Spot The Bot

Семантические траектории текстов естественного языка

Выполнил: Амрин Айдар, БПМИ-204

Научный руководитель: Громов Василий Александрович[1]

[1]Доцент физико-математических наук, профессор и заместитель руководителя Департамента Анализа Данных и Искусственного Интеллекта ФКН

8 июня 2022 г.

HSE University Faculty of Computer science

Содержание

Содержание

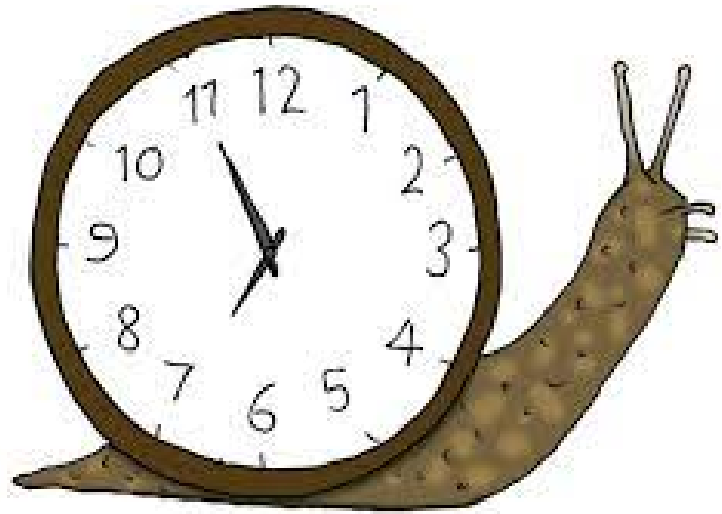
Актуальность проблемы

Актуальность проблемы

Глубинные нейронные сети

На данный момент, самым распространенным решением поставленной задачи является – обучение нейронных сетей. Однако, у такого подхода есть несколько заметных минусов:

1. Требуются большие объемы данных
2. Обучение нейросети является долгим и ресурсо затратным процессом
3. Модель нужно обучать под какой-то конкретный, заранее известный нам, генератор текстов



Содержание

Цель и задачи работы

Основной целью данной работы является – построение универсального алгоритма, который не опирается на внутреннюю структуру какого-либо генератора и показывает удовлетворительные результаты в самых разных случаях. Желательно, чтобы итоговый алгоритм был легко имплементируемым и не затратным в плане ресурсов.

Цель и задачи работы

Задачи поставленные в рамках курсовой работы:

- Анализ основных статей и алгоритмов связанных с темой работы
- Сбор литературных текстов на казахском языке и построение корпуса
- Моделирование взаимоотношений внутри текстовых данных путем построения семантического пространства
- Научиться осуществлять преобразование word-to-vec (text to time series)
- Построение эффективного алгоритма идентификации ботов, основанный на вычислении характеристики хаотических временных рядов
- Проведение серии экспериментов для валидации результатов

Содержание

Семантическое пространство

Пусть $\mathcal{M}$ – множество состоящее из M различных слов, а $\mathcal{N}$ – множество валидных последовательностей этих слов. Хотим построить отображение из дискретных множеств $\mathcal{M}, \mathcal{N}$ в непрерывное векторное пространство $\mathcal{L}$.
От пространства $\mathcal{L}$ мы требуем, чтобы оно отражало “скрытые” зависимости между различными словами и документами.

TF-IDF матрица

Для построения этого семантического пространства, каждому слову мы присваиваем некий вес, так называемую *tf-idf* статистику (от англ. *tf* – term frequency, idf – inverse document frequency):

$$\text{tf-idf}(t, d) = \frac{f(t, d)}{\sum\limits_{t' \in D} f(t', d)} \cdot \log \frac{|\mathcal{N}|}{|\{d \in \mathcal{N} : t \in d\}|}, \tag{1}$$

где функция $f : \mathcal{M} \times \mathcal{N} \to \mathbb{N}$ определена как $f(t, d)$ – количество вхождений слова t в документ d.
Таким образом, получаем матрицу $W \in \mathbb{R}^{M \times N}$, $W_{ij} = \text{tf-idf}(t_i, d_j)$, где $\mathcal{M} = \{t_1, t_2, \ldots, t_M\}$ и $\mathcal{N} = \{d_1, d_2, \ldots, d_N\}$. Таким образом, каждое слово из $\mathcal{M}$ может ассоциироваться с вектор строкой размерности N, а каждый документ из $\mathcal{N}$ с вектор столбцом размерности M.

Сингулярное разложение

Пусть дано $r \leq \text{rank}(W)$, тогда $W \approx \hat{W} = U\Sigma V^T$ будем называть SVD порядка r матрицы W, где $U \in \mathbb{R}^{M \times r}$, $V \in \mathbb{R}^{N \times r}$ – матрицы с ортонормированными столбцами, а $\Sigma \in \mathbb{R}^{r \times r}$ – диагональная матрица, на диагонали которой стоят числа

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$$

Строки матрицы U обозначим через $u_i, 1 \leq i \leq M$, и строки V обозначим через $v_j, 1 \leq j \leq N$. Если Θ_k и $\Xi_k (1 \leq k \leq r)$ – столбцы U и V соответственно, то столбцы U (аналогично V) задают ортонормированный базис $\text{span}(u_1, u_2, \ldots, u_M)$ (аналогично $\text{span}(v_1, v_2, \ldots, v_N)$). Несложно заметить, что тогда

$$W^{(j)} = \sum_{k=1}^{r} (v_j \Sigma)_k \Theta_k, \tag{2}$$

$$W_{(i)} = \sum_{k=1}^{r} (u_i \Sigma)_k \Xi_k^T. \tag{3}$$

Статистическая мера сложности

Что такое сложность? Какие системы мы называем сложными, а какие нет?

Статистическая мера сложности

Что такое сложность? Какие системы мы называем сложными, а какие нет?

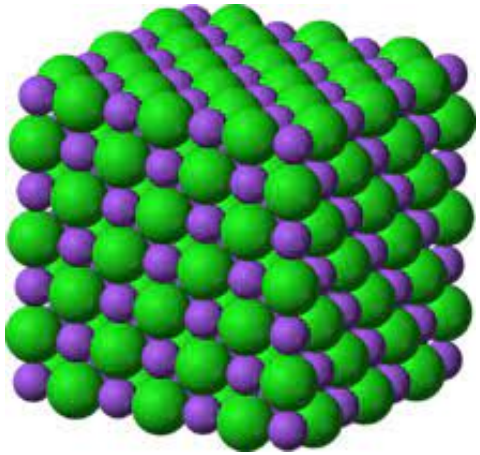


Рис.: Кристальная структура

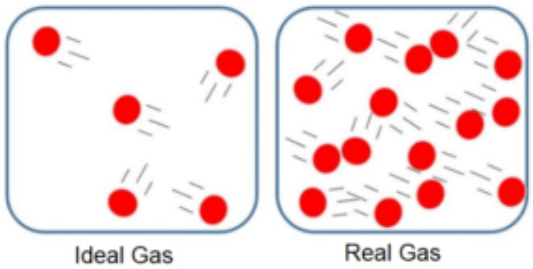


Рис.: Идеальный газ

Идеальный кристалл – модель кристалла с идеальной, совершенной симметрической структурой. Изолированный идеальный газ – равновероятно находится в любом из доступных состояний. Мы считаем, что у идеального газа нет никакого внутреннего порядка.

Статистическая мера сложности

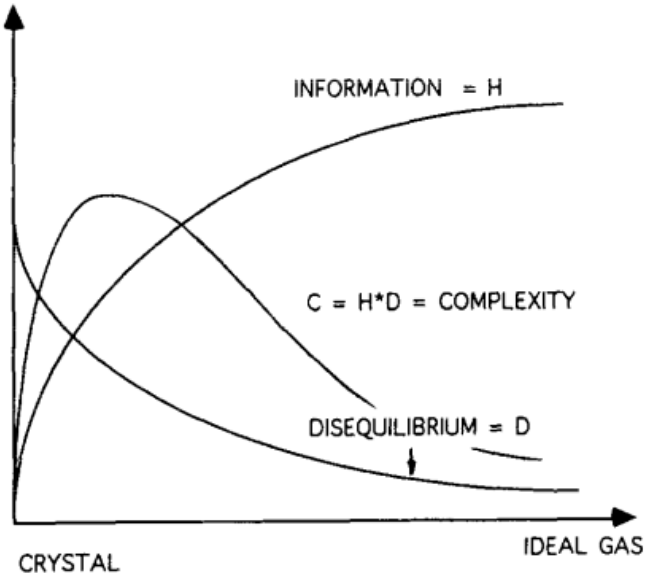


Рис.: интуитивное понятие сложности

Введем меру “сложности” как $C = H \cdot D$, где H – информация содержащаяся в системе и D – расстояние до равновероятного распределения, т.е. disequilibrium.

Информация и disequilibrium

Предположим, что система может находиться в одном из N возможных состояний $\{x_1, x_2, \ldots, x_N\}$ и пусть $\{P(x_1), P(x_2), \ldots, P(x_N)\} = \{p_1, p_2, \ldots, p_N\}$ – это соответствующие вероятности (с условием $\sum_{i=1}^{N} p_i = 1$ и $p_i > 0$). Тогда в качестве меры "информации" возьмем нормированную энтропию Шеннона:

$$H(P) = -\frac{1}{\ln N} \sum_{i=1}^{N} p_i \ln(p_i), \tag{4}$$

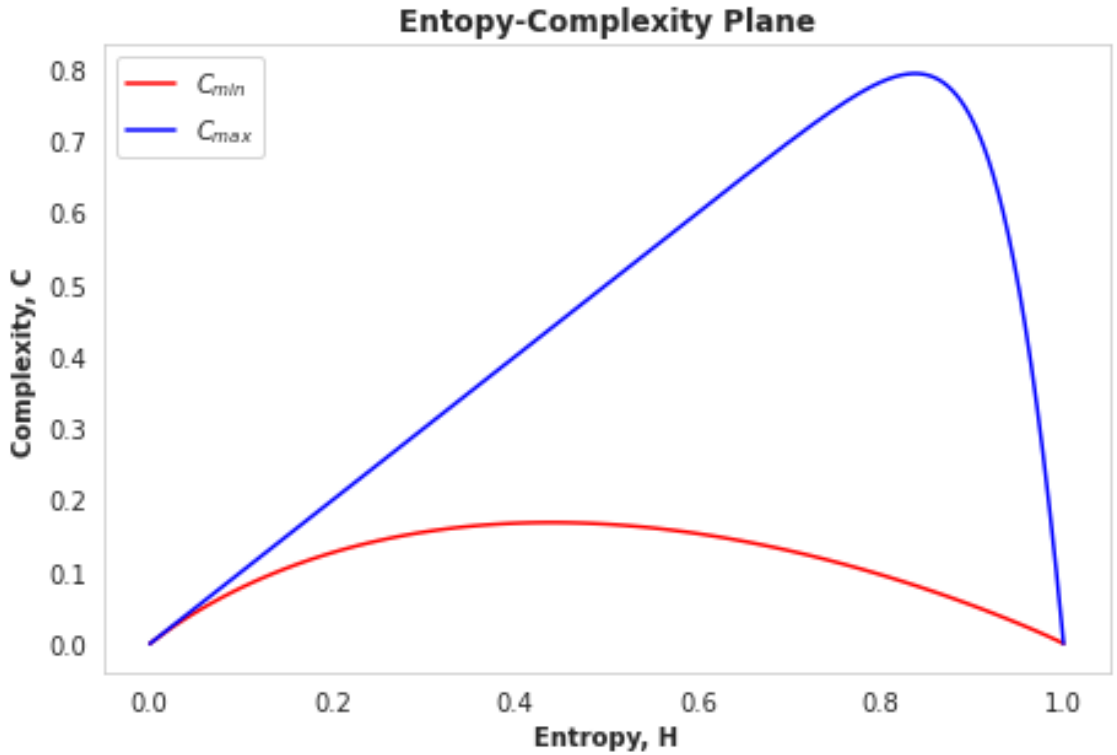
а в качестве "disequilibrium" возьмем дивергенцию Йенсена-Шеннона:

$$D(P) = Q_0 \cdot \left(H\left(\frac{P + P_e}{2}\right) - H\left(\frac{P}{2}\right) + H\left(\frac{P_e}{2}\right)\right), \tag{5}$$

где P_e – равновероятное распределение, а Q_0 – коэффициент нормировки.

Плоскость энтропия-сложность

Для каждого значения энтропии H можем определить C_{min} и C_{max} как минимальная и максимальная сложность какого-либо распределения, имеющего энтропию равную H, и таким естественным образом и получаем плоскость энтропия-сложность.



Содержание

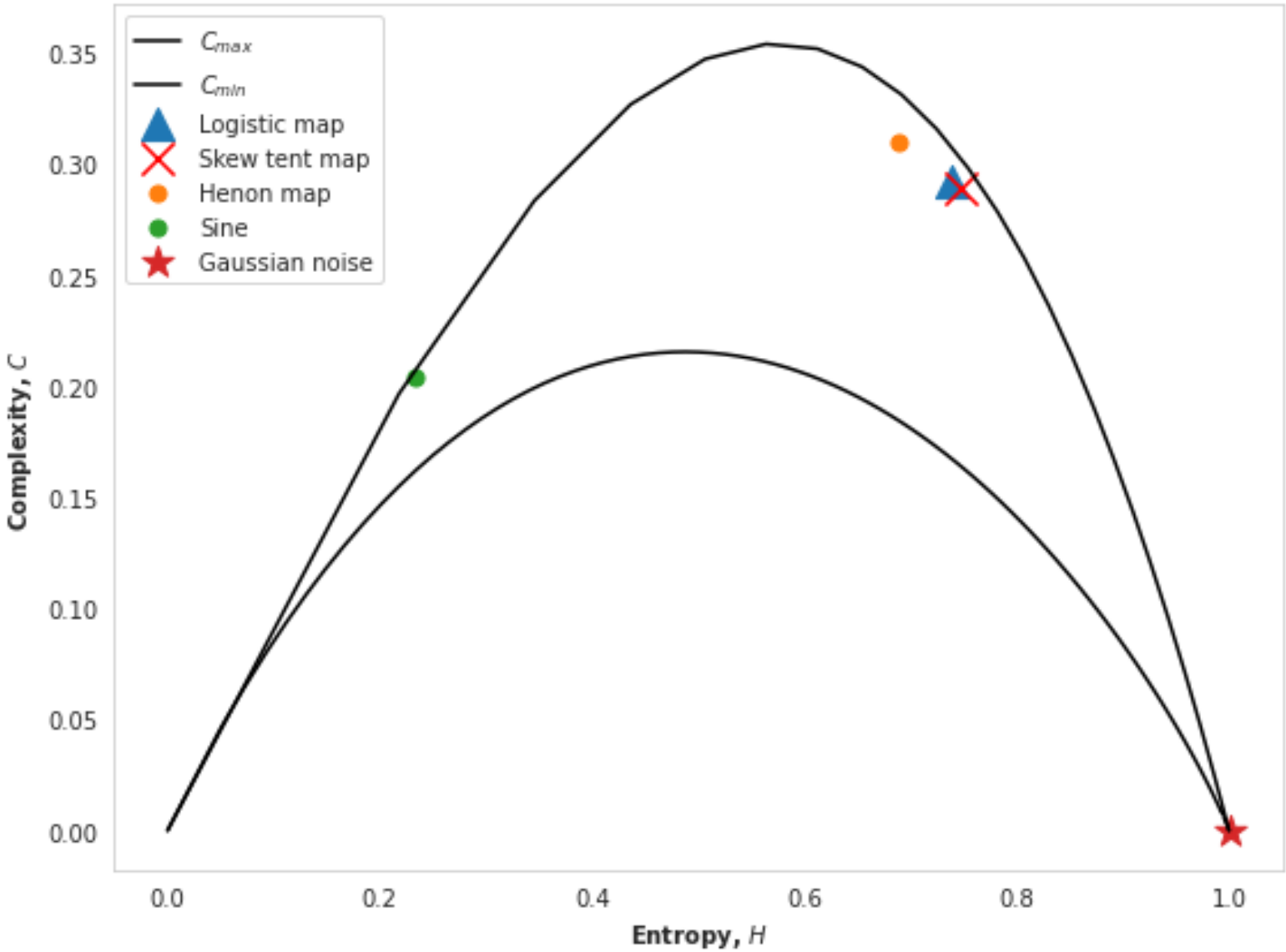
Построение казахского словаря

- Сбор данных (парсинг сайтов с литературными текстами, википедии)
- Нормализация текстовых данных
- Токенизация
- Удаление стоп-слов
- Приведение слов к их каноническим формам (леммам)
- Построение $TF - IDF$ матрицы и применение SVD (ранга r).

В итоге получаем словарь – отображение слов казахского языка в их векторное представление.

Содержание

Хаотичные vs. стохастические vs. детерминированные процессы



Построение вероятностного распределения по многомерному временному ряду

Используем модификацию алгоритма Bandt-Pompe для многомерных временных рядов. Согласно этому методу, для данного вектора $x \in \mathbb{R}^k$ мы строим бинарный вектор $y \in \{0,1\}^{k-1}$ такой, что $y_i = 1 \iff x_i \geq x_{i+1}$. Такой вектор будем называть порядковым паттерном. Чтобы обобщить этот концепт на многомерные временные ряды, мы сравниваем вектора в лексикографическом порядке покоординатно. И тогда уже строим вероятностное распределение на этих порядковых паттернах, в надежде на выявление важных свойств порядковой структуры данного временного ряда.

Построение вероятностного распределения по многомерному временному ряду

1. Фиксируется число $n \in \mathbb{N}$ и рассматриваются $N - n + 1$ векторов $s_i = (x_{i-(n-1)}, \ldots, x_{i-1}, x_i)$, где x_i – элементы многомерного временного ряда (т.е. векторы).
2. Для каждого s_i ставим ему в соответствие порядковый паттерн
3. Для каждого порядкового паттерна π ставим ему в соответствие число

$$\frac{1}{N-n+1} \cdot \left|\left\{ n \leq i \leq N \,\middle|\, \text{где } s_i \mapsto \pi \right|\right| \tag{6}$$

Описание разработанного алгоритма

Итоговый алгоритм выглядит следующим образом:

- Отображение данного текста в многомерный временной ряд (используя заранее построенные словарь казахского языка)
- Построение вероятностного распределения соответствующего данному временному ряду
- Отображение на плоскость энтропия-сложность:
  - Если находится ближе к зашумленной зоне, то классифицируем текст как сгенерированный ботом
  - Если находится ближе к хаотичной зоне, то классифицируем текст как человеческий

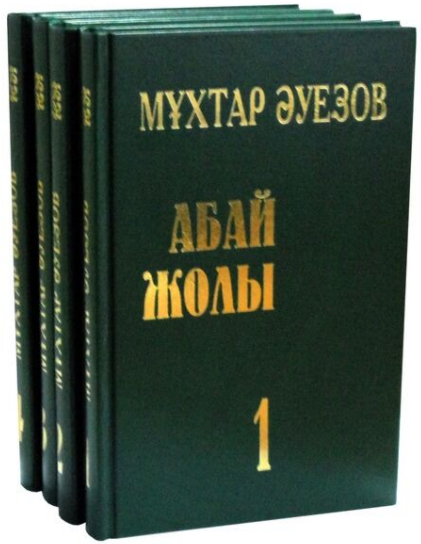
Содержание

Результаты экспериментов

Для проверки гипотезы о хаотичности человеческих текстов использовались известные произведения казахской литературы, такие как “Путь Абая” (3 тома), “Дикое яблоко” и “Улпан”. В качестве бота испольлзовался *char-based* генератор, использующий LSTM и обученный на литературных текстах.

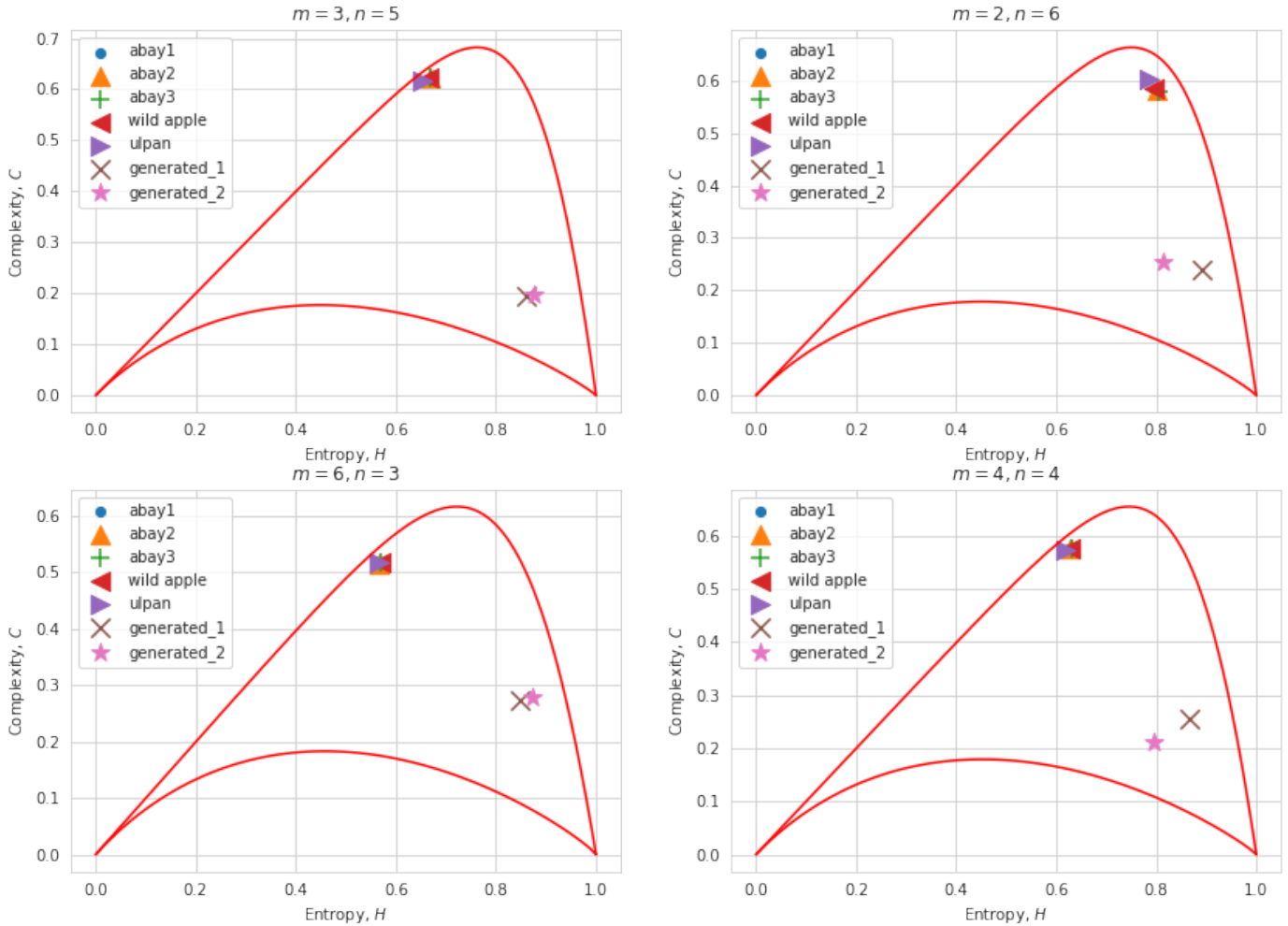
```
model.eval()

print(evaluate(
    model,
    char_to_idx,
    idx_to_char,
    temp=0.3,
    prediction_len=100,
    start_text='Отан үшін отқа түс күймейсің'
    )
)
```

```
Отан үшін отқа түс күймейсіңіз, ал бірақ бір кезде де болар еді.
```

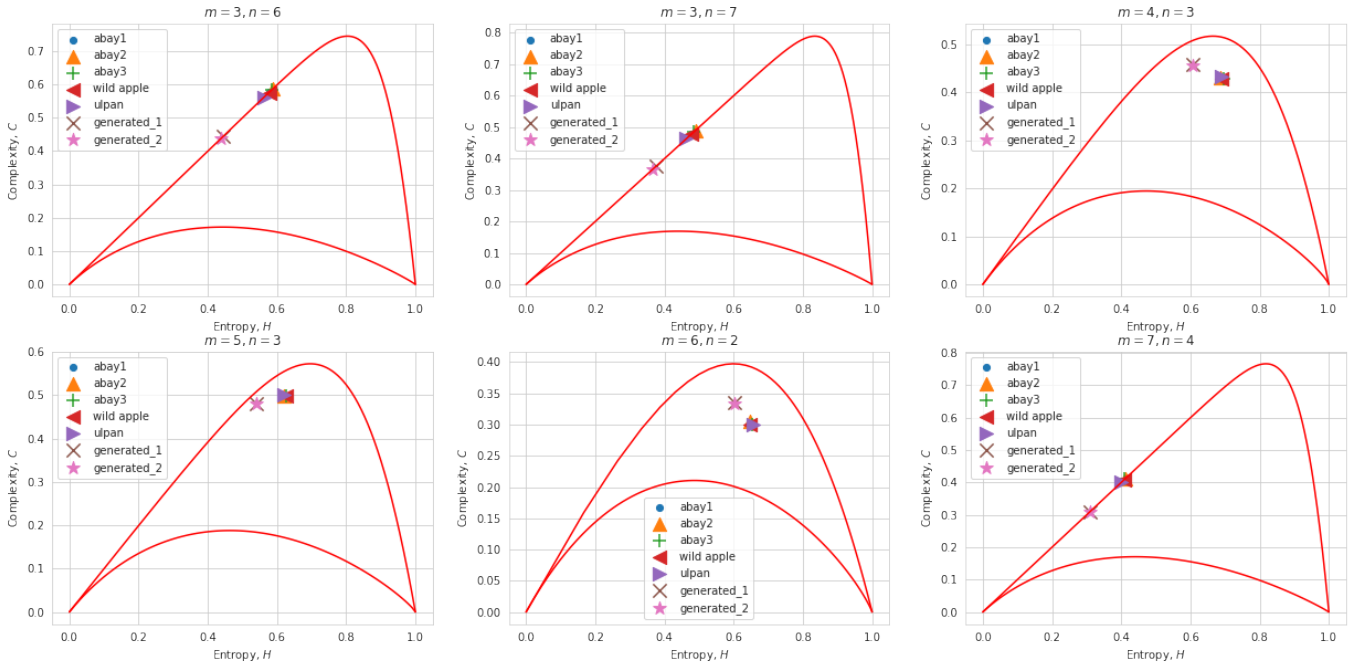


Результаты экспериментов

здесь m – это размерность векторов (ранг SVD приближения матрицы $TF - IDF$), а n – размерность эмбеддингов из метода Bandt-Pompe.



Результаты экспериментов



Таким образом, получаем, что для казахского языка в качестве параметров нашего алгоритма отлично подходят $(m, n) \in \{(3, 5), (2, 6), (6, 3), (4, 4)\}$.

Содержание

Список источников

[1] P.W. Anderson. *Physics today*, 1991.

[2] C. Bandt and B. Pompe. A natural complexity measure for time series. *Physical Review Letters*, 88, 2002.

[3] Jerome R. Bellegarda. *Latent Semantic Mapping: Principles and Applications.* Synthesis lectures on speech and audio processing. Morgan & Claypool, 2007.

[4] Carl Echart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrica,* 1:211-218, 1936

[5] O.A. Rosso, M.T. Martin, A. Plastino. Statistical complexity and disequilibrium. *Physics Letters A*, 311:126-132, 2003.

[6] M.T. Martin A. Plastino O.A. Rosso, H.A. Larrondo and M.A. Fuentes. Distinguishing noise from chaos. *Physical Review Letters*, 2007.

[7] C.E. Shannon and W. Weaver. The mathematical theory of communication. *University of Illinois Press, Urbana, IL*, 1949.