Latent Semantic Analysis for German literature investigation

Preslav Nakov

Sofia University, Rila Solutions
27, Acad. G. Botchev Str, Sofia, Bulgaria
preslav@rila.bg, preslav@rocketmail.com

Abstract. The paper presents the results of experiments of usage of LSA for analysis of textual data. The method is explained in brief and special attention is pointed on its potential for comparison and investigation of German literature texts. Two hypotheses are tested: 1) the texts by the same author are alike and can be distinguished from the ones by different person; 2) the prose and poetry can be automatically discovered.

1 Introduction

One of the most interesting text features is the text variance. There are always several ways to express the same thought and the authors are forced to choose between different syntactic constructions, synonyms and terminology according to the intended audience and the impact the text must produce. Furnas, Landauer, Gomez and Dumais have shown in [8] that people use the same words to describe the same subject 10-20% of the time.

Authors make their choices according to both the specific text intention and their own subjective preferences. These (denoted as *style*) are consistent (along the text or all the author's oeuvres) and easy to discover for humans but very hard to describe and measure. Researchers in statistical stylistics have concentrated at word-based statistics (word length, word length distribution, long words count, type/token ratios, see [15]), text-based statistics (sentence length, clause complexity, see [12,14]) and statistics based on specific items (pronouns counts, presence/absence of contractions/amplifiers, relative frequency of specific verbs: e.g. *seem*, *appear* etc., see [2,11]). We go different way: Our purpose here is to study the possibilities of using a classic semantic analysis method without additional tuning for automatically discovery of the texts from the same author and to discriminate between prose and poetry.

2 Latent Semantic Analysis

The *Latent Semantic Analysis (LSA)* is a powerful statistical technique for indexing, retrieval and analysis of textual information used in different fields of the human cognition during the last decade. The method is fully automatic and does not use any preliminary constructed dictionaries, semantic networks, knowledge bases, conceptual hierarchies, grammatical, morphological nor syntactic analysers, etc. The general idea is that there exists a set of latent dependencies between the words and their contexts (phrases, paragraphs and texts). They both are represented in the same semantic space. The identification and proper treatment of the latent dependency permits LSA to deal successfully with the synonymy and partially with the polysemy, which are the major problems with the word-based approaches.

LSA is a two-stage process and includes education and analysis of the indexed data. During the education phase LSA performs an automatic document indexing. The process starts with the construction of a matrix X whose columns are associated with documents, and the rows with terms (words or key-phrases). The cell (i,j) contains the occurrence frequency of term i in document j. The matrix X is then submitted to *singular value decomposition* (*SVD*) which gives as a result three matrices T, D (orthonormal) and S (diagonal), such that $X=TSD^t$. Most of the rows and columns of T, S and D are removed in a way that the matrix $X'=T'S'D'$ is the least squares best-fit approximation of X. This results in the compression of the source space in much smaller one where we have only a limited number of significant factors (generally between 50 and 400). Thus, each term or document is associated a vector (column in the $D'S'$ matrix) of reduced dimensionality, e.g. 100. It is possible to perform a sophisticated SVD, which speeds up the process by directly finding the truncated matrices T', S' and D', see [1].

The second phase is the analysis phase. Most often this includes the study of the proximity between a couple of documents, a couple of words or between a word and a document. A simple mathematical transformation permits to obtain the vector for a non-indexed text. The proximity degree between two documents can be calculated as the dot product between their normalised LSA vectors. The usage of other measures is also possible, e.g.: Euclidean and Manhattan distances, Minkowski measures, Pearson's coefficient etc. [3,13,16,21].

3 Experiments

The experiments were performed on German literature texts we collected on the Web mainly from the following sites [22,23,24]. The file contents were carefully investigated and all index and biographic files were removed. The remaining files were pre-processed and the HTML tags were discarded. The remaining files were pre-processed and the HTML tags were discarded together with the headers, footers and editors' comments leaving just the plain text. We thus obtained the following oeuvres, grouped by author:

- Theodor Fontane: *Effie Briest (181), Der Stechlin (241), Der Schach von Wuthenow (92)*
- Johann Wolfgang von Goethe: *Faust (104), Die Wahlverwandtschaften (152), Die Leiden des jungen Werthers (67), Wilhelm Meisters Lehrjahre (182)*
- Georg Wilhelm Friedrich Hegel: *Phänomenologie des Geistes (90)*
- Heinrich Heine: *Buch der Lieder (245), Neue Gedichte (870)*

- Franz Kafka: *Erzählungen (29), Der Prozeß (126)*
- Karl May: *Vinnetou I, II, III (147)*

Since LSA tries to capture the mutual dependencies between the words and their contexts it is of crucial importance to provide contexts of reasonable size. Usually, when indexing small documents they are passed as they are since it is best to work on the whole document. It is clearly not the case here and we decided to split the documents into chunks of almost equal size. We experimented with different chunk sizes: 1,2,4,8 KB and finally chose 4 KB as most appropriate. The files for each novel were concatenated (if more than one) and split into almost equal parts of approximately 4 KB (In fact the chunk size varies since we do not split the sentences). There are two exceptions: the Goethe and Heine poems were left intact, one poem per file as they originally were.

The words met in just one document were removed since they cannot contribute to the proximity, thus reducing the total different non-stop word forms considered from 72141 to 34191. Someone would argue that these are exactly these words which are characteristic of a text and hence carry important information. This is true in case we want to assign a text a subject category but are of no particular sense when we want to measure similarity between texts using cosine measure.

After the frequency matrix X (2191 × 34191) was built, we divided each row by its entropy and just then performed SVD. [3,5,6,7,13,16] No other words (e.g. stop words) were removed since their frequency of usage could be characteristic for a specific author and thus contribute to the document proximity (Nevertheless the entropy weights them appropriately). No word stemming was performed for the same reasons.

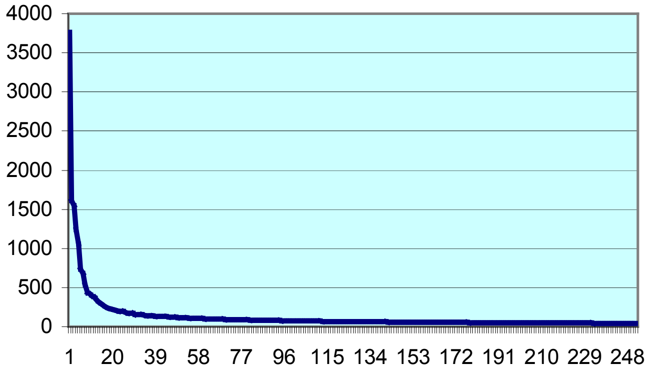

Fig. 1. Choice of similarity measure

A crucial moment when using LSA is the correct choice of dimensionality. Figure 1 shows the top 250 singular values sorted in descending order. The curve goes straight down and then flattens. Intuitively we have to cut the singular values just in the place where the curve behaviour changes. If we cut further we lose important information and if we keep more values we start modelling the noise.

We performed two different space reductions: to space with dimensionality 25 and 100. For each of these cases we calculated the dot product between the normalised vectors for all the document couples. The corresponding correlation matrices (2191×2191) were almost identical. Fig. 2 and 3 show the results in 5 different colours for the five correlation intervals: 87,5-100%, black colour; 75-87,5%, dark grey; 62,5-75%, grey; 50-62,5%, light grey; 0-50%, white.

A well-known feature of LSA is that it maps the semantically related texts next to each other in the vector space. Thus, in our particular case the chunks from the same oeuvre are expected to be more similar to each other than those from different oeuvres.

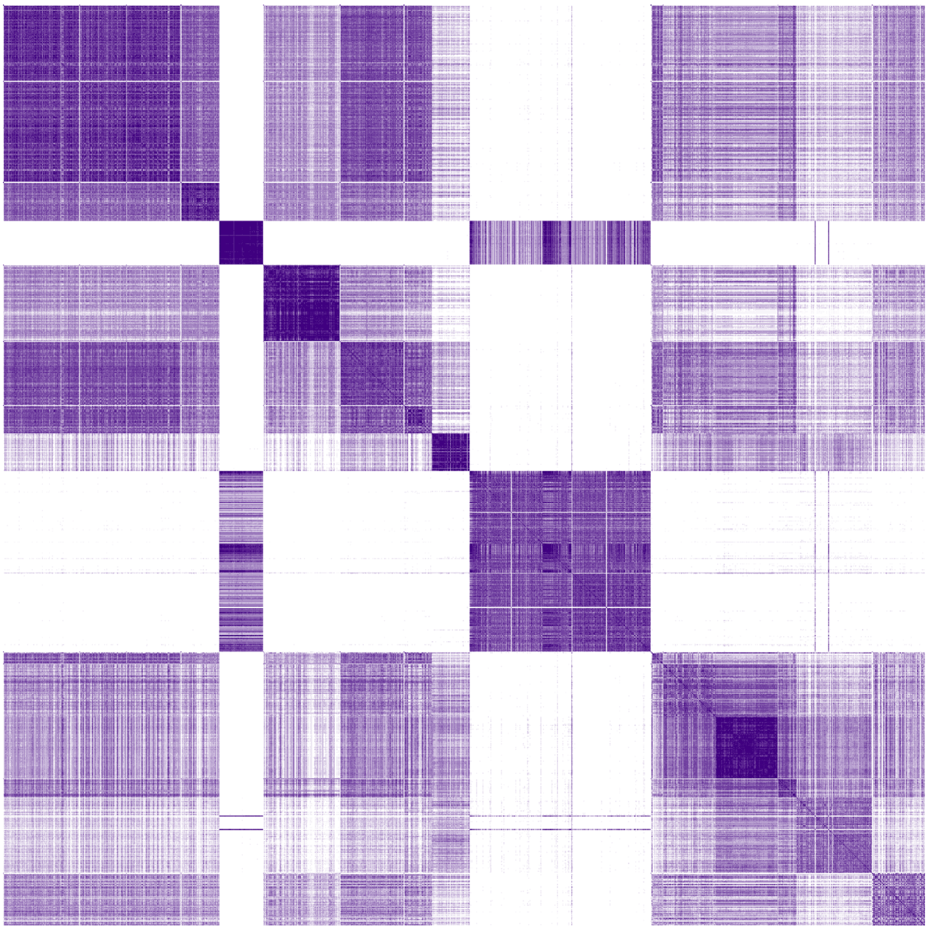

Fig. 2. Correlation matrix map: (vector space dimension 100)

There are several dark squares on the main diagonal, which are clearer on figure 2 than on figure 3 because of the choice of a bit more appropriate vector dimensionality. Let’s look at Fig.2 in more details from the upper left towards the down right corner skipping the region in the black square. There are several squares (clusters) and Fontane forms the first one we can see. This is a monolith black block with some artificially added white lines whose function is to separate *Effie Briest* (181)*, Der Stechlin* (241), and *Der Schach von Wuthenow* (92), which follow in this order (file counts put in parentheses). The latter is distinguishable and forms a small internal sub-square. The other two oeuvres can be distinguished only thanks to the artificial white line. Thus, the oeuvres by Fontane form a good cluster.

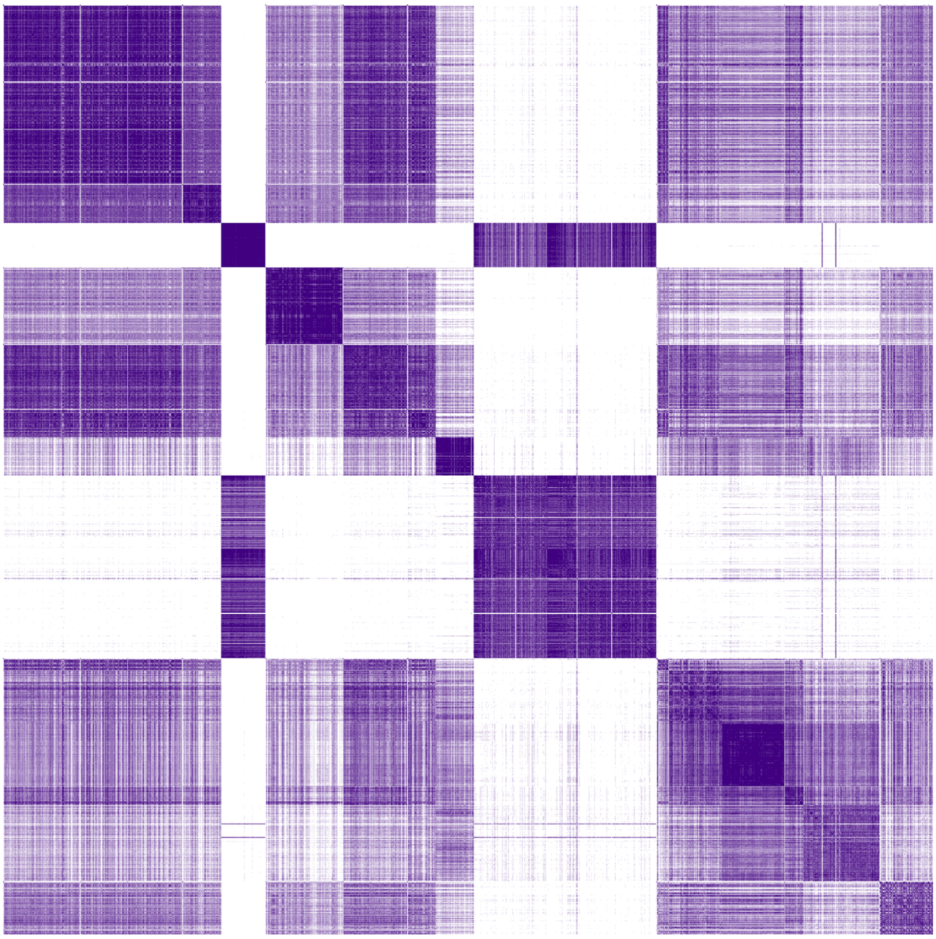

Fig. 3. Correlation matrix map: (vector space dimension 25)

The following small black square represents the Goethe's *Faust* (104). The smooth one just after it has three sub-squares, the first of which corresponds to *Die Wahlverwandtschaften* (152). The second one is split again into two smaller ones corresponding to *Die Leiden des jungen Werthers* (67) and *Wilhelm Meisters Lehrjahre* (152). While each of the last 3 oeuvres forms a well-separated cluster it is normal to expect lower proximity level between texts from different oeuvres. What remains unclear is why *Faust*, although being written by the same author, is so different. The answer is because of the dictionary differences between prose and poetry.

At the end of the previous square there is a small dark cluster not very well separated from the previous ones and representing the Hegel's *Phänomenologie des Geistes* (90). Looking at figure 3 we can see that although it is very similar to the last three Goethe texts their internal proximity is a bit higher. Thus, we can separate the texts by the two authors. We attribute the high proximity level between the Goethe's and Hegel's prose to the fact that they lived in the same historical époque. The same applies in fact to Fontane: looking at the big smooth out-of-diagonal squares we can see the relatively high correlation between the three authors.

The following almost monolith black square contains the Heine's poems: *Buch der Lieder* (245) and *Neue Gedichte* (870). Looking at the out of diagonal elements we can see the high proximity level with the Goethe's poem *Faust*. This must be due mainly to the similar poetic structure. Looking at figure 2 we can see that we are still able to distinguish Goethe's and Heine's poems.

Let us look now at the last big smooth square. It contains the oeuvres by Kafka, May, Keller, Nietzsche and Rilke, in that order. Although it seems difficult for LSA to separate some of them we can distinguish each author from the picture. The darker sub-rectangle in the middle is the Karl May's *Vinnetou* (147). The texts before May belong to Kafka's *Der Prozeß* (126) and *Erzählungen* (29). The *Der Prozeß* (126) can be discovered as a very small (just 29 files) sub-rectangle at the beginning of the smooth one.

Looking carefully one can see a bit bigger small (just 45 files) sub-rectangle just after May: this is the Keller's *Romeo und Julia auf dem Dorfe* (45). Then follow two easier to discover and much bigger sub-rectangles corresponding to Nietzsche's *Also sprach Zarathustra* (179) and Rilke's *Die Aufzeichnungen des Malte Laurids Brigge* (126).

Here we can see again a high semantic proximity between the whole last smooth rectangle and Fontane, Goethe's prose and Hegel. Investigating the texts one can see that all these are prose while the Goethe's *Faust* and Heine's oeuvres are poetry. Thus, we discovered a very good separation between the poetry and the prose among all the texts considered. This gives a strong support for our second hypothesis.

What about the first hypothesis, it seems to hold when comparing the different authors but the judgement whether a particular text belongs to a particular author seems much more difficult. This result is a bit surprising for us since the other researcher experiments on English and Bulgarian (as well as recent experiments on Russian) text collections showed much stronger support for the first hypothesis [19,20]. The problem may be the mixture of prose and poetry we made here: they seem to be very different one from another for this particular text collection, which results in rise of the inner group correlation level harming the support for the first hypothesis. Another problem may be the disproportion of the chunk counts for the different authors.

4 Discussion

As we saw, the different dimensionality reductions reveal different kinds of correlation between the texts. The higher dimensionality matrices show that the texts from the same author are more alike and tend to form separate clusters. When we perform a further reduction we obtain just two classes: the prose and the poetry.

This is consistent with our previous experiments for English, Russian and Bulgarian literature. In general the highest dimensionality matrices show that the texts from the same oeuvre are more alike and tend to form separate clusters. In case the dimensionality is high enough some internal clusters can be discovered inside the same oeuvre. When a further reduction is performed the oeuvres by the same author lose their differences and each author tends to obtain its own cluster (in fact two clusters must be expected if the author is represented by both prose and poetry, as happened above). When we perform a further reduction we obtain just two classes: the prose and the poetry.

5 Conclusion

The experiments performed show that in the general case the selected German authors can be distinguished using LSA but it seems to be hard for some of the authors. On the other hand the selected texts give a strong support for the hypothesis that the prose and poetry can be automatically discovered.

6 Future work

Additional experiments on new (possibly different language) corpora with new authors have to be performed in order to justify the results obtained and to better study all the factors (e.g. époque, internal oeuvre/author substructure) influencing the text proximity when using LSA.

An interesting possibility is to combine and compare the semantic proximity with the traditional stylistic statistics methods used by [4,10,11,12,14,15].

References

1. Berry M., Do T., O'Brien G., Krishna V., and Sowmini Varadhan, SVDPACKC (Version 1.0) User's Guide. April 1993.
2. Biber D.A typology of English Texts. Linguistics, 27, pp. 3-43. 1989.
3. Deerwester S., Dumais S., Furnas G., Laundauer T., Harshman R. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Sciences, 41 (1990), pp. 391-47.
4. Diab M., Schuster J., Bock P., A Preliminary Statistical Investigation into the impact of an N-Gram Analysis Approach based on Word Syntactic Categories toward Text Author

Classification, Proc. Of 6th International Conference on Artificial Intelligence Applications, Cairo, Egypt, 1998.

5. Dumais, S. T. (1993) LSI meets TREC: A status report. In: D. Harman (Ed.), The First Text REtrieval Conference (TREC1). National Institute of Standards and Technology Special Publication 500-207, (pp. 137-152).
6. Dumais, S. T. (1994) Latent Semantic Indexing (LSI) and TREC-2. In: D. Harman (Ed.), The Second Text REtrieval Conference (TREC2), National Institute of Standards and Technology Special Publication 500-215 , (pp. 105-116).
7. Dumais, S. T. (1995) Using LSI for information filtering: TREC-3 experiments. In: D. Harman (Ed.), The Third Text REtrieval Conference (TREC3) National Institute of Standards and Technology Special Publication , in press 1995.
8. Furnas G., Landauer T., Gomez L. and Dumais T. Statistical semantics: Analysis of the Potential Performance of Keyword Information Systems. Bell Syst.Tech.J., 62, Number 6, pp. 1753-1806, 1986.
9. Harman, D. How effective is suffixing? In Journal of The American Society of Information Science. Vol. 42, No 1. 1991.
10. Jiang, J. Using Latent Semantic Indexing for Data Mining. Department of Computer Science, University of Tennessee, December 1997.
11. Karlgen J., Douglas C. Recognizing Text Genres with Simple metrics Using Discriminant Analysis. Proceedings of COLING 94, Kyoto, pp. 1071-1075.
12. Klare G.The Measurement of Readability. Ames: Iowa University Press. 1963.
13. Laudauer T., Foltz P., Laham D. Introduction to Latent Semantic Analysis. Discourse Processes, 25, pp. 259-284.
14. Lorge I. The Lorge Formula for Estimating Difficulty of Reading Materials. New York: Teachers College Press, Columbia University, 1959.
15. Losee R. Text Windows and Phrases Differing by discipline, Location in Document, and Syntactic Structure.Information Processing & Management 32(Nov): 747-67. 1996
16. Nakov P. Getting Better Results with Latent Semantic Indexing. In Proceedings of the Students Presentations at ESSLLI-2000, pp. 156-166, Birmingham, UK, August 2000.
17. Nakov, P. Web-personalisation using extended Boolean operations with Latent Semantic Indexing. Proc. AIMSA-2000, Varna, Bulgaria, Lecture Notes in Artificial Intelligence 1904, Springer 2000, pp. 189-198.
18. Nakov P. Latent Semantic Analysis of Textual Data. In Proceedings of CompSysTech’2000, Sofia, Bulgaria. June 2000.
19. Nakov P. Latent Semantic Analysis for Bulgarian literature. In Proceedings of Spring Conference of Bulgarian Mathematicians Union. Borovetz. Bulgaria. 2001.
20. Nakov P. Latent Semantic Analysis for Russian literature investigation. In Proceedings of the 120 years Bulgarian Naval Academy Conference. Varna. Bulgaria. 2001.
21. http://lsa.colorado.edu
22. http://www.phil-fak.uni-duesseldorf.de/germ5/service/download/index.html
23. http://gutenberg.aol.de
24. http://www.phil.uni-erlangen.de/~p2gerlw/ressourc/archiv.html