	her father	-3.1	*we weten*	we know
3.4	*een boek*	a book	-3.1	*enige wat*	only thing
3.2	*de winkel*	the shop	-3.1	*en dus*	and so
3.1	*elke keer*	each time	-3.0	*in godsnaam*	in god's name

weight	literary features, style		weight	non-literary features, style	
21.8	*! WW*	! VERB ,	-13.8	*nu toe*	until now
20.5	*u ,*	you (FORMAL) ,	-13.4	*en dus*	and so
18.0	*haar haar*	her her	-13.4	*achter me*	behind me
16.5	*SPEC :*	NAME :	-13.2	*terwijl ik*	while I
15.4	*worden ik*	become I	-13.1	*tot nu*	until now

- Some features relate to topics of genres (*war, murder*)
- 'Timeless' character of literary features (*book, letter*) vs. modern technology in non-literary books (*computer, mobile phone*).
- More markers of colloquial language in non-literary texts, in both the content and style features

Conclusions

Perceived literariness can be predicted, to a surprisingly large extent.

Literariness is easier to predict than general quality.

Both content features and style features perform well.

The regression results show that the model does not just recognize genres, but also accounts for literariness differences *within* genres.

How to describe the nature of literariness still an open question.

References

This work is part of The Riddle of Literary Quality, a project supported by the Royal Netherlands Academy of Arts and Sciences through the Computational Humanities Program.

Paper Andreas van Cranenburgh, Corina Koolen (2015). Identifying Literary Novels with Bigrams. Workshop on Computational Linguistics for Literature, Denver, Colorado, USA.

Project The Riddle of Literary Quality; cf. http://literaryquality.huygens.knaw.nl

Painting Christine Bittremieux (2007), Untitled. Oil on canvas. 60×80 cm. www.bittremieux.nl

Corpus

Legend: literary novel, thriller, literary thriller.

Appel, René Van twee kanten
Auel, Jean Marie Het lied van de grotten
Avallone, Silvia Staal
Baldacci, David De provocatie, De rechtvaardigen, De zesde man, Die zomer, Familieverraad, Geniaal geheim, In het geheim, Onschuldig, Rechteloos, Verlos ons van het kwaad
Bernlef, J. De een zijn dood, Geleende levens
Binet, Laurent HhhH
Blum, Jenna In tweestrijd
Clancy, Tom In het vizier, Op leven en dood, De ogen van de vijand
Coben, Harlan Levenslijn
Cronin, Justin De oversteek
Cussler, Clive Dodenschip, Wassende maan
Dijkzeul, Lieneke Koude lente, Verloren zoon
Dis, Adriaan van Tikkop
Donoghue, Emma Kamer
Dorrestein, Renate De stiefmoeder
Durlacher, Jessica De held
Enquist, Anna De verdovers
Evans, Nicholas De vergeving
Forbes, Elena Sterf met mij
Forsyth, Frederick De cobra
Fragoso, Margaux Tijger, tijger
French, Nicci Blauwe maandag, Medeplichtig
Galen, Alex van Süskind
George, Elizabeth Een duister vermoeden
Gerritsen, Tess De Mefisto Club, Het aandenken, Het stille meisje, Koud hart, Sneeuwval, Verdwijn
Gilbert, Elizabeth Eten, bidden, beminnen
Giordano, Paolo De eenzaamheid van de priemgetallen
Giphart, Ronald IJsland
Gudenkauf, Heather In stilte gehuld
Hannah, Sophie Kleine meid, Moederziel
Harbach, Chad De kunst van het veldspel
Hart, Maarten 't Verlovingstijd
Haynes, Elizabeth Waarheen je ook vlucht
Heijden, A.F.Th van der Tonio
Hill, Lawrence Het negerboek
Hoag, Tami Dieper dan de doden
Hodgkinson, Amanda Britannia Road 22
Hosseini, Khaled Duizend schitterende zonnen
Irving, John De laatste nacht in Twisted River, In een mens
James, Erica Schaduwleven, Zussen voor altijd
Janssen, Roel De tiende vrouw
Japin, Arthur De overgave, Vaslav
Kepler, Lars Hypnose
Kluun Haantjes
Koch, Herman Het diner, Zomerhuis met zwembad
Koryta, Michael Begraven
Krauss, Nicole Het grote huis
Kroonenberg, Yvonne De familielightblues
Lackberg, Camilla IJsprinses
Lapidus, Jens Bloedlink, Snel geld, Val dood
Larsson, Stieg De vrouw die met vuur speelde, Gerechtigheid, Mannen die vrouwen haten
Launspach, Rik 1953 (De Storm)
Lewinsky, Charles De verborgen geschiedenis van Courtillon, Het lot van de familie Meijer
Läckberg, Camilla Steenhouwer, Vuurtorenwachter
Mastras, George Tranen over Kashmir
McCoy, Sarah De bakkersdochter
McFadyen, Cody Tijd om te sterven
McNab, Andy Oorlogswond
Moor, Marente de De Nederlandse maagd
Moor, Margriet de De schilder en het meisje
Mortier, Erwin Godenslaap
Nesbo, Jo De schim, De sneeuwman, Het pantserhart
Noort, Saskia Afgunst, De eetclub, De verbouwing, Koorts
Patterson, James De affaire, Hitte
Pauw, Marion Daglicht, Jetset
Pick, Alison Donderdagskind
Picoult, Jodi Negentien minuten
Roberts, Nora Erfbetoon
Robotham, Michael Gebroken
Rose, Karen Moord voor mij
Rosenboom, Thomas Zoete mond
Rosenfeldt, Hjorth Wat verborgen is
Rosnay, Tatiana de Het appartement, Het huis waar jij van hield
Ruiz Zafón, Carlos De gevangene van de hemel, De schaduw van de wind
Sansom, Christopher John Winter in Madrid
Scholten, Jaap Kameraad Baron
Siebelink, Jan Het lichaam van Clara, Oscar
Slaughter, Karin Genadeloos, Genesis, Gevallen, Onaantastbaar, Verbroken
Stevens, Chevy Vermist
Terlouw, Jan / Terlouw, Sanne Hellehonden
Trussoni, Danielle Het uur van de engelen
Verhoef, Esther Alles te verliezen, Close-up, Déjà vu, Erken mij, Tegenlicht
Verhulst, Dimitri De laatste liefde van mijn moeder
Vermeer, Suzanne Après-ski, Cruise, De suite, De vlucht, Zomertijd
Visser, Judith Stuk
Vlugt, Simone van der Blauw water, Het laatste offer, In mijn dromen, Op klaarlichte dag
Watson, S.J. Voor ik ga slapen
Winter, Leon de Recht op terugkeer, VSV of daden van onbaatzuchtigheid
Yalom, Irvin D. Het raadsel spinoza