**Литература**

1. Канеман Д. Думай медленно…Решай быстро / Д. Канеман ; пер. с англ. – М. : изд. АСТ, 20915. – 653 с.

2. Гаванде Атул. Чек-лист. Система предотвращения ошибок / А. Гаванде ; (пер. с англ.). – М. : изд. Альпина Паблишер, 2017. – 352 с.

## СУЧАСНИЙ СТАН ЛІНГВО-СТАТИСТИЧНИХ МЕТОДІВ СЕМАНТИЧНОГО АНАЛІЗУ ТЕКСТІВ

*Зембицька М. В.*[1], *Горошко А. В.*[2]
*Хмельницький національний університет, e-mail*: [1]*zembitska@i.ua*, [2]*iftomm@.ukr.net*

Квантитативні методи, зокрема вивчення статистичних властивостей текстів є предметом досліджень значної частини робіт з лінгвістики [1–8]. Так у роботах багатьох авторів описані результати апроксимації частоти зустрічальності слів відомими параметричними законами імовірнісних розподілів. Аналіз цих робіт показує, що на сьогодні не знайдено «загального» закону розподілу, який би «задовільно» описував частоти для різних слів у різноманітних текстах різними мовами. Головними причинами є індивідуальні прояви досліджуваних об’єктів.

Однією з найважливіших задач лінгво-статистичного аналізу є автоматична семантична кластерізація текстів. Найбільшого поширення набули методи, у яких тексти представляються у вигляді векторів у багатомірному просторі ознак. У найпростішому випадку кожна ознака відповідає наявності в тексті однієї з словоформ, які зустрічаються в тексті. При цьому компонента може дорівнювати нулю або одиниці, а у складніших випадках за кількістю випадків зустрічальності терміну в тексті формується вектор частот. Такі вектори можуть нормуватись. Слід зазначити, що алгоритми кластерізації оперують матрицями, тому через занадто велику розмірність простору ознак їх застосування проблематичне. Для зниження розмірності використовують відомі методи [9–12].

Для даних цілей перспективним видається EM-алгоритм (Expectation Maximization). Цей алгоритм оперує імовірнісною моделлю відношення документу до відповідного кластеру. Базується на представленні реалізації багатомірної випадкової величини.

Для зниження розмірності може бути використаний метод головних компонент PCA (Principal Component Analysis). Тут простір змен-

шеної розмірності будується на власних векторах коваріаційної матриці, яка відповідає декільком найбільшим власним числам.

Отже, запропоновані методи і алгоритми принципово можуть використовуватись в задачах статистичної обробки текстів, зокрема англомовних, але питання їх практичного застосування потребує грунтовних досліджень.

**Література**

1. Régnier M. A. Unified Approach to Word Occurence Probabilities / M. A. Régnier // Discrete Applied Mathematics, 2000. – Vol. 104, issue 1–3. – P. 259–280.

2. Blake C. A. Comparison of Document, Sentence, and Term Event Spaces / C. A. Blake // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006. – Pages: 601–608.

3. Rennie J. A Better Model for Term Frequencies / J. Rennie // 2005.

4. Zipf G. Human behaviour and the principle of least effort / G. Zipf // An introduction to human ecology, 1949. 1st edn., Addison Wesley.

5. Baeza-Yates R. Modern Information Retrieval / R. Baeza-Yates, B. Ribeiro-Neto // Addison Wesley, 1999.

6. Reinert G. Probabilistic and Statistical Properties of Words: An Overview / G. Reinert, S. Schbath, M. Waterman // Journal of Computational Biology, 2000. – Vol. 7, number 1/2. – Pp. 1–46.

7. Schbath S. An Overview on the Distribution of Word Counts in Markov Chains / S. Schbath // Journal of Computational Biology, 2000. – Vol. 7, number 1/2. – Pp. 193–201.

8. Gotoh Y. Statistical Language Modelling / Y. Gotoh, S. Renals // Lecture Notes in Computer Science, 2003. Springer. – Vol. 2705. – P. 78–105.

9. Régnier M. Rare events and conditional events on random strings / M. Régnier, A. Denise // Discrete Mathematics and Theoretical Computer Science, 2004. – Vol. 6, n°2. – P. 191–214.

10. Church K. Poisson Mixtures / K. Church, W. Gale // Journal of Natural Language Engineering, 1995.

11. McCallum A. A Comparison of Event Models for Naive Bayes Text Classification / A. McCallum, K. Nigam // In AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41–48. Technical Report WS-98-05. AAAI Press. 1998.

12. Горошко А. В. Застосування методу головних компонент для усіченої оцінки найменших квадратів під час розв'язання оберненої задачі ідентифікації ексцентриситетів ротора / А. В. Горошко // Вісник Хмельницького національного університету. Технічні науки. – 2015. – № 6. – С. 49–53.