КЛАССИФИКАЦИЯ ТЕКСТОВЫХ СООБЩЕНИЙ

ВЫПОЛНИЛИ:
КРЫЛОВА МАРИЯ БПМИ209
КИРДЯШОВА АЛИСА БПМИ208

ЦЕЛЬ И ЗАДАЧИ

ЦЕЛЬ: СРАВНЕНИЕ АРХИТЕКТУР ДЛЯ ПРЕДСКАЗАНИЯ ТОКСИЧНОСТИ СООБЩЕНИЙ .

ЗАДАЧИ:

- **Создать базовую модель**
- **Сравнить способы токенизации**
- **Сравнить переводы текста в векторы**
- **Проанализировать эмбеддинги**
- **Улучшить базовую модель**

ОПИСАНИЕ ДАННЫХ

НАБОР ДАННЫХ СОДЕРЖИТ ТЕКСТ, КОТОРЫЙ МОЖЕТ БЫТЬ СОЧТЕН НЕПРИСТОЙНЫМ, ВУЛЬГАРНЫМ ИЛИ ОСКОРБИТЕЛЬНЫМ.

- **Размер исходных данных:** 1780823

- **Тип данных объектов:** str

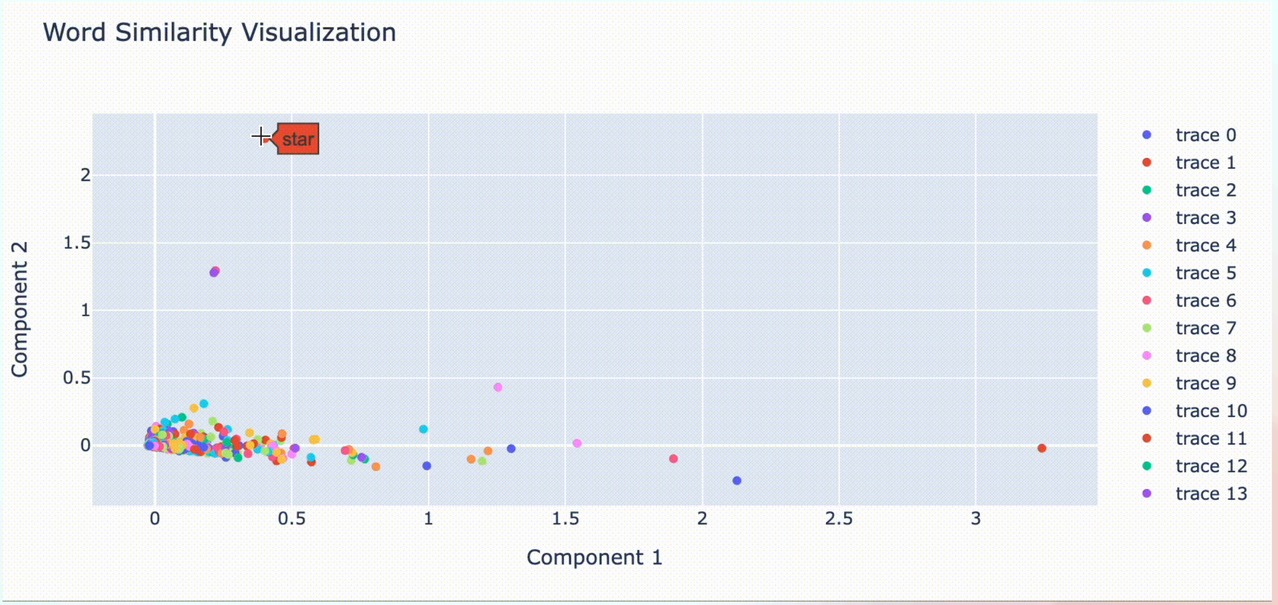
- **Тип таргета:** float

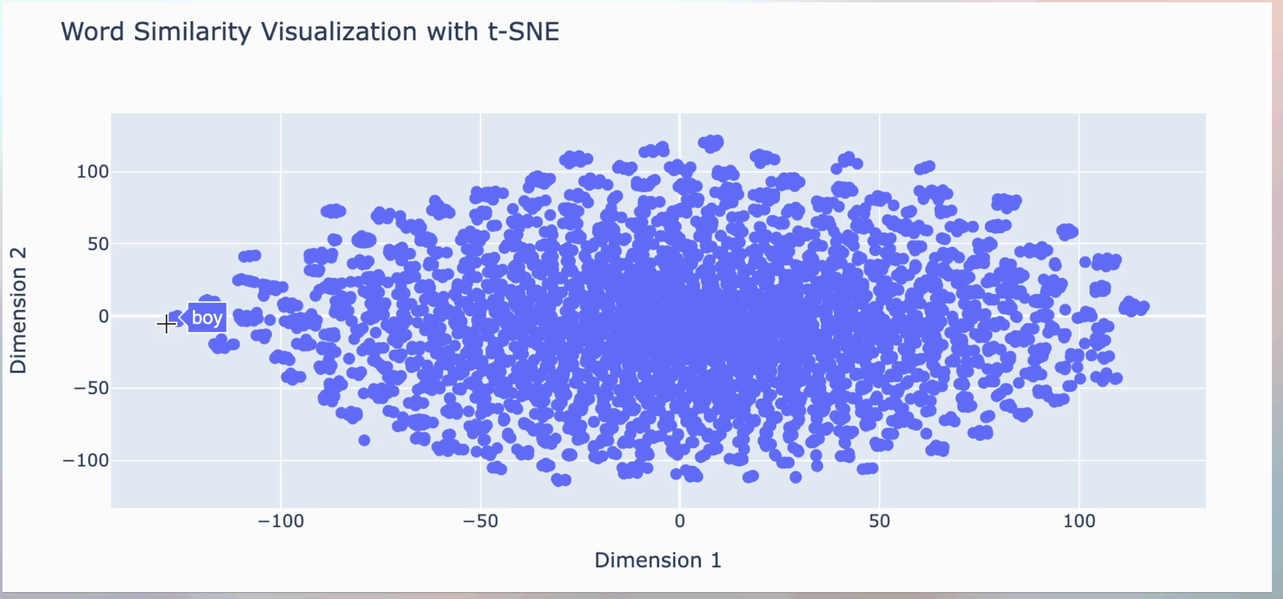
- **Пример:**

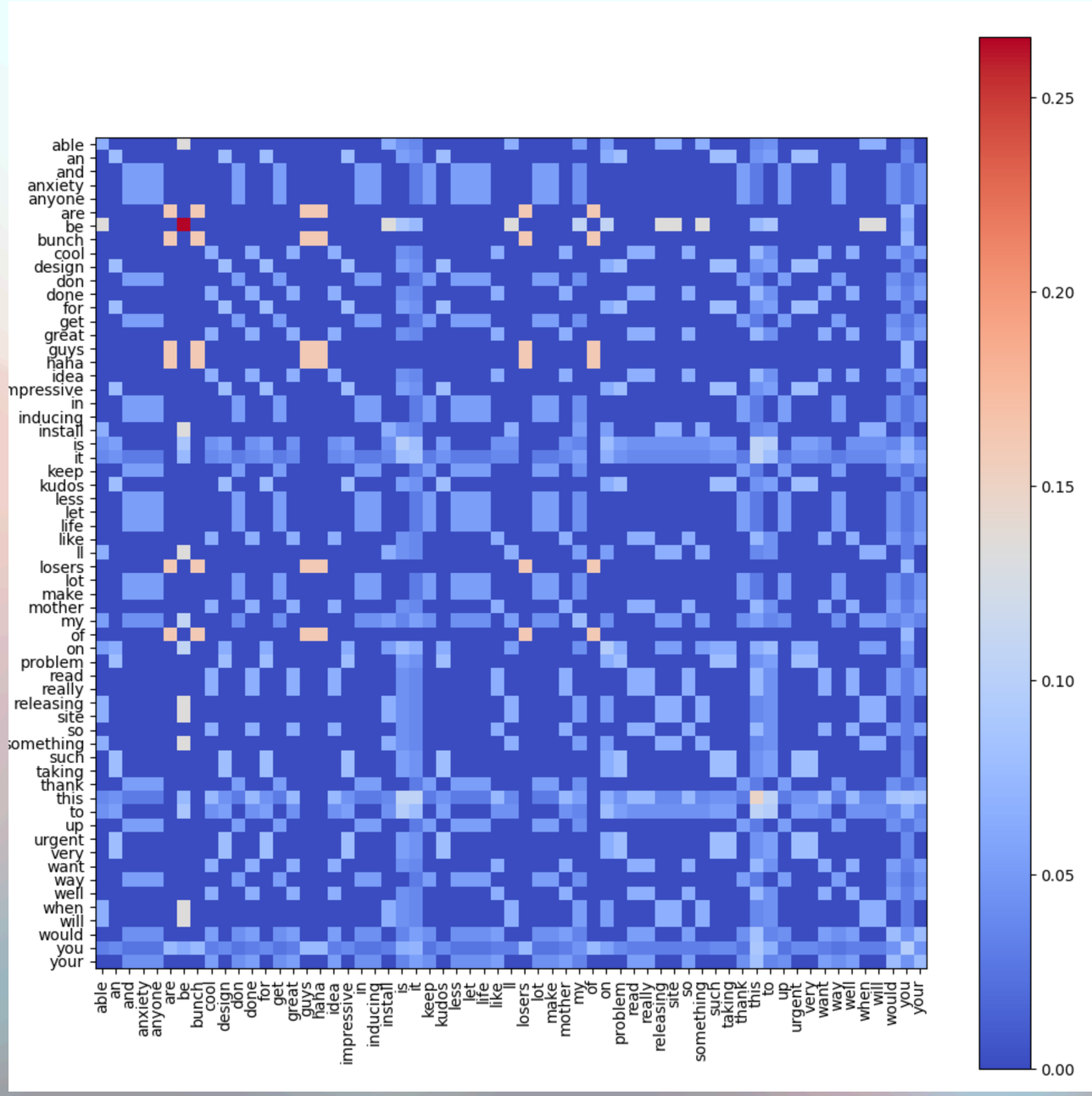
  Continue to stand strong LGBT community. Yes, indeed, you'll overcome and you have.

  Toxicity Labels: All 0.0

АНАЛИЗ ТЕКСТОВЫХ ДАННЫХ


Word Similarity Visualization
star
trace 0
trace 1
trace 2
trace 3
trace 4
trace 5
trace 6
trace 7
trace 8
trace 9
trace 10
trace 11
trace 12
trace 13
Component 2
Component 1
0
0.5
1
1.5
2
2.5
3


Word Similarity Visualization with t-SNE
boy
Dimension 2
Dimension 1
100
50
0
−50
−100

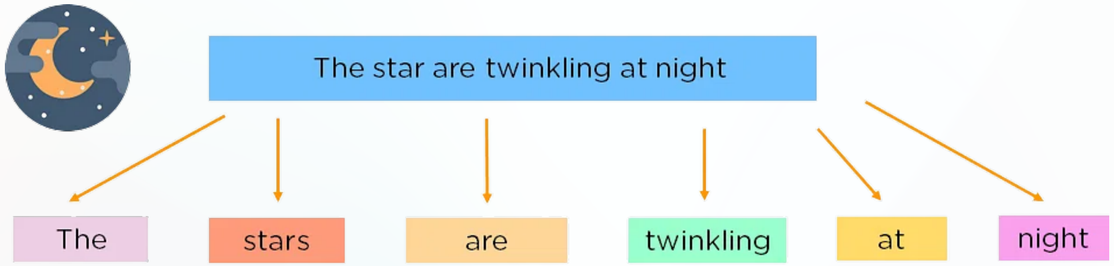

able
an
and
anxiety
anyone
are
be
bunch
cool
design
don
done
for
get
great
guys
haha
idea
impressive
in
inducing
install
is
it
keep
kudos
less
let
life
like
ll
losers
lot
make
mother
my
of
on
problem
read
really
releasing
site
so
something
such
taking
thank
this
to
up
urgent
very
want
way
well
when
will
would
you
your
0.25
0.20
0.15
0.10
0.05
0.00

ИСПОЛЬЗУЕМЫЕ МЕТОДЫ

ТОКЕНИЗАТОРЫ

ТОКЕНИЗАТОР – ИНСТРУМЕНТ ДЛЯ АВТОМАТИЧЕСКОГО ИЛИ ПОЛУАВТОМАТИЧЕСКОГО РАЗДЕЛЕНИЯ ТЕКСТА НА ТОКЕНЫ, Т.Е. НА СЛОВА И ДРУГИЕ ЦЕПОЧКИ СИМВОЛОВ, КОТОРЫЕ МЫ ХОТИМ СЧИТАТЬ МИНИМАЛЬНЫМИ ЛИНЕЙНЫМИ ЕДИНИЦАМИ ТЕКСТА.

- **Посимвольная:** процесс разделения текста на буквы-компоненты
- **Пословная:** процесс разделения исходного текста на слова-компоненты.
- **Лемматизация:** процесс, который использует словарь и морфологический анализ, чтобы привести слово к его канонической форме - лемме.
- **Стемминг:** эвристический процесс, отрезающий от корня слов, что приводит к потере словообразовательных суффиксов



ПРИМЕР РАБОТЫ ТОКЕНИЗАТОРОВ

Комментарий	посимвольная	пословная	лемматизация	стемминг
This is such an urgent design problem; kudos to you for taking it on. Very impressive!	T', 'h', 'i', 's', ' ', 'i', 's', ' ', 's', 'u', 'c', 'h', ' ', 'a', 'n', ' ', 'u', 'r', 'g', 'e', 'n', 't', ' ', 'd', 'e', 's', 'i', 'g', 'n', ' ', 'p', 'r', 'o', 'b', 'l', 'e', 'm', ';', ' ', 'k', 'u', 'd', 'o', 's', ' ', 't', 'o', ' ', 'y', 'o', 'u', ' ',	This', 'is', 'such', 'an', 'urgent', 'design', 'problem', ';', 'kudos', 'to', 'you', 'for', 'taking', 'it', 'on', '.', 'Very', 'impressive', '!'	this', 'be', 'such', 'an', 'urgent', 'design', 'problem', ';', 'kudo', 'to', 'you', 'for', 'take', 'it', 'on', '.', 'very', 'impressive', '!'	thi', 'is', 'such', 'an', 'urgent', 'design', 'problem', ';', 'kudo', 'to', 'you', 'for', 'take', 'it', 'on', '.', 'veri', 'impress', '!'
haha you guys are a bunch of losers	h', 'a', 'h', 'a', ' ', 'y', 'o', 'u', ' ', 'g', 'u', 'y', 's', ' ', 'a', 'r', 'e', ' ', 'a', ' ', 'b', 'u', 'n', 'c', 'h', ' ', 'o', 'f', ' ', 'l', 'o', 's', 'e', 'r', 's'	haha', 'you', 'guys', 'are', 'a', 'bunch', 'of', 'losers'	haha', 'you', 'guy', 'be', 'a', 'bunch', 'of', 'loser'	haha', 'you', 'guy', 'are', 'a', 'bunch', 'of', 'loser'
The ranchers seem motivated by mostly by greed; no one should have the right to allow their animals destroy public land	T', 'h', 'e', ' ', 'r', 'a', 'n', 'c', 'h', 'e', 'r', 's', ' ', 's', 'e', 'e', 'm', ' ', 'm', 'o', 't', 'i', 'v', 'a', 't', 'e', 'd', ' ', 'b', 'y', ' ', 'm', 'o', 's', 't', 'l', 'y', ' ', 'b', 'y', ' ', 'g', 'r', 'e', 'e', 'd', ';', ' ', 'n', 'o', '	The', 'ranchers', 'seem', 'motivated', 'by', 'mostly', 'by', 'greed', ';', 'no', 'one', 'should', 'have', 'the', 'right', 'to', 'allow', 'their', 'animals', 'destroy', 'public', 'land',	the', 'rancher', 'seem', 'motivate', 'by', 'mostly', 'by', 'greed', ';', 'no', 'one', 'should', 'have', 'the', 'right', 'to', 'allow', 'their', 'animal', 'destroy', 'public', 'land'	the', 'rancher', 'seem', 'motiv', 'by', 'mostli', 'by', 'greed', ';', 'no', 'one', 'should', 'have', 'the', 'right', 'to', 'allow', 'their', 'anim', 'destroy', 'public', 'land'

ЭМБЕДДИНГИ

ЭМБЕДДИНГИ – ВЕКТОРНЫЕ ПРЕДСТАВЛЕНИЯ СЛОВ, ФРАЗ ИЛИ ДОКУМЕНТОВ В ЧИСЛОВОМ ВИДЕ.

- **GloVe**

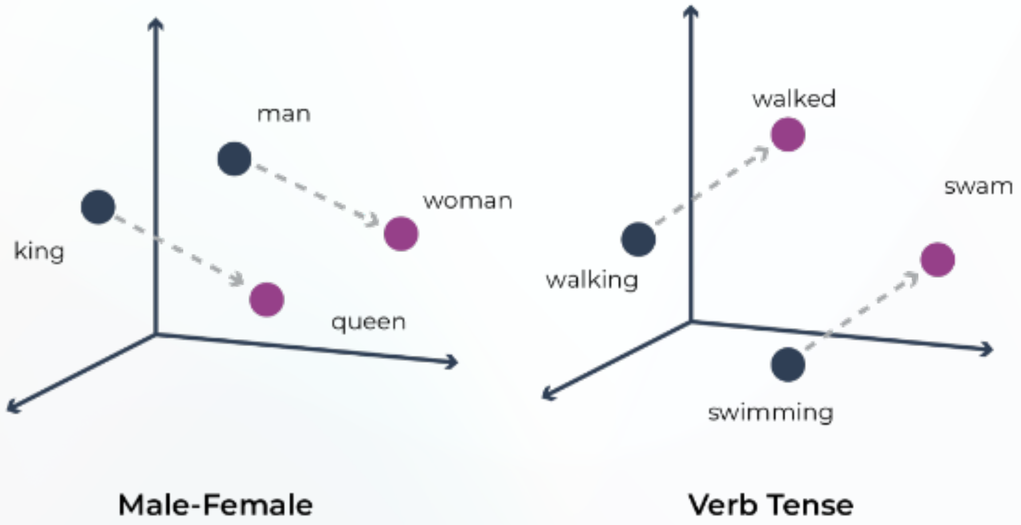
  Архитектура: TF-IDF

  Контекст: глобальная статистика совстречаймости

- **Word2Vec**

  Архитектура: CBOW, Skip-gram

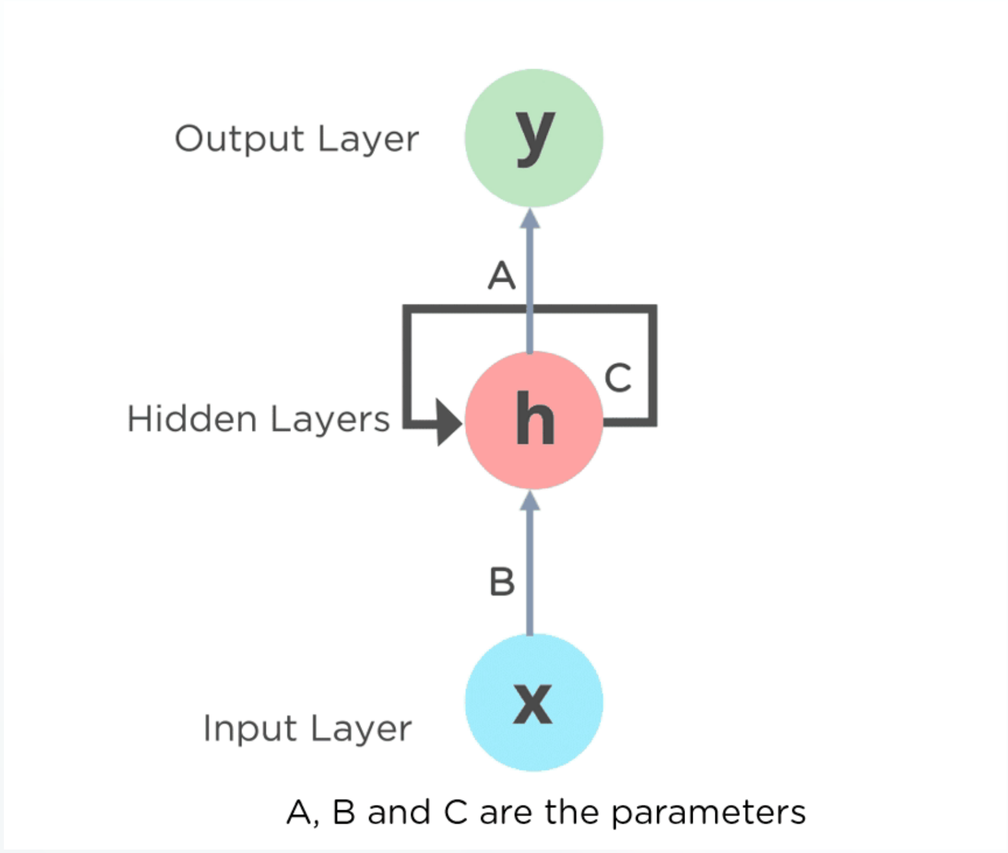
  Контекст: локальный



RNN АРХИТЕКТУРА

РЕКУРРЕНТНЫЕ НЕЙРОННЫЕ СЕТИ – ВИД НЕЙРОННЫХ СЕТЕЙ, ГДЕ СВЯЗИ МЕЖДУ ЭЛЕМЕНТАМИ ОБРАЗУЮТ НАПРАВЛЕННУЮ ПОСЛЕДОВАТЕЛЬНОСТЬ.

	Пословная	Лемматизация	Стемминг
GloVe	0.218	0.216	0.212
W2V	0.119	0.106	0.117

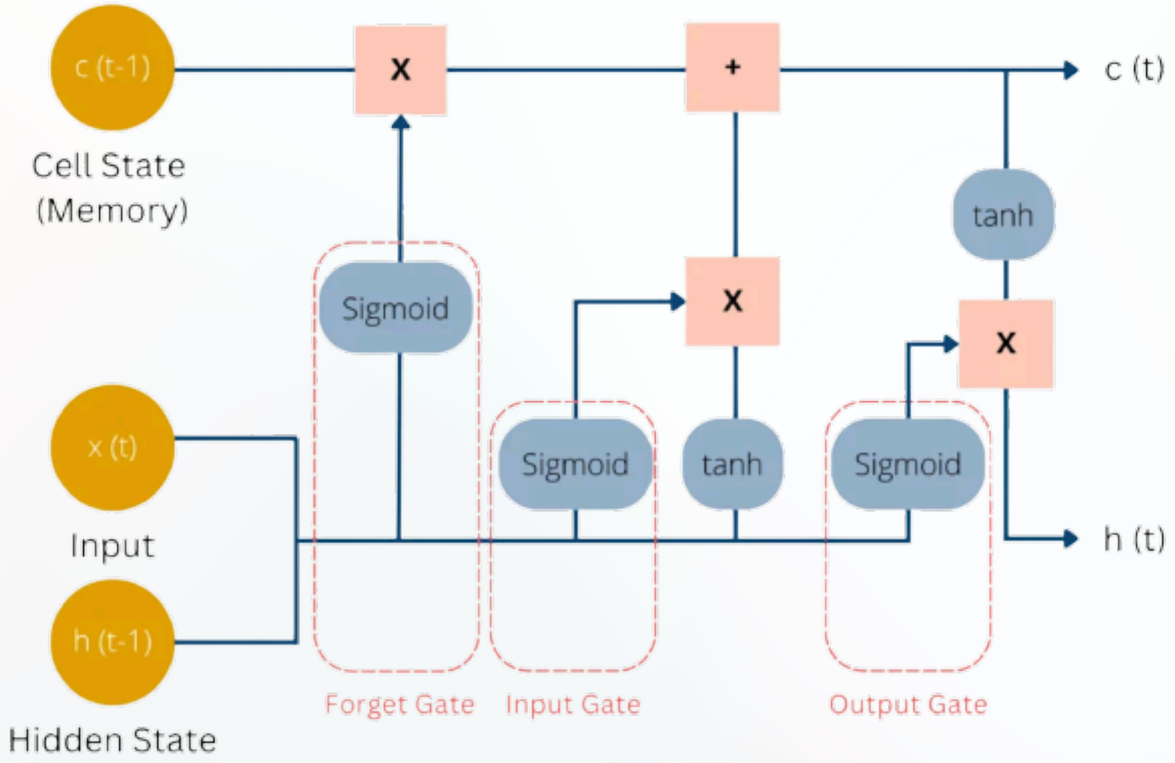


Посимвольный эмбэддинг: 0.134

LSTM АРХИТЕКТУРА

LSTM - ОСОБАЯ РАЗНОВИДНОСТЬ АРХИТЕКТУРЫ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ, СПОСОБНАЯ К ОБУЧЕНИЮ ДОЛГОВРЕМЕННЫМ ЗАВИСИМОСТЯМ.

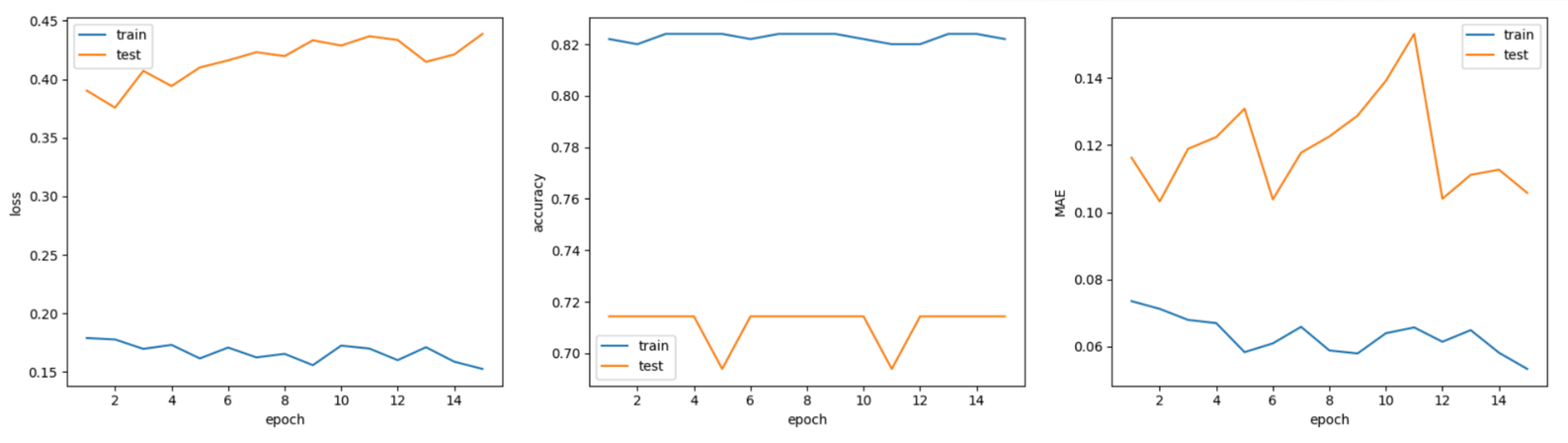
	GloVe	W2V
Пословная	0.219	0.110
Лемматизация	0.221	0.110
Стемминг	0.221	0.111

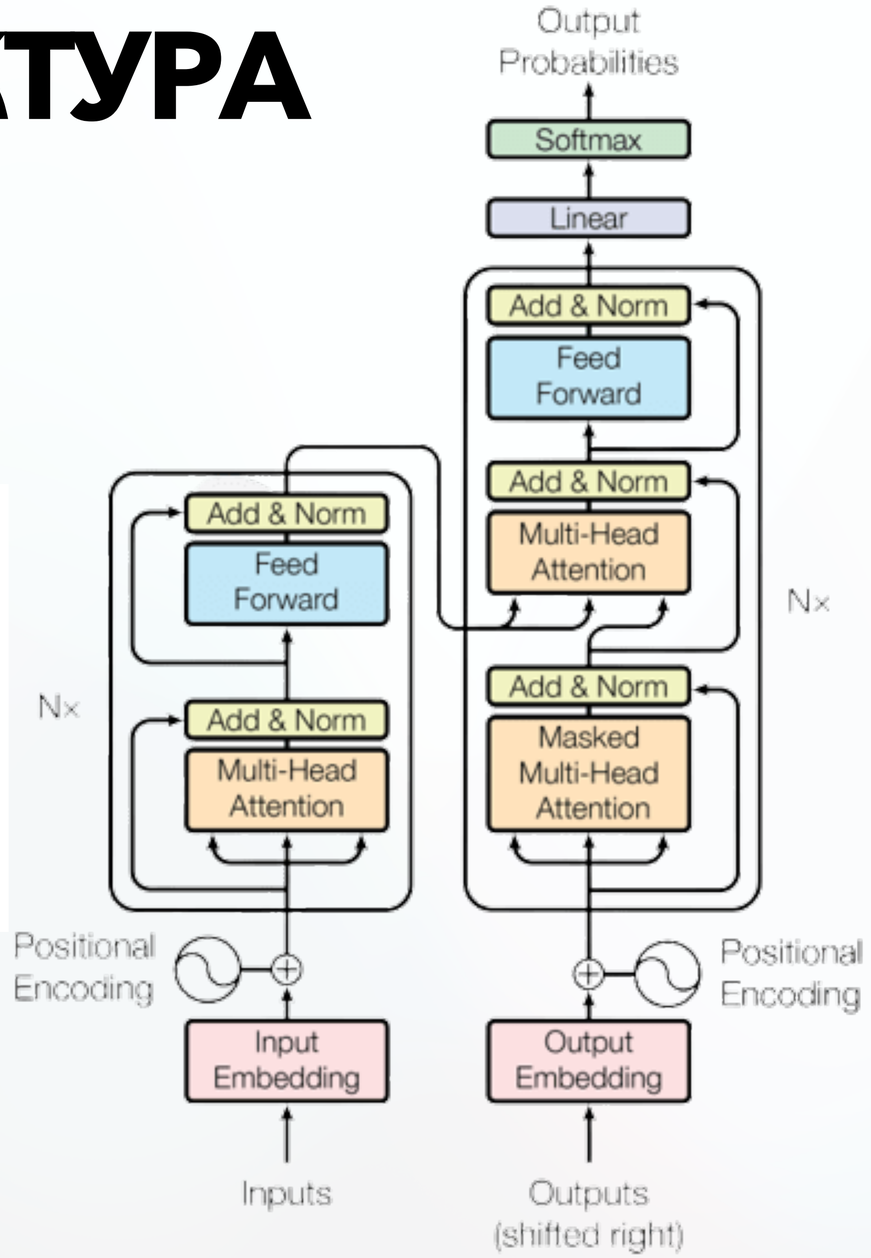


Посимвольный эмбэддинг: 0.126

TRANSFORMER АРХИТЕКТУРА

TRANSFORMER - ЭТО АРХИТЕКТУРА МОДЕЛИ ГЛУБОКОГО ОБУЧЕНИЯ, ОСНОВАННАЯ НА МЕХАНИЗМЕ ВНИМАНИЯ, И ПОЗВОЛЯЮЩАЯ ЭФФЕКТИВНО ОБРАБАТЫВАТЬ ПОСЛЕДОВАТЕЛЬНОСТИ ДАННЫХ





W2V stemming: 0.105

ДАЛЬНЕЙШИЕ ПУТИ РАЗВИТИЯ

- Усложнение архитектур моделей
- Побить benchmark в kaggle соревновании
- Попробовать другие архитектуры нейронный сетей