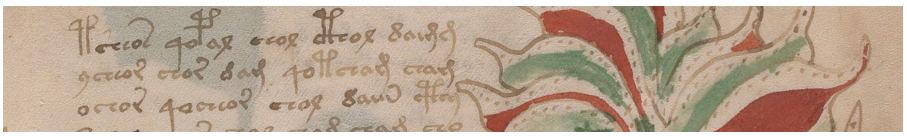
Odkrywanie ukrytej struktury danych z obserwacji

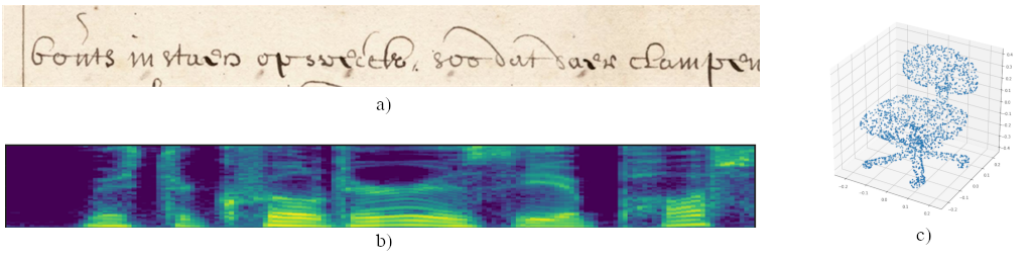
Wyobraźmy sobie, że dostaliśmy kopię pewnego tajemniczego dokumentu, takiego jak manuskrypt Wojnicza ([1]; Rys. 1). Zawiera on wiersze powtarzających się symboli, które przypominają tekst w obcym, a być może nawet nieistniejącym języku. Po przeanalizowaniu dokumentu możemy sformułować pierwsze hipotezy dotyczące rozróżnianych liter. Zauważamy też powtarzające się ciągi liter, układające się w całości, które nazwiemy słowami. Ale oczywiście nie możemy być pewni naszych decyzji. W dokumencie występuje wiele różnych charakterów pisma i nie jest oczywiste, jak utworzyć z nich wspólny alfabet. Czy da się stworzyć inteligentny system, który mógłby skorzystać z dostępności dużej ilości danych i automatycznie zinterpretować powtarzające się wzorce?



Rysunek 1: Fragment strony manuskryptu Wojnicza. Wyraźnie widać tekst, jest on jednak zapisany nieznanym nam alfabetem, a jego znaczenie jest wciąż zagadką.

Wszyscy wiemy, że istnieją metody automatycznego rozpoznawania pisma (OCR), które zamieniają zeskanowany obraz tekstu na ciąg znaków. Gdybyśmy chcieli z nich skorzystać, napotkamy na problem: metody te (jak i wiele innych współczesnych metod uczenia maszynowego) do prawidłowego działania wymagają ogromnej ilości opisanych danych. Aby je wykorzystać w tym przypadku, należałoby rozpocząć od mozolnego przepisywania wielu stron tekstu [2] – ale tu jest to niemożliwe: wszakże nadal nie potrafimy przeczytać tego alfabetu. Należy zatem szukać innej drogi.

Niniejszy projekt ma na celu pomóc w tym zadaniu (i wielu innych, w pewnym stopniu analogicznych). Autor wniosku wraz ze współpracownikami pracuje obecnie nad algorytmami, które analizując zeskanowane, odręczne manuskrypty, będą w stanie odkrywać strukturę tekstu i automatycznie konstruować hipotezy o znaczeniu znaków, słów, czy nawet o gramatyce badanego fragmentu. Pragniemy, aby tworzone przez nas algorytmy były w stanie wykorzystać naszą wiedzę o dziedzinie, w której się poruszamy, nawet jeżeli wiedza ta jest bardzo ograniczona. Poszukujemy sposobów, aby móc wprowadzać do algorytmów takie informacje, jak liczba liter w alfabecie, czy średnia szerokość pisanej litery. Chcemy również wymuszać pewne statystyczne właściwości odkrywanego języka, na przykład dopasowanie rozkładu słów do ciężkoogonowego rozkładu Zipfa.



Rysunek 2: Przykłady danych, na których pracujemy: a) pismo odręczne [3], b) mowa [4], c) chmury punktów [5].

W celu zapewnienia, że zastosowanie naszych algorytmów nie ograniczy się tylko do pisma odręcznego, przetestujemy ich działanie na innych typach danych, na których do tej pory pracowaliśmy ([3–5]; Rysunek 2). Spodziewamy się, że nasze badania przyniosą zarówno teoretyczne, jak i praktyczne rezultaty. Od strony teoretycznej chcemy zrozumieć, co jest potrzebne, aby maszyny umiały automatycznie wnioskować o strukturze danych: odkrywać podstawowe jednostki w nich obecne i relacje między nimi. Od strony praktycznej, będziemy starali się zapewnić, że tworzone przez nas reprezentacje danych pozwolą wykorzystać metody uczenia maszynowego bez konieczności żmudnego etykietowania dużych zestawów próbek danych.

1 Bibliografia

[1] "Voynich manuscript," Dec. 2019.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[3] J. Chorowski *et al.*, "Unsupervised Neural Segmentation and Clustering for Unit Discovery in Sequential Data," in *Workshop on Perception as Generative Reasoning, NeurIPS 2019*, Vancouver, Canada, Dec. 2019.

[4] J. Chorowski *et al.*, "Unsupervised Speech Representation Learning Using WaveNet Autoencoders," *IEEE/ACM TASLP*, vol. 27, no. 12, pp. 2041–2053, Dec. 2019.

[5] M. Stypułkowski *et al.*, "Conditional Invertible Flow for Point Cloud Generation," in *Published in Sets & Partitions Workshop at NeurIPS 2019*, Vancouver, Canada, Dec. 2019.