

СОВРЕМЕННЫЕ МЕТОДЫ ИЗВЛЕЧЕНИЯ КЛЮЧЕВОЙ ИНФОРМАЦИИ ИЗ НОРМАТИВНЫХ ДОКУМЕНТОВ¹

*М.А. Милкова, И.В. Неволин,
Д.П. Пигорев*

DOI: 10.33293/1609-1442-2021-2(93)-101-114

В статье демонстрируется подход к устранению сложностей, возникающих при анализе правовых документов в рамках экономических и междисциплинарных исследований. В условиях роста объема и постоянного обновления информации и (или) появления новой области исследований наиболее целесообразным на первом этапе является получение общей структуры всей коллекции документов, некая семантическая компрессия информации. Цель работы – продемонстрировать возможности применения методов анализа естественного языка для анализа нормативных документов, регламентирующих вопросы продовольствия и питания, в частности связанные с предупреждением развития железодефицитной анемии (ЖДА). Подход включает выделение ключевой информации объемных текстов (ключевых слов и предложений) на основе графового алгоритма TextRank. Важным звеном, способствующим познанию, является также визуализация семантических связей между словами внутри документов. По нашему мнению, именно комбинация

© Милкова М.А., Неволин И.В., Пигорев Д.П.,
2021 г.

Милкова Мария Александровна, научный сотрудник ЦЭМИ РАН, Москва, Россия; m.a.milkova@gmail.com. ORCID 0000-0002-9393-1044

Неволин Иван Викторович, к.э.н., ведущий научный сотрудник ЦЭМИ РАН, Москва, Россия; i.nevolin@cemi.rssi.ru

Пигорев Дмитрий Павлович, к.э.н., научный сотрудник ЦЭМИ РАН, Москва, Россия; dpigorev@ya.ru.

¹ Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проект № 19-57-80003).

смысловой компрессии и визуализации информации как крупный план текстовых документов, а также возможности дальнейшей детализации путем линейного чтения и анализа являются наиболее актуальным подходом в условиях перенасыщения информации и дефицита внимания. Особенно актуально активное внедрение методов текстовой аналитики для систем, которые не борются за внимание потребителей. Удобство именно таких систем существенно отстает при извлечении значимой информации. Приемы улучшения понимания больших объемов нормативных документов принесут существенную пользу аналитикам, ведущим экономические, юридические или мультидисциплинарные исследования.

Ключевые слова: цифровая экономика, извлечение ключевых терминов, резюмирование, извлечение ключевых предложений, TextRank, bm-25, графовый алгоритм, анемия, железодефицитная анемия.

Классификация JEL: D80, D83.

ВВЕДЕНИЕ

Вопрос о возрастающей сложности получения знаний в условиях перенасыщения информации, дефицита внимания (Simon, 1971), социального ускорения и клипового мышления является актуальным практически во всех областях человеческой деятельности. Речь идет не только о необъятном объеме информации, представленной в сети Интернет, но и о восприятии научных статей (Недумов, Кузнецов, 2018), юридических документов (Pagallo et al., 2018), патентных данных (Милкова, 2020a) и литературных произведений (Моретти, 2016).

Таким образом, актуальным вопросом является внедрение подходов смысловой компрессии информации – извлечение наиболее значимой информации и преобразование ее в структурированную форму с целью облегчения понимания и дальнейшего использования и анализа. Отметим, что методы извлечения информации как формы представления знаний получают стремительное развитие в рамках работ поисковых систем. Построение семантических сетей и графов знаний для

веб-страниц позволяет предоставлять наиболее точные ответы на запросы пользователей. Однако причина такого быстрого развития² достаточно очевидна – это конкуренция за внимание посетителей (с целью его монетизации). Аналогичная тенденция наблюдается и в других системах (рекомендательные, вопросно-ответные и др.) – социальное ускорение (т.е. ускорение общественной жизни), дефицит внимания, спрос на получение быстрого ответа стимулируют развитие методов анализа естественного языка. Стоит отметить, что системы, не борющиеся за внимание потребителей, существенно отстают в удобстве извлечения значимой информации. Так, справочно-правовые системы Консультант Плюс, Гарант, обладая возможностями для быстрого поиска специалистом нужной статьи или документа, не предоставляют удобных инструментов для исследователей, которые вторгаются в новую для себя область. Открытыми остаются следующие вопросы: какой документ является ключевым? Какие смежные документы актуальны? С чего начать изучение?

В рамках данной работы представляет интерес извлечение ключевой информации из текстов нормативных документов – резюмирование текстов (*summarization*). Подчеркнем, что важность линейного (обычного) прочтения текстов при этом несколько не принижается. Однако в условиях, когда объем информации слишком велик, постоянно обновляется и (или) область исследования является новой, наиболее целесообразным становится подход, позволяющий быстро получать общее представление обо всех текстах на основе ключевой информации в противовес линейному чтению только основных (или по какой-либо причине продвигаемых) текстов. Отметим также, что подход на основе резюмирования является начальным этапом в понимании исследуемого вопроса, за кото-

рым зачастую следует дальнейшее линейное чтение, но уже с пониманием общей структуры всей коллекции документов.

Применение анализа текстов в законодательной деятельности имеет огромный потенциал. Правовая аналитика (*legal analytics*, LA), объединяющая науку о данных, машинное обучение, методы обработки естественного языка и статистические методы, использует обширный и разнообразный массив юридических документов для вывода новых закономерностей и юридических знаний (Ashley, 2017). В частности, столкнувшись с большими объемами документов, которые необходимо изучить, юристы и исследователи получают большую пользу от резюмирования документов (Moens, 2007). Правовая аналитика позволяет делать видимыми скрытые шаблоны в данных и документах, выявлять невидимые особенности структуры и функционирования правовых институтов. Важной составляющей являются удобство использования и легкий доступ к юридическим знаниям, результаты правовой аналитики должны быть представлены в понятной форме для людей, которые не являются экспертами в правовых вопросах, без изменения формулировки самого сообщения (Pagallo et al., 2018).

В последние годы интерес к вопросам правовой аналитики отмечается и в российских публикациях (Исаков, 2018). Отечественные ученые отмечают сложности в понимании стандартов в нормативных документах в рамках анализа экономических проблем (см., например, исследование инновационных рисков в области продуктов питания (Варшавский, 2015), обеспечение экологической безопасности (Гусев, 2019)).

1. ПОСТАНОВКА ЗАДАЧИ ИССЛЕДОВАНИЯ

Предметом исследования в данной работе являлось извлечение значимой информации

² Конференции, посвященные искусственному интеллекту в 2020 г., отмечают две доминирующие области: анализ естественного языка и компьютерное зрение (Милкова, 2020б).

из российских нормативных документов, регламентирующих вопросы обеспечения достаточного количества микронутриентов, в частности железа. Последствия неполноценного питания³ имеют скрытый, но разрушительный характер, незаметно ослабляя потенциальные возможности развития людей, общества и национальной экономики. Одним из важнейших пищевых веществ является железо, различные формы недостаточности которого – основные и наиболее распространенные нарушения питания в мире. По данным ВОЗ, от дефицита железа страдает более 30% населения мира⁴. Распространенность железодефицитной анемии (ЖДА)⁵ в разных странах различна и зависит от уровня экономического развития, продовольственной безопасности, особенностей питания, доступности медицинской помощи и других факторов (Чернов, Тарасова, 2013). В России по обобщенным данным у разных групп населения ЖДА отмечается в 21–26% случаев⁶. В ряде исследований показано, что анемия приводит к длительным и необрати-

мым последствиям: вызывает задержку психомоторного развития и нарушения когнитивных функций у детей, снижает уровень энергии и способность полноценно функционировать у взрослых (Тарасова, 2011). По некоторым оценкам, анемия в детском возрасте вызывает снижение заработной платы в зрелом возрасте на 2,5% (Horton, Ross, 2003). Таким образом, вопросы питания в целом и железодефицитной анемии в частности являются актуальными в контексте обсуждения негативных экономических последствий.

В рамках исследования вопроса обеспечения полноценного и здорового питания важная роль отводится социально-институциональным изменениям. В том числе актуальна разработка инструментов, нацеленных на поддержку принятия решений и выработку политики в области питания. В частности, важным звеном является разработка подхода, способствующего пониманию ключевых концепций имеющихся нормативных документов, позволяющего емко представлять большие объемы текстовой информации, извлекать интересующие данные и отсеивать ненужную на начальных этапах познания информацию.

Для достижения данных целей была сформирована выборка из справочно-правовой системы Консультант Плюс. Собранная база содержит 24 документа и обобщает разнородную информацию: нормы потребления для различных категорий граждан, список рекомендованных продуктов, особенности налогообложения сельскохозяйственных товаропроизводителей, распределение субсидий на стимулирование подотраслей агропромышленного комплекса и др. (перечень документов и их характеристик приведен в Приложении). Рассматриваемые документы, как правило, достаточно объемные и содержат различного рода нормативы, поэтому задача емкого извлечения ключевой информации оправдана. Кроме того, некоторые нормативы часто обновляются, поэтому актуальной задачей является автоматическое фиксирование изменений ключевых параметров. Резюмируя вышесказанное: наша задача состоит

³ Неполноценное питание понимается не только как нехватка калорий, но также как дефицит микронутриентов, избыточный вес и ожирение.

⁴ Дефицит микроэлементов [Текст электронный] // Всемирная организация здравоохранения. URL: <https://www.who.int/nutrition/topics/ida/ru/> (дата обращения: 12.04.2021).

⁵ Дефицит железа является наиболее распространенной причиной анемии, однако ее может вызвать и недостаток других видов питательных микроэлементов (включая фолат (+соли фолиевой кислоты), витамин B12 и витамин A), острые и хронические заболевания.

Концентрации гемоглобина для диагностики анемии и оценки ее тяжести [Текст электронный] // Всемирная организация здравоохранения. URL: https://www.who.int/vmnis/indicators/haemoglobin_ru.pdf (дата обращения 12.04.2021).

⁶ The Global Prevalence of Anemia in 2011 [Текст электронный] // World Health Organization. 2011. URL: https://apps.who.int/iris/bitstream/handle/10665/177094/9789241564960_eng.pdf (access date: 12.04.2021).

в выделении ключевых терминов и ключевых предложений документов, получении структуры и семантических связей как внутри документа, так и между ними.

Также извлечение ключевой информации из текстов необходимо для решения таких задач, как составление словарей (тезаурусов, онтологий, глоссариев) и высокоуровневое описание содержания документа (Saad et al., 2008), выявление тематики текста (Rizoiu, Velcin, 2011).

2. МЕТОДИКА ИССЛЕДОВАНИЯ

В настоящее время существует широкий спектр методов извлечения ключевых терминов (см., например, обзор (Firoozeh et al., 2019)), которые можно разделить на статистические (например, на основе вычисления метрик $TF \times IDF$ ⁷, Salton, Buckley, 1988), основанные на машинном обучении (Wu et al., 2007), структурные (лингвистические и графовые) (Beliga et al., 2015) и гибридные – синтезирующие различные подходы (Teresa et al., 2006). Своего рода продолжением методов извлечения ключевых слов являются методы резюмирования текстов – методы выделения ключевых предложений, сосредоточивающих в себе наибольший объем информации. Подходы

к резюмированию текста могут быть классифицированы по типу используемой техники суммаризации (*extractive summarization* – выделение из текста наиболее значимых, возможно, не связанных между собой единиц; *abstractive summarization* – генерация связного краткого содержания), по числу обрабатываемых документов (отдельный документ или многодокументный анализ), а также по зависимости или независимости от языка.

В нашей работе извлечение ключевых слов и предложений основано на графовом методе. Данный подход оказывается эффективным для гомогенных по функциональному стилю *корпусов текстов* (таких как научные статьи и нормативные документы) (Schmedding et al., 2018; Ванюшкин, Гращенко, 2016).

Графовые алгоритмы представляют собой подход к ранжированию *элементов текстовых документов*, например, для извлечения ключевых слов или предложений. Основная идея алгоритма – преобразование текста в граф, содержащий рассматриваемые элементы в качестве вершин. Ребра графа определяются в соответствии с отношениями между элементами. Как правило, отношения определяются в зависимости от близости узлов. В случае когда узлами являются слова, близость определяется в соответствии с совместной встречаемостью внутри заданного окна, однако может отражать и семантическое отношение. В случае когда вершины – это предложения, близость может определяться в зависимости от числа общих слов в предложении или как семантическая близость.

Граф может быть как ориентированным (направленность ребер отражает последовательность элементов в тексте), так и неориентированным (ребра демонстрируют наличие отношения); как взвешенным (ребра имеют вес характеризующий, например, частоту совместной встречаемости), так и невзвешенным. Для вершин графа вычисляется мера центральности, отражающая степень участия вершины в процессе распространения информации.

Существуют различные графовые алгоритмы, отличающиеся методом определе-

⁷ $TF-IDF$ – наиболее распространенная метрика определения веса слова (словосочетания). TF – *term frequency* – частота встречаемости слова (словосочетания) в документе; IDF – *inverse document frequency* – обратная документная частота – обратное число документов в коллекции, в которых встретилось слово (словосочетание): $IDF(w_i) = \log N/n(w_i)$. Учет IDF уменьшает вес широкоупотребительных слов: чем выше частота присутствия слова в конкретном документе по сравнению с другими, тем больше общее значение $TF-IDF$. Используются различные варианты метрики $TF-IDF$. В классическом случае это $TF \times IDF$, однако могут использоваться и более сложные комбинации этих мер.

ния отношений между вершинами, способами вычисления меры центральности вершин (наиболее полный обзор графовых методов представлен в (Beliga et al., 2015)). В целом графовые алгоритмы не зависят от естественного языка, а также не предполагают использования лингвистических ресурсов.

Для данной работы был выбран алгоритм TextRank (Mihalcea, Tarau, 2004)⁸ как наиболее часто используемый и развиваемый графовый подход (Bougouin et al., 2013).

2.1. Извлечение ключевых терминов

TextRank для извлечения ключевых слов строит взвешенный неориентированный граф $G = (V, E)$, где V – множество вершин (уникальных слов); E – множество ребер. Две вершины соединяются ребром, если слова совместно встречаются внутри окна из N слов, $N \in [2, 10]$. Вес ребра $w(e_{ij})$ предлагается определять по формуле (Милкова, 2018):

$$w(e_{ij}) = \begin{cases} 1 - \frac{d(v_i, v_j) - 1}{(N-1)(2d(s(v_i), s(v_j)) + 1)}, & \text{если } d(v_i, v_j) \in (0, N); \\ 0, & \text{если } d(v_i, v_j) \geq (0, N), \end{cases} \quad (1)$$

где $d(v_i, v_j)$ – модуль разности позиций слова v_i и v_j ; $d(s(v_i), s(v_j))$ – модуль разности позиций между предложениями, в которых находятся слова.

Вес, или значимость слова, $w(v_i)$ определяется значимостью слов, с которыми оно связано:

$$w(v_i) = (1-d) + d \sum_{v_j \in Adj(v_i)} \frac{w(e_{ij})}{\sum_{v_k \in Adj(v_j)} w(e_{jk})} w(v_j), \quad (2)$$

⁸ TextRank является приложением алгоритма PageRank (используемого Google Search для ранжирования веб-сайтов в результатах поисковой выдачи, см. (Page et al., 1998)) для обработки естественного языка.

где $Adj(v_i)$ – множество вершин, связанных с вершиной v_i ; $d \in (0, 1)$ – коэффициент затухания (*damping factor*), обычно устанавливается равным 0,85 (Brin, Page, 1998; Mihalcea, Tarau, 2004).

Начальное значение $w(v_i)$ для каждой вершины равно 1, алгоритм повторяется до достижения порогового уровня сходимости. В соответствии с итоговыми значениями $w(v_i)$ вершины графа ранжируются, и отбираются вершины с наибольшим значением $w(v_i)$.

Отдельно отметим, что извлечение ключевых слов с помощью графовых алгоритмов позволяет визуализировать связи между словами, что, согласно многим исследованиям, положительно влияет на познавательные функции (Heer et al., 2010; Patterson et al., 2014; Zhang et al., 2017)⁹.

2.2. Извлечение ключевых предложений

Извлечение ключевых предложений в работе Михальцеа и Тарау (Mihalcea, Tarau, 2004) также является результатом обработки графа. Пусть S, R – два предложения. $f(s_i, R)$ – частота встречаемости каждого слова i предложения S в предложении R . Мера схожести между двумя предложениями $Sim(S, R)$ в основополагающей работе (Там же) определяется по формуле

$$Sim(S, R) = \frac{f(s_i, R)}{\log(|S|) + \log(|R|)}. \quad (3)$$

Однако формула (3) не совсем адекватно описывает присутствие общеупотребительных слов. Несмотря на то что высокочастотные слова можно убрать из анализа на этапе предобработки текстов, адекватным подходом в определении близости является присвоение более высоких значений связям между предложениями, схожими нечастотными словами. Данная идея нашла отражение в алгоритме

⁹ Однако данное влияние отмечается в контексте среды, для которой визуализация была разработана, но она не обязательно будет эффективной вне этого контекста (Patterson et al., 2014).

BM-25¹⁰ (Robertson, Zagaroza, 2009). Мера близости по алгоритму BM-25 определяется следующим образом:

$$BM25(S, R) = \sum_{i=1}^n IDF(s_i) \frac{f(s_i, R)(k_1 + 1)}{f(s_i, R) + k_1 \left((1-b) + b \frac{|R|}{avgDL} \right)}, \quad (4)$$

где k_1 , b , ε – параметры, значения которых, согласно множеству экспериментов, как правило, лежат в промежутках $k_1 \in (1, 2; 2)$; $b \in (0, 5; 0, 8)$; $\varepsilon \in (0; 0, 5)$ (Robertson, Zagaroza, 2009; Barrios et al., 2015); $avgDL$ – средняя длина предложений коллекции; $IDF(s_i)$ – обратное число предложений, в которых встречается слово s_i . Используется сглаженная формула (на основании данных вероятностной модели Робертсона существует предположение, что более надежные оценки получаются, если вводить некий псевдосчет частоты 0,5 (Robertson, Sparck, 1976; Robertson, Zagaroza, 2009)), отрицательные значения логарифма заменяются на среднее IDF с коэффициентом ε :

$$IDF(s_i) = \begin{cases} \log \frac{N - n(s_i) + 0,5}{n(s_i) + 0,5}, & n(s_i) > N/2; \\ \varepsilon avgIDF, & n(s_i) \leq N/2, \end{cases} \quad (5)$$

где N – общее число предложений; $n(s_i)$ – число предложений, в которых встретилось слово s_i .

Таким образом, представление документа в виде ключевых слов (включая визуализацию связей между словами) позволяет составить представление о содержании документа в самом первом приближении, извлеченные предложения – познакомиться чуть более подробно с ключевыми положениями документа. В этой связи можно провести некую аналогию с метаинформацией для научных публикаций (ключевые слова и аннотация к статье).

¹⁰ BM – сокращение от *best matching* – алгоритм, разрабатываемый с целью улучшения работы критерия $TF-IDF$. Среди прочих алгоритмов BM-25 показал наилучшие результаты. В перечне других алгоритмов он был 25-м по счету, за что и получил устоявшееся сокращенное название BM-25.

3. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Анализ нормативных документов проводился в программной среде Python¹¹.

Получение представления о документах реализовано с помощью алгоритма TextRank для извлечения ключевых слов и словосочетаний¹². Итоговые значения параметров алгоритма приведены в табл. 1.

Для краткости в данной статье мы приводим результаты только для одного из анализируемых документов (№ 31 «О мерах по профилактике заболеваний, обусловленных дефицитом микронутриентов, развитию производства пищевых продуктов»); результаты резюмирования для всех документов приведены в репозитории¹³.

Пример графа связей между ключевыми словами показан на рис. 1.

Как видно из рис. 1, основной акцент в документе делается на аспектах производства хлебобулочных изделий (самая сильная связь между словами).

¹¹ Файлы данных, исходный код, результаты выложены в де/репозитории GitHub // Веб-сервис GitHub. URL: https://github.com/behavioral-econ-codes/docs_summarization/ (дата обращения: 12.04.2021).

¹² Несмотря на наличие готовых библиотек, позволяющих извлекать ключевые слова и предложения с помощью алгоритма TextRank, их гибкость существенно ограничена. В исследовании синтезировались некоторые базовые библиотеки, модификации базовых библиотек и собственные процедуры. Использовались следующие библиотеки Python: NLTK; модификация библиотеки mystem для предобработки текста; pagerank – для запуска алгоритма PageRank на основе построенного графа; библиотеки matplotlib; networkx – для визуализации связей между ключевыми словами; gensim – для реализации алгоритма BM-25 для резюмирования.

¹³ Веб-сервис GitHub. URL: https://github.com/behavioral-econ-codes/docs_summarization/blob/main/results/summary_marks.xlsx?raw=true (дата обращения: 12.04.2021).

Таблица 1
Параметры алгоритма TextRank для извлечения
ключевых слов и предложений

Задача	Мера близости	Параметры
Извлечение ключевых слов	Внутри окна $N = 3; (1)$	Число извлекаемых слов – 20
Извлечение ключевых предложений	BM-25; (4)	$k_1 = 1,5; b = 0,75;$ $\varepsilon = 0,25.$ Число извлекаемых слов – не более 250

В табл. 2 приведен список извлеченных ключевых слов, а также ключевых предложений для Постановления № 31.

Точность извлечения ключевых предложений оценивалась экспертно, каждому предложению в резюмировании ставилась оценка: «очень полезное», «полезное», «бесполезное» или «вводит в заблуждение» (Schmedding et al., 2018). Результаты оценивания точности резюмирования для всей коллекции представлены в табл. 3.

В целом на основании ключевых аспектов корпуса документов (перечень документов

дан в Приложении) можно охарактеризовать следующим образом политику страны в области продовольствия и питания. Во-первых, научными и лечебными учреждениями выявлен ряд нарушений питания населения РФ (в том числе населения из групп повышенного риска: беременных и кормящих женщин, детей раннего возраста), которые ведут к ухудшению состояния здоровья и развитию заболеваний. Отмечается важность обеспечения полноценным питанием беременных женщин, кормящих матерей, детей в возрасте до трех лет, а также детей и женщин с нарушением состояния здоровья, из малоимущих семей (разработаны рекомендуемые наборы продуктов), в том числе через специальные пункты питания (на основе ключевых предложений документа № 15-3/691-04). Ключевые моменты в области профилактики заболеваний, обусловленных дефицитом микронутриентов, лежат в области обогащения хлеба и хлебобулочных изделий микронутриентами и витаминами группы В, развития функционального и специализированного хлебопечения (на основе ключевых предложений Постановления № 31). Таким образом, наряду со стан-

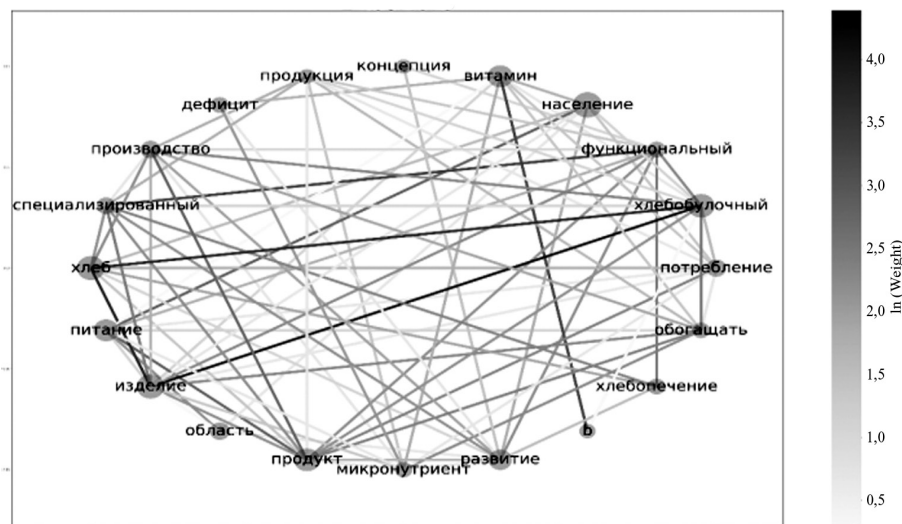


Рис. 1. Граф семантических связей между ключевыми терминами Постановления № 31 «О мерах по профилактике заболеваний, обусловленных дефицитом микронутриентов, развитию производства пищевых продуктов»

Таблица 2

Ключевые термины и ключевые предложения Постановления № 31

Название документа	Ключевые термины	Ключевые предложения
Постановление № 31 «О мерах по профилактике заболеваний, обусловленных дефицитом микронутриентов, развитию производства пищевых продуктов»	Хлеб и хлебобулочное изделие, функциональное и специализированное хлебопечение, функциональное и специализированное назначение, питание населения, витамин В, продукты питания, производство продуктов, обогащение продуктов, обогащение хлебобулочных изделий, обогащение микронутриентами, область питания, дефицит микронутриентов и витаминов, развитие концепции, потребление микронутриентов	Основами государственной политики Российской Федерации в области здорового питания населения на период до 2020 г., утвержденными распоряжением Правительства Российской Федерации от 25.10.2010 № 1873-р, предусмотрено увеличение доли производства продуктов массового потребления, обогащенных витаминами и минеральными веществами, включая массовые сорта хлебобулочных изделий, а также молочные продукты до 40–50% общего объема производства. 1. В настоящее время по инициативе Российской гильдии пекарей и кондитеров при участии Федеральной службы по надзору в сфере защиты прав потребителей и благополучия человека, Федерального государственного бюджетного учреждения «Научно-исследовательский институт питания» Российской академии медицинских наук разработана Концепция обеспечения санитарно-эпидемиологического благополучия населения путем развития функционального и специализированного хлебопечения в Российской Федерации до 2020 года (Хлеб – это здоровье). 2. Рекомендовать юридическим лицам и индивидуальным предпринимателям, осуществляющим производство пищевых продуктов, принять меры по расширению ассортимента и выпуску продукции, обогащенной йодом путем использования в составе рецептур йодированной соли и других эффективных, доступных источников йода, производству продукции из цельнозернового зерна, а также обогащенной другими необходимыми макро- и микронутриентами (пищевые волокна, витамины группы В, фолиевая кислота, железо и др.), исключению глютена из состава отдельных видов хлеба и хлебобулочных изделий специализированного назначения. Концепция развития функционального и специализированного хлебопечения в Российской Федерации до 2020 г. (Хлеб – это здоровье) направлена на снижение микронутриентной недостаточности посредством потребления хлеба и хлебобулочных изделий функционального и специализированного назначения. Таким образом, анализ объемов производства и потребления хлеба, изменений структуры ассортимента вырабатываемых изделий, их пищевой ценности свидетельствует об уменьшении поступления микронутриентов (витаминов и минеральных веществ) с одним из массовых продуктов питания, каковым является хлеб и хлебобулочные изделия, вследствие чего этот продукт утрачивает свою роль основного источника витаминов группы В и ряда минеральных веществ (кальция, железа, йода) в питании населения России

Таблица 3
Оценки точности результатов резюмирования

Категория	Число, %
Очень полезное	92(63)
Полезное	44(30)
Бесполезное	11(7)
Вводит в заблуждение	0(0)

дартами потребления, существует *политика производства обогащенных продуктов*. Данный факт свидетельствует о том, что государство не может обеспечить граждан сбалансированной диетой, однако предлагаются меры для восполнения их нехватки на индивидуальном уровне.

Также существуют нормы питания в больницах, родильных домах, школьных и до-

школьных организациях (МР 28-1/10/2-1994, МР 2.3.1.2432-08.2.3.1), что задает некоторый стандарт потребления. С экономической точки зрения стандарт влияет на рынок через государственные закупки, создавая гарантированный спрос со стороны учреждений соответствующего профиля. Полученные данные могут быть использованы, например, как входные для инициализации агентоориентированных моделей для оценки эффективности различных продовольственных сценариев.

Другим аспектом является анализ документов, устанавливающих налоговые льготы сельскохозяйственным производителям и пищевым компаниям, регулирующих цены на продовольственном рынке и деятельность производителей, регламентирующих распределение талонов на питание (если имеются), школьное питание. Проведенный анализ определяет условия освобождения от исполнения обязанностей налогоплательщика (НК РФ, ст. 145), отмечает наличие двух видов производителей в зависимости от режима налогообложения (переход на Единый сельскохозяйственный налог (ЕСХН) или общая система налогообложения) (НК РФ, гл. 26.1, ст. 346.1–346.8). На основе данной информации исследователь может, например, в рамках построения агентоориентированной модели предусмотреть различия производителей по различным параметрам. Важно отметить также развитие системы поддержки фермеров и развитие сельской кооперации (Постановление № 16), выделение субсидий на развитие малых форм хозяйствования, развитие приоритетных подотраслей агропромышленного комплекса (№ 380-ФЗ).

ОБСУЖДЕНИЕ И ВЫВОДЫ

Для поддержки принятия решений в области питания (равно как и в любой другой области) необходимо уделять отдельное внимание

вопросам правовой аналитики. Как правило, число, объем, а также частота обновления нормативных документов достаточно велики, что осложняет их понимание для исследователей. Тем не менее стремительное развитие методов анализа естественного языка и компьютерной лингвистики позволяет в значительной степени использовать их в рамках междисциплинарных исследований.

Оправданность использования графового метода TextRank для извлечения ключевой информации обусловлена его независимостью от естественного языка (возможно расширение в сторону включения нормативных документов на других языках), возможностью визуализации семантических связей между ключевыми терминами. Важным аспектом работы алгоритма является способ определения весов ребер (или меры близости вершин – слов или предложений). В данной работе для определения меры близости предложений использовался алгоритм BM-25, позволяющий присваивать более высокий вес связям между предложениями, схожими нечастотными словами. Отметим, что в области компьютерной лингвистики и технических наук, существует огромный спектр конкурирующих между собой алгоритмов для определения как близости элементов, так и способов смысловой компрессии информации (Aries et al., 2019). Выбранный в работе подход оправдан для гомогенных по функциональному стилю корпусов текстов, к которым относятся нормативные документы.

Отметим также, что алгоритм построения и обработки графов TextRank может быть модифицирован путем задействования более сложных конструкций, например использования графов не только слово–слово или предложение–предложение, но и слово–предложение (например, (Ying et al., 2017)) или в качестве вершин – тем, формируемых на основе отобранных слов-кандидатов (Bougouin et al., 2013) и др. В случае если коллекция документов большая, подход может быть расширен методами кластеризации, тематическим моделированием.

В целом методы смысловой компрессии не отменяют дальнейшее линейное чтение интересующих текстов, однако с учетом знаний ключевых положений всех документов. Известно, что когнитивные вычисления могут быть заменены простыми перцептивными выводами, которые способны улучшить понимание, память и принятие решений (Heer, 2010). В данном контексте информационный дизайн (в виде визуализации связей) – попытка создать внешние представления для усиления познания (Card et al., 1999).

Таким образом, именно комбинация смысловой компрессии и визуализации информации как «крупный план» текстовых до-

кументов, а также возможности дальнейшей детализации путем линейного чтения и анализа являются наиболее актуальным подходом в условиях перенасыщения информации и дефицита внимания. Особенно актуально активное внедрение методов аналитики текстов для систем, которые не борются за внимание потребителей. Именно такие системы существенно отстают в удобстве извлечения значимой информации. Подходы к улучшению понимания больших объемов нормативных документов позволят принести существенную пользу исследователям в рамках экономических, юридических или мультидисциплинарных исследований.

ПРИЛОЖЕНИЕ

Перечень и характеристики нормативных документов, регламентирующих вопросы питания и продовольствия в отношении ЖДА

	Номер	Название	Дата введения	Число слов	Число таблиц
1	МР 2.3.1.2432-08. 2.3.1.	Методические рекомендации «Нормы физиологических потребностей в энергии и пищевых веществах для различных групп населения Российской Федерации»	18.12.2008	9417	9
2	28-1/10/2-1994	Методические рекомендации «Рекомендуемые нормы лечебного питания (среднесуточных наборов основных пищевых продуктов) для беременных и кормящих женщин в родильных домах (отделениях) и детей различных возрастных групп в детских больницах (отделениях) Российской Федерации»	24.03.2017	6156	15
3	№ 148	Постановление «О дополнительных мерах по профилактике заболеваний, обусловленных дефицитом железа в структуре питания населения»	16.09.2003	1147	0
4	№ 31	Постановление «О мерах по профилактике заболеваний, обусловленных дефицитом микронутриентов, развитию производства пищевых продуктов функционального и специализированного назначения»	14.06.2013	9815	9
5	№ 28	Постановление «Об усилении надзора за производством и оборотом пищевых продуктов»	29.08.2006	2272	0
6	СанПиН 2.4.5.2409-08	Санитарно-эпидемиологические требования к организации питания обучающихся в общеобразовательных учреждениях, учреждениях начального и среднего профессионального образования	23.07.2008	12 591	6
7	СанПиН 2.3.2.2804-10	Об утверждении СанПиН 2.3.2.2804-10 «Дополнения и изменения № 22 к СанПиН 2.3.2.1078-01 “Гигиенические требования безопасности и пищевой ценности пищевых продуктов”»	27.12.2010	9403	0
8	МР 2.4.5.0107-15. 2.4.5.	Методические рекомендации «Гигиена. Гигиена детей и подростков. Детское питание. Организация питания детей дошкольного и школьного возраста в организованных коллективах»	12.11.2015	2669	4

Окончание приложения

	Номер	Название	Дата введения	Число слов	Число таблиц
9	СанПиН 2.3.2.1940-05	Продовольственное сырье и пищевые продукты организация детского питания	01.07.2005	5138	0
10	СанПиН 2.3.2.1078-01	Гигиенические требования безопасности и пищевой ценности пищевых продуктов	06.11.2001	100 640	27
11	СанПиН 2.3.2.2509-09	Дополнение № 14 к СанПиН 2.3.2.1078-01 «Гигиенические требования безопасности и пищевой ценности пищевых продуктов»	05.05.2009	1852	0
12	СанПиН 2.3.2.241-08	Дополнения и изменения № 10 к СанПиН 2.3.2.1078-01 «Гигиенические требования безопасности и пищевой ценности пищевых продуктов»	16.06.2008	7125	0
13	СанПиН 2.3.2.2650-10	Дополнения и изменения № 18 к санитарно-эпидемиологическим правилам и нормативам СанПиН 2.3.2.1078-01 «Гигиенические требования безопасности и пищевой ценности пищевых продуктов»	28.06.2010	4370	0
14	№ 15-3/691-04	Рекомендуемые наборы продуктов для питания беременных женщин, кормящих матерей и детей до 3-х лет	15.05.2006	2259	0
15	№ 06-731	О формировании культуры здорового питания обучающихся, воспитанников	12.04.2012	20 233	4
16	№ 117-ФЗ	НК РФ, гл. 26.1. (ст. 346.1–346.8) «Система налогообложения для сельскохозяйственных товаропроизводителей (единый сельскохозяйственный налог)»	05.08.2000	10 584	0
17	№ 117-ФЗ	НК РФ, ст. 284 (п. 1.3) «Налоговые ставки»	05.08.2000	5596	0
18	№ 117-ФЗ	НК РФ, ст. 145 «Освобождение от исполнения обязанностей налогоплательщика»	05.08.2000	2666	0
19	№ 380-ФЗ	ФЗ «О федеральном бюджете на 2020 год и на плановый период 2021 и 2022 годов» (фрагмент). Распределение субсидий бюджетам субъектов Российской Федерации	02.12.2019	2994	3
20	№ 16	В рамках национального проекта «Малое и среднее предпринимательство и поддержка индивидуальной предпринимательской инициативы». Проект «Создание системы поддержки фермеров и развитие сельской кооперации»	24.12.2018	590	0
21	№ 717	Государственная программа развития сельского хозяйства и регулирования рынков сельскохозяйственной продукции, сырья и продовольствия	14.07.2012	71 782	0
22	№ 264-ФЗ	ФЗ «О развитии сельского хозяйства»	22.12.2006	4729	0
23	№ 1528	Постановление «Правила предоставления из федерального бюджета субсидий российским кредитным организациям, международным финансовым организациям и государственной корпорации развития “ВЭБ.РФ” на возмещение недополученных ими доходов по кредитам, выданным сельскохозяйственным товаропроизводителям (за исключением сельскохозяйственных кредитных потребительских кооперативов), организациям и индивидуальным предпринимателям, осуществляющим производство, первичную и (или) последующую (промышленную) переработку сельскохозяйственной продукции и ее реализацию, по льготной ставке»	29.12.2016	6253	0
24	№ 20	Доктрина продовольственной безопасности Российской Федерации	21.01.2020	3791	0

Список литературы / References

- Ванюшкин А.С., Гращенко Л.А. (2016). Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. № 19. С. 85–93. [Vanyushkin A.S., Grashchenko L.A. (2016). Keyword extraction methods and algorithms. *New information technologies in automated systems*, no. 19, pp. 85–93 (in Russian).]
- Варшавский А.Е. (2015). Инновационные риски в области продуктов питания для России в условиях глобализации и либерализации рынков // Экономическая наука современной России. № (4). С. 91–108. [Varshavskij A.E. (2015). Innovative risks in the field of food products for Russia in the context of globalization and liberalization of markets. *Economics of Contemporary Russia*, no. 4, pp. 91–108 (in Russian).]
- Гусев А.А. (2019). Экономические и институциональные основы обеспечения экологической безопасности // Экономическая наука современной России. № 1 (84). С. 70–81. [Gusev A.A. (2019). Economic and institutional framework for ensuring environmental safety. *Economics of Contemporary Russia*, no. 1 (84), pp. 70–81 (in Russian).] DOI: 10.33293/1609-1442-2019-1(84)-70-81
- Исаков В.Б. (2018). Правовая аналитика как информационный процесс // Право и информация: вопросы теории и практики: сб. матер. междунар. науч.-практ. конф. / науч. ред. Н.А. Шевелёва. 271 с. [Isakov V.B. (2018). Legal analytics as an information process. Law and information: theory and practice: collection of articles. mater. international scientific-practical conf. / scientif. ed. N.A. Shevelyova (sc. ed). 271 s. (in Russian).]
- Милкова М.А. (2018). Извлечение ключевых терминов направления «Цифровая экономика»: графоориентированный подход // Цифровая экономика. № 4 (4). С. 57–65. [Milkova M.A. (2018). Extraction of key terms of the digital economy: A graph-oriented approach. *Digital Economy*, no. 4 (4), pp. 57–65 (in Russian).]
- Милкова М.А. (2019). Теория подталкивания и ее искажения в информационной среде // Цифровая экономика. № 4 (8). С. 21–26. [Milkova M.A. (2019). Nudge theory and its distortions in the information environment. *Digital Economy*, no. 4 (8), pp. 21–26 (in Russian).]
- Милкова М.А. (2020a). Инновационный подход к поиску информации на примере патентного анализа плана импортозамещения // Экономическая наука современной России. № 1 (88). С. 143–157. [Milkova M.A. (2020a). An innovative approach to information retrieval using the example of patent analysis of an import substitution plan. *Economics of Contemporary Russia*, no. 1 (88), pp. 143–157 (in Russian).] [https://doi.org/10.33293/1609-1442-2020-1\(88\)-143-157](https://doi.org/10.33293/1609-1442-2020-1(88)-143-157)
- Милкова М.А. (2020b). OpenTalks.AI: Конференция 20–21 февраля 2020 года // Цифровая экономика. № 1 (9). С. 76–79. [Milkova M.A. (2020b). OpenTalks.AI Conference February 20–21, 2020. *Digital Economy*, no. 1 (9), pp. 76–79 (in Russian).]
- Моретти Ф. (2016). Дальнее чтение / пер. с англ. А. Вдовин, О. Собчук, А. Шели; научн. ред. И. Кушнарева. М.: Издательство Института Гайдара. 352 с. [Moretti F. (2016). Further reading. Transl. from Engl. by A. Vdovin, O. Sobchuk, A. Sheli. Sc. ed. I. Kushnareva. Moscow Gaidar Institute Publishing House. 352 p. (in Russian).]
- Недумов Я.Р., Кузнецов С.Д. (2018). Исследовательский поиск научных статей // Труды ИСП РАН. Т. 30. Вып. 6. С. 171–198. [Nedumov Ya.R., Kuznetsov S.D. (2018). Research Search of Scientific Articles. *Proceedings of ISP RAS*, vol. 30, iss. 6, pp. 171–198 (in Russian).]
- Тарасова И. (2011). Железодефицитная анемия у детей и подростков // Вопросы современной педиатрии. № 10 (2). С. 40–48. [Tarasova I. (2011). Iron deficiency anemia in children and adolescents. *Questions of modern pediatrics*, no. 10 (2), pp. 40–48 (in Russian).]
- Чернов В.М., Тарасова И.С. (2013). Профилактика железодефицитной анемии: обоснование, принципы проведения, эффективность // Поликлиника. № 4. С. 9–12. [Chernov V.M., Tarasova I.S. (2013). Prevention of iron deficiency anemia: rationale, principles of conduct, effectiveness. *Polyclinic*, no. 4, pp. 9–12 (in Russian).]
- Aries A., Zegour D., Hidouci W.K. (2019). Automatic text summarization: What has been done and what has to be done // arXiv:1904.00688

- Ashley K. (2017). Legal text analytics. *Artificial intelligence and legal analytics: New tools for law practice in the digital age*. Cambridge: Cambridge University Press. 446 p.
- Barrios F., López F., Argerich L., Wachenchauzer R. (2015). Variations of the similarity function of TextRank for automated summarization. *Anales de las 44JAIIO. Jornadas Argentinas de Informática. Argentine Symposium on Artificial Intelligence*, 2015.
- Beliga S., Martincic-Ipsic S., Meštrović A. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences*, no. 39 (1).
- Bengio Y., Ducharme R., Vincent P. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, no. 3, pp. 1137–1155.
- Bird S., Klein E., Loper E. (2009). Natural language processing with Python: Analyzing text with the natural language toolkit. Sebastopol: O'Reilly. 504 p.
- Bougouin A., Boudin F., Daille B. (2013). TopicRank: Graph-based topic ranking for keyphrase extraction. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya, Japan, pp. 543–551.
- Brin S., Page L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, no. 30 (1–7).
- Card S.K., Mackinlay J.D., Shneiderman B. (1999). Readings in information visualization: Using vision to think. San Francisco: Morgan Kaufmann.
- Feldman R., Sanger J. (2007). The text mining handbook: Advanced approaches in analyzing unstructured data. Cambridge: Cambridge University Press. 410 p.
- Firoozeh N., Nazarenko A., Alizon F., Daille B. (2019). Keyword extraction: Issues and methods. *Natural Language Engineering*, no. 26 (3), pp. 259–291.
- Horton S., Ross J. (2003). The economics of iron deficiency. *Food Policy*, no. 28, pp. 51–75.
- Heer J., Bostock M., Ogievetsky V. (2010). A tour through the visualization zoo. *Queue*, no. 8 (5), 20.
- Mihalcea R., Tarau P. (2004). TextRank: Bringing order into texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–414.
- Moens M.F. (2007). Summarizing court decisions. *Information Processing and Management*, vol. 43, pp. 1748–1764.
- Pagallo U., Palmirani M., Casanovas P., Sartor G., Villata S. (2018). Introduction: Legal and Ethical Dimensions of AI, NorMAS, and the Web of Data. Pagallo U., Palmirani M., Casanovas P., Sartor G., Villata S. (eds). *Lecture Notes in Artificial Intelligence*. Cham.: Springer.
- Page S., Brin S., Motwani R., Winograd T. (1998). The PageRank Citation Ranking: Bringing order to the web. Technical Report. Stanford: Stanford University.
- Patterson L.E., Blaha L.M., Grinstein G.G., Liggett K.K., Kaveney D.E., Sheldon K.C., Havig P.R., Moore J.A. (2014). A human cognition framework for information visualization. *Computers & Graphics*, no. 42, pp. 42–58.
- Pazienza M.T., Pennacchiotti M., Zanzotto F.M. (2006). Terminology Extraction: An analysis of linguistic and statistical approaches. Sirmakessis S. (ed.) *Knowledge Mining. Studies in Fuzziness and Soft Computing*, vol. 185. Berlin: Springer, Heidelberg. P. 255–279.
- Rizoiu M.-A., Velcin J. (2011). Topic extraction for ontology learning. *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*. Wilson Wong, Wei Liu, Mohammed Bennamoun (eds.). Hershey: IGI Global. P. 38–60.
- Robertson S.E., Sparck K.J. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, no. 27(3), pp. 129–146.
- Robertson S., Zagaroza H. (2009). The probabilistic relevance framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389.
- Saad S., Salim N., Omar N. (2008). Keyphrase extraction for Islamic Knowledge ontology. *International Symposium on Information Technology*. ITSIm, 26–28 Aug., Kuala Lumpur, Malaysia.
- Salton G., Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, no. 5 (24), pp. 513–523.
- Schmedding F., Klügl P., Baehrens D., Simon C., Simon K., Tomanek K. (2018). EuroVoc-Based Summarization of European Case Law. Pagallo U., Palmirani M., Casanovas P., Sartor G., Villata S. (eds). *AI Approaches to the Complexity of Legal Systems*. AICOL 2015, AICOL 2016, AICOL 2016, AICOL 2017, AICOL 2017. *Lecture Notes*

- in *Computer Science*, vol. 10791. Springer, Cham. https://doi.org/10.1007/978-3-030-00178-0_13
- Simon H.A. (1971). Designing organizations for an information-rich world. In: M. Greenberger (ed.). *Computers, communications, and the public interest*. Baltimore: The Johns Hopkins Press.
- Wu C., Marchese M., Jiang J., Ivanyukovich A., Liang Y. (2007). Machine learning-based keywords extraction for scientific literature. *Journal of Universal Computer Science*, vol. 13, no. 10, pp. 1471–1483.
- Ying Y., Qingping T., Qinzhen X., Ping Z., Panpan L. (2017). A graph-based approach of automatic keyphrase extraction. *International Congress of Information and Communication Technology (ICICT 2017). Procedia Computer Science*, no. 107, pp. 248–255.
- Zhang A.X., Verou L., Karger D. (2017). Wikum: Bridging discussion forums and wikis using recursive summarization. *ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2017)*.

Рукопись поступила в редакцию 31.12.2020 г.

MODERN METHODS OF EXTRACTING KEY INFORMATION FROM REGULATORY DOCUMENTS

M.A. Milkova, I.V. Nevolin, D.P. Pigorev

DOI: 10.33293/1609-1442-2021-2(93)-101-114

Maria A. Milkova, Central Economics and Mathematics Institute of Russian Academy of Sciences, Moscow, Russia; m. a.milkova@gmail.com. ORCID 0000-0002-9393-1044
Ivan V. Nevolin, Central Economics and Mathematics Institute of Russian Academy of Sciences, Moscow, Russia; i.nevolin@cemi.rssi.ru
Dmitriy P. Pigorev, Central Economics and Mathematics Institute of Russian Academy of Sciences, Moscow, Russia; dpigorev@ya.ru

Acknowledgment. This study was supported by the Russian Foundation for Basic Research (project no. 19-57-80003).

This article is an attempt to comprehend the difficulties and propose approaches to eliminate them when analyzing legal documents in the framework of economic and interdisciplinary research. The utmost goal is to seek incorporating advances in computational linguistics and natural language analysis into the discourse of the digital economy in order to develop methods involved in decision-making and strategy development based on the analysis of textual information. In conditions when the amount of information is too large, is constantly updated and / or the area of study is new, the most expedient at the first stage is to obtain the general structure of the entire collection of documents, some kind of semantic compression of information. The practical part contains the development of an approach for the analysis of regulations governing food and nutrition issues, in particular, related to the prevention of the development of iron deficiency anemia (IDA). The approach includes the extraction of key information of voluminous texts (keywords and key sentences) based on the TextRank graph algorithm. An important link contributing to cognition is also the visualization of semantic relationships between words within documents. In our opinion, it is the combination of semantic compression and visualization of information as a “close-up” of text documents, as well as the possibility of further detailing by linear reading and analysis, which are the most relevant approach in conditions of information overload and attention deficit. The active introduction of text analytics methods for systems that are not involved in attention markets, which lag significantly behind in the convenience of extracting meaningful information, is especially important. Approaches to improve the understanding of large volumes of regulations will be of significant value to researchers in economic, legal or multidisciplinary research.

Keywords: digital economy, key terms, key term extraction, summarization, key sentences extraction, TextRank, bm-25, graph algorithm, anemia, iron-deficiency anemia.

Classification JEL: D80, D83.

Manuscript received 31.12.2020