DOI 10.34185/1562-9945-5-124-2019-18
УДК 681.3.06(075.8)

К.Ю. Островська, Н.О. Кислова, Г.Ю. Станчиць, І.В.Стовпченко

ІНТЕЛЕКТУАЛЬНА СИСТЕМА АНАЛІЗУ ЯКОСТІ ТЕКСТУ З ВИКОРИСТАННЯМ МАШИННОГО НАВЧАННЯ

Анотація. Робота присвячена проектуванню інтелектуальної системи аналізу якості тексту з використанням машинного навчання, а саме розробці програмного продукту, що дозволяє оцінювати якість текстів по ряду критеріїв.
Метою роботи є створення програмного продукту, який дозволив би проводити якісний аналіз текстів відповідно до низки критеріїв.
Представлені результати тестування і дослідження даного програмного продукту, в результаті чого були визначені найбільш ефективні моделі з отриманих в процесі розробки програми.
Ключові слова: аналіз текстів на якість, тональність, рекламні, лінгвістика, машинного навчання, класифікація текстів.

Пошукова оптимізація - це процес підготовки і організації контенту на Інтернет-сторінках або сайтах для збільшення його потенційної релевантності за певними ключовими словами в певній пошукової системі.

Пошуковими системами не проводиться аналіз текстів на якість; періодично вводяться нові фільтри, але пріоритетними критеріями аналізу є унікальність контенту, його грамотність, якість подачі інформації, її достовірність, що SEO компанії, неякісно виконують свою роботу, нерідко використовують в своїх інтересах. Такий аналіз не дозволяє повністю виключити неякісні й марні матеріали. Відповідно, сайти, містять подібні матеріали, пропускаються пошуковими фільтрами.

Новизна роботи полягає в здатності розроблених алгоритмів служити фільтром для подібних текстів.

Метою роботи було створення програмного продукту, який дозволив би проводити якісний аналіз текстів відповідно до низки критеріїв.

Для досягнення поставленої мети, необхідно було визначити критерії оцінки, які здатні найбільш повно відобразити якість текстів.

В рамках роботи поставлена задача визначення цих критеріїв. Необхідно було розробити систему перетворення тексту в певний набір чисел, придатний для машинного навчання.

Актуальність роботи полягає в можливості застосування даної системи для зменшення кількості рекламних текстів в мережі Інтернет: алгоритми, розроблені в процесі створення даного програмного продукту, можуть бути застосовані в фільтрах пошукових систем, що дозволить блокувати рекламні і неякісні Інтернет-сторінки в результатах пошуку в випадках, якщо перед користувачем стоїть завдання знаходження певного рекламного контента.

Постановка задачі класифікації текстів. Класифікація - один з розділів машинного навчання. Класифікувати об'єкт - значить, вказати номер (або найменування класу), до якого належить даний об'єкт.

Дана задача є задачею класифікації текстів, для вирішення якої доцільно застосувати машинне навчання в оцінкою текстів за показниками рекламності і тональності, так як для їх обчислення не існує універсальних формул.

У машинному навчанні завдання класифікації відноситься до розділу навчання з учителем.

Ознакою називається результат вимірювання деякої характеристики об'єкта.

Ознака - це відображення $f : X \to D_f$, де D_f - множина допустимих значень ознаки. Залежно від природи цієї множини ознаки діляться на наступні типи:

- бінарна ознака: $D_f = \{0;1\}$;
- номінальна ознака: D_f - кінцева множина;
- порядкова ознака: D_f - кінцева впорядкована множина;
- кількісна ознака: $D_f = R$.

У разі оцінки тексту за критеріями рекламність і тональність буде застосовуватися признаковий номінальний опис.

Постановка задачі в рамках машинного навчання.

Нехай X - множина описів об'єктів, Y - кінцева множина номерів (імен, міток) класів. Існує невідома цільова залежність - відображення

$$y^* : X \to Y, \quad (1)$$

значення якої відомі тільки на об'єктах кінцевої навчальної вибірки

$$X^m = \{(x_1, y_1), \ldots (x_m, y_m)\}. \quad (2)$$

Потрібно відновити залежність

$$\alpha : X \to Y, \quad (3)$$

здатну класифікувати довільний об'єкт.

В рамках даного завдання значеннями будуть значення показників рекламності і тональності текстів. Системи машинного навчання не можуть працювати безпосередньо зі словами, пропозиціями чи текстами: текст необхідно уявити в деякому числовому вигляді.

Критерії для оцінки текстів:

1. Рекламність і тональність тексту.
2. Орфографічна і синтаксична коректність тексту.
3. Інформативність тексту.
4. Водність тексту.

За визначеними критеріями будемо оцінюватися тексти.

Водність:

$$V = \frac{W_v}{W_t}$$

де V - водність тексту, W_v - число стоп-слів, W_t - загальна кількість слів у тексті.

Цей критерій може набувати значень від «0» до «1». Значення «0» означає, що в тексті відсутні стоп-слова, «1» - що текст повністю складається з стоп-слів, але на практиці, не буває текстів, повністю складаються з стоп слів.

Орфографічна коректність:

$$orf\_err = \frac{N_{or}}{W_t}$$

де orf_err - значення критерію орфографічна коректність, N_{or} – число орфографічних помилок (неправильних слів), W_t - загальна кількість слів у тексті.

Цей критерій може приймати значення від «0» до «1», де «0» - значення, при якому всі слова в тексті правильно написані, при «1» все слова некоректні.

Синтаксична коректність:

$$syn\_err = \frac{N_{sr}}{W_t}$$

де syn_err - значення критерію синтаксична коректність, N_{sr} – число синтаксичних помилок, W_t - загальна кількість слів у тексті аналогічно критерієм орфографічна правильність

Інформативність:

$$inf = \frac{W_{var}}{W_t}$$

де inf - значення критерію інформативність, W_{var} - число розрізняються слів, W_t - загальне число слів в тексті. Можливі значення в діапазоні від «0» до «1». «0» відповідає тексту, який складається з одного повторюваного слова, «1» - тексту, в якому всі слова різні.

При значеннях оцінки, менших або рівних 0.3, показник інформативність прирівняється до 0, при значеннях, більших 0.8 - до 1. В інтервалі від 0.3 до 0.8 значення без змін. Дані порогові значення застосовуються в більшості інтернет-сервісів з оцінки якості текстів.

Показник рекламність (таблиця 1), тональність (таблиця 2):

Таблиця 1

Опис показника рекламність

Значення виходу	Опис
0	не рекламний текст (до 5% реклами)
1	повністю рекламний текст, (більше 80% реклами)

Таблиця 2

Опис критерію тональність

Значення виходу	Опис
0	Негативна
0.5	Нейтральна
1	Позитивна

Сама по собі тональність не грає практично ніякої ролі в оцінці якості тексту, але вона підсилює рекламний ефект тексту.

Алгоритми оцінок за критеріями:

1. Передобробка тексту.

Вхідними даними для алгоритму є текстовий файл, в який користувач повинен попередньо скопіювати статтю, яку необхідно проаналізувати. Відбувається розбір тексту на слова, а також приведення їх до початкової формі за допомогою бібліотеки Pymorphy2.

Результатом роботи даного блоку є частотний словник тексту, а також список всіх слів.

2. Синтаксична перевірка.

Відбувається аналіз необробленого тексту на наявність синтаксичних помилок за допомогою бібліотеки LanguageTool.

Результат роботи блоку - оцінка синтаксичної коректності тексту.

3. Орфографічна перевірка.

На вході даного блоку - список всіх слів, складових текст. Відбувається аналіз тексту на наявність орфографічних помилок за допомогою бібліотеки PyEnchant.

Результат роботи блоку - оцінка орфографічною коректності тексту.

4. Обчислення показника водності.

На вході даного блоку - список слів, отриманий в блоці, що виконує предобработку тексту. Відбувається аналіз тексту за критерієм водність з допомогою бібліотеки pymorphy2.

Результат роботи блоку - оцінка водності тексту.

5. Оцінка інформативності.

На вході даного блоку - частотний словник. Відбувається аналіз тексту на інформативність.

Результат роботи блоку – оцінка інформативності тексту.

6. Навчання моделей для обчислення показників тональності і рекламні.

Алгоритм навчання моделей Word2Vec і Random forest винесено в окремий блок, так як в основній частині програми не відбувається нав-

чання моделей - вони доступні відразу після завантаження файлів, в які були збережені.

Модель Word2Vec приймає на вхід список слів, приведених до початковій формі (леми), на якому і навчається. В отриманій моделі кожному слову відповідає вектор певної довжини, яка задається при навчанні, що відображає зв'язки, виражені косинусной близькістю, з іншими словами.

Алгоритм навчання Random forest приймає на вхід список текстів, представлений в Word2Vec форматі: кожен текст ділиться на список слів, кожне слово замінюється відповідним вектором з Word2Vec уявлення, потім вектора, отримані в рамках одного тексту, підсумовуються і результат ділиться на кількість слів.

Таким чином, навчальною базою для класифікатора є список векторів.

7. Обчислення показника рекламних.

На вході даного блоку - текст, який перетворюється до певного формату, після чого оцінюється класифікатором.

8. Оцінка тональності.

Алгоритм оцінки тональності тексту повторює алгоритм обчислення показника рекламних, але в ній використовуються інші моделі Word2Vec і Random forest.

9. Обчислення загальної оцінки.

Підсумовуються оцінки за всіма критеріями. На етапі розробки алгоритму прийнято:

$$\text{sum} = (\omega_1 \times \text{V} + \omega_2 \times \text{orf\_err} + \omega_3 \times \text{syn\_err} + \omega_4 \times \text{inf} + \\ + \omega_5 \times \text{abs}(\text{ton} - 0.5) \times \text{rekl}) / \sum_{i=1}^{5} \omega_i \quad (4)$$

де orf_err - результат оцінки за критерієм орфографічна коректність, syn_err - синтаксична коректність, inf - інформативність, ton - тональність, rekl - рекламність; ω_i - вагові коефіцієнти.

Дані коефіцієнти відображають значущість того чи іншого критерію в загальній оцінці.

Таблиця 3

Значимість критеріїв в загальній формулі (4)

Рекламність + тональність	↓
Інформативність	
Водність, орфографічна коректність, граматична коректність	

Результати по різним Word2Vec моделям, критерій тональність. Корпус текстів веб-сервісу Wikipedia.

Дана модель складається з 392 339 лем, що становлять 600 мільйонів слів.

Навчання класифікатора, заснованого на Word2Vec моделі, на текстах Вікіпедії, а також відгуків з сайту IMDB (кожен класифікатор навчався по три рази).

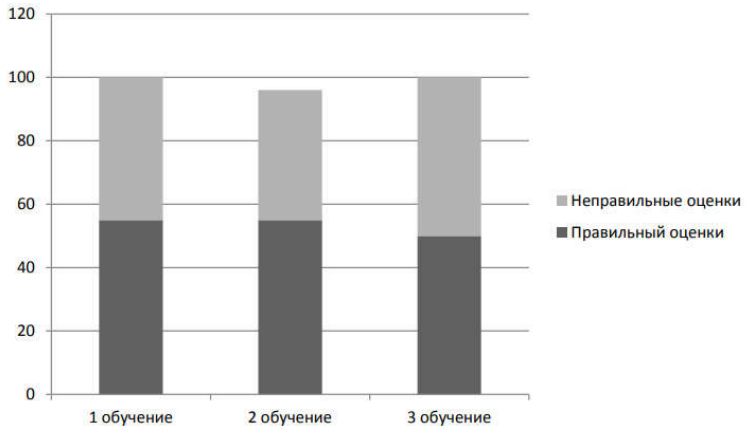

Рисунок 1 - Тестування класифікатора (Word2Vec модель - Вікіпедія, класифікатор - Вікіпедія, IMDB)

В результаті першого навчання була отримана модель, яка оцінює тексти з точністю 55% за тестовою вибіркою, в результаті другого - 55%, третього - 50%. Це показує, що в результаті навчання класифікатора типу Random Forest кожен раз виходить абсолютно нова модель.

Навчання класифікатора, заснованого на Word2Vec моделі, на текстах з Інтернет - Енциклопедія Сучасної України, а також відгуків з сайту IMDB.

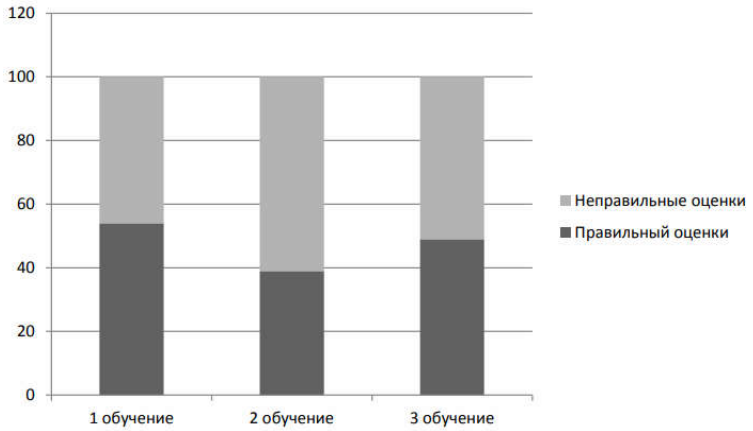

Рисунок 2 - Тестування класифікатора (Word2Vec модель - Вікіпедія, Інтернет - Енциклопедія Сучасної України, IMDB)

В результаті першого навчання була отримана модель, яка оцінює тексти з точністю 54% за тестовою вибіркою, в результаті другого - 40%, третього - 49%.

Корпус з відгуків IMDB. Дана Word2Vec модель навчена на 100 тисячах відгуків фільмів з сайту IMDB, складається з 98 745 лем, містить близько 100 мільйонів слів.

Навчання класифікатора, заснованого на Word2Vec моделі, на текстах Вікіпедії, а також відгуків з сайту IMDB (кожен класифікатор навчався по три рази).

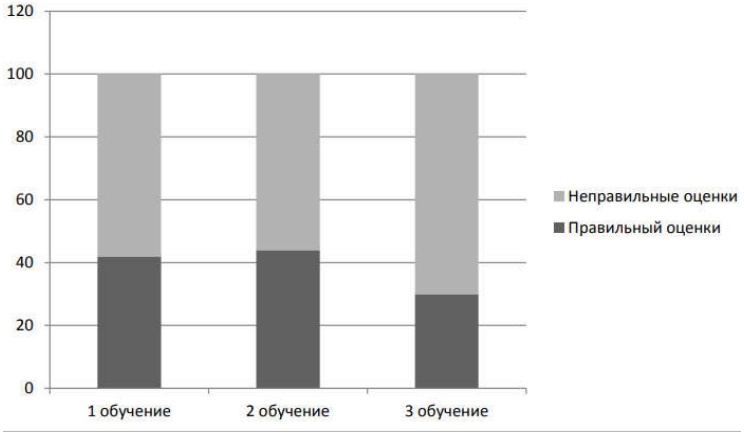

Рисунок 3 - Тестування класифікатора (Word2Vec модель - IMDB, класифікатор - Вікіпедія, IMDB)

В результаті першого навчання була отримана модель, яка оцінює тексти з точністю 42% за тестовою вибіркою, в результаті другого - 44%, третього - 30%.

Навчання класифікатора, заснованого на даній Word2Vec моделі, на текстах з Інтернет - Енциклопедія Сучасної України, а також відгуків з сайту IMDB.

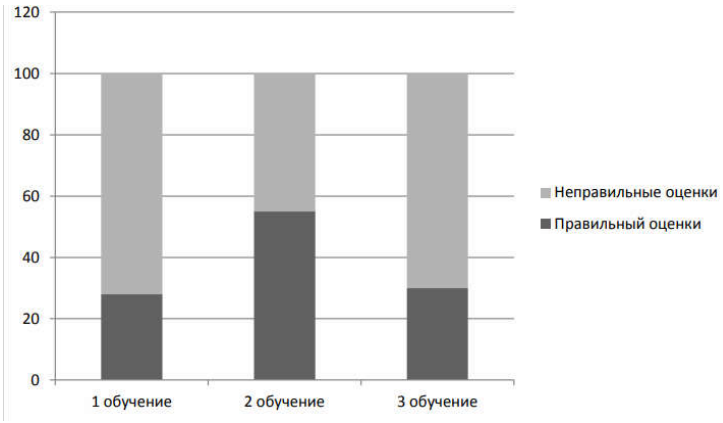

Рисунок 4 - Тестування класифікатора (Word2Vec модель - IMDB, Інтернет - Енциклопедія Сучасної України, IMDB)

В результаті першого навчання була отримана модель, яка оцінює тексти з точністю 28% за тестовою вибіркою, в результаті другого - 55%, третього -30%.

Корпус текстів української мови. Дана Word2Vec модель навчена на 100 тисячах відгуків фільмів з сайту IMDB, складається з 184 973 лем, містить близько 250 мільйонів слів.

Навчання класифікатора, заснованого на даній Word2Vec моделі, на текстах Вікіпедії, а також відгуків з сайту IMDB (кожен класифікатор навчався по три рази).

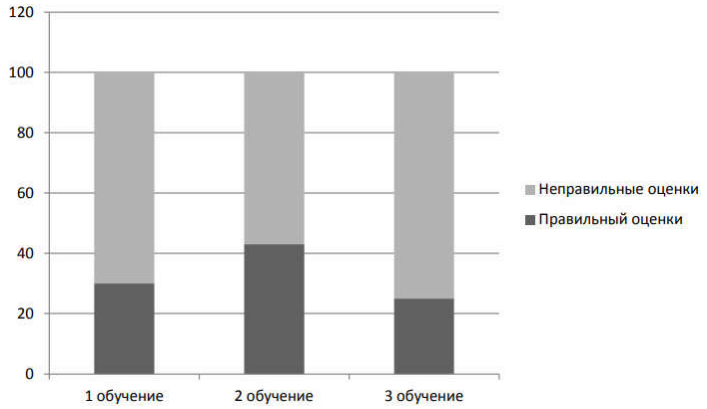

Рисунок 5 - Тестування класифікатора (Word2Vec модель - КТУМ, класифікатор - Вікіпедія, IMDB)

В результаті першого навчання була отримана модель, яка оцінює тексти з точністю 30% за тестовою вибіркою, в результаті другого - 43%, третього - 25%.

Навчання класифікатора, заснованого на даній Word2Vec моделі, на текстах з Інтернет - Енциклопедія Сучасної України, а також відгуків з сайту IMDB.

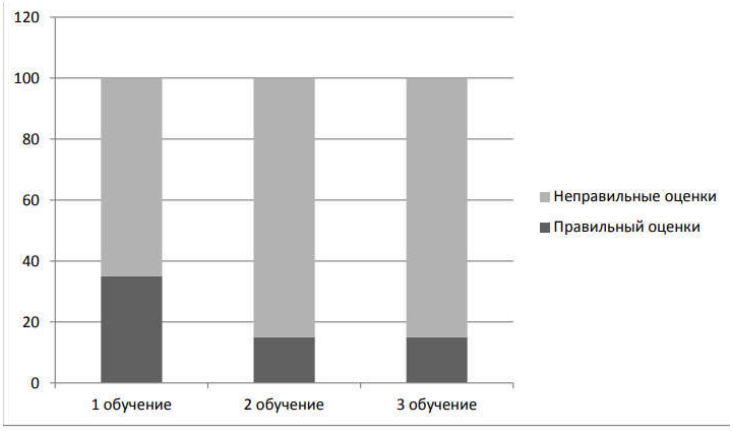

Рисунок 6 - Тестування класифікатора (Word2Vec модель - КТУМ, Інтернет - Енциклопедія Сучасної України, IMDB)

В результаті першого навчання була отримана модель, яка оцінює тексти з точністю 35% за тестовою вибіркою, в результаті другого - 15% третього - 15%.

Порівняння результатів. Класифікатор був 18 раз навчений на різних Word2Vec моделях.

Таблиця 4

Порівняння результатів навчання класифікатора

Модель Word2Vec	Навчальна вибірка для класифікатора	Точність отриманої моделі, %		
		1 навчання	2 навчання	3 навчання
Wikipedia	IMDB + Wikipedia	55	55	50
	IMDB + Інтернет - Енциклопедія Сучасної України	54	40	49
IMDB	IMDB + Wikipedia	42	44	30
	IMDB + Інтернет - Енциклопедія Сучасної України	28	55	30
Корпус текстів української мови (КТУМ)	IMDB + Wikipedia	30	43	25
	IMDB + Інтернет - Енциклопедія Сучасної України	35	15	15

 ISSN 1562-9945

З таблиці видно, що найбільшу точність має класифікатор, навчений на Word2Vec моделі, навченої на текстах з інтернет-енкціклопедії Вікіпедія, при цьому навчальною вибіркою для класифікатора також були тексти з Вікіпедії, і відгуки до фільмів з сайту IMDB.

Максимальна точність вийшла при другому навчанні (55%).

Результати по різним Word2Vec моделям, критерій рекламність. Для класифікатора, що виробляє оцінку тексту за показником рекламність, була складена навчальна вибірка з текстів, що включають в себе статті з енциклопедій (Wikipedia, Інтернет - Енциклопедія Сучасної України, Вікіпедія), а також великого набору рекламних оголошень з ряду інтернет-ресурсів відповідної тематики (OLX.ua, Prom.ua).

Корпус текстів веб-сервісу Wikipedia. Дана модель складається з 392 339 лем, що становлять 600 мільйонів слів.

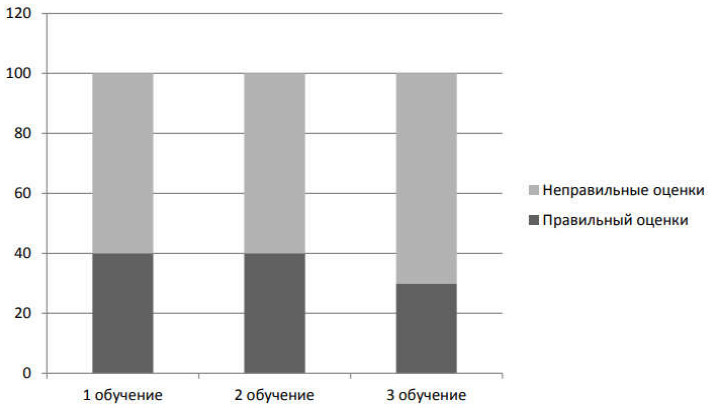

Рисунок 7 - Тестування класифікатора (Word2Vec модель - Вікіпедія)

В результаті першого навчання була отримана модель, яка оцінює тексти з точністю 40% за тестовою вибіркою, в результаті другого - 40%, третього - 30%. Це показує, що в результаті навчання класифікатора типу Random Forest кожен раз виходить абсолютно нова модель.

Корпус з відгуків IMDB. Word2Vec модель навчена на 100 тисячах відгуків фільмів з сайту IMDB, складається з 98 745 лем, містить близько 100 мільйонів слів.

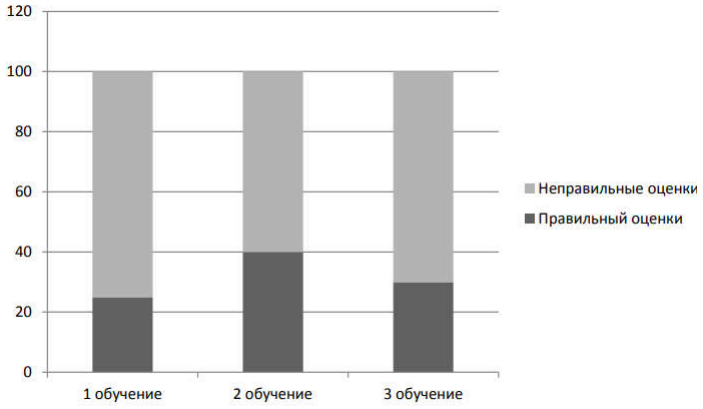

Рисунок 8 - Тестування класифікатора (Word2Vec модель - IMDB)

В результаті першого навчання була отримана модель, яка оцінює тексти з точністю 25% за тестовою вибіркою, в результаті другого - 40%, третього - 30%.

Корпус текстів української мови. Дана Word2Vec модель навчена на 100 тисячах відгуків фільмів з сайту IMDB, складається з 184 973 лем, містить близько 250 мільйонів слів.

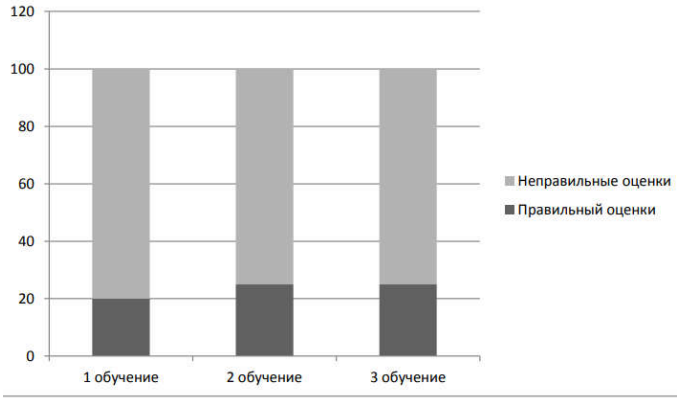

Рисунок 9 - Тестування класифікатора (Word2Vec модель - КТУМ)

В результаті першого навчання була отримана модель, яка оцінює тексти з точністю 30% за тестовою вибіркою, в результаті другого - 43%, третього -25%.

Таблиця 5

Порівняння результатів навчання класифікатора

Модель Word2Vec	Точність отриманої моделі, %		
	1 навчання	2 навчання	3 навчання
Wikipedia	40	40	30
IMDB	25	40	30
Корпус текстів української мови (КТУМ)	20	25	25

З таблиці видно, що найбільшу точність має класифікатор, навчений на Word2Vec моделі, навченої на текстах з інтернет-енкціклопедії Вікіпедія (40%).

Висновки. В результаті було проведено аналіз існуючих систем аналізу тексту, описані переваги та недоліки кожної з таких систем, виділені ключові алгоритми, взяті за основу алгоритмів оцінки за деякими з критеріями. Також, були сформульовані критерії для оцінки текстів в рамках даного проекту та обґрунтовано їх вибір, були розроблені алгоритми обчислення оцінок за цими критеріями. В ході розробки були вивчені і застосовані алгоритми машинного навчання і багатокритеріальних задач прийняття рішень.

Програмний продукт, що розроблено, був протестований на різних навчальних вибірках класифікатора і моделі представлення текстів, був проведено аналіз отриманих результатів і обрані найбільш точні моделі. З огляду на, що можливості для навчання моделей машинного навчання були обмежені продуктивністю комп'ютера, отримана точність в 60% - досить хороший результат, який показує, що при навчанні на більшій вибірці, дана система буде здатна оцінювати якість текстів з високою точністю.

ЛІТЕРАТУРА / ЛІТЕРАТУРА

1. Журавлев Ю.И., Рязанов В.В., Сенько О.В. «Распознавание». Математические методы. Программная система. Практические применения. -М.: Фазис, 2006.
2. Мальковский М.Г., Большакова Е.И. Интеллектуальная система контроля качества научно-технического текста // Интеллектуальные системы - 1997 С.149-155.
3. Пазельская А., Соловьев А. Метод определения эмоций в текстах. // Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011» - М.: Изд-во РГГУ, 2011.– С.510-523.
4. Посилання в мережі Інтернет:
URL: http://datareview.info/article/sovremennyie-metodyianaliza-tonalnosti-teksta/
5. Посилання в мережі Інтернет:
URL: https://code.google.com/archive/p/word2vec/
6. Посилання в мережі Інтернет:
URL: http://www.script-coding.com/Browse.html

7. Посилання в мережі Інтернет: Документация PyMorphy2.
URL: https://pymorphy2.readthedocs.io/en/0.2/user/index.html
8. Посилання в мережі Інтернет: Документация gensim Word2Vec.
URL: https://radimrehurek.com/gensim/models/word2vec.html
9. Посилання в мережі Інтернет:
URL: http://scikit-learn.org/stable/testimonials/testimonials.html

REFERENCES

1. Zhuravlev Yu.I., Ryazanov V.V., Senko O.V. «Raspoznavanie». Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya. -M.: Fazis, 2006.
2. Malkovskij M.G., Bolshakova E.I. Intellektualnaya sistema kont-rolya kachestva nauchno-tekhnicheskogo teksta // Intellektualnye sistemy - 1997 C.149-155.
3. Pazelskaya A., Solovev A. Metod opredeleniya emoczij v tekstakh. // Kompyuternaya lingvistika i intellektualnye tekhnologii: «Dialog-2011» - M.: Izd-vo RGGU, 2011.– S.510-523.
4. Posilannya v merezhi Internet: URL: http://datareview.info/article/sovremennyie-metodyianaliza-tonalnosti-teksta/
5. Posilannya v merezhi Internet:
URL: https://code.google.com/archive/p/word2vec/
6. Posilannya v merezhi Internet:
URL: http://www.script-coding.com/Browse.html
7. Posilannya v merezhi Internet: Dokumentacziya PyMorphy2.
URL: https://pymorphy2.readthedocs.io/en/0.2/user/index.html
8. Posilannya v merezhi Internet: Dokumentacziya gensim Word2Vec.
URL: https://radimrehurek.com/gensim/models/word2vec.html
9. Posilannya v merezhi Internet:
URL: http://scikit-learn.org/stable/testimonials/testimonials.html

Received 22.11.2019.
Accepted 25.11.2019.

Інтелектуальна система аналізу якості тексту з використанням машинного навчання

Робота присвячена проектуванню інтелектуальної системи аналізу якості тексту з використанням машинного навчання, а саме розробці програмного продукту, що дозволяє оцінювати якість текстів по ряду критеріїв.

Метою роботи є створення програмного продукту, який дозволив би проводити якісний аналіз текстів відповідно до низки критеріїв.

Представлені результати тестування і дослідження даного програмного продукту, в результаті чого були визначені найбільш ефективні моделі з отриманих в процесі розробки програми.

Intelligent text quality analysis systemusing machine learning

The work is devoted to the design of an intelligent text quality analysis system using machine learning, namely the development of a software product that allows one to evaluate the quality of texts by a number of criteria.

The aim of the work is to create a software product that would allow a qualitative analysis of texts in accordance with a number of criteria.

The results of testing and research of this software product are presented, as a result of which the most effective models from those obtained during the development of the program were determined.

Островська Катерина Юріївна – к.т.н., доцент кафедри інформаційних технологій та систем НМетАУ.

Кислова Надія Олександрівна – магістр кафедри інформаційних технологій та систем НМетАУ.

Станиць Георгій Юрійович – старший викладач кафедри інформаційних технологій та систем НМетАУ.

Стовпченко Іван Володимирович – старший викладач кафедри інформаційних технологій та систем НМетАУ.

Островская Екатерина Юрьевна - к.т.н., доцент кафедры информационных технологий и систем НМетАУ.

Кислова Надежда Александровна - магистр кафедры информационных технологий и систем НМетАУ.

Станчиц Георгий Юрьевич - старший преподаватель кафедры информационных технологий и систем НМетАУ.

Стовпченко Иван Владимирович - старший преподаватель кафедры информационных технологий и систем НМетАУ.

Ostrovskaya Ekaterina - Ph.D., associate professor of the Department of Information Technologies and Systems NMetAU.

Kislova Nadezhda - Master of the Department of Information Technologies and Systems NMetAU.

Stanchits Georgy - Senior Lecturer, Department of Information Technology and Systems NMetAU.

Stovpchenko Ivan - Senior Lecturer, Department of Information Technology and Systems NMetAU.