Ads Text Generation

Inna Wendell

innafomina@ucla.edu

Russian in the World

Russian is native for 154 million people

265 million people speak Russian

Where it was spoken:

Armenia Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Moldova, the Russian Federation, Tajikistan, Turkestan, Ukraine, and Uzbekistan

https://www.rbth.com/lifestyle/328250-how-many-people-speak-russian

Advertisement Market

In 2016, advertisement market in Russia 5.49 billion dollars

Online advertisement the fastest growing venue

Online ad revenue grew by 21%

2.08 billion dollars

https://www.rbc.ru/technology_and_media/21/03/2017/58cfddda9a7947de08281065

Current State Of Affairs

Ads are written by humans

Ads are a genre with many clichés

Potential for automation: no need to write an ad from scratch

Our Goals

- Help customers save time by automating ad writing:
  - Select the best headline from the existing list of headlines
  - First experiments in generating new text from scratch

Dataset

- Obtained from: https://www.kaggle.com/kotobotov/context-advertising
- The database was collected in October 2016 - January 2017
- Contains ads from regions of Russia, Ukraine, Belarus, and Kazakhstan
- Variety of product types and services advertised
- Originally 799999 rows but many duplicates
- After preprocessing: 191257 unique rows

	atitle	atext	adomain
0	Клуб активного отдыха «0.67»	Детский пейнтбол. Спортивный пейнтбол. Тактиче...	0-67.relax.by
1	Антигравитационный чехол IPhone 5	Успейте купить антигравитационный чехол для IP...	0-antigravity.ru
2	Антигравитационный чехол купить!	Антигравитационный чехол для телефона купить з...	0-antigravity.ru
3	Беспроцентный заем от Moneyveo	Без справок! Получите до 3 000 грн. на карту п...	0-credit.moneyveo.ua
4	Беспроцентный заем сотруднику	Акция! Получите Кредит Онлайн под 0%. Без Спра...	0-credit.moneyveo.ua

Stage I

Finding the closest match from existing ad texts / ad headlines:

a) A client types an ads headline: our model finds the best matching ads texts

b) A client types an ads text: our model finds the best matching ads headlines

Unsupervised Approach

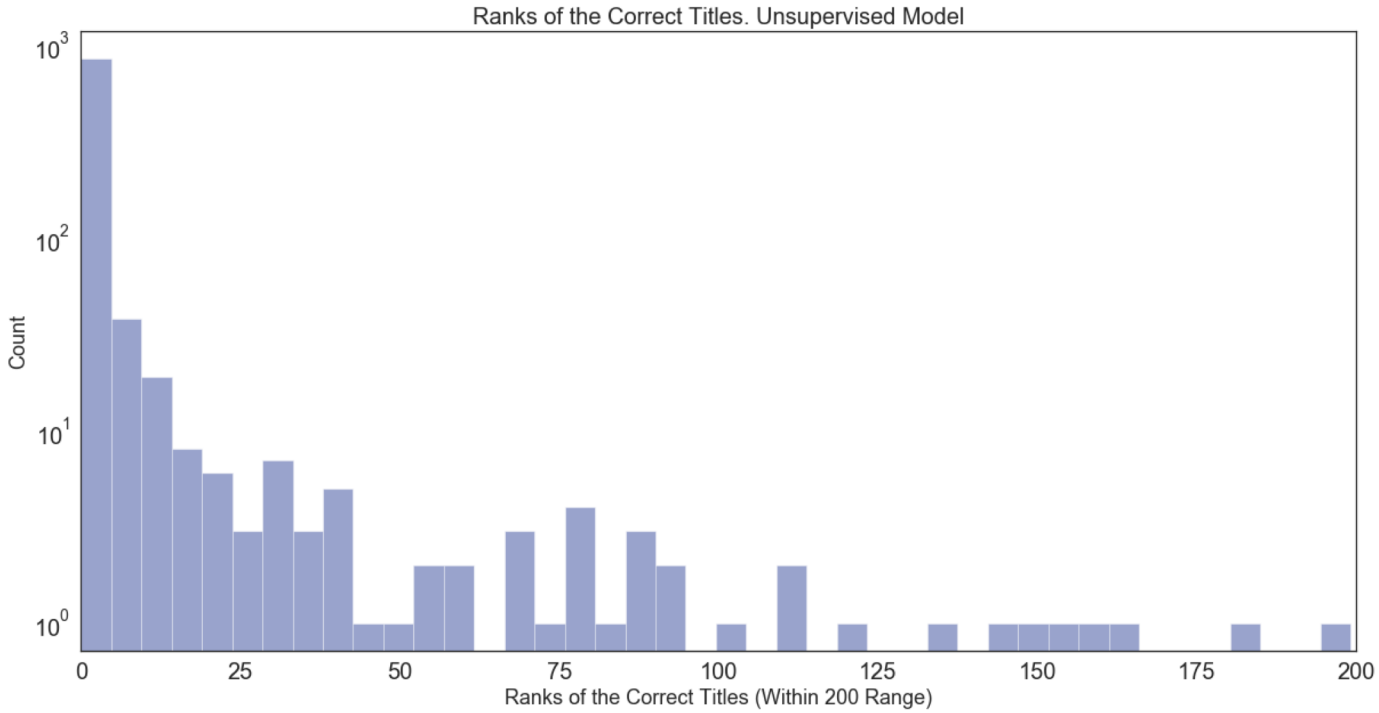
- Lemmatize with Mystem parser
- Word2Vec
- From word vectors to document vectors
- Best matches based on cosine similarity

Ranking

- Take a random sample of a 1000 rows.
- Convert each ad text and each ad headline into document vectors
- Calculate cosine similarity scores for each ad text with 1000 potential ad headlines.
- Then, we will compute the rank of the correct headline.
- This procedure, we will repeat for each ad text in our sample.

- Eventually, we will be able to compute the mean rank of the correct headline and the proportion of the correct headline in the top three results. This will give us a way to evaluate our unsupervised solution.

Unsupervised Model: Results



Mean rank of the correct headline: 16.98

Proportion of the correct headlines in the top three results: 0.837

Stage II. Supervised Approach

Prepare

- We split our dataset into half: correct and incorrect ads text –ads headline pairs (second half is shuffled)

Train

- Train a binary classifier to identify correct /incorrect pairs (0/1)

Evaluate

- Evaluate not only based on accuracy but also the rank of the correct pair (out of all other possible pair combinations in a sample of 1000)

Data

Features: concatenated ads text and ads headline vectors with PCA applied

Outcome: 0/1 (probability of an outcome)

Models

- Logistic Regression
- Random Forest
- XGBoost
- Recurrent Neural Network (RNN):
  - different preprocessing pipeline
  - each wordvector is indexed in an embeddings matrix, which is used in an embeddings layer

RNN

- Processes “sequences explicitly as sequences”
- In our case, we can feed to it a sequence of word vectors, one word vector at a time
- Often applied to solve NLP problems
- Has a ‘temporal’ or ‘memory’ aspect to it
- “The hidden layer from the previous timestep provides a form of memory, or context, that encodes earlier processing and informs the decisions to be made at later points in time.”

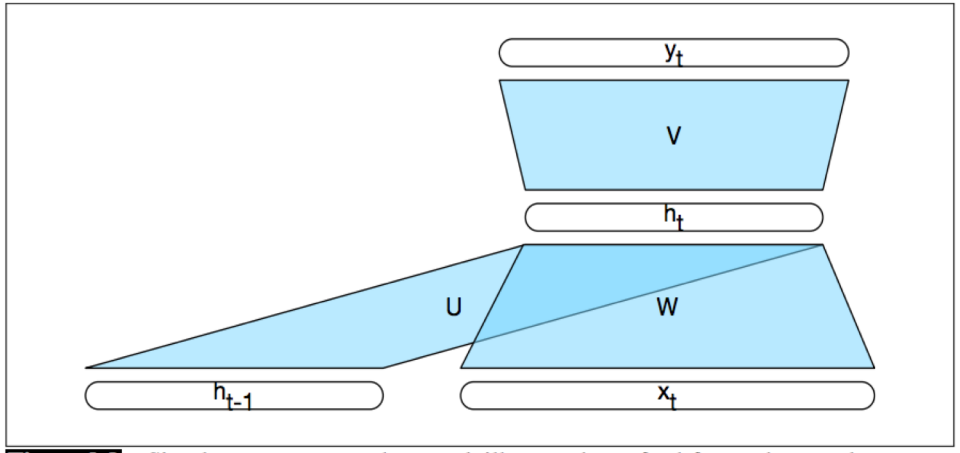


Figure 9.3 Simple recurrent neural network illustrated as a feed-forward network.

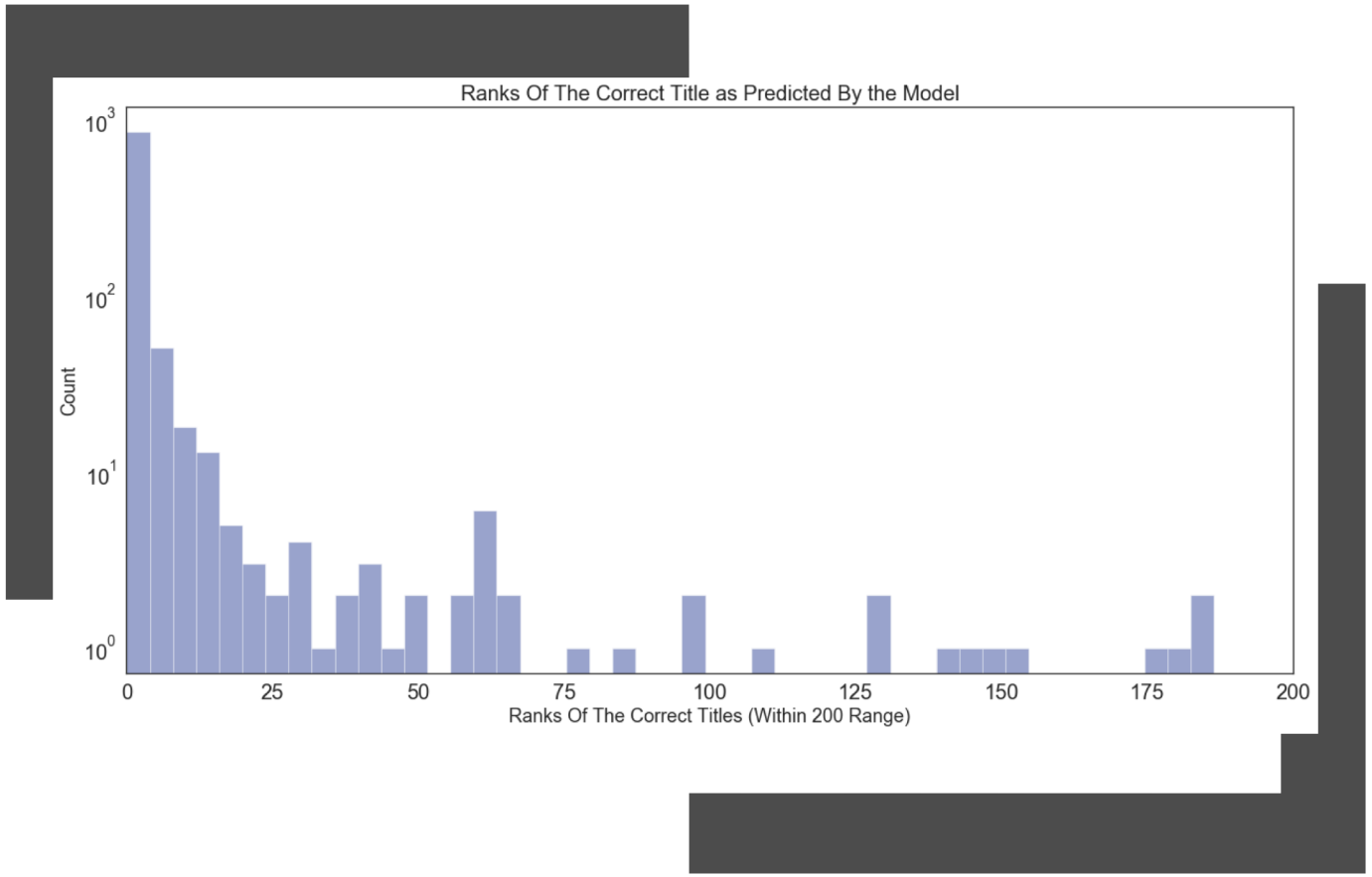
Daniel Jurafsky and James H. Martin. *Speech and Language Processing*, 3rd ed. draft. (Stanford University / University of Colorado at Boulder, 2018), 177-179. Accessed October 23, 2018. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Models Performance

Models	Training Set Accuracy	Test Set Accuracy	Test Set Precision	Test Set Recall	Test Set F1 Score	Confusion Matrix
Logistic Regression	0.5087	0.50	0	0	0	[28528 0 28526 0]
Random Forest	0.8868	0.93	0.94	0.92	0.93	[26843 1685 2311 26215]
XGBoost	0.9302	0.967	0.972	0.963	0.968	[27737 791 1042 27484]
RNN	0.9956	0.9640	0.965	0.9625	0.964	[27696 993 1073 27616]

Ranking Evaluation (sample of a 1000 rows)

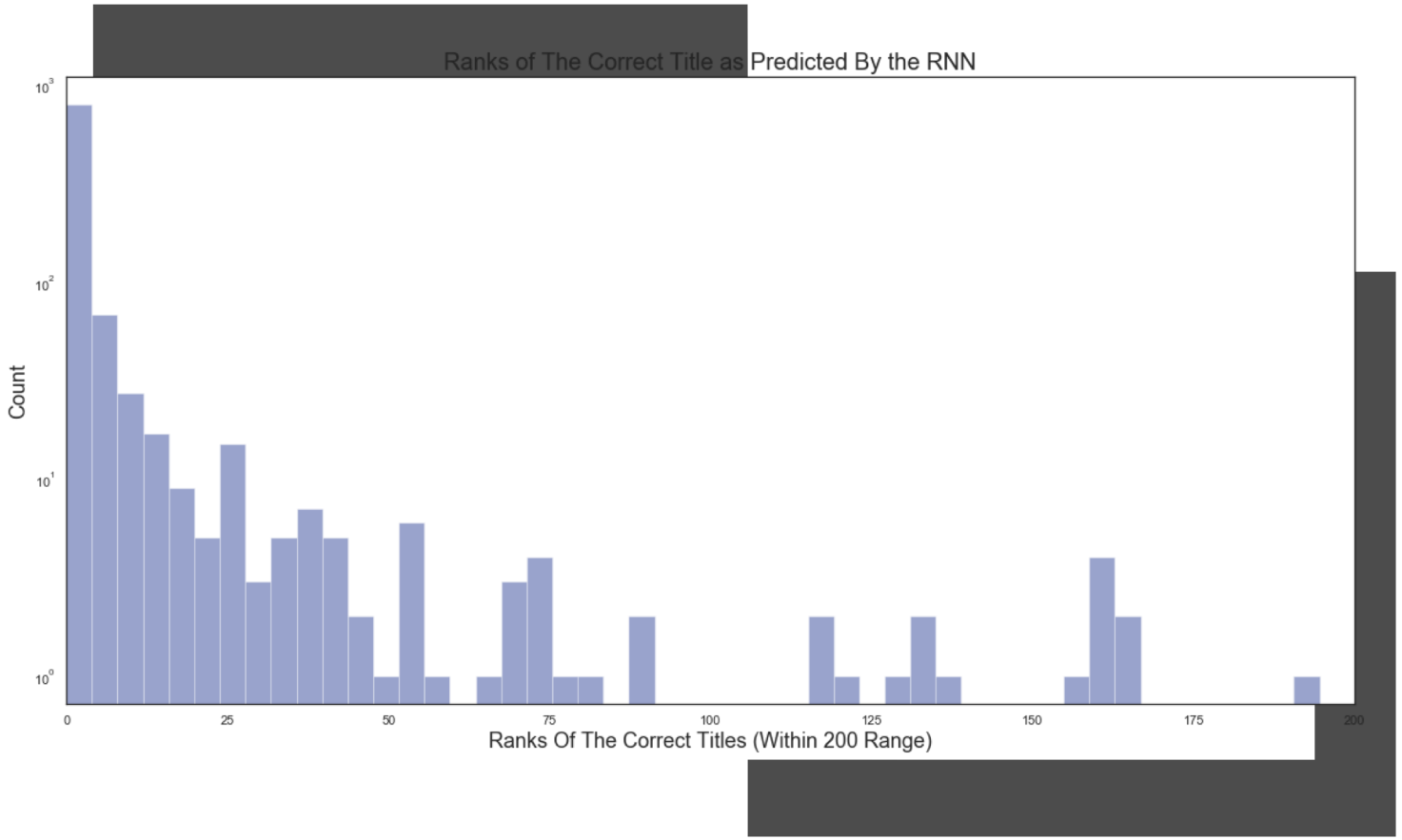
XGBoost



Mean rank of the correct headline: 10.35

Proportion of the correct headlines in the top 3 results: 0.855

RNN



Mean rank of the correct headline: 10.835

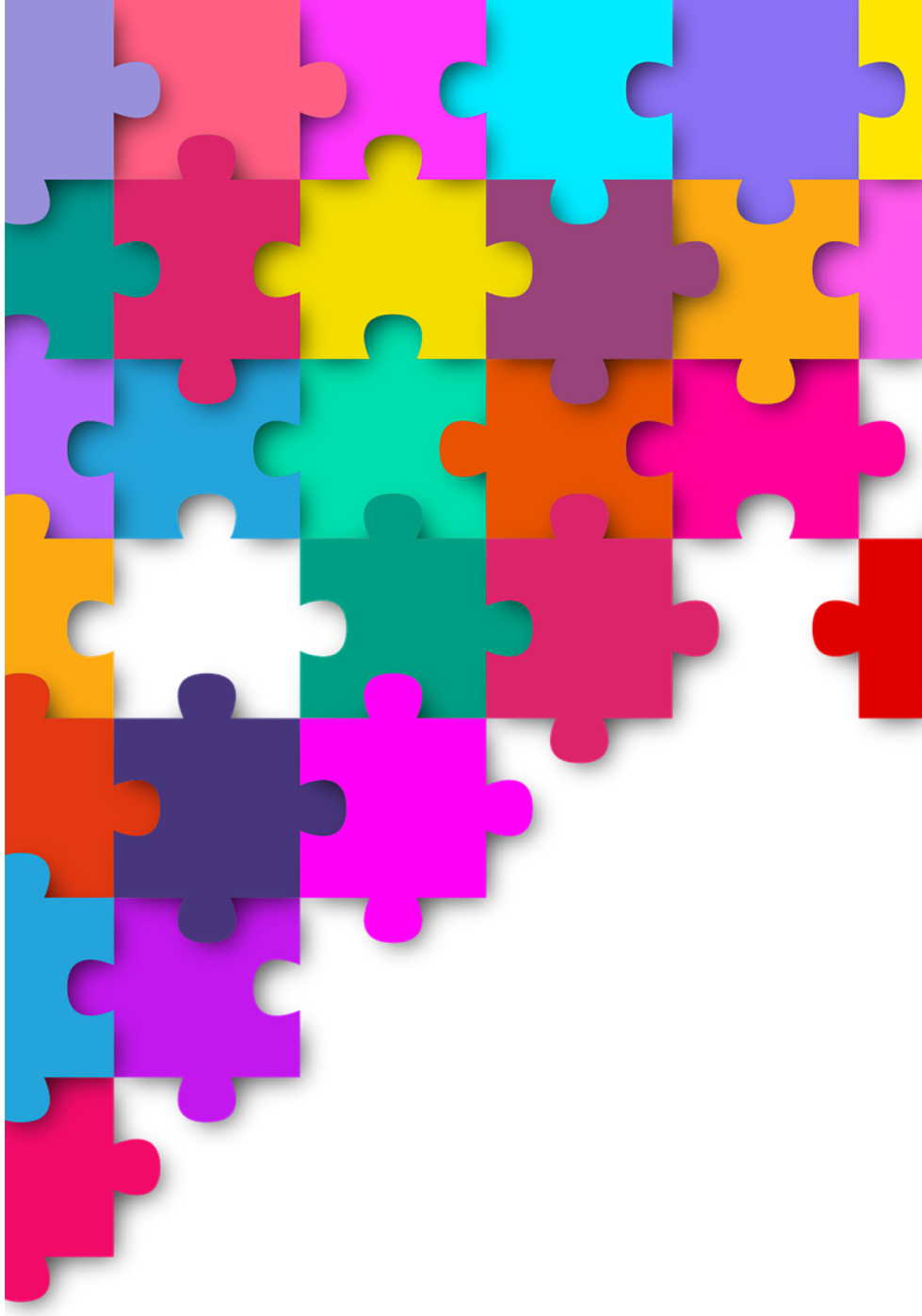
Proportion of the correct headlines in the top 3 results: 0.79

Generate completely new texts from scratch

Select best matches from existing headlines/ texts

Ads Text Generator with RNN

1) Preprocessing: no lemmatization because we want to preserve correct grammatical forms
2) Training word2vec model
3) Creation of unique indices for the most frequent words in the vocabulary
4) Creation of an embedding matrix that keeps all the word vector values by index to be used in our Embedding layer



Creation of predictors and labels:

I love data science

[1, 2, 3, 4]

predictors	label
[1, 2]	[3]
[1, 2, 3]	[4]

We also need to pad all predictors to the same length:

[0, 0, 1, 2]
[0, 1, 2, 3]

Some Examples

“windows” / “окна”

- “окна в хабаровске от производителя скидка на все 5% бесплатная доставка сравните цены всех интернет магазинов”
- ”Windows in Khabarovsk (directly) from a manufacturer 5% off free delivery compare prices of all online stores”)

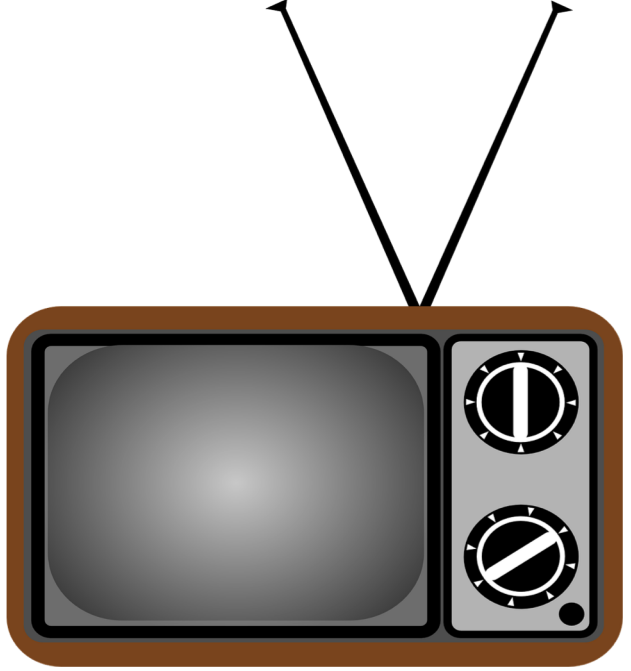
*Grammatically correct (Russian has complex grammar, makes sense (learned some clichés of the genre)

“Blow dryer” / “фен”

- 'фен philips низкая цена гарантия большой выбор быстрая доставка гарантия жми сравните цены всех интернет магазинов’
- ‘blow dryer phillips low price warranty large selection fast delivery warranty click (here) compare prices of all online stores

*Learned some clichés of the genre but overuses the phrase ‘compare prices of all online stores’, uses ‘warranty’ twice

Some Examples



TV-sets / телевизоры

- 'телевизоры сравните цены всех интернет магазинов сравните цены всех интернет магазинов сравните цены всех интернет'
- 'TV-sets compare prices of all online stores compare prices of all online stores compare prices of all online stores

Clearly just repeats the same combination (seems to be a default, 'go-to' phrase when the model did not learn enough examples of texts advertising this particular product

"hotel" / "отель"

- 'отель в чите wifi уютно удобно скидка 42% документы жмите *от производителя hp hp 15 в 220 вольт сравните цены'*
- 'hotel in Chita wifi cozy comfortable 42% off with papers (provided*?) from a manufacturer hp hp 15 220 volt compare prices*

First half of the ad makes sense but then the rest of the text mixes expressions typical of a different product type

Future Goals

Refine	Refine the model and find metrics to measure its performance
Experiment	Generate ads based on product key words or product descriptions
Sell	Sell our generator to Yandex

Special thanks to my mentor Vincent Lonij,
whose kind heart made my work possible