Тема роботи: Виявлення текстової подібності назв продуктів із використанням штучних нейроних мереж

Студент:
Степаненко Є. Ю.
Група: КА-53
Науковий керівник:
ас. Кухарєв С.О.

Постановка задачі

1. Виконати аналіз методів проектування та реалізації систем виявлення текстової подібності

2. Виконати дослідження ефективності штучних нейронних мереж в задачах обробки тексту для виявлення текстової подібності

3. Застосувати моделі за найркащою ефективністю для виявлення співставлення

    3.1 Використати методи збіжності тексту як аналізу зображень (згорткова мережа)

    3.2 Обчислити оцінки точності моделі та визначити степінь адекватності моделей

4. Виробити рекомендації щодо подальшого вдосконалення розробленої системи.

Предмет та ціль роботи

Предмет: застосування нейронних мереж для перевірки текстової подібності

Об’єкт: deep text matching

Ціль: Розробити систему для оцінки продуктових співставлень за текстовою подібністю

Актуальність

Ведення електронної торгівлі через Інтернет призводить до утворення масиву даних, що може бути використаний для збільшення продажів, удосконалення організації продажу продовольчих та непродовольчих товарів й оптимізації діяльності та бізнес процесів компаній, або окремих магазинів.

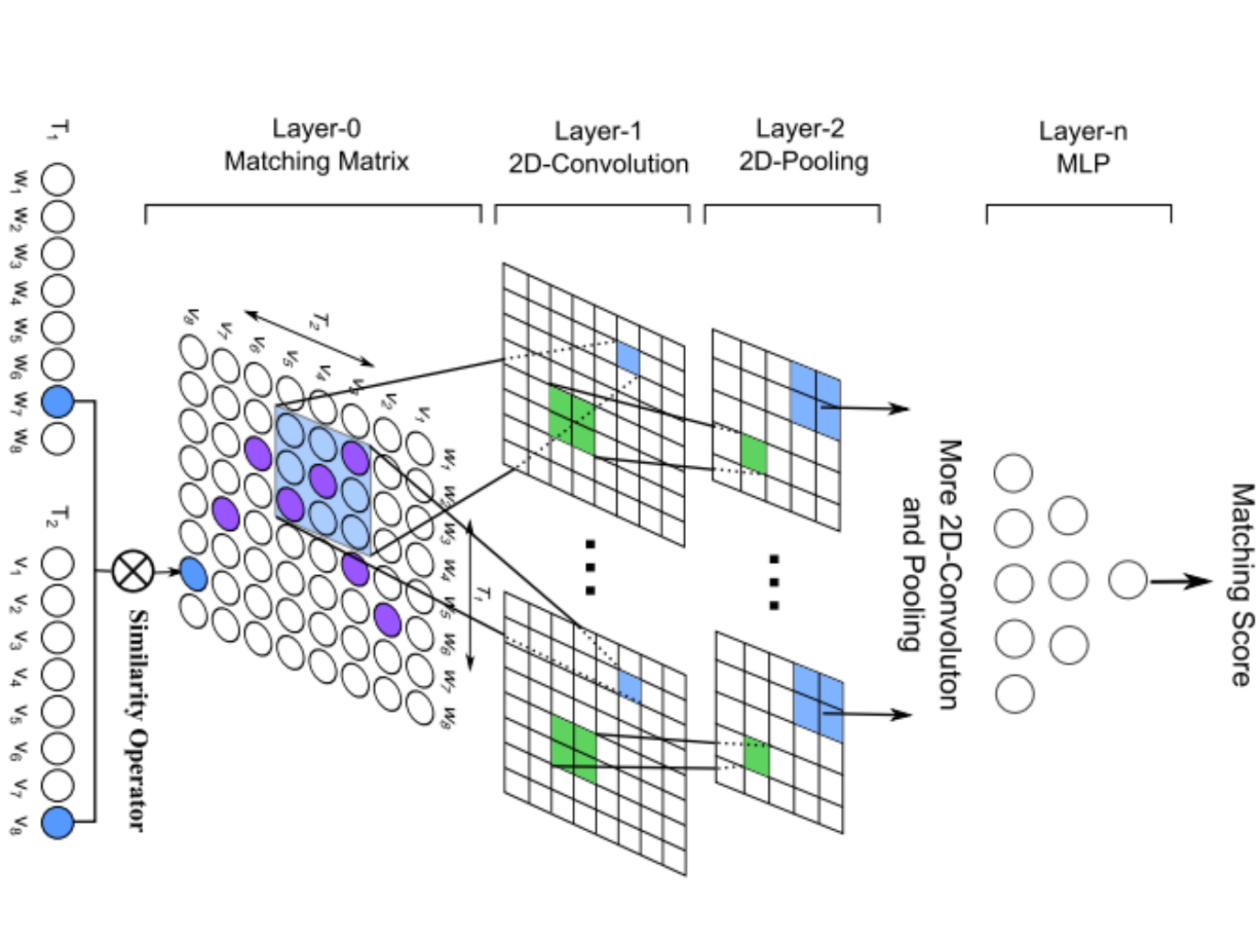
Актуальність полягає в вирішенні проблеми продуктових співставлень що є затратною по ресурсам для SaaS рішень

Дані що використовувались

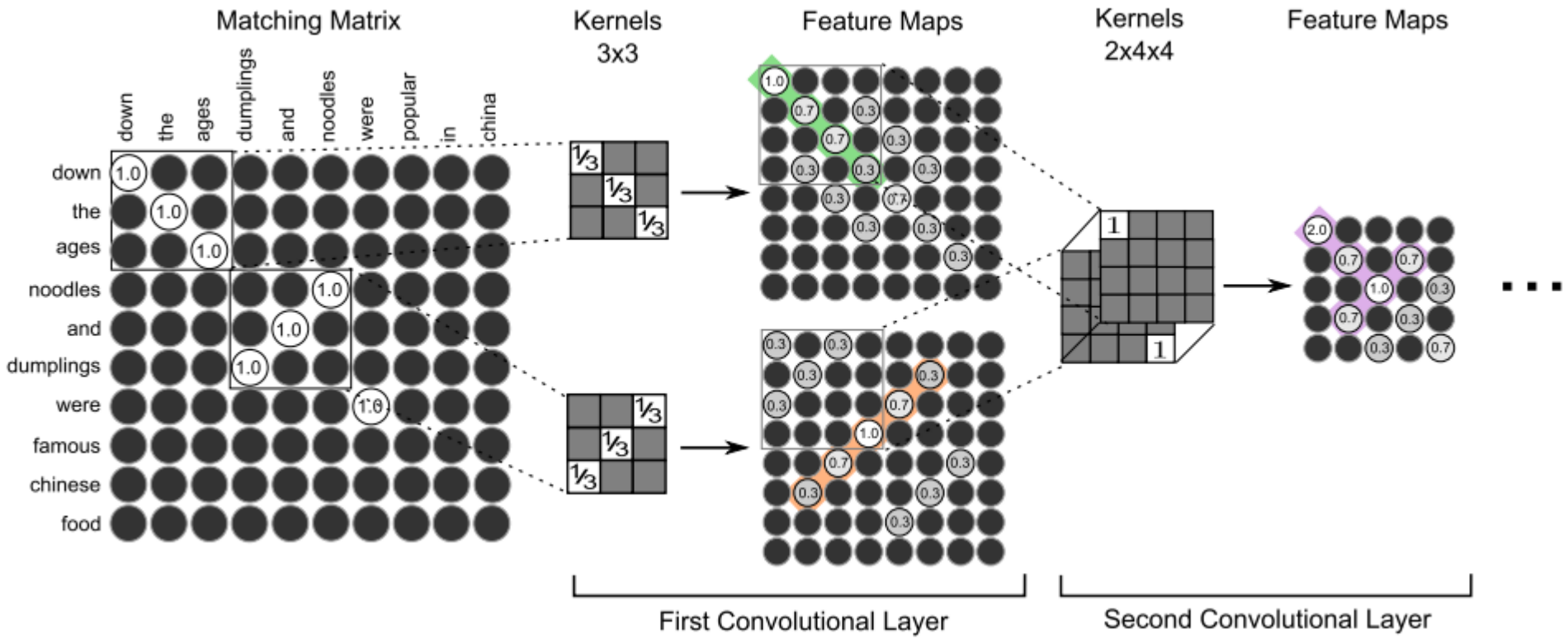
	client_title	competitor_title	check
0	OrganicTai Скраб для тела на основе соли Андам...	Juicy Cleanse Скраб для тела на основе соли с ...	0
2	Сверло по металлу FIT, 6 x 93 мм. 34460	Сверло по металлу FIT 34460	1
3	Игрушки из дерева Игра-шнуровка Яблоко, 462000...	Деревянная игрушка Мир деревянных игрушек Шнур...	1
4	Астигматические контактные линзы Johnson & Joh...	Линзы контактные Johnson & Johnson 1-Day Acuvu...	0
5	Стабилизатор напряжения Powerman "AVS 2000 D",...	Стабилизатор напряжения Powerman AVS 2000D	1

1.5 млн. пар продуктових назв

Структура моделі



Огляд моделі



Реалізація

Мова Python

Фреймворк MatchZoo для deep learning

Plotly та Cufflinks для візуалізації

Numpy та Pandas для обробки даних

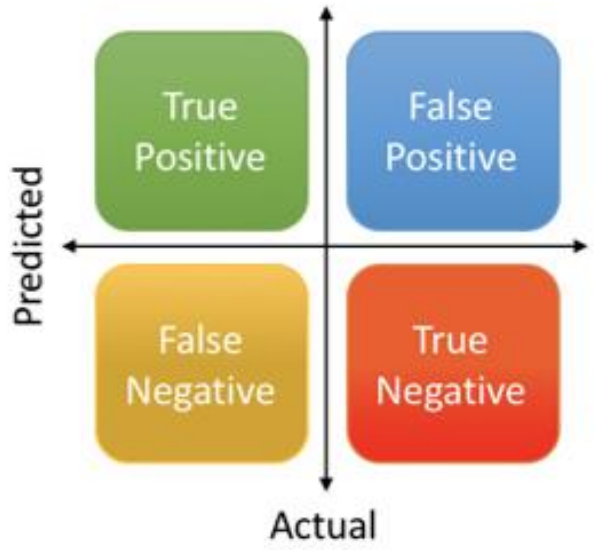
Jupyter Notebook як середа розробки

Критерії адекватності

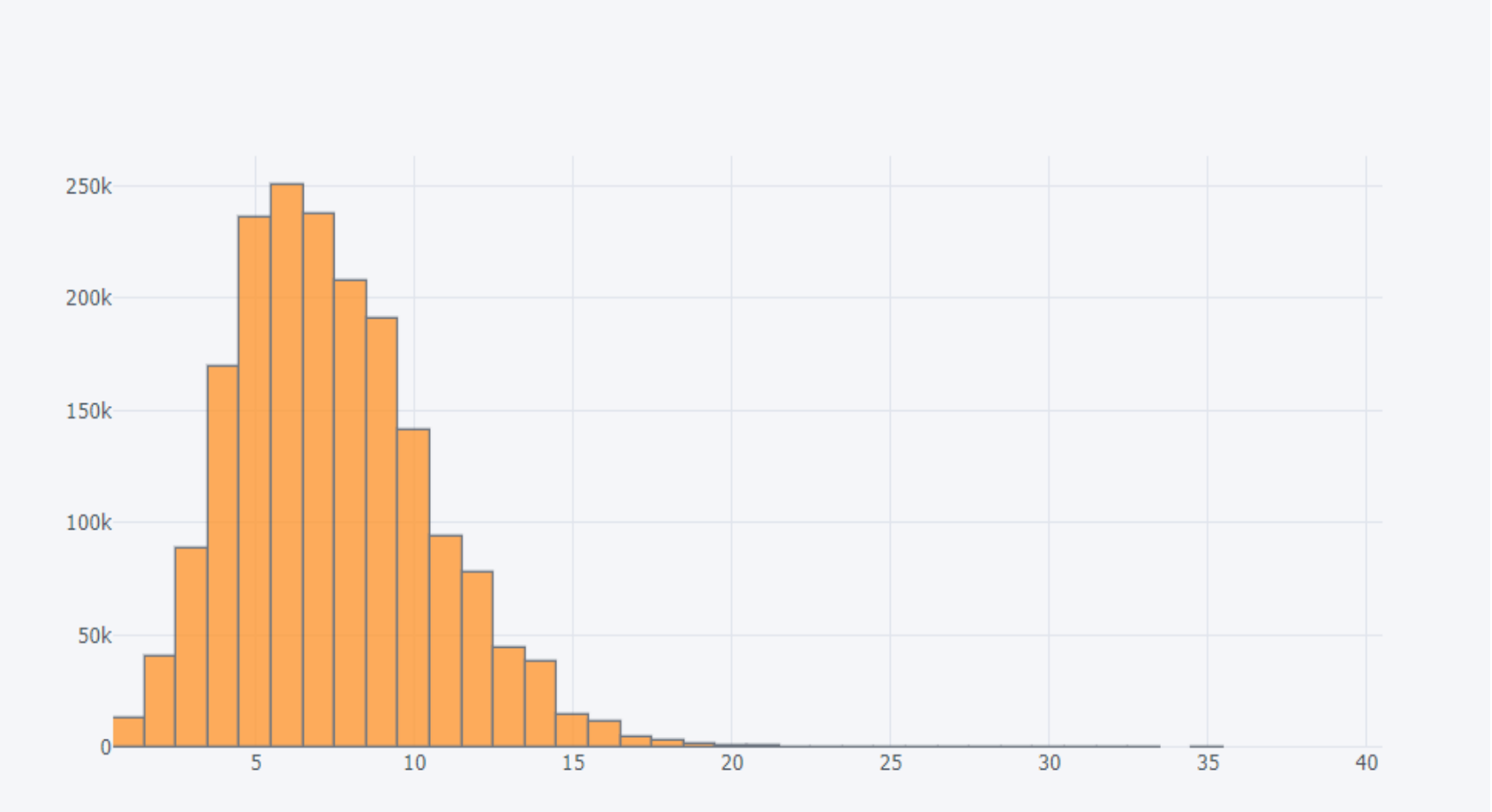
$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



Препроцессинг датасету



Результати роботи (оригінальні назви товарів)

Threshold	True positive	True negative	False positive	False negative	Precision	Recall	Accuracy	F1
0.2	1123	1002599	4685	483713	0.193354	0.002316	0.672682	0.004578
0.3	166	1006845	439	484670	0.27438	0.000342	0.674886	0.000684
0.4	59	1007119	165	484777	0.263393	0.000122	0.674998	0.000243
0.5	20	1007210	74	484816	0.212766	4.13E-05	0.675033	8.25E-05
0.6	15	1007232	52	484821	0.223881	3.09E-05	0.675044	6.19E-05
0.7	8	1007250	34	484828	0.190476	1.65E-05	0.675052	3.30E-05
0.8	6	1007268	16	484830	0.272727	1.24E-05	0.675062	2.47E-05
0.9	3	1007274	10	484833	0.230769	6.19E-06	0.675064	1.24E-05
1	2	1007280	4	484834	0.333333	4.13E-06	0.675068	8.25E-06

Результати роботи (назви з перекладом)

Threshold	True positive	True negative	False positive	False negative	Precision	Recall	Accuracy	F1
0.2	419624	450806	455394	17001	0.47956	0.961063	0.648208	0.639845
0.3	384489	643057	263143	52136	0.593684	0.880593	0.765212	0.709221
0.4	319596	779795	126405	117029	0.716581	0.731969	0.818715	0.724193
0.5	244164	851161	55039	192461	0.816048	0.559208	0.815687	0.663644
0.6	175235	883308	22892	261390	0.884458	0.40134	0.788296	0.552137
0.7	120666	897040	9160	315959	0.929444	0.276361	0.757884	0.426042
0.8	81944	902628	3572	354681	0.95823	0.187676	0.733209	0.313877
0.9	55977	904671	1529	380648	0.973411	0.128204	0.715393	0.226567
1	38862	905467	733	397763	0.981488	0.089005	0.703241	0.16321

Порівняння конфігурацій моделі

Activation	Threshold	Embedding layer dim	Optimization	Precision	Recall	F1 Score
Sigmoid	0.7	100	rmsprop	0.929	0.276	0.426
Sigmoid	0.9	100	adam	0.898	0.115	0.2
Relu	0.7	100	rmsprop	0.917	0.291	0.442
Sigmoid	0.6	300	rmsprop	0.936	0.6	0.73

Результати роботи на тестовому датасеті

Threshold	True positive	True negative	False positive	False negative	Precision	Recall	Accuracy	F1
0.2	108358	748	225984	617	0.324093	0.994338	0.325004	0.488851
0.4	107700	1580	225152	1275	0.323567	0.9883	0.325522	0.487521
0.6	106769	3215	223517	2206	0.323262	0.979757	0.327619	0.48613
0.8	105237	6234	220498	3738	0.323076	0.965699	0.332048	0.484171
1	102258	11661	215071	6717	0.322246	0.938362	0.339341	0.479742

Висновки

- Виконано обробку датасету з 1.5 мільйонів пар продуктових назв та навчання моделі
- Перевірено ефективність модель шляхом кросс-валідації
- Розроблено програмниймодуль для навчання та тестування нейронних мереж, який показав задовільні результати та відповідає потребам production-ready систем.

Подальші дослідження

- Валідація токенізованих вхідних даних для підвищення репрезентабельності результатів роботи моделі.
- Збільшення датасету та його попередня фільтрація
- Розробка виділеного пайплайну препроцесингу даних.

Дякую за увагу!