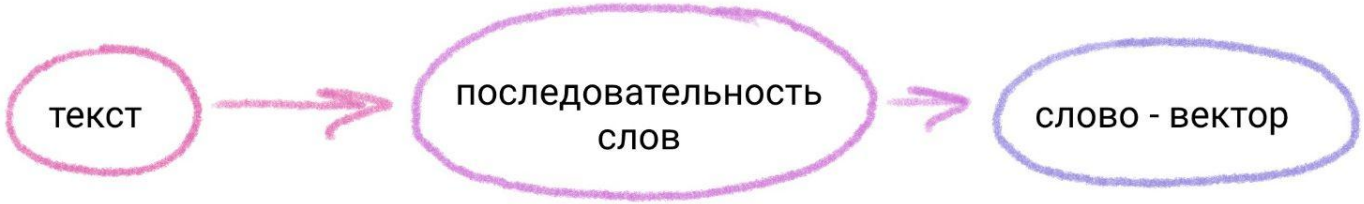
Spot the bot: семантические траектории текстов естественного языка

научный руководитель: Громов Василий Александрович
студент: Фальчикова Вероника Евгеньевна

Постановка задачи



Многомерное векторное пространство - семантическое пространство

Предположение: в семантическом пространстве траектории людей и ботов не совпадают - существуют области в семантическом пространстве, где бывают только люди или только боты.

Актуальность и релевантные работы

Большинство исследований - обучение с учителем.

Проблемы:

- модель подстраивается под конкретный генератор текста
- качество разметки выборки.

Проведенные исследования

Сбор датасетов

Литературный

- ~10,000 текстов по 1,000 слов
- 100,000 уникальных слов

Сгенерированный

- LSTM-бот
- Фрагмент 100 слов: 10 -> 90 предсказанных

1. Очистка + лемматизация текстов
2. Составление словаря слово-вектор с помощью SVD разложения матрицы TF-IDF

$$\mathrm{tf}(t,d) = \frac{n_t}{\sum_k n_k}$$ → число вхождений слова *t* в документ *d*

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{\, d_i \in D \mid t \in d_i \,\}|}$$ → инверсия частоты, с которой слово *t* встречается в документах коллекции *D*

Проведенные исследования

Сбор датасетов

Литературный

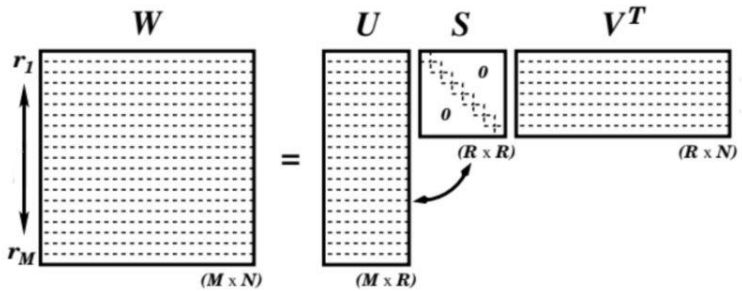
- ~10,000 текстов по 1,000 слов
- 100,000 уникальных слов

Сгенерированный

- LSTM-бот
- Фрагмент 100 слов:

10 -> 90 предсказанных

1. Очистка + лемматизация текстов
2. Составление словаря слово-вектор с помощью SVD разложения матрицы TF-IDF



Проведенные исследования

Сбор датасетов

Литературный

- ~10,000 текстов по 1,000 слов
- 100,000 уникальных слов

Сгенерированный

- LSTM-бот
- Фрагмент 100 слов: 10 -> 90 предсказанных

Очистка + лемматизация текстов

Составление словаря слово-вектор с помощью SVD разложения матрицы TF-IDF

Составление датасета из векторов n-грамм

Проведенные исследования

Сбор датасетов

Литературный

- ~10,000 текстов по 1,000 слов
- 100,000 уникальных слов

Сгенерированный

- LSTM-бот
- Фрагмент 100 слов: 10 -> 90 предсказанных

1. Очистка + лемматизация текстов
2. Составление словаря слово-вектор с помощью SVD разложения матрицы TF-IDF
3. Составление датасета из векторов n-грамм

Проведенные исследования

Метрики качества кластеризации

- 10 синтетических датасетов

Метрика/Номер датасета	1	2	3	4	5	6	7	8	9	10
RMSSTD						+				
RS						+				
Hubert										
CH	+	+		+	+	+		+	+	
I	+	+		+	+	+	+	+		+
Dunn's		+			+	+				
Silhouette	+	+			+	+		+		+
DB	+	+		+	+	+		+		+
Xie-Beni		+			+	+				+
SD	+	+	+		+	+				+
S_Dbw		+							+	
CVNN		+		+		+				+

Проведенные исследования

Метрики качества кластеризации

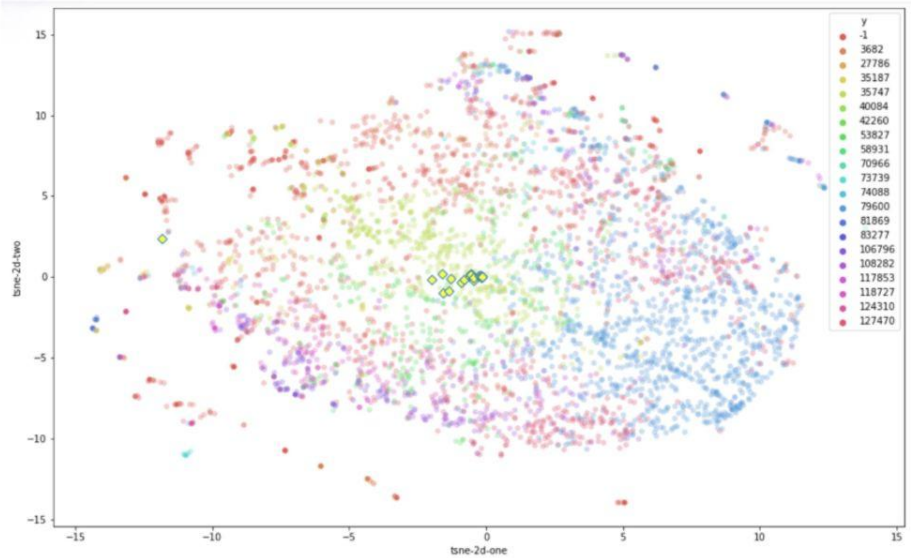
- 10 синтетических датасетов
- Чаще всего правильный результат давали:
  I-индекс
  Дэвис-Болдин
  Силуэт

Метрика/Номер датасета	1	2	3	4	5	6	7	8	9	10
RMSSTD						+				
RS						+				
Hubert										
CH	+	+		+	+	+		+	+	
I	+	+		+	+	+	+	+		+
Dunn's		+			+	+				
Silhouette	+	+			+	+		+		+
DB	+	+		+	+	+		+		+
Xie-Beni		+			+	+				+
SD	+	+	+		+	+				+
S_Dbw		+							+	
CVNN		+		+		+				+

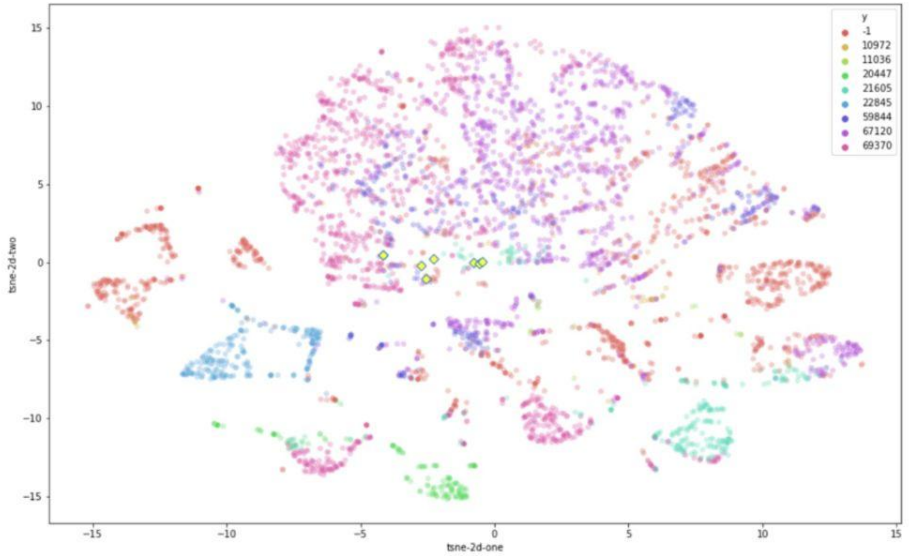
Проведенные исследования

Подпространственная кластеризация PROCLUS

	Best bot [I Index]	Best human [I Index]	Best bot [S & DB]	Best human [S & DB]	2nd best human [DB]
Среднее расстояние до центра	2.7	2.2	2.1	2.6	1.8
Среднее расстояние до медоида	3.2	3.2	3.0	2.7	2.4
Средний размер кластера	6 917	33 842	20 802	72 666	5 767

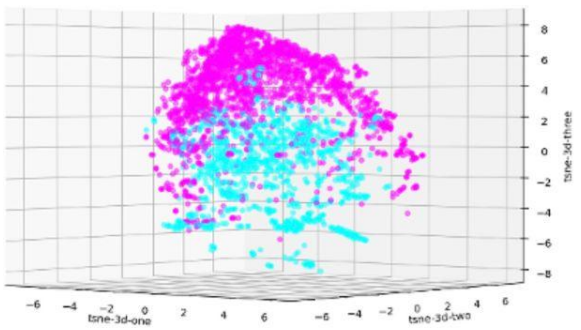


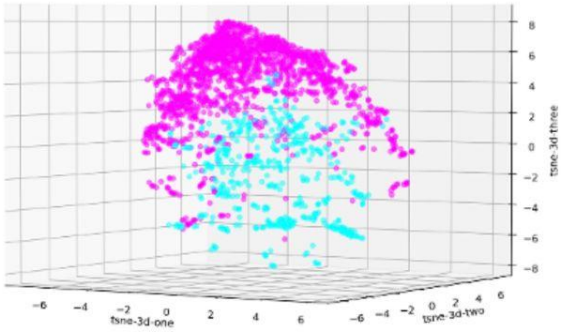
художественные тексты

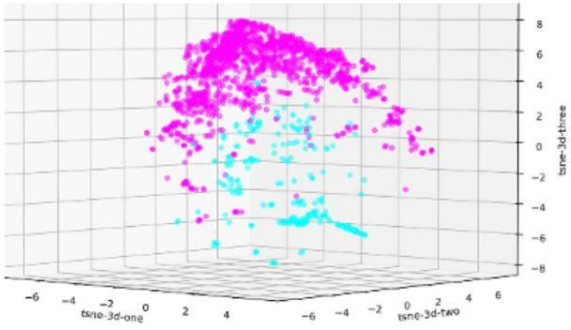


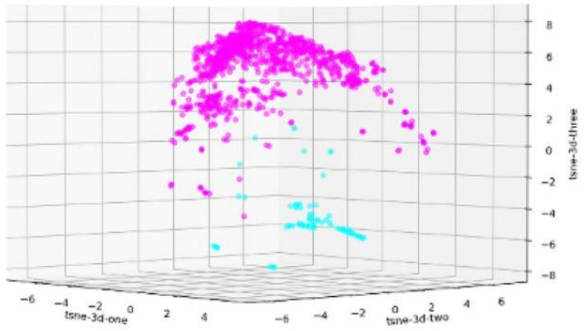
сгенерированные тексты

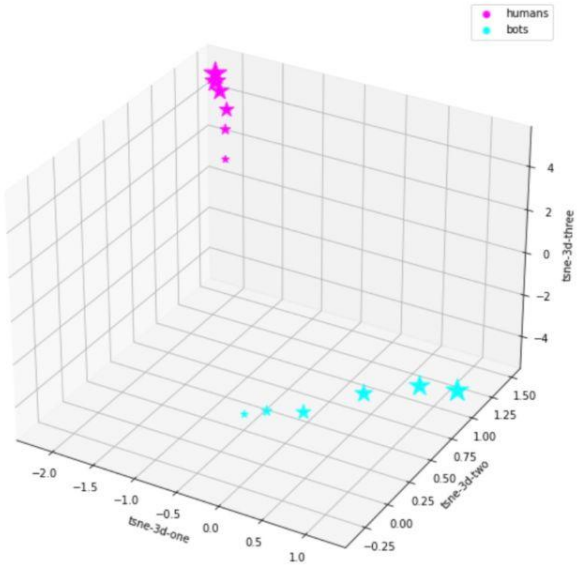
Результаты







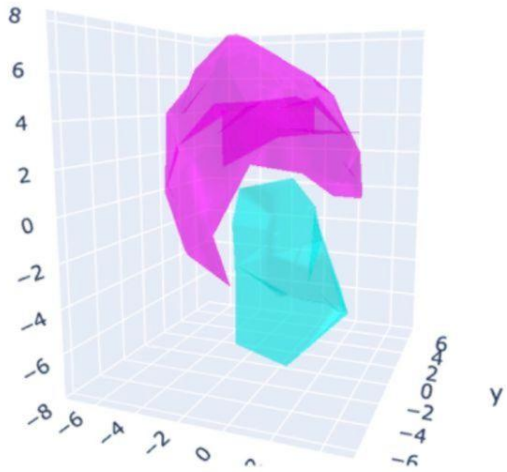




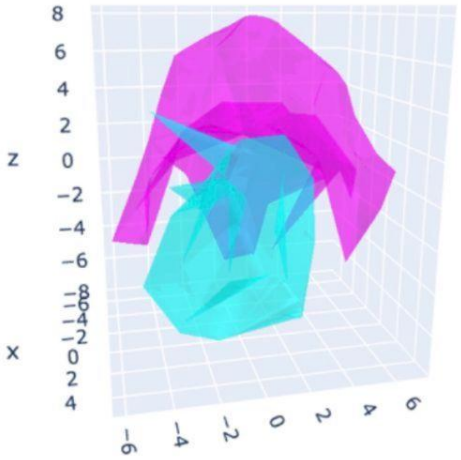
центроиды

симметрическая разность при *alpha* = 0.1, 0.3, 0.5, 0.65, 0.8, 0.9, 1.

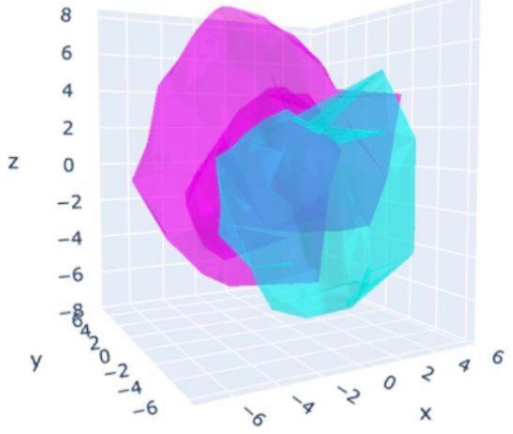
Результаты



alpha = 1.0



alpha = 0.9



alpha = 0.8

Существуют такие области в семантическом пространстве, где бывают только люди, и такие, в которых бывают только боты.