МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2001
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2001
Proceedings of Thirtieth Spring Conference of
the Union of Bulgarian Mathematicians
Borovets, April 8–11, 2001

LATENT SEMANTIC ANALYSIS FOR BULGARIAN LITERATURE

Preslav Ivanov Nakov

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 of Bulgarian literature texts. Two hypotheses are tested:
• The texts from the same author are alike and can be automatically discovered;
• The texts belonging to different periods can be distinguished automatically.

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 is based on the general idea that there exists a set of latent dependencies between the words and their contexts (phrases, paragraphs and texts). Their identification and proper treatment permits LSA to deal successfully with the synonymy and partially with the polysemy.

LSA starts with the construction of a term to document occurrence frequency matrix, which is then submitted to *singular value decomposition* (*SVD*). As a result each term or document is associated a vector of low dimensionality (e.g. 100). The proximity between two documents can be calculated as the dot product between their normalised vectors. (see [1,2,3] for details)

Application to Bulgarian literature texts. The experiments were performed on Bulgarian literature texts we found in the *Virtual Library for Bulgarian Literature* at: http://slovo.orbitel.bg ([4]). We selected all the 3032 available texts for the following authors grouped by period (the text counts are in parentheses):

• *Bulgarian Renaissance (XVIII–XIX c.):* Paisiy Hilendarski (15), Sofroniy Vrachanski (3), Rayko Zhinzifov (6), Dobri Chintulov (17), Gueorgui S. Rakovski (18), Petko R. Slavejkov (25), Lyuben Karavelov (47), Konstantin Miladinov (6), Stefan Stambolov (34) and Hristo Botev (48); (10 authors)

• *Period between the Liberation and the First World War (1878–1918)*: Ivan Vazov (753), Konstantin Velichkov (40), Stoyan Mihaylovski (17), Aleko Konstantinov (33), Anton Strashimirov (10), Kiril Hristov (32), Pencho Slaveykov (60), Mara Belcheva (4), Peyo Yavorov (185), Petko Todorov (7), Simeon Radev (35), Teodor Trayanov (193), Dimcho Debelyanov (141), Hristo Yasenov (33), Dimitar Boyadzhiev (40) and Ekaterina Nencheva (16); (16 authors)

• *Period between the two World Wars:* Geo Milev (73), Hristo Smirnenski (56), Elisaveta Bagryana (116), Yordan Yovkov (49), Elin Pelin (104), Chudomir (67), Nikola

Vaptzarov (58), Alexander Vutimsky (40), Vessela Vassileva (14), Sirak Skitnik (10), Vesselin Hantchev (3); (11 authors)

• *Contemporary literature (after the Second World War):* Yordan Radichkov (17), Damyan Damyanov (43), Radoy Ralin (70), Guencho Stoev (9), Gueorgui Danailov (47), Evtim Evtimov (42), Penyo Penev (38), Hristo Fotev (119), Nikolay Kanchev (102), Gueorgui Konstantinov (29) and Petya Dubarova (108). (11 authors)

Experiments. 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 removed, together with all stop-words from a pre-selected list for the Bulgarian Language including all the: adverbs, conjunctions, interjections, numerics, particles, prepositions, pronouns and auxiliary verbs. No word stemming was performed. (Although our previous experiments show it is beneficial for the highly inflexional Bulgarian language. [5])

For the first experiment we left the texts from the Bulgarian Renaissance out and kept only the ones from the last 3 periods (2813 texts by 38 authors). We stripped out all words that occur just in one document, since they cannot contribute to the proximity, reducing the total different non-stop word forms considered from 116307 to 54325. After the frequency matrix X (2813×54325) was built, we divided each row by its entropy and just then performed SVD. [1,2,5]

We performed two different space reductions: to space with dimensionality 100 and 200. For each of these cases we calculated the dot product between the normalized vectors for all the document couples. The corresponding correlation matrices (2813×2813) are shown on Fig. 1 and 2 in 5 different colors for the five correlation intervals: 87,5–100%, black color; 75–87,5%, dark gray; 62,5–75%, gray; 50–62,5%, light gray; 0–50%, white.

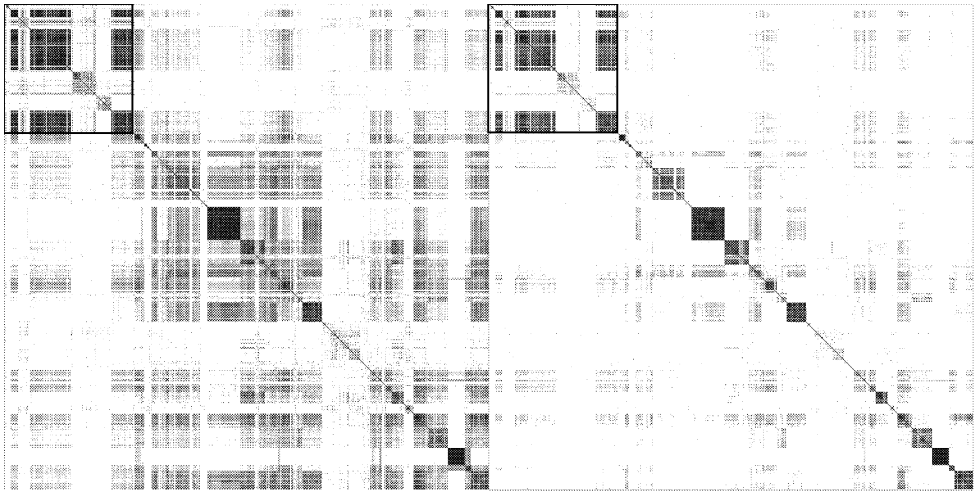

Fig.1. Last periods (vector space dimension 100)

Fig.2. Last periods (vector space dimension 200)

There are several dark squares on the main diagonal, which are clearer on figure 2 (because of the choice of more appropriate vector dimensionality; see [5] for details

on how to choose the correct 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 small clusters and the first bigger (and a bit smooth) one we can see is formed by the Yavorov texts (185 texts). The next big well-distinguished ones are the texts by Trayanov (193), then by Debelyanov (a bit smooth – 141). The following authors (Boyadhiev, Nencheva, Geo Milev and Smirnenski) although forming some clusters are not distinguished very well. The immediately following very clear cluster is formed by Elissaveta Bagryana. Then follow several authors represented by few texts and the last 5 bigger and clear clusters are due to Ralin, Evtimov–P.Penev (they form a common cluster but Evtimov can still be distinguished), Fotev, Kanchev and Dubarova.

So, the texts by the same author tend to form a well-separated cluster in the correlation matrix and thus be automatically discovered. This can be easily explained by the specificity of the style and the vocabulary used. The phenomenon was well demonstrated by the authors with higher number of texts.

It is possible that two authors have a similar writing style and are difficult to discriminate. We saw this above on moderately well presented authors (Evtimov:42 and Penev:38). Our previous experiments on a corpus of English religious and sacred texts (see [5]) show that books by authors that are guaranteed to be different can be impossible to be distinguished: using LSA one can distinguish the Old from the New Testaments but not the New Testament and the Book of Mormons!

Thus, there are several other factors influencing the text proximity using LSA. Let us look now at the top left corner of the correlation matrix (Fig.2 and Fig.3) which was not considered above. This is a large enough quantity of 753 texts all by the same author: Ivan Vazov. There are several well-formed clusters inside among with some clear out-diagonal dependencies. Why some of the texts seem to be different from the others? This is because of the diversity of texts by Vazov used: novels, narrations, descriptions, poems etc. The further investigation shows that the biggest dark regions of interdependencies show proximity between the poems by Vazov. From left to right there are 3 big black clusters formed by the cycles: the small cluster "Stihove za malki detza", then a big cluster including "Epopeya na zabravenite", "Gusla", "Italiya", "Pesni za Macedoniya", "Nebe", and a third one formed by "Priaporec i gusla", "Polya i gori", "V lonoto na Rila", "Skitnishki pesni", "Slivnica" and "Tagite na Balgariya". The three smooth regions from left to right are formed by the narrations "Draski i sharki" and the two novels "Nova zemja" and "Pod igoto". The specific language style of the poems by the same author groups them together while distinguishing from the other texts by the same author. The more free narrative style groups together the texts from the same novel (although smoothly) but not the ones from all the novels as this was with the poems.

Let us now proceed to the second hypothesis. We have to check whether the texts belonging to different literature periods can be automatically distinguished. We tried to see the things clearer by investigating the periods by couples. The neighbor periods were not interesting because we are aware that the borders are artificial and not well separated. Figure 4 shows the results for the comparison between the first (the Renaissance, which was not present on the previous figures) and the last period. These periods are supposed to be the most different since they are the most distant in time. Unfortunately, no clear clusters formed by the periods are found and several inter-period dependencies are discovered. The other couples gave comparable results.

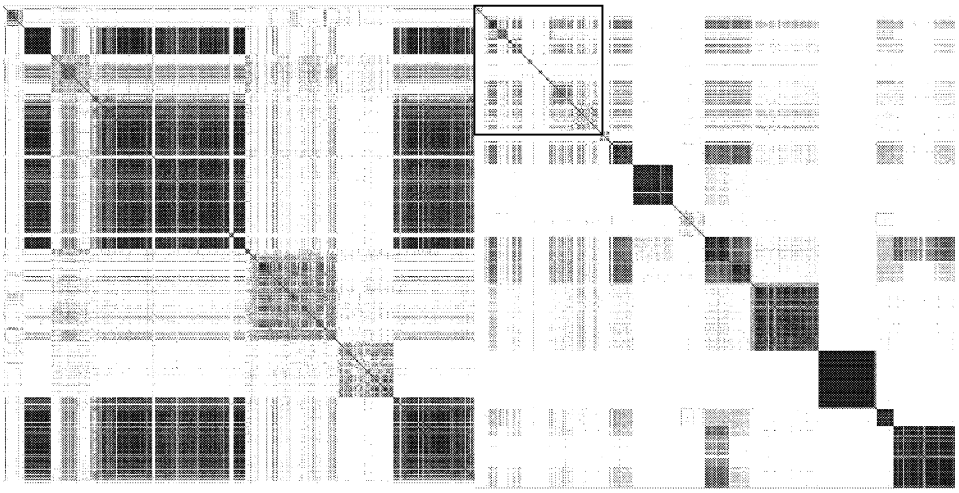

Fig.3. Vazov only(vector space 100)

Fig.4. Renaissance and after II World War (200)

Conclusion. The experiments performed show that in the general case the selected Bulgarian authors can be effectively distinguished using LSA but there are some exceptions due to several other factors that must be taken into account. The hypothesis that the different time periods can be discovered automatically fails here although it may work for other corpora with possibly more distant periods.

Further work. Additional experiments on new (possibly different language) corpora have to be performed in order to justify the results obtained and to better understand the factors influencing the text proximity when using LSA.

REFERENCES

[1] S. Deerwester, S. Dumais, G. Furnas, T. Laundauer, R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Sciences*, **41** (1990), 391.

[2] T. Laudauer, P. Foltz, D. Laham. Introduction to Latent Semantic Analysis. Discourse Processes, vol. 25, 259–284.

[3] LSA 1990-99, see `http://lsa.colorado.edu`

[4] Slovoto. Virtual Library for Bulgarian Literature, see `http://slovo.orbitel.bg`

[5] P. Nakov. Getting Better Results with Latent Semantic Indexing. In: Proceedings of the Students Presentations at ESSLLI-2000, Birmingham, UK, August 2000, 156–166.

Preslav Ivanov Nakov
27 Acad. G. Bontchev Str.
Rila Solutions
1113 Sofia, Bulgaria
e-mail: `preslav@rila.bg`

ЛАТЕНТЕН СЕМАНТИЧЕН АНАЛИЗ НА БЪЛГАРСКА ЛИТЕРАТУРА

Преслав Иванов Наков

Представени са резултатите от използването на ЛСА за анализ на текстови данни. Същността на метода е изложена накратко и специално внимание е обърнато на потенциала му при сравняване на български литературни текстове. Тествани са две хипотези:

• Текстовете от един и същ автор са сходни и могат да бъдат откривани автоматично.

• Текстовете, принадлежащи на различни периоди могат да бъдат различавани.