TIMKoD - Lab 1 - Przybliżenie języka naturalnego

28 lutego 2018

Opis pliku z zadaniami

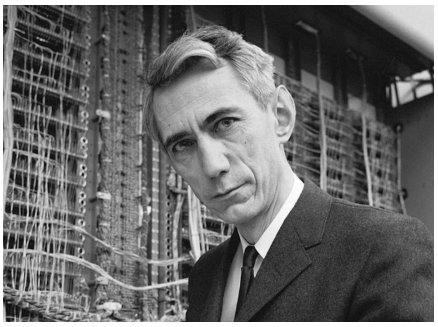
Wszystkie zadania na zajęciach będą przekazywane w postaci plików `.pdf`, sformatowanych podobnie do tego dokumentu. Zadania będą różnego rodzaju. Za każdym razem będą one odpowiednio oznaczone:

- Zadania do wykonania na zajęciach oznaczone są symbolem $\triangle$ – nie są one punktowane, ale należy je wykonać w czasie zajęć.
- Punktowane zadania do wykonania na zajęciach oznaczone są symbolem $\diamond$ – należy je wykonać na zajęciach i zaprezentować prowadzącemu, w wypadku nie wykonania zadania w czasie zajęć lub nieobecności, zadania staje się zadaniem do wykonania w domu ($\star$).
- Zadania do wykonania w domu oznaczone są symbolem $\star$ – są one punktowane, należy je dostarczyć w sposób podany przez prowadzącego i w wyznaczonym terminie (zwykle przed kolejnymi zajęciami).
- Zadania programistyczne można wykonywać w dowolnym języku programowania, używając jedynie biblioteki standardowej dostępnej dla tego języka.

Wprowadzenie

Z powodu braku pierwszego wykładu zamieszczamy słowo wstępu do przedmiotu:

Krótka historia



Rysunek 1: Claude Shannon (1916 - 2001)

Mało która dyscyplina ma swój początek w tak konkretnym miejscu i czasie jak teoria informacji, dlatego szkoda by było nie zawrzeć tutaj skróconej historii jej powstania.

W latach 20-tych ubiegłego wieku zaczęto zastanawiać się nad położeniem transatlantyckiego kabla telefonicznego (pierwszy telegram wysłano transatlantyckim kablem telegraficznym już w 1858, finalizacja kabla telefonicznego nastąpiła dopiero w 1956). Zetknięto się z serią praktycznych problemów dotyczących tego, jakie powinny być właściwości kabla, by był on w stanie poradzić sobie z dziennym obciążeniem oraz jak zakodować tajne rządowe połączenia. Informacja i jej własności były w tym okresie bliżej niezdefiniowane.

Wiadome było jednak to, że za pomocą sygnału elektrycznego można przesłać wiadomość – telegraf, dźwięk, obraz – oraz, że ten skomplikowany sygnał można rozbić na sumę funkcji okresowych o różnych częstotliwościach. Sieć komunikacyjna mogła przenosić sygnały tylko o określonym zakresie tych częstotliwości, definiowanym jako szerokość pasma. (ówczesne sieci telefoniczne działy w zakresie od 200 Hz do 3200 Hz). W celu przesłania bardziej złożonej (ciekawszej) wiadomości należało wygenerować bardziej złożone funkcje.

Harry Nyquist, pracujący dla Bell Labs (założone przez AT&T w 1925 roku; pracownikom tego laboratorium przypisuje się wynalezienie m.in. tranzystora, systemu Unix, języka C i właśnie teorii informacji), pokazał w 1928 roku, że przepustowość kanału komunikacyjnego ogranicza ilość "inteligencji" (jak sam to nazwał), która może zostać przesłana z daną szybkością. Implikacją tego było stwierdzenie, że sygnał ciągły może być zastąpiony przez sygnał dyskretny w ramach danej szerokości pasma i nikt nie będzie w stanie wskazać różnicy.

W tym samym roku Ralph Hartley również pracujący dla Bell Labs, rozwija idee wprowadzone przez Nyquista. Proponuje on nowy sposób myślenia o informacji w ramach fizycznych systemów ich przesyłania, który zakłada zi-

gnorowanie kwestii interpretacji wiadomości, a skupia się na ilości możliwych wiadomości, które będzie mógł rozróżnić odbiorca. Ostatecznie przedstawia on wzór na ilość przesłanej informacji w systemach dyskretnych:

$$H = \log(m^n),$$

gdzie H to ilość informacji, m jest ilością "liter" w alfabecie – ilością dyskretnych wartości, którą posługuje się system (np. jeśli rozróżniamy tylko stany "on"/"off" to $m = 2$), natomiast n to długość wiadomości.

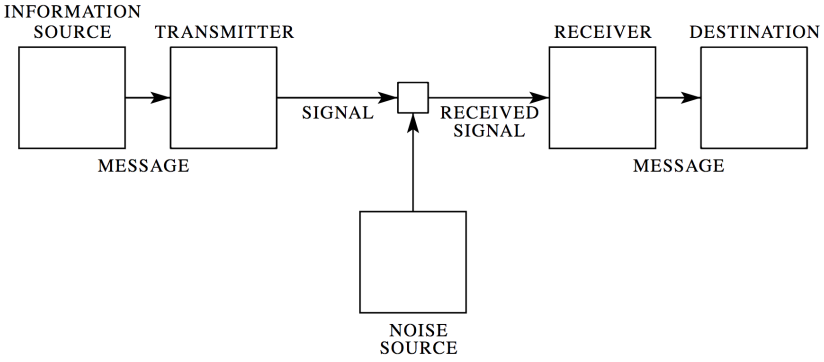
Jeśli będziemy przesyłać losowe litery, ilość możliwych wiadomości rośnie wykładniczo względem długości naszej wiadomości. Zarówno Nyquistowi, jak i Hartleyowi bardziej intuicyjne wydawało się, że ilość informacji powinna rosnąć liniowo z długością komunikatu, tak by 32-literowy telegram zawierał 2 razy więcej informacji od 16-literowego, stąd zastosowany logarytm we wzorze Hartleya.

Prace Nyquista i Hartleya miały silny wpływ na innego pracownika Bell Labs, Clauda Shannona, który idzie o krok dalej i jednoznacznie stwierdza, że znaczenie wiadomości może być zignorowane, a istotną rzeczą określającą wiadomość jest to, że została ona wybrana ze zbioru możliwych wiadomości. Proponuje on też uniwersalny schemat systemu komunikacji, składający się z:

- źródła informacji, które produkuje wiadomość,
- transmitera, który koduje wiadomość do formatu, który można wysłać jako sygnał,
- kanału, który jest medium, przez które przekazywany jest sygnał,
- źródła szumu, które reprezentuje zniekształcenia jakim ulega sygnał w drodzę do odbiornika,
- odbiornika, który dekoduje wiadomość,
- miejsce przeznaczenia, które jest adresatem wiadomości.

W 1948, Shannon publikuje pracę "A Mathematical Theory of Communication"[1] uznawaną za początek teorii informacji.

[1] `http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf`



Rysunek 2: Diagram ogólnego systemu komunikacji zaproponowany przez Shannona

Cel zajęć

Na tych zajęciach chcemy się skupić na dyskretnym źródle informacji. Możemy myśleć o nim jako generatorze wiadomości, który podaje nam symbol po symbolu. Źródło wybiera kolejny symbol z jakimś prawdopodobieństwem, które jest najczęściej zależne od poprzednio podanych symboli, poza najprostszymi przypadkami.

Możemy więc traktować dyskretne źródło informacji jako proces stochastyczny. I odwrotnie, każdy proces stochastyczny, który generuje dyskretną sekwencję zdarzeń ze skończonego zbioru, może być traktowany jako źródło informacji.

Języki naturalne (np. angielski) możemy traktować jako dyskretne źródło informacji (oczywiście nie tylko języki są źródłem informacji; inne źródła to na przykład skwantyzowany dźwięk, obraz czy video).

By zobrazować tą idee, poniższe ćwiczenia skupiają się na tworzeniu sztucznych źródeł informacji, które wraz ze wzrostem skomplikowania będą generować wiadomości coraz bardziej zbliżone do języka angielskiego. Przy okazji pokażemy, jak wiedza statystyczna na temat źródła pozwala na zastosowanie odpowiedniego kodowania, które pozwala zredukować czas nadawania wiadomości lub rozmiar kodu.

Przygotowanie do zajęć

- Do wykonania zadań potrzebne będą korpusy tekstowe, które można pobrać z
  `http://www.cs.put.poznan.pl/kjasinska/lectures/timkod/data/lab1`
- Pliki są znormalizowane, zawierają jedynie 26 małych liter alfabetu łacińskiego, cyfry i spacje.
- Przygotuj funkcję do wczytywania pliku do pamięci.

1 Przybliżenie zerowego rzędu △

Treść

Zadanie polega na wygenerowaniu tak zwanego przybliżenia zerowego rzędu dla języka angielskiego. Czyli ciągu losowych znaków (26 liter alfabetu + spacja), w którym symbole mają jednakową szansę na wystąpienie (1/27) niezależnie od poprzednio wygenerowanych symboli.

Wskazówka: kolejne zadania również dotyczą generowania innych przybliżeń, zaprojektowanie kodu w oparciu o generator, który przyjmuje źródło informacji jako argument, może zaoszczędzić pisania sporej ilości podobnego kodu.

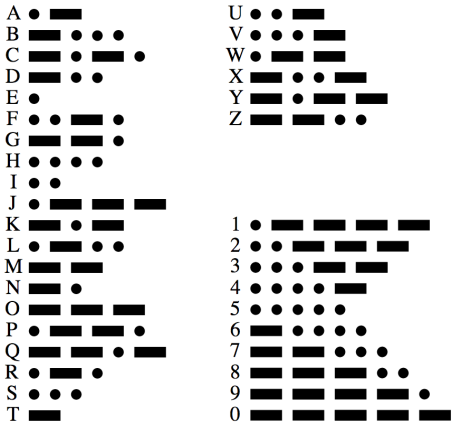
Jaka jest średnia długość słowa w tym przybliżeniu?

2 Częstość liter △

Treść

Czy różne znaki występują jednakowo często w prawdziwym tekście? Oblicz częstość wystąpienia poszczególnych znaków w angielskim tekście. Wskazówka: nie musisz zliczać wszystkich znaków.

Jakie znaki są najbardziej prawdopodobne, a które najmniej? Zwróć uwagę na kody przypisane poszczególnym literą w kodzie Morse'a, czy widzisz jakąś zależność, z czego wynika?



Rysunek 3: Międzynarodowy kod Morse'a

3 Przybliżenie pierwszego rzędu △

Treść

Używając wyliczonych prawdopodobieństw w poprzednim zadaniu, wygeneruj nowy ciąg znaków – będzie to przybliżenie pierwszego rzędu.

Jaka jest średnia długość słowa w tym przybliżeniu? Czy jest ona zbliżona do faktycznej średniej tego korpusu?

4 Prawdopodobieństwo warunkowe liter △

Treść

Czy różne znaki występują jednakowo często po sobie? Oblicz prawdopodobieństwo wystąpienia poszczególnych znaków po każdym z 2. najczęściej występujących znaków w korpusie.

Przykład

Rozważmy następujący korpus: "beekeepers keep bees in a beehive".

W tym zdaniu najczęstszą literą jest litera "e". W przybliżeniu pierwszego rzędu wyliczonym na podstawie tego korpusu "e" miałoby prawdopodobieństwo pojawienia się jako kolejny znak wynoszące aż 12/32. Jednak żadne ze słów w tym korpusie nie zaczyna się od litery "e" – po spacji, których jest 5, nigdy nie występuje litera "e".

To pokazuje, że bardziej skomplikowaną strukturę możemy otrzymać, jeśli zamiast wybierać kolejne symbole niezależnie, prawdopodobieństwo wyboru kolejnej litery będzie zależne od poprzedzających ją liter. Możemy stworzyć tak zwane źródło Markova pierwszego rzędu, w którym kolejne znaki są generowane z następującym prawdopodobieństwem:

$$P(j|i) = P(i,j)/P(i),$$

gdzie $P(i)$ jest prawdopodobieństwem litery i w tekście, $P(i,j)$ jest prawdopodobieństwem wystąpienia bigramu (pary występujących po sobie liter) (i,j), a $P(j|i)$ jest prawdopodobieństwem warunkowym wystąpienia litery j zaraz po literze i.

Z racji, że Shanon w 1948 nie mógł łatwo użyć komputera do policzenia tych liczności, zaproponował on więc rodzaj metody Monte Carlo by przybliżyć źródło Markova. Brał książkę, którą otwierał na losowej stronie, wybierał losową linijkę i losową pozycję w linii i przesuwał się w prawo tak długi, aż nie znalazł pierwszej litery (i) i brał następującą po niej (j). Proces powtarzał od początku by uzyskać kolejny symbol.

5 Przybliżenia na podstawie źródła Markova *10pt* ◇

Treść

Wygeneruj przybliżenie języka angielskiego na podstawie źródła Markova pierwszego rzędu (źródła, gdzie prawdopodobieństwo następnego symbolu zależy od 1. poprzedniego).

Następnie zrób to samo dla źródła Markova trzeciego rzędu (źródła, gdzie prawdopodobieństwo następnego symbolu zależy od 3. poprzednich).

Na koniec wygeneruj przybliżenie źródła Markova piątego rzedu. Zaczynij od ciągu znaków zawierającego już słowo 'probability'.

Jaka jest średnia długość wyrazu w tych przybliżeniach?