TIMKoD – Lab 3 – Entropie warunkowe języków naturalnych

28 marca 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.

Cel zajęć

Do tej pory zajmowaliśmy się przybliżaniem języka naturalnego stosując coraz to bardziej złożone źródła informacji. Na dzisiejszych zajęciach zajmujemy się entoropią warunkową na przykładzie języka naturalnego.

Przygotowanie do zajęć

- Do wykonania zadań potrzebne będą korpusy tekstowe, które można pobrać z
  http://www.cs.put.poznan.pl/kjasinska/lectures/timkod/data/lab3
- Pliki są znormalizowane, zawierają jedynie 26 małych liter alfabetu łacińskiego, cyfry i spacje.
- Przygotuj funkcję do wczytywania pliku do pamięci (skopiuj z poprzednich zajęć).

1 Entropia – powtórka △

Treść

Wzór na entropię wyrażoną w bitach (podstawa logarytmu = 2) zmiennej losowej X o wartościach ze zbioru $\{x_1, \ldots, x_n\}$:

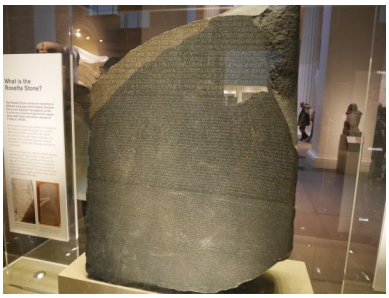
$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \,,$$

gdzie $p(x)$ to prawdopodobieństwo zajścia zdarzenia x.

Jaka będzie entropia przybliżenia zerowego rzędu dla przybliżenia języka angielskiego generowanego na poziomie znaków? Przypomnienie: 26 liter + cyfry + spacja, mające jednakowe prawdopodobieństwo wystąpienia.

Policz entropię występowania znaków w języku angielskim na podstawie kodu z pierwszych zajęć.

Skąd wiemy, że hieroglify to pismo i jak szukamy sygnałów od obcych cywilizacji?



Rysunek 1: Kamień z Rosetty (Muzeum Brytyjskie)

Odnalezienie kamienia z Rosetty miało przełomowe znaczenie dla odczytania egipskich hieroglifów. Miało to miejsce w 1799 r. podczas wyprawy Napoleona do Egiptu.

Kamienna tablica zawiera ten sam tekst zapisany na trzy sposoby: pismem hieroglificznym, demotycznym (pismo codziennego użytku starożytnych egipcjan) oraz po grecku. Znalezisko pozwoliło już w 1822 rozszyfrować większą część egipskiego pisama hieroglificznego.

Odnalezienie podobnych tłumaczeń pozwoliło odczytać część innych starożytnych języków. Język Mezopotamski został odczytany w 1857, a język Majów w 1952. Istnieją jednak inne starożytne języki, które pozostają dla nas niezrozumiałe do teraz.

Rysunek 2: Tabliczna z czasów cywilizacji Indusu

Głównym pytaniem, które było stawiane w przypadku nieodczytanych do dzisiaj starożytnych skryptów, jest: czy odnalezione piktogramy są w ogóle rodzajem pisma, które niesie ze sobą jakieś znaczenie?

Przykładem może być badanie nad pismem cywilizacji doliny Indusu (znaną też jako kultura Mohendżo-Daro), która istniała między 3300 a 1300 r. p.n.e. Główną pozostałością po niej są małe gliniane tabliczki zawierające po kilka symboli. Archeolodzy długo spierali się, czy owe symbole są pismem. Odpowiedź uzyskali dopiero dzięki teorii informacji, a dokładnie entropii warunkowej, której użycie do badania języka zaproponował Shannon w 1950 r.

Entropia

Entropia mierzy ilość informacji. Można ją interpretować jako niepewność wystąpienia danego zdarzenia elementarnego. Jeśli zdarzenie następuje z prawdopodobieństwem 1 to entropia wynosi 0 ponieważ nie ma niepewności. Entropia osiąga wartość najwyższą gdy wszystkie zdarzenia elementarne mają równe prawdopodobieństwa.

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \,,$$

Entropia warunkowa

Entropia warunkowa mierzy ile wynosi entropia zmiennej losowej Y, jeśli znamy wartość innej zmiennej losowej X. Często oznacza się ją jako $H(Y|X)$. Wzór na entropię warunkową to:

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x, y) \log p(y|x) \,,$$

gdzie $p(x, y)$ to prawdopodobieństwo łączne zdarzenia x i y, a $p(y|x)$ to prawdopodobieństwo zdarzania y pod warunkiem zdarzenia x. Intuicyjnie entropia warunkowa mówi nam ile dodatkowej informacji dostajemy na podstawie zmiennej Y, jeśli znamy zmienną X.

Przekładając powyższą definicję na analizę języka, $p(x, y)$ jest prawdopodobieństwo łącznym wystąpienia kolejno n-gramu x oraz znaku/słowa y, a $p(y|x)$ jest prawdopobieństwem wystąpienia y po x. W kolejnym zadaniu będziemy mówić o entropii warunkowej rzędu n, gdzie rząd n oznacza długość n-gramu x.

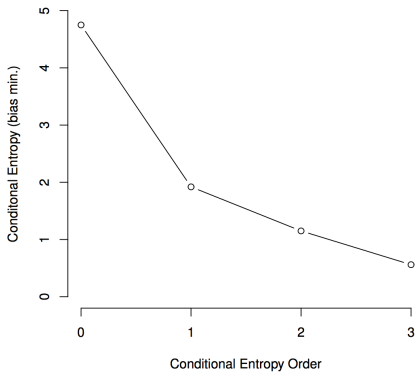
Język delfinów i poszukiwanie cywilizacji pozaziemskich

Badając entropię warunkową starożytnego języka cywilizacji Indusu, jak i wszystkich znanych nam języków, odkryjemy, że entropia warunkowa różnych języków jest zbliżona do siebie i, co najważniejsze, maleje znacząco z każdym rzędem.

Jest to efekt spodziewany. Wyobraźmy sobie, że naszym zadaniem jest zgadnąć słowo wybrane z książki. W takiej sytuacji możemy tylko próbować strzelać (ewentualnie wskazać słowo z największym prawdopodobieństwem). Jeśli jednak zamiast tego zostanie podane słowo, a zadaniem będzie zgadnąć słowo występujące po nim, możemy już znacznie zawęzić zbiór możliwych słów. Zadanie stanie staje się jeszcze łatwiejsze, jeśli zostaną nam podana dwa słowa. Im mamy dłuższą sekwencję, tym zbiór możliwych słów maleje lub pewne słowa stają się znacznie bardziej prawdopodobne. Wzrasta przewidywalność, czyli jak już wiemy, entropia musi maleć.

Podobną analizę przeprowadzono na odgłosach wydawanych przez wiele gatunków zwierzęta, by ocenić złożoność ich komunikacji. Okazało się, że podobny wzorzec możemy dostrzec u gatunków zwierząt oznaczających się wysoką inteligencją, jak np. delfiny, które, jak możemy wnioskować na podstawie teorii informacji, nie wydają z siebie losowych pisków, ale posługują się logicznym językiem.

Badania nad językami delfinów zainteresowały astronoma i astrofizyka Franka Draka. Zainicjował on w 1959 r. trwający do dzisiaj projekt SETI (Search for Extraterrestrial Intelligence), mający na celu znalezienie sposobu kontaktu z pozaziemskimi cywilizacjami poprzez poszukiwanie sygnałów radiowych i świetlnych pochodzących z przestrzeni kosmicznej. Dzieje się to, upraszczając, przez podoną analizę entropii odbieranych sygnałów i szukanie w nich wzorców, którymi oznaczają się języki naturalne oraz np. wspomniane języki delfinów.



Rysunek 3: Entropia warunkowa języka butlonosów

Źrodła

- https://en.wikipedia.org/wiki/Entropy_(information_theory)
- https://en.wikipedia.org/wiki/Conditional_entropy
- http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf
- www.mdpi.com/1099-4300/16/1/526/pdf
- https://en.wikipedia.org/wiki/Rosetta_Stone
- https://en.wikipedia.org/wiki/Indus_Valley_Civilisation
- https://pl.wikipedia.org/wiki/Search_for_Extraterrestial_Inteligence

2 Rozpoznawanie czy dany tekst to język *10pt*◇

Treść

Wylicz entropie znaków i słów oraz ich entropie warunkowe kolejnych rzędów dla próbki języka angielskiego (plik norm_wiki_en.txt). ***(2pt)***

Następnie wylicz entropie znaków i słów oraz ich entropie warunkowe kolejnych rzędów dla próbki języka łacińskiego (plik norm_wiki_lo.txt). ***(2pt)***

Możesz również dokonać analizy dla próbek innych języków:

- esperanto (plik norm_wiki_eo.txt),
- estońskiego (plik norm_wiki_et.txt),
- somalijski (plik norm_wiki_so.txt),
- haitański (plik norm_wiki_ht.txt),
- navaho (plik norm_wiki_nv.txt),

Korzystając z zaobserwowanych wartości entropii warunkowej odpowiedź na pytanie, czy następujące pliki zawierają język naturalny: ***(6pt)*** (po ***(1pt)*** za każdy dobrze rozpoznany plik)

- sample0.txt,
- sample1.txt,
- sample2.txt,
- sample3.txt,
- sample4.txt,
- sample5.txt.

Uwagi do zadania

- niektóre z podanych języków posiadają dodatkowe litery, które w unormowanych próbkach zostały sprowadzone do ich najbliższych odpowiedników dostępnych w alfabecie łacińskim, tak by pliki zawierały jedynie 26 małych liter alfabetu łacińskiego, cyfry i spacje,
- w celu ułatwienia sprawdzenia poprawności zadania entropie wyrażaj w bitach,
- podczas sprawdzania zadania będą sprawdzane wyniki, poprawność odpowiedzi oraz jej uzasadnienie w oparciu o otrzymane wyniki,
- zadanie może być podchwytliwe i podchwytliwych pytań można się spodziewać od prowadzących przy sprawdzaniu.