ИКОНОМИЧЕСКИ УНИВЕРСИТЕТ – ВАРНА

Катедра "Информатика"

Борис Иванов Банков

СОФТУЕРНА СИСТЕМА ЗА АВТОМАТИЗИРАНА ОБРАБОТКА НА НЕСТРУКТУРИРАНИ ДАННИ ОТ СОЦИАЛНИТЕ МРЕЖИ

АВТОРЕФЕРАТ

на дисертация за присъждане на образователна и научна степен "доктор"

по докторска програма

"Информатика"

Научен консултант: доц. д-р Снежана Сълова

Варна, 2018

Защитата на дисертационния труд ще се състои на 25.06.2018 г. от 13:00 часа в зала 1 на Икономически университет – Варна на заседание на Научно жури, назначено със Заповед № 06-1777 от 02.05.2018 г. на Ректора на Икономически университет – Варна.

Материалите по защитата са на разположение на интернет страницата на Икономически университет – Варна, https://www.ue-varna.bg.

Борис Иванов Банков

СОФТУЕРНА СИСТЕМА ЗА АВТОМАТИЗИРАНА ОБРАБОТКА НА НЕСТРУКТУРИРАНИ ДАННИ ОТ СОЦИАЛНИТЕ МРЕЖИ

АВТОРЕФЕРАТ

на дисертационен труд

за присъждане на образователна и научна степен "доктор"

по докторска програма „Информатика“

в професионално направление „4.6. Информатика и компютърни науки“

НАУЧЕН КОНСУЛТАНТ:

доц. д-р Снежана Динева Сълова

НАУЧНО ЖУРИ:

доц. д-р Тодорка Борисова Атанасова, ИУ-Варна

доц. д-р Снежана Динева Сълова, ИУ-Варна

проф. д-р Аврам Моис Ескенази, ИМИ-БАН

доц. д-р Димитрина Полимирова – Николова, НЛКВ-БАН

доц. д-р Веселина Господинова Жечева, БСУ

РЕЗЕНЗЕНТИ:

проф. д-р Аврам Моис Ескенази

доц. д-р Тодорка Борисова Атанасова

ВАРНА, 2018

I. Обща характеристика на дисертационния труд

1. Актуалност на проблема

Социалните мрежи, обемът и скоростта, с която данните се генерират в Интернет са основните предизвикателства пред софтуерните решения, които са насочени към извличане на знания. Скритият слой от информация в големите текстови масиви, резултат от публикации в социалните мрежи, е ценен ресурс за организации от всяка сфера и бранш. Можем да считаме, че със засилващото се взаимодействие на потребителите в онлайн пространството, нараства и необходимостта от софтуерен продукт, които позволява автоматизирана обработка на неструктурираните данни, извличани от социалните мрежи.

2. Теза

Основна теза на дисертационния труд е, че обработката на неструктурирани данни от социалните мрежи дава възможност за разкриване на нова и полезна за бизнеса информация.

3. Цел и задачи на изследването

Цел на дисертационния труд е да се създаде модел на софтуерна система за обработка на неструктурирани данни, извлечени от социалните мрежи и да се предложи концепция за нейната реализация. С оглед реализиране на поставената цел е необходимо да бъдат решени следните **основни задачи**:

1. Да се анализира състоянието на теоретичните изследвания и да се идентифицират нерешени проблеми в областта на автоматизираната обработка на неструктурирани данни, извлечени от социалните мрежи.
2. Да се създаде модел на софтуерна система за обработка на неструктурирани данни, извлечени от социалните мрежи, като се

дефинира нейната структура, подходът и технологиите за изграждането ѝ.

3. Да се разработи концепция за реализиране и внедряване на софтуерната система за обработка на неструктурирани данни.
4. Да се апробират основните функционалности на предложената софтуерна система за обработка на неструктурирани данни.

4. Обект и предмет на изследване

Обект на изследване са социалните мрежи като източник на данни, а **предмет на изследване** е обработката на неструктурирани текстови данни чрез приложение на математически методи за категоризация на текст.

5. Методология на изследването

При изследването са приложени сравнителният, системният и комплексният подходи, методите на логическия и статистическия анализ, на моделиране и проектиране на информационни системи, на алгоритмизация и др.

6. Апробация

По темата на дисертацията са публикувани две статии и два доклада. Разработен е концептуален модел и са разгледани отделните модули на системата за автоматизирана обработка на неструктурирани данни. За апробация на системата е избрана Медийна група Черно море, а като подходяща социална платформа е определена Twitter. За реализацията на предлаганата система са избрани съвременни софтуерни инструменти и технологии и са представени основните етапи от нейната разработка. Системата е апробирана, чрез провеждане на експеримент в периода 01.04 – 16.04.2018 г.

II. Структура на дисертационния труд

Дисертационният труд има общ обем 188 страници и се състои от въведение, изложение в три глави, заключение, списък на използваната литература от 105 литературни и 47 интернет източника, 14 приложения и списък на публикациите по дисертационния труд. В основния текст са включени 16 таблици и 52 фигури.

Съдържание на дисертационния труд

Глава II. Модел на софтуерна система за автоматизирана обработка на неструктурирани данни от социалните мрежи

2.1. Обхват на разработваната система

2.1.1. Цели и задачи

2.1.2. Концептуален модел на системата

2.2. Модул „Извличане и съхранение“

2.2.1. Специфика на извличане на данни от социалните мрежи

2.2.2. Организация на съхранението на данните в отделните модули на системата

2.3. Модул „Първична обработка“

2.3.1. Разпознаване на думи и предварително почистване на потребителските съобщения

2.3.2. Измерване на честота на срещане на думи

2.3.3. Определяне на значимостта на потребителските съобщения

2.4. Модул „Модел на данните“

2.4.1. Матрица на термините

2.4.2. Намиране на семантично сходство между термините чрез невронна мрежа

2.5. Модул „Клъстеризация“

2.5.1. К-средни и клъстеризация на постоянна емисия от данни

2.5.2. Алгоритъм за определяне на активни и неактивни клъстер

2.6. Модул „Вторична обработка“

2.6.1. Обработка чрез онтологичен подход

2.6.2. Определяне на части на речта чрез BulNet

2.7. Модул „Визуализация“

III. Кратко съдържание на дисертационния труд

Глава I. Концепции и модели за обработка на неструктурирани данни в дигитална среда

В **първа глава** са представени теоретичните постановки в областта на работата с неструктурирани данни, като акцент е поставен върху социалните мрежи като източник на данни. Доказана е необходимостта от разработването на софтуерна система, която позволява обработката на неструктурирани данни на български език с цел разкриване на скрит слой от ценна информация в големия обем от потребителски съобщения, произлизащи от платформите за социално взаимодействие в Интернет.

В първия параграф са разгледани основните понятия и дефиниции за **данни**, **информация** и **познание** на автори като Nicolas Henry[1], Allen Newell[2], Stuart Russell[3] и Chaim Zins[4]. Представени са термините извличане на данни, извличане на информация и извличане на знания, с оглед тяхната употреба в чуждестранната литература и е уточнено използването им в контекста на дисертационния труд. Разгледана е същността и разликите между **структурирани** и **неструктурирани** данни, както и съвкупността от процеси, обединени с термина **обработка**.

Анализирани са научните постановки на William Inmon [5] за предизвикателствата при софтуерната реализация на системи за управление на неструктурираните данни. Разгледани са източниците на неструктурирани данни в организациите, според фирменият отдел, който ги генерира. Развитието на средствата за масова комуникация в Интернет

[1] Henry, N. Knowledge management: a new concern for public administration. Public Administration Review, 1974, pp.189-196.

[2] Newell, A., Simon, H. Computer science as empirical inquiry: Symbols and search. Communications of the ACM, vol. 19, No. 3, pp. 113-126.

[3] Russell, S., Norvig, P. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited, 2016, pp. 234.

[4] Zins, C. Conceptual approaches for defining data, information, and knowledge. Journal of the Association for Information Science and Technology, vol.58, no. 4, 2007, pp. 480.

[5] Inmon, W., Nesavich, A., Tapping into unstructured data, 2007.

допринася за обособяването на нови източници на неструктурирани данни според вида на уеб ресурсите. Считаме, че те са:

- уеб ресурси със свободен или отворен достъп – такива са например световните информационни агенции като Yahoo! News и HuffingtonPost и енциклопедии като Britannica и Wikipedia;
- уеб ресурси с директен или поверителен достъп – в тази категория попадат приложения за управление на имейли Gmail и Outlook и приложенията за чат групи Skype и Discord;
- уеб ресурси със смесен достъп – такива са социалните мрежи като Facebook, Twitter, YouTube, Instagram и др.

Обект на авторски анализ в тази точка са социалните мрежи и техните характеристики. Според съдържанието на потребителските съобщения, публикациите разделяме на текстови, графични и мултимедийни. В края на първи параграф са разгледани най-популярните и разпространени платформи за социално взаимодействие в Интернет, като е представена йерархичната структура на публикациите в Facebook, YouTube, Instagram и Twitter (вж. Таблица 1).

Таблица 1

Наименование на публикациите в социалните мрежи по ниво в йерархията

Социална мрежа / Наименование на публикация	Facebook		YouTube		Instagram	Twitter
Първо ниво	статус	пост	видео	пост	пост	туит
Второ ниво	коментар	коментар	коментар	коментар	коментар	отговор
Трето ниво	отговор	отговор	отговор	-	отговор	отговор

Въведена е актуална класификация на потребителските съобщения според съдържанието и нивото им в тази йерархия за съответната платформа. В Таблица 2, Таблица 3 и Таблица 4 са представени видовете съдържание на публикациите съответно на първо, второ и трето ниво. С отметка са отбелязани тези, които са достатъчни за създаването на публикация, с хикс – забранени или без възможност за използване, с плюс – достъпни, но недостатъчни за създаване на публикация, а с думи са посочени конкретни изключения.

Таблица 2

Съдържание на публикациите в социалните мрежи от първо ниво

Социална мрежа / Съдържание на публикация		Facebook		YouTube		Instagram	Twitter
Първо ниво		статус	пост	видео	пост	пост	туит
Текст		✔	✔	+	✔	+	✔
Графика	Идеограма	✔	✔	+	✔	+	✔
	Пиктограма	✔	✔	x	x	+	✔
Мултимедия		✔	✔	видео	хипервръзка, изображение	изображение, видео	✔

Таблица 3

Съдържание на публикациите в социалните мрежи от второ ниво

Социална мрежа / Съдържание на публикация		Facebook		YouTube		Instagram	Twitter
Второ ниво		коментар	коментар	коментар	коментар	коментар	отговор
Текст		✔	✔	✔	✔	✔	✔
Графика	Идеограма	✔	✔	✔	✔	✔	✔
	Пиктограма	✔	✔	x	x	✔	✔
Мултимедия		✔	✔	хипервръзка	хипервръзка	x	✔

Таблица 4

Съдържание на публикациите в социалните мрежи от трето ниво

Социална мрежа / Съдържание на публикация		Facebook		YouTube		Instagram	Twitter
Трето ниво		отговор	отговор	отговор	-	отговор	отговор
Текст		✔	✔	✔	-	✔	✔
Графика	Идеограма	✔	✔	✔	-	✔	✔
	Пиктограма	✔	✔	x	-	✔	✔
Мултимедия		✔	✔	хипервръзка	-	x	✔

Във **втори параграф** от дисертационния труд са разгледани подходите и методите за обработка на неструктурирани данни. Проследено е развитието на технологиите за текстов анализ и са

представени различните проблеми, свързани с управлението на неструктурирани данни и тяхната интеграция в структурирана среда според William Inmon и Ramana Rao[6].

Представени са класификацията и клъстеризацията като примери за подходи, които се използват в машинното обучение за моделиране на образец на данните. Изследвани са разработките на Aggarwal[7] и Zhao[8] в областта на клъстеризацията на текстови данни, извлечени от социалните мрежи. По-детайлно е разгледан клъстерният анализ, като акцент е поставен върху алгоритъма к-средни и неговото приложение в комбинация с модела на векторно пространство за клъстеризация на текстови данни.

В края на втори параграф се изследва приложението на невронните мрежи при текстовия анализ. За целта са представени моделите Continuous Bag-of-Words (CBOW) и Skip-gram на Tomas Mikolov[9] за обучение на невронна мрежа, чрез които се определят контекст и семантично сходство между термини в множество от неструктурирани данни.

В **третия параграф** се отделя внимание на софтуерните приложения и инструменти за работа с неструктурирани данни. Изследвани са основните предизвикателства на софтуерната обработка на неструктурираните данни, тяхното правилно съхранение и интеграция в структурирана среда. Представени са етапите в развитието на системите за работа със текстови документи и техните особености.

В края на параграфа са разгледани публикации и изследвания на пазара на софтуерни инструменти за текстов анализ. Извършено е проучване на възможността за разпознаване и работа с текст на български език, като от изброените инструменти само един предлага тези функции.

---

[6] Rao, R., From Unstructured Data to Actionable Intelligence, IT Pro, 2003, vol. 5, pp. 29.

[7] Aggarwal, C., Zhai, C., Mining Text Data. Springer Science & Business Media, 2012. pp. 299 и pp. 313

[8] Zhao, Z., Feng, S., Wang, Q., Huang, J., Williams, G., Fan, J. Topic oriented community detection through social objects and link analysis in social networks. Knowledge-Based Systems, vol. 26, 2012, pp. 164.

[9] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. 2013, pp. 3111-3119.

Направено е заключение, че съществува необходимост от **софтуерна система, чрез която да се прилагат съвременните технологии за обработка на данни от неструктурирани интернет източници върху текстови масиви на български език. Поради естеството и формата на потребителските съобщения в социалните мрежи, предлаганата система цели да осигури подходящ алгоритъм за първична обработка на данните.**

Глава II. Модел на софтуерна система за автоматизирана обработка на неструктурирани данни от социалните мрежи

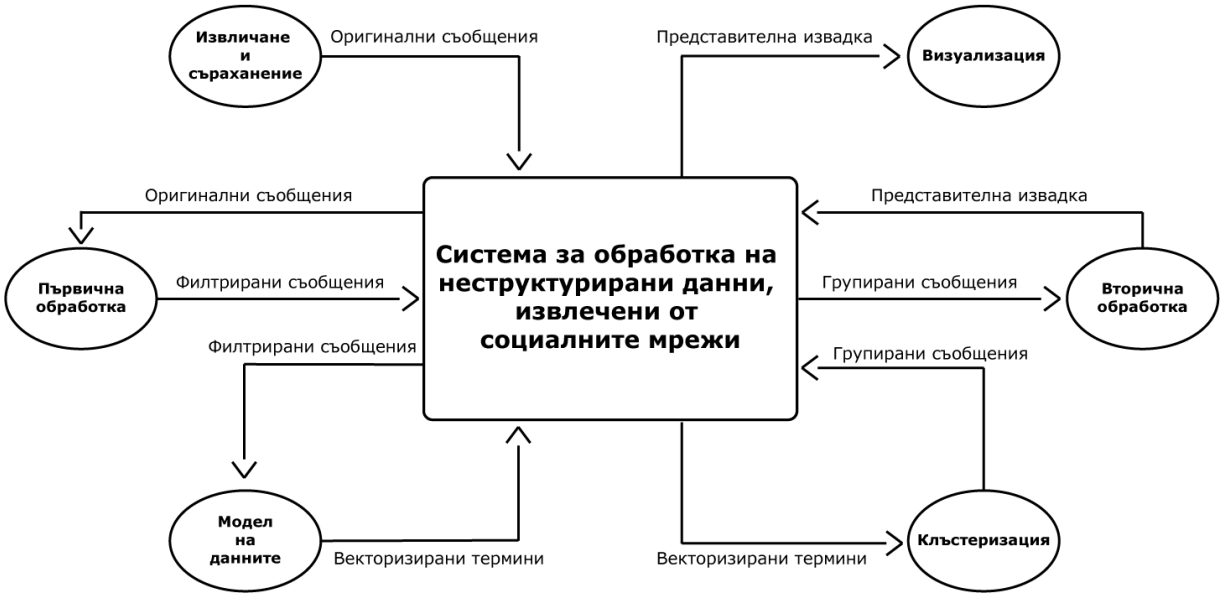
Втора глава е насочена към разработването на концептуален модел на предлаганата система, придружен от обосновка на основните функции на отделните компоненти или модули на системата.

В **първи параграф** са определени обхватът и изискванията към разработваната система. Дефинирана е основната цел на системата, която **е обработването на текстови масиви на български език, което да доведе до откриването на скрит слой информация в многомерното пространство от потребителски съобщения.**

За постигане на целта са предложени набор от задачи:

- **извличане и съхранение на неструктурирани данни от социалните мрежи;**
- **първична обработка на входните данни;**
- **изграждане на модел на данните;**
- **клъстеризация;**
- **вторична обработка на изходните данни;**
- **визуализация на резултатите**.

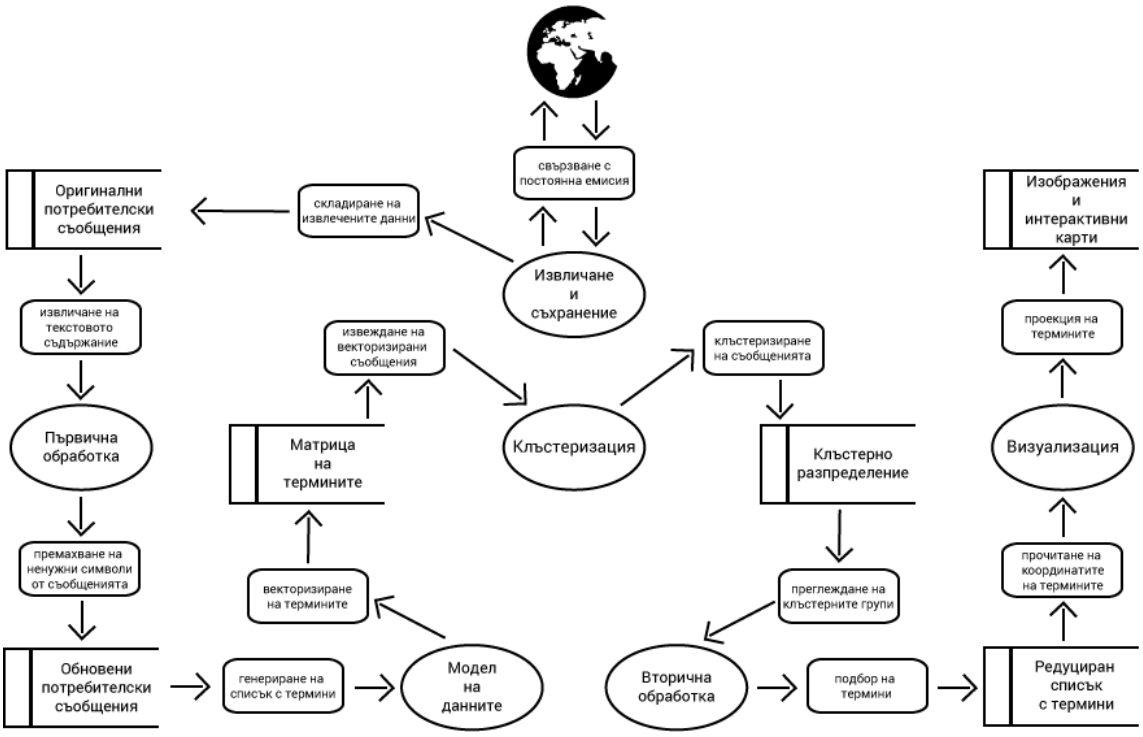
От задачите произтичат 6 модула или подсистеми, които отговарят за изпълнението на основните дейности по обработката на текстовите данни. На Фиг. 1 е представен концептуален модел на предлаганата система, чрез който в абстрактна форма се демонстрират авторовите идеи.



Фиг. 1. Концептуален модел на системата за обработка на неструктурирани данни, извлечени от социалните мрежи

Във **втория параграф** са разгледани основните функции на модул **„Извличане и съхранение“**. Дисертационният труд е насочен към на извличането на данни в реално време (*live stream*), чрез постоянна емисия потребителски съобщения, предоставена от съответната социална мрежа посредством приложен интерфейс. В този параграф се демонстрира нуждата от използването на инвертирано индексиране на заявки при потребителско търсене, поради големия обем и интензитет на постъпване на данни.

Модул „Извличане и съхранение“ включва и описание на основните структури от данни, които се генерират при работата на отделните модули. В края на параграфа е изведена диаграма на потоците от данни от първо ниво в системата (вж. Фиг. 2).



Фиг. 2. Диаграма на потоците от данни от първо ниво

Трети параграф включва основната функционалност от модул **„Първична обработка“** по подготовката на данните за текстов анализ. Правилата при първичната обработка са съобразени със същността и особеностите на публикациите в социалните мрежи. Разгледани са правилата за определяне на думи в компютърна среда, като се доразвива идеята, че думата е „последователност от буквени символи между два интервала“[10].

В модула е предложен алгоритъм за първична обработка на потребителските съобщения, чрез разпознаване на думи и квазидуми, пречистване на съдържанието на потребителските съобщения, премахване на ненужни пунктуационни знаци, премахване на символи, които не присъстват в кирилицата и латиницата, филтриране на съобщения, съдържащи само хипервръзки и др.

[10] Паскалева, Е. Компютърна морфология. Ресурси и инструменти. София, 2007, с. 11.

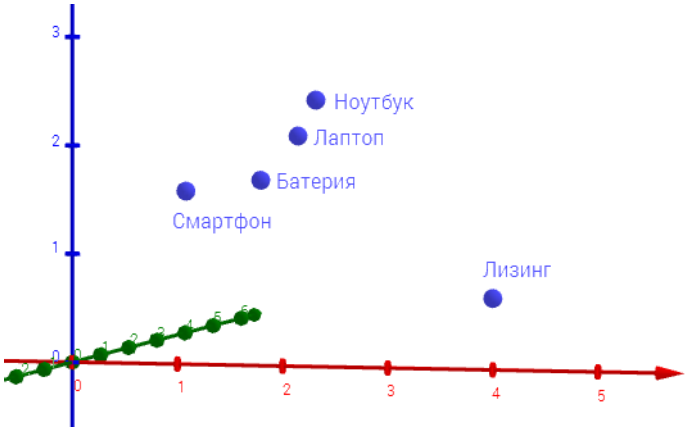
Тук се определя и честотата на срещане на думи чрез индексът tf-idf (Term Frequency – Inverse Document Frequency или честота на термина – обратна честота на документа).

В края на параграфа е предложен алгоритъм за определяне на значимостта и информационната стойност на потребителските съобщения. Ако се приеме, че за дадена социална мрежа има информация относно броя харесвания n_1, броя коментари n_2 и броя споделяния n_3, то теглата им, които се определят от потребителя, са ω_1, ω_2, ω_3, Значимостта на дадено съобщение се пресмята по формула (1) като сума от произведението на отделните тегла на тези характеристики и тяхното количество:

$$\omega = 1 + \sum_{i=1}^{k} \omega_i * n_i \qquad (1)$$

В **четвърти параграф** са представени дейностите от работата на модул **„Модел на данните“**. Той служи за преобразуването на потребителските съобщения във вектори и предсказване на семантично сходни думи, чрез използването на невронна мрежа.

В началото на параграфа са представени особеностите при изграждане на матрица от термини на база потребителските съобщения, извлечени от социалните мрежи. Тук се основаваме на идеята на Tomas Mikolov за определянето на думи-вектори (*word embeddings*), които в многомерното пространство трябва да бъдат в близост, ако са семантично сходни или далече едни от други, ако са твърде различни (вж. Фиг. 3).

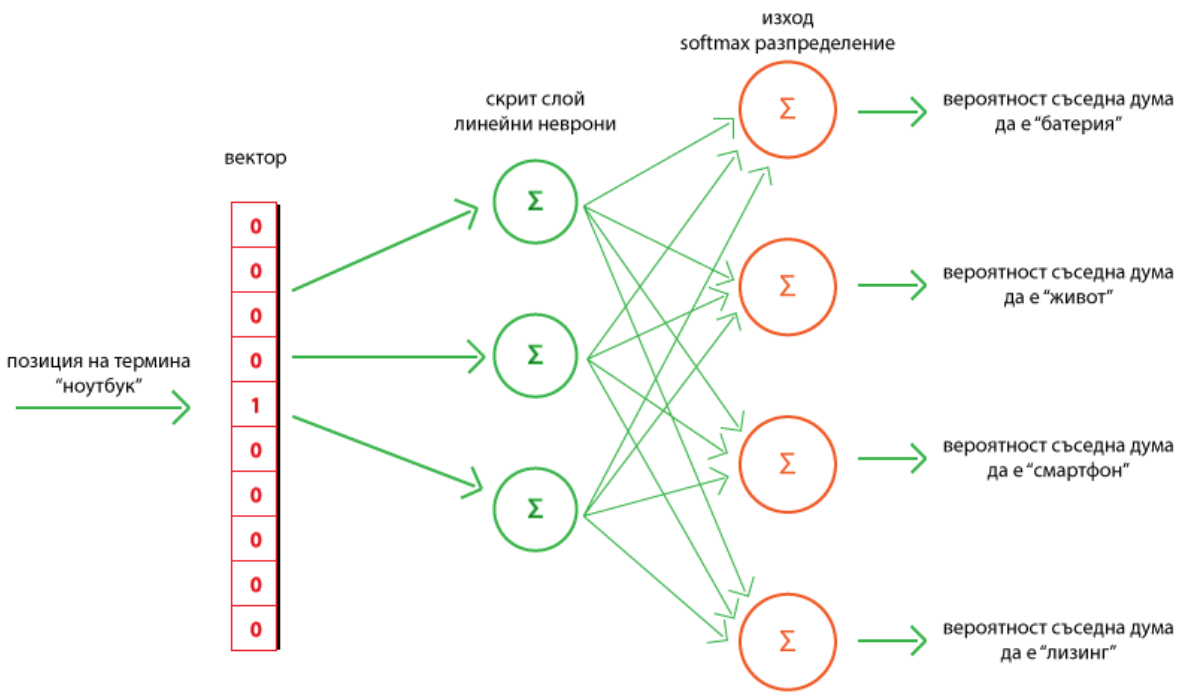


Фиг. 3. Примерно разпределение на термини в триизмерно пространство

За намирането на сходство между думите в текста се прилага технологията **word2vec**, която чрез логистична регресия прави предположение за избран термин, кои са вероятните съседни думи или контекста, в който се намира той (вж. Фиг. 4).

Фиг. 4. Определяне на контекст, чрез съседни думи

За реализацията на модела на данните word2vec се използва права, равнинно свързана невронна мрежа с обратно разпространяващ се Skip-gram. В контекста на обработката на текстови данни думата-вектор и съседите-класификатори се използват за обучение на невронна мрежа, чийто резултат е предположение за елементите от множеството до колко са семантично еднакви или близки (вж. Фиг. 5).



Фиг. 5. Архитектура на права равнинно свързана невронна мрежа

Изграждането на модел, предсказващ контекста на дадена дума е съпътстван и от изчисление на **softmax** функция и по конкретно **sampled softmax**. Чрез **softmax,** изходът от невронна мрежа се свежда сумарно до единица, като елементите на матрицата са в диапазона {0,1} и представляват вероятностно разпределение на дискретни независими променливи (вж. Фиг. 6)

Фиг. 6. Преобразуване на вероятностно разпределение със softmax

В **пети параграф** са представени функциите на модул **„Клъстеризация“**. Той включва процесите по категоризиране и обединение на семантично близки потребителски съобщения. Тук се прилагат алгоритми за намиране на разстоянието между векторите,

определят се клъстерни центрове и се групират обекти от многомерното пространство.

Централна част от авторско изследване са алгоритмите за клъстеризация на постоянна емисия от данни на автори като Shi Zhong[11], Charu Aggarwal и Philip Yu[12] и Marcel Ackermann[13]. На база на техни разработки е предложен алгоритъм за определяне на активни и изключването на неактивни клъстери, с оглед динамичността на дискусиите в социалните мрежи.

Алгоритъмът използва величините **време, значимост и брой вектори**, с които се изпълняват и проверяват набор от правила за отделните клъстери. Така за клъстер ***A*** времето намира изражение в три променливи:

- t_s – времето, изразено като сумата от периоди на неактивност на клъстер ***A***;
- t_a – времето, изразено като сумата от периоди на активност на клъстер ***A***;
- T – представлява периодът на „живот“ на клъстера, сума на t_s и t_a.

Приемаме за клъстер ***A***, че общата **реална** значимост на векторите в него се изразява с ω, а броят им е n. С $\frac{\omega}{n}$ се намира средната значимост за една точка от клъстер ***A***. С променлива N изразяваме общия брой обекти, постъпили във векторното пространство по време на периода на живот T. При $\frac{N}{T} * t_a$, което дава максималния допустим брой точки, с произведението от формула (2) може да се изчисли **максималната допустима значимост ω'** на клъстер ***A*.**

$$\omega' = \frac{N}{T} * t_a * \frac{\omega}{n},\ T \neq 0, n \neq 0 \qquad (2)$$

[11] Zhong, S. Efficient Streaming Text Clustering. Neural Networks, vol. 18, no. 5-6, 2005, pp. 790-798.

[12] Aggarwal, C., Yu, P., A Framework for Clustering Massive Text and Categorical Data Streams, SIAM Conference on Data Mining, 2006, pp. 479-483.

[13] Ackermann, M., Märtens, M., Raupach, C., Swierkot, K., Lammersen, C., Sohler, C. StreamKM++: A clustering algorithm for data streams. Journal of Experimental Algorithmics (JEA). vol. 17, 2012, pp. 2-4.

С помощта на ω' може да бъде определено **време за възстановяване t_r**, през което се поддържа клъстер $\boldsymbol{A}$, след като t_s и t_a са се изравнили. За целта се използва идеята на Aggarwal и Yu за *half-life* (период, за който значимостта на клъстера намалява наполовина), но е предложено въвеждането и използването на величината максимална допустима значимост $\frac{1}{2} * \omega'$. Ако реалната значимост ω намалява равномерно, **то периодът от изравняване на t_s и t_a до момента на изключване на клъстер A е времето за достигане от $\boldsymbol{\omega}$ до $\frac{1}{2} * \boldsymbol{\omega'}$.** За да се контролира скоростта, с която намалява значимостта е използван фактор на разпад от $\frac{1}{\gamma}$, където γ представлява темпото, с което $\boldsymbol{\omega}$ спада за единица време. Факторът на разпад се определя от потребителя и приема стойност между 0 и 1. **За изчисление на времето за възстановяване t_r се предлага формула** (3):

$$\boldsymbol{t_r} = \left(\boldsymbol{\omega} - \frac{1}{2} * \boldsymbol{\omega'}\right) * \frac{1}{\gamma}, \gamma \neq 0 \qquad (3)$$

В края на параграфа е използван пример за онагледяване на процеса по определяне на времето за възстановяване.

Шести параграф от втора глава на дисертационния труд е насочен към функциите на модул „**Вторична обработка“**. Тук се изпълняват дейностите по извличането на подходящи думи от векторните образования, които семантично описват клъстерните групи.

В точката е разгледано приложението на онтологии за анализ на контекста, в които се намират текстовите данни. Автори като Andreas Hotho[14] изследват възможността за извличане на синонимни множества от лексикална база като WordNet за определяне на йерархия от концепции, на които отговарят резултатите от клъстеризация.

[14] Hotho, A., Staab, S., Maedche, A. Ontology-based Text Clustering, KI, 16,4, 2002, pp. 48-54.

В края на параграфа се предлага използването на BulNet и системата Хидра (Hydra) за подбор на термини, които семантично описват клъстерните групи. Кандидат-термините преминават през морфологичен анализ, като крайната цел е да се изберат определен брой съществителни имена, между които няма синонимна връзка, за да се даде по-широка представа за информацията в клъстера.

Фиг. 7. Процес на подбор на термини

Седми параграф рамкира същността на последния модул от разработваната система – модул **„Виузализация“.** Тук се използват съвременни технологии за преобразуването на обекти от многомерно пространство в точки върху двумерна равнина. За целта е представен алгоритъмът t-SNE, чрез който се съхраняват локални структури и се поддържат разнородни обекти на голямо разстояние едни от други.

С модула се генерират изображения и интерактивни карти на векторното пространство, в които са нанесени термините, описващи семантично съдържанието на клъстерните групи.

Глава III. Софтуерна реализация на системата за автоматизирана обработка на неструктурирани данни от социалните мрежи в Медийна група Черно море

Трета глава представя избраната организация, в която се внедрява системата, определяне на подходяща социална мрежа за апробация, анализ на технологични средства и инструменти, основните моменти от разработката и провеждането на експеримент в периода 01.04 – 16.04.2018 г.

В **първи параграф** от трета глава са разгледани дейността и нуждите на Медийна група Черно море. Основната цел на организацията е подготвянето на репортажи и излъчването на най-актуалните новини за град Варна и региона. Работата със социалните мрежи е част от задълженията на служителите в организацията. Отразяването на събития и верификацията на източници се случват изключително динамично. Направен е извод, че съществува необходимост от софтуерна система за мониторинг на социалните мрежи, с цел автоматизиране на процеса по извличане и анализ на публикуваните потребителски съобщения.

Параграфът продължава с анализ на възможностите на социалните мрежи Facebook, YouTube, Instagram и Twitter за предоставяне на достъп до публичните емисии от потребителски съобщения. От разгледаните приложни интерфейси за връзка с тези платформи, най-подходящ за апробация на системата е Track API на Twitter.

Вторият параграф представя задълбочено изследване и избор на софтуерни средства за разработване на функционалните възможности на отделните модули. Взетите решения при определяне на подходящ инструментариум са базирани на популярността и разпространението на избраната технология, достъпността й и наличието на лиценз за свободно използване.

За модул „Извличане и съхранение“ са разгледани библиотеки за свързване с приложния интерфейс на платформата Twitter, написани на

различни програмни езици. Избрани са скриптовият сървърен език **PHP** и библиотеката **twitterauth**. Параграфът продължава с анализ на системи за управление на бази от данни и е определена **MySQL** като достатъчно, надеждно и скалируемо решение с отворен код.

Друг компонент от „Извличане и съхранение“, които спомага за комуникацията между отделните модули, относно състоянието съхраняваните данни, е системата за разпределено управление на съобщения. В анализа са включени 5 софтуерни продукта, като най-подходящ за целите на разработваната система е **Apache Kafka**.

Програмното осигуряване на модул „Първична обработка“ се изпълнява с PHP и авторски списък от регулярни изрази, чрез които се пречистват и филтрират извлечените потребителски съобщения. В допълнение са разгледани разработки, написани на PHP за определяне на естествения език, използван в публикациите. За нуждите на системата е предложена библиотеката **language detection**.

За нуждите на модулите „Модел на данните“ и „Клъстеризация“ е направено сравнение между най-използвани програмни езици за извършване на машинно обучение – Python и R. Повечето проучвания не дават категорично заключение за това кой от двата има преднина в тази сфера. Според изследвания IEEE Spectrum и IBM [15] за най-търсени специалисти с познания в сферата на машинното обучение Python е на първо място сред най-разпространените и употребявани програмни езици за 2017 година. С цел улесняване на процеса по разработка е избран **Python**, поради факта, че в модул „Вторична обработка“ се използва системата Hydra, която е написана същия език.

Построяването на модела от данните и обучението на невронната мрежа са функции, за чието изпълнение е избрана библиотеката

[15] The Most Popular Languages for Data Science and Analytics – 2017, http://makemeanalyst.com/most-popular-languages-for-data-science-and-analytics-2017/, 04.04.2018.

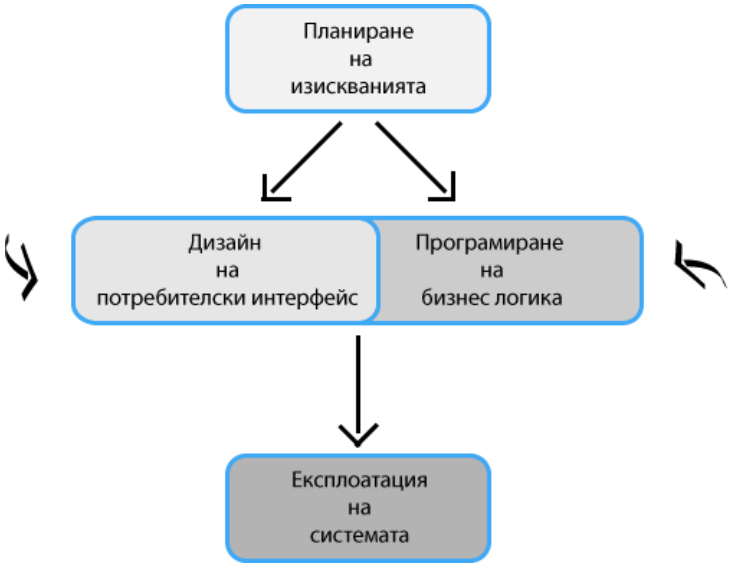
TensorFlow. Тя позволява създаване на структури от графи, където върхът или възелът (*node*) представлява математическа операция, а ребрата или дъгите са входните и изходните данни (*tensors*). В един tensor може да съхранява едно число или многомерен масив от стойности. Чрез тази структура става възможно децентрализираното изпълнение на операциите. Изграждането на модел на данните например може да се извършва от един или няколко централни процесора. Визуализацията на сложни математически функции се поема от един или няколко графични процесора.

Като подходящо средство за извършване на клъстеризацията е посочен пакетът от инструменти с отворен код **scikit-learn** за Python. В него са включени най-популярните библиотеки NumPy и SciPy, които съдържат алгоритми за клъстеризация от семейството на k-means, функционалност за работа с многомерни структури от данни и тяхната визуализация.

Разработването на модул „Вторична обработка“ включва интеграция на услугите на системата **Hydra**. Чрез нея правилно могат да се определят кандидат-термини, описващи същността на клъстерните групи. Тя е платформено независима програма за създаване и валидизиране на лексикално-семантични мрежи.

Дейностите по визуализация на резултатите от клъстеризацията са обвързани с изчисление на разстоянието между обектите от многомерното пространство и тяхната проекция в двумерна равнина. За съхранение на локалните структури се използва алгоритъмът **t-SNE**. Направено е проучване на съвременните библиотеки за визуализация на данни и са разгледани техните предимства и недостатъци. Възможността да се доразвиват графики в реално време е добра предпоставка да бъде избрана библиотеката **Bokeh**.

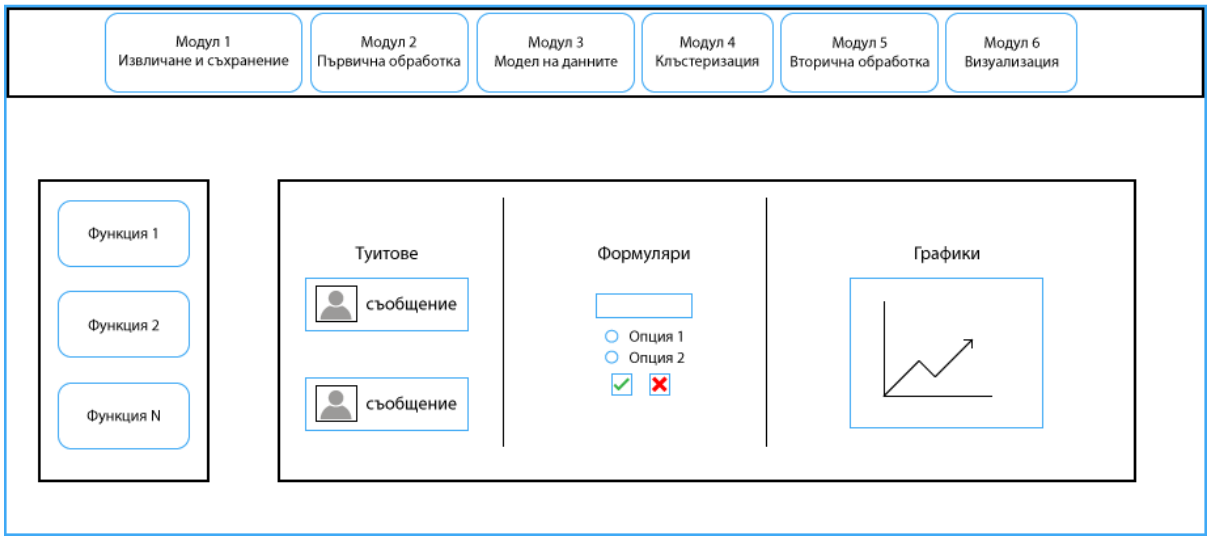
В **трети параграф** се проследяват основните етапи от разработката на предлаганата система. Избрана е избрана методологията на Бързата разработка на приложения (Rapid Application Development – RAD), предложена от James Martin[16] (вж. Фиг. 8). При тази методология етапът на планиране е сравнително кратък, с цел по-бързата реализация на софтуерния продукт. Етапите на дизайн и програмиране се извършват с участие на бъдещите потребители на системата, като се създават модели и прототипи на входни и изходни екрани.



Фиг. 8. RAD жизнен цикъл на софтуерна разработка

Първо е представен обобщен модел на потребителския интерфейс (вж. Фиг. 9), а след това за всеки модул е демонстриран потребителски интерфейс и са описани основните опции и настройки на съответния екран.

[16] Martin, J. Rapid Application Development. Macmillan. 1991.



Фиг. 9. Модел на потребителския интерфейс на разработваната система

За допълнително онагледяване на програмната реализация са публикувани набор от приложения към дисертационния труд, съдържащи класове, методи и процедури за всеки един от модулите на системата.

Четвърти параграф представя резултатите от направена апробация на системата в периода 01.04 – 16.04.2018 г. През тези дни в диапазона 08:00 – 24:00 часа се изтеглят потребителски съобщения от платформата Twitter. Извлечените данни и резултатите от работата на отделните модули са достъпни в Интернет на адрес https://phd.deepcloud.eu.

Апробацията е изпълнена на самостоятелна компютърна конфигурация със следните спецификации:

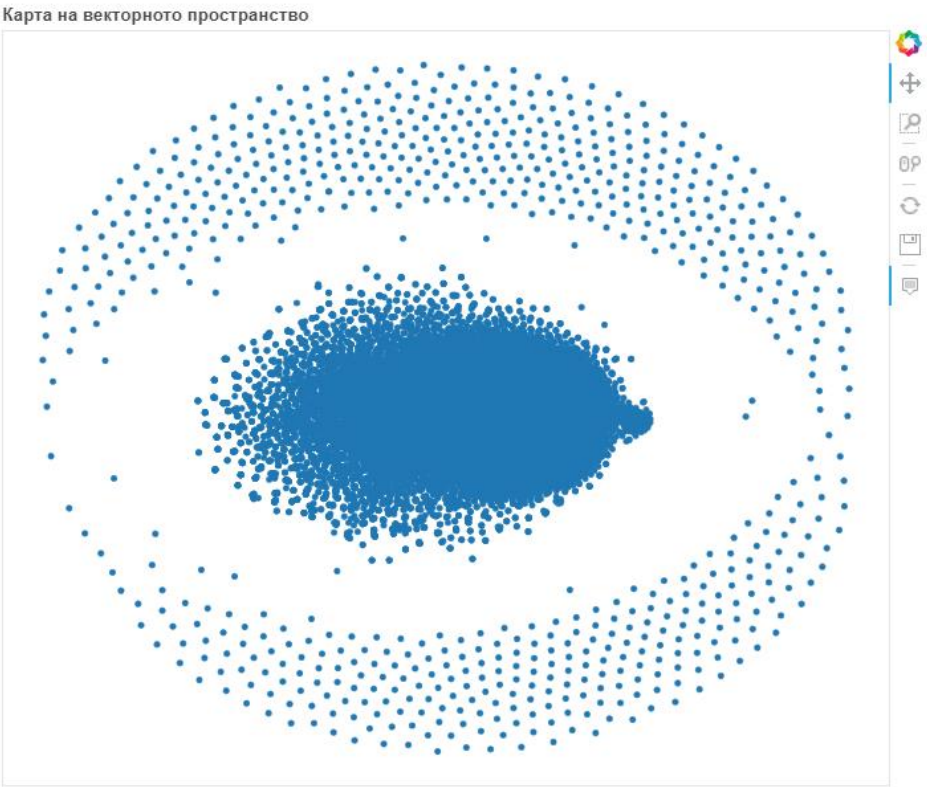
- Процесор Intel Core i5 – 4570 @ 3.2 GHz (4 CPUs);
- RAM памет 16 GB;
- Графична карта GeForce GTX 660 4GB;
- Операционна система Windows 7 Professional x64.

Извлечените потребителски съобщения са достъпни в оригиналния си вид в JSON файлове, разделени по дни и са записани в база от данни. Налични са също и файлове с публикациите след тяхната първична обработка, като оригиналния им брой е 182146, а след пречистване остават

126168. Сред тях освен на български език има съобщения на сръбски, македонски и руски, но в относително малки количества.

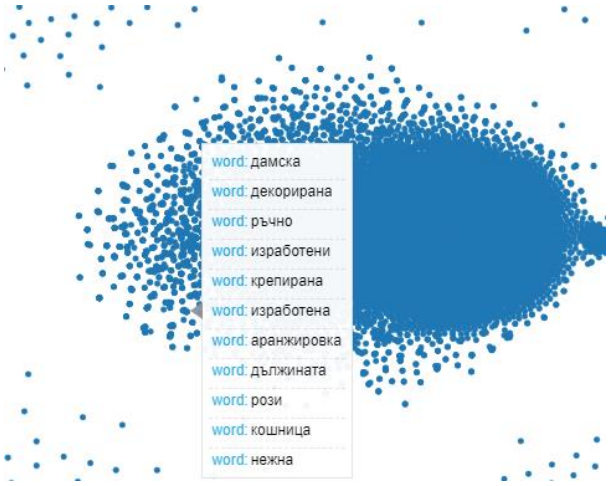
Изграден е модел на данните, като при обучението се пропускат термини с по-малко от 10 срещания в цялото множество. За определяне на контекста на думата е избран диапазон от 10 съседни класификатора от ляво и от дясно на нея. При изпълнение на програмния код се използват 4 процесорни ядра. Времето за обучение възлиза на 45 минути при посочената компютърна конфигурация и крайният файл е с големина 60MB. Той съдържа координатите на термините в многомерното векторно пространство и може да се използва за намиране на семантично сходство с други примерни съобщения.

Етап от апробацията е визуализирането на данните. За тази цел е използван моделът на данните и са отчетени струпвания от тематично близки термини (вж. Фиг. 10).

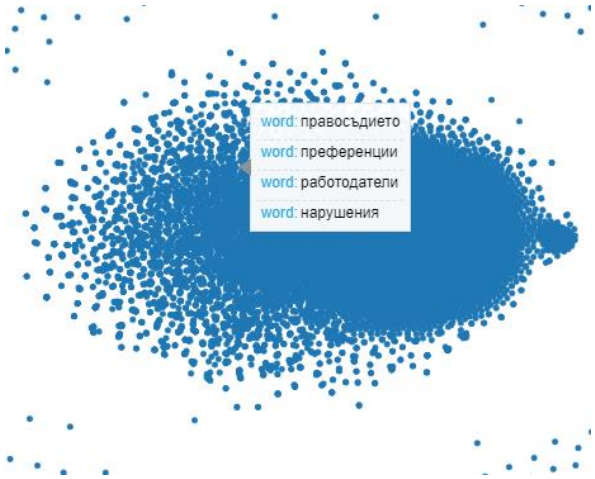


Фиг. 10. Визуализация на модела на данните

Централната структура съдържа разнообразни тематични области. Така например в левия ъгъл се намират съобщения на тема подаръци за жените (вж. Фиг. 11), докато в горната част присъстват термини от правно и административно естество (вж. Фиг. 12).



Фиг. 11
Струпвания от термини



Фиг. 12
Струпвания от термини

На база на извлечените данни са определени 10 клъстерни групи и е включено най-доброто разпределение на дискусионните области:

- Cluster 0: отравяне дъщеря Сергей Великобритания Юлия
- Cluster 1: община Плевен деца болен отново
- Cluster 2: мляко козунаци яйца продукти алкохол
- Cluster 3: военен самолет жертви Алжир падна
- Cluster 4: дългата линия света маршрут борят
- Cluster 5: фенове днес продават хип билети
- Cluster 6: Play Приложение Кръстословици Google Бързи
- Cluster 7: Обявиха фестивала Мездра любовта болки
- Cluster 8: край няма Благоевград своя катастрофа
- Cluster 9: алкохолизма лечение сочи цена области

Параграфът завършва са насоки за развитие и подобрение на функционалността при визуализацията на данните.

В заключение разработването на софтуерна система за автоматизирана обработка на неструктурирани текстови данни на български език, извлечени от социалните мрежи е съобразено със съвременните технологии в областта на машинното обучение и с нуждите на Медийна група Черно море. Ползите за организацията се свързват с възможността за анализ на потребителски съобщения и откриване на зараждащи се тематични дискусии в платформата Twitter. Предложеният модел на софтуерната система може да бъде адаптиран и апробиран към други социални мрежи.

IV. Справка за приносите на дисертационния труд

В дисертационния труд е проведено изследване на същността на неструктурираните данни, извлечени от социалните мрежи и методите, подходите и моделите за тяхната обработка. С оглед на обхвата и представените резултати в дисертацията се счита, че са постигнати следните научни и приложни приноси:

1. Изследвани са източниците, методите, моделите и технологиите за обработка на неструктурираните данни в дигитална среда и е доказана необходимостта от автоматизирана обработка на неструктурирани текстови данни на български език.
2. Анализирани са неструктурираните данни в социалните мрежи и е въведена класификация на различните градивни единици на потребителските съобщения за четири от най-популярните платформи за социално взаимодействие в Интернет.
3. Създаден е модел на софтуерна система за автоматизирана обработка на неструктурирани данни, който е съобразен със спецификата и характера на българския език, използван в публикациите в социалните мрежи.
4. Разработен е алгоритъм за определяне на активни клъстери, при клъстеризация на постоянна емисия от данни, чрез въвеждане на променлива, която отчита времето за възстановяване на неактивни клъстери на база на тяхната значимост.
5. Разработена е концепция за реализиране и частично внедряване на предлаганата софтуерна система в организация от медийната сфера.

V. Публикации по дисертационния труд

Научни статии

1. Банков, Б. Извличане на топ тенденции от дискусиите на български език в Twitter. Известия на Съюза на учените – Варна. Серия „Икономически науки“, Съюз на учените – Варна, 2017, 2, с. 254-259.
2. Bankov, B. An Approach for Clustering Social Media Text Messages, Retrieved from Continuous Data Streams. Science. Business. Society: International Scientific Journal, Sofia: Scientific Technical Union of Mechanical Engineering INDUSTRY 4.0 et. al., 3, 2018, 1, pp. 6-9.

Научни доклади

1. Банков, Б. Възможности за извличане на данни в реално време от платформата Twitter. Предизвикателства пред информационните технологии в контекста на „Хоризонт 2020“, Академично издателство „Ценов“, 1, 2016, с. 313-318.
2. Банков, Б., Куюмджиев, И. Приложение на алгоритми за изчисление на сходство и вариация на текстови низове. Икономиката в променящия се свят - национални, регионални и глобални измерения (ИПС-2017), Варна: Наука и икономика, 1, 2017, с. 562-565.