An exploratory comparison between Latent Semantic Analysis and human based tagging for similarity of stanza in a small corpus of Dutch poems.

Michael Janus

Jheronimus Academy of Data Science, ‘s Hertogenbosch, The Netherlands

Introduction

A common, straight-forward way to classify texts is by tagging them with keywords which indicate their topic(s). However, such method requires human intervention for the labelling, which can be error-prone and subjective to opinion.

Alternatively, Latent Semantic Analyses (LSA), as first described by Deerwester *et al.* [1], is a computational linguistic technique used in natural language processing to analyze the contextual relationships between documents.

In this study, LSA is applied on a small corpus of Dutch poems and assessed if LSA can replace the human based tagging that is currently used for categorizing in this corpus.

Methodology

The dataset

The corpus consists of 231 short poems (*stanza),* ranging from 4 to 10 lines and were written by amateur poets who submitted their contributions to a Dutch rhyming web site [2]. Upon approval of each submission, the web site’s moderator has tagged each poem with keywords that describe its topic(s) in order to facilitate searching. However, as the process of tagging has been spread out over time and subject to human judgement, the tags are not per se complete or consistent.

Pre-processing

Before the analysis, the corpus was cleaned from HTML tags and then vectorized with TF-IDF [3], an encoding that normalizes the frequency of tokens in a document with respect to the rest of the corpus. Also the associated tags have also been vectorized with TF-IDF, but independent of the stanza, forming its own TF-IDF vector space. No stop-word removal, stemming or lemmatization was applied on the stanza or tags.

Modelling & Similarity Assessment

A LSA model was created by applying Single Value Decomposition (SVD) on the vectorized corpus. From this model, the pair-wise similarities between poems are measured by cosine distance. In addition, the cosine distance of the TF-IDF vectors without SVD was measured. For comparison, a similar LSA model was created for the human based tags. Finally, an existing LSA model - trained on a large corpus of the Dutch language - has been used to assess the cosine distance between poems.

All results are compared to the original, tag based categorization.

Evaluation method

As there is no guarantee that the human based tags are complete or consistent, there is no real ground truth or golden standard for this data set. Hence, as this study is a first exploration, the results are assessed by human judgement and compared to the tag based categorization. To limit the search space, only a number of poems have been hand-picked to assess a result in detail.

Toolsets used

All the analyses were performed in Python, using the machine learning toolkit Sklearn [3], except for the assessment with the pre-trained LSA model, for which the web tool “Linguistic tools for Cognitive Science” of the University of Tilburg [6] were used.

Results

For a first impression of the LSA model applied on the corpus, the first two SVD components of each poem are shown in figure 1. A quick scan of this graph does not reveal significant grouping of some sort, but the density of the data points makes it hard to assess. Hence, similar plots were created for pairs of topics, as depicted in figure 2. But also in these plots, no clear separation of topics is observed.

To assess the similarities of the applied techniques, the results for a selection of poems is shown in Table 1. From this table, it is evident that nor the LSA model trained on the verses nor the pre-trained Dutch LSA model was capable of finding similar poems. Only the LSA model on the tags had some success in finding close matches.

Figure 1. *A plot of the first two SVD components of each poem, only showing their title.*

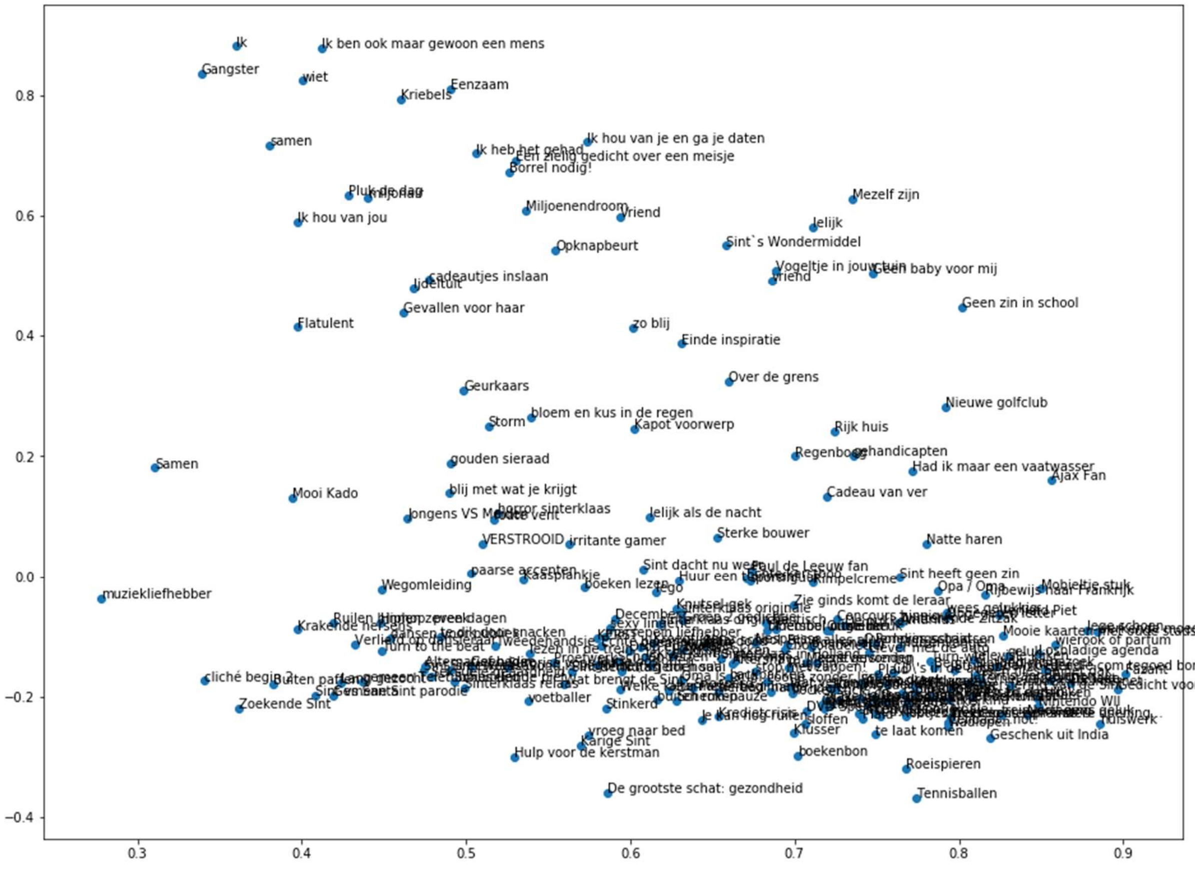


Figure 2. *Plots of the first two SVD components of poems in two different topics (only showing their title); a.) topics 'love' and 'money (love and money); b.) topics 'school' and 'sport'; c.) topics 'music' and 'driving cars'*

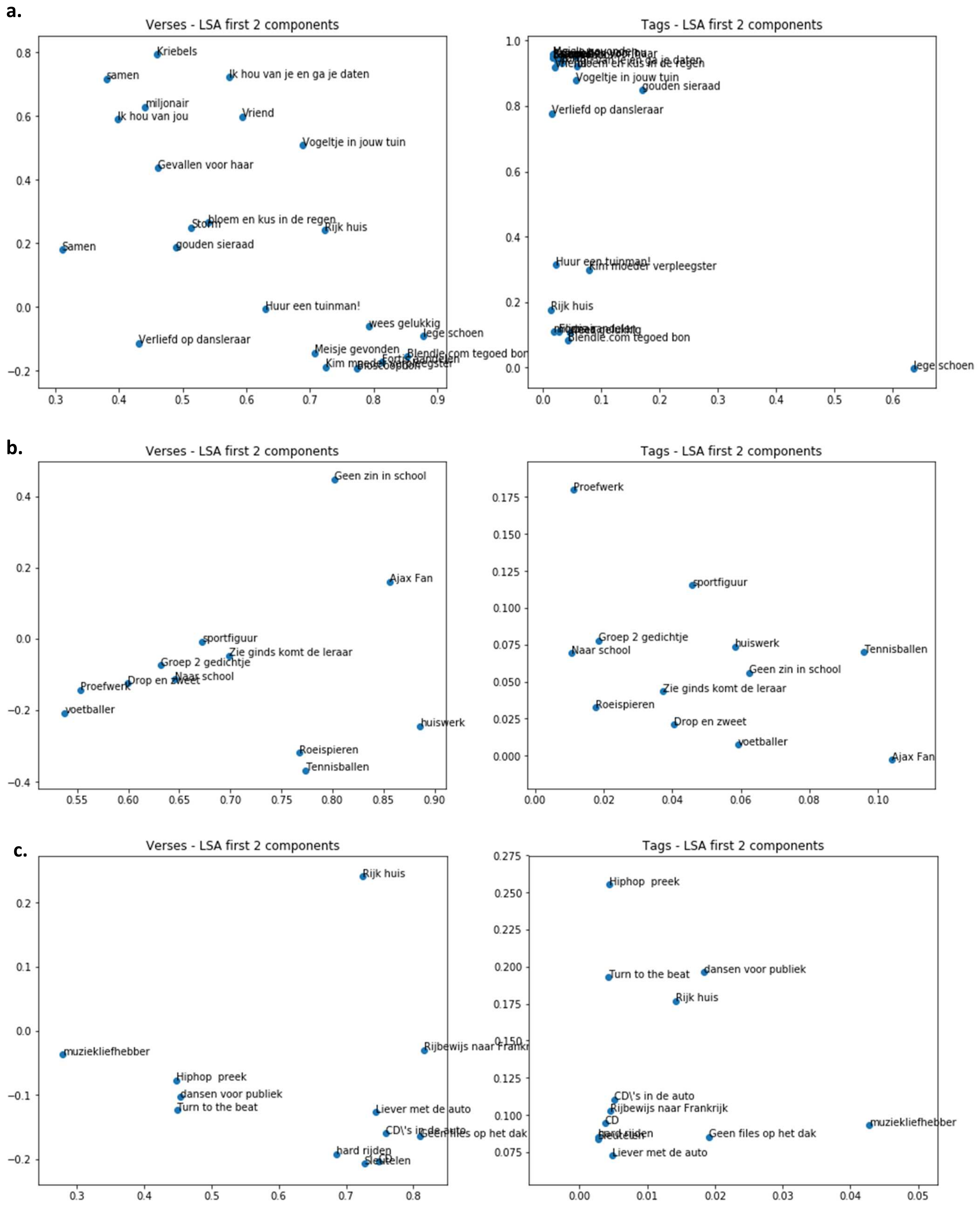
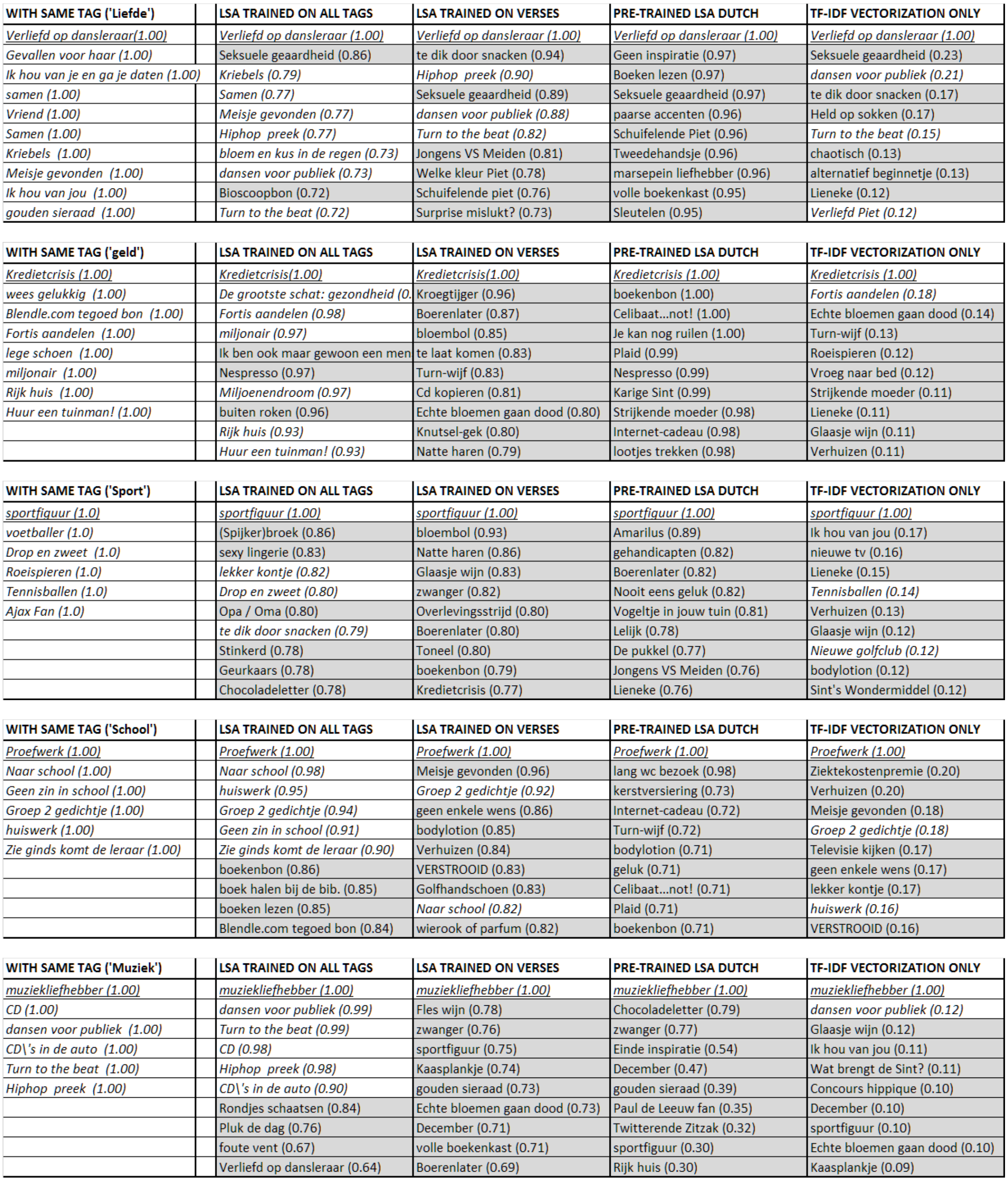


Table 1. *Results of finding 'closest' poems using different techniques for a number of hand-picked poems with distinct topics. For each technique, only the titles of the first ten best matches are shown with their similarity score between parenthesis. Titles of poems that are indeed a match (based on human judgement) are printed in italics; erroneous matches have been marked with a gray background.*



WITH SAME TAG ('Liefde')		LSA TRAINED ON ALL TAGS	LSA TRAINED ON VERSES	PRE-TRAINED LSA DUTCH	TF-IDF VECTORIZATION ONLY
Verliefd op dansleraar(1.00)		*Verliefd op dansleraar (1.00)*	*Verliefd op dansleraar (1.00)*	*Verliefd op dansleraar (1.00)*	*Verliefd op dansleraar (1.00)*
Gevallen voor haar (1.00)		Seksuele geaardheid (0.86)	te dik door snacken (0.94)	Geen inspiratie (0.97)	Seksuele geaardheid (0.23)
Ik hou van je en ga je daten (1.00)		*Kriebels (0.79)*	*Hiphop preek (0.90)*	Boeken lezen (0.97)	*dansen voor publiek (0.21)*
samen (1.00)		*Samen (0.77)*	Seksuele geaardheid (0.89)	Seksuele geaardheid (0.97)	te dik door snacken (0.17)
Vriend (1.00)		*Meisje gevonden (0.77)*	*dansen voor publiek (0.88)*	paarse accenten (0.96)	Held op sokken (0.17)
Samen (1.00)		*Hiphop preek (0.77)*	*Turn to the beat (0.82)*	Schuifelende Piet (0.96)	*Turn to the beat (0.15)*
Kriebels (1.00)		*bloem en kus in de regen (0.73)*	Jongens VS Meiden (0.81)	Tweedehandsje (0.96)	chaotisch (0.13)
Meisje gevonden (1.00)		*dansen voor publiek (0.73)*	Welke kleur Piet (0.78)	marsepein liefhebber (0.96)	alternatief beginnetje (0.13)
Ik hou van jou (1.00)		Bioscoopbon (0.72)	Schuifelende piet (0.76)	volle boekenkast (0.95)	Lieneke (0.12)
gouden sieraad (1.00)		*Turn to the beat (0.72)*	Surprise mislukt? (0.73)	Sleutelen (0.95)	*Verliefd Piet (0.12)*

WITH SAME TAG ('geld')		LSA TRAINED ON ALL TAGS	LSA TRAINED ON VERSES	PRE-TRAINED LSA DUTCH	TF-IDF VECTORIZATION ONLY
Kredietcrisis (1.00)		*Kredietcrisis(1.00)*	*Kredietcrisis(1.00)*	*Kredietcrisis (1.00)*	*Kredietcrisis (1.00)*
wees gelukkig (1.00)		*De grootste schat: gezondheid (0.*	Kroegtijger (0.96)	boekenbon (1.00)	*Fortis aandelen (0.18)*
Blendle.com tegoed bon (1.00)		*Fortis aandelen (0.98)*	Boerenlater (0.87)	Celibaat...not! (1.00)	Echte bloemen gaan dood (0.14)
Fortis aandelen (1.00)		*miljonair (0.97)*	bloembol (0.85)	Je kan nog ruilen (1.00)	Turn-wijf (0.13)
lege schoen (1.00)		Ik ben ook maar gewoon een men	te laat komen (0.83)	Plaid (0.99)	Roeispieren (0.12)
miljonair (1.00)		Nespresso (0.97)	Turn-wijf (0.83)	Nespresso (0.99)	Vroeg naar bed (0.12)
Rijk huis (1.00)		*Miljoenendroom (0.97)*	Cd kopieren (0.81)	Karige Sint (0.99)	Strijkende moeder (0.11)
Huur een tuinman! (1.00)		buiten roken (0.96)	Echte bloemen gaan dood (0.80)	Strijkende moeder (0.98)	Lieneke (0.11)
		Rijk huis (0.93)	Knutsel-gek (0.80)	Internet-cadeau (0.98)	Glaasje wijn (0.11)
		Huur een tuinman! (0.93)	Natte haren (0.79)	lootjes trekken (0.98)	Verhuizen (0.11)

WITH SAME TAG ('Sport')		LSA TRAINED ON ALL TAGS	LSA TRAINED ON VERSES	PRE-TRAINED LSA DUTCH	TF-IDF VECTORIZATION ONLY
sportfiguur (1.0)		*sportfiguur (1.00)*	*sportfiguur (1.00)*	*sportfiguur (1.00)*	*sportfiguur (1.00)*
voetballer (1.0)		(Spijker)broek (0.86)	bloembol (0.93)	Amarilus (0.89)	Ik hou van jou (0.17)
Drop en zweet (1.0)		sexy lingerie (0.83)	Natte haren (0.86)	gehandicapten (0.82)	nieuwe tv (0.16)
Roeispieren (1.0)		*lekker kontje (0.82)*	Glaasje wijn (0.83)	Boerenlater (0.82)	Lieneke (0.15)
Tennisballen (1.0)		*Drop en zweet (0.80)*	zwanger (0.82)	Nooit eens geluk (0.82)	*Tennisballen (0.14)*
Ajax Fan (1.0)		Opa / Oma (0.80)	Overlevingsstrijd (0.80)	Vogeltje in jouw tuin (0.81)	Verhuizen (0.13)
		te dik door snacken (0.79)	Boerenlater (0.80)	Lelijk (0.78)	Glaasje wijn (0.12)
		Stinkerd (0.78)	Toneel (0.80)	De pukkel (0.77)	*Nieuwe golfclub (0.12)*
		Geurkaars (0.78)	boekenbon (0.79)	Jongens VS Meiden (0.76)	bodylotion (0.12)
		Chocoladeletter (0.78)	Kredietcrisis (0.77)	Lieneke (0.76)	Sint's Wondermiddel (0.12)

WITH SAME TAG ('School')		LSA TRAINED ON ALL TAGS	LSA TRAINED ON VERSES	PRE-TRAINED LSA DUTCH	TF-IDF VECTORIZATION ONLY
Proefwerk (1.00)		*Proefwerk (1.00)*	*Proefwerk (1.00)*	*Proefwerk (1.00)*	*Proefwerk (1.00)*
Naar school (1.00)		*Naar school (0.98)*	Meisje gevonden (0.96)	lang wc bezoek (0.98)	Ziektekostenpremie (0.20)
Geen zin in school (1.00)		*huiswerk (0.95)*	*Groep 2 gedichtje (0.92)*	kerstversiering (0.73)	Verhuizen (0.20)
Groep 2 gedichtje (1.00)		*Groep 2 gedichtje (0.94)*	geen enkele wens (0.86)	Internet-cadeau (0.72)	Meisje gevonden (0.18)
huiswerk (1.00)		*Geen zin in school (0.91)*	bodylotion (0.85)	Turn-wijf (0.72)	*Groep 2 gedichtje (0.18)*
Zie ginds komt de leraar (1.00)		*Zie ginds komt de leraar (0.90)*	Verhuizen (0.84)	bodylotion (0.71)	Televisie kijken (0.17)
		boekenbon (0.86)	VERSTROOID (0.83)	geluk (0.71)	geen enkele wens (0.17)
		boek halen bij de bib. (0.85)	Golfhandschoen (0.83)	Celibaat...not! (0.71)	lekker kontje (0.17)
		boeken lezen (0.85)	*Naar school (0.82)*	Plaid (0.71)	*huiswerk (0.16)*
		Blendle.com tegoed bon (0.84)	wierook of parfum (0.82)	boekenbon (0.71)	VERSTROOID (0.16)

WITH SAME TAG ('Muziek')		LSA TRAINED ON ALL TAGS	LSA TRAINED ON VERSES	PRE-TRAINED LSA DUTCH	TF-IDF VECTORIZATION ONLY
muziekliefhebber (1.00)		*muziekliefhebber (1.00)*	*muziekliefhebber (1.00)*	*muziekliefhebber (1.00)*	*muziekliefhebber (1.00)*
CD (1.00)		*dansen voor publiek (0.99)*	Fles wijn (0.78)	Chocoladeletter (0.79)	*dansen voor publiek (0.12)*
dansen voor publiek (1.00)		*Turn to the beat (0.99)*	zwanger (0.76)	zwanger (0.77)	Glaasje wijn (0.12)
CD\'s in de auto (1.00)		*CD (0.98)*	sportfiguur (0.75)	Einde inspiratie (0.54)	Ik hou van jou (0.11)
Turn to the beat (1.00)		*Hiphop preek (0.98)*	Kaasplankje (0.74)	December (0.47)	Wat brengt de Sint? (0.11)
Hiphop preek (1.00)		*CD\'s in de auto (0.90)*	gouden sieraad (0.73)	gouden sieraad (0.39)	Concours hippique (0.10)
		Rondjes schaatsen (0.84)	Echte bloemen gaan dood (0.73)	Paul de Leeuw fan (0.35)	December (0.10)
		Pluk de dag (0.76)	December (0.71)	Twitterende Zitzak (0.32)	sportfiguur (0.10)
		foute vent (0.67)	volle boekenkast (0.71)	sportfiguur (0.30)	Echte bloemen gaan dood (0.10)
		Verliefd op dansleraar (0.64)	Boerenlater (0.69)	Rijk huis (0.30)	Kaasplankje (0.09)

Conclusion and discussion

For this corpus of Dutch stanza, applying basic computational linguistic techniques is not sufficient to capture similarities between poems. Human-based tagging – even with limited consistency – clearly outperformed the LSA based techniques. Only LSA on these same tags gave limited success, but would still require the human based tagging.

Perhaps the poems in the corpus are too small and results can be improved by using stemming and lemmatization prior to LSA. In a next study, such improved model should be analyzed. Furthermore, defining a golden standard would help to obtain more quantitative results, though human judgement may remain required for final acceptance of the results.

References

1. Deerwester, S., et al, Improving Information Retrieval with Latent Semantic Indexing, Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, 1988, pp. 36–40.

2. Mick’s Rijmwoordenboek – Rijmzoeker (coupletten); www.rijmwoordenboek.nl , www.rijmzoeker.nl

3. Scikit-learn - Machine Learning in Python. www.scikit-learn.org

4. Spärck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Journal of Documentation. **28**: 11–21. CiteSeerX 10.1.1.115.8343. doi:10.1108/eb026526.

5. Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich. "Introduction to Information Retrieval". Cambridge University Press.

6. Linguistic tools for Cognitive Science (alpha) - Cosine Distance Tool; University of Tilburg, https://weblingtools.uvt.nl/cosinedist