TI – zad 1

Kamil Kowalczyk, 136742, L1, czwartek 13:30

Zad 1

Średnia długość słowa dla pliku posiadającego 10 milionów znaków wyniosła: **27,03**
Ilość słów w pliku obliczono zliczając ciągi znaków oddzielone jednym, lub więcej niż jednym znakiem spacji.

```
ckb oi
ffthszuwbzhcecqwrcggciaxajghmyiqnlpboaihvxrjqvegeykuyhlytjiqxkdoziaua
vvpfoeleuzrnemturxftjsxjxqdueuuwderigq
vhhjqsfoywvfprfymobpjzqgmvezzfirokpmf
```

Zad 2

Poniżej przedstawione są symbole, oraz ich prawdopodobieństwa wyliczone na postawie pliku *norm_wiki_sample.txt*, posortowane malejąco po wartości prawdopodobieństwa.

```
' ': 0.17059199786151394
'e': 0.0935363350304724
'a': 0.07209938398958711
't': 0.06629621943432631
'i': 0.06095500939341498
'n': 0.05965627210307295
'o': 0.05811617655523373
'r': 0.05432303318740922
's': 0.053081113336332086
'h': 0.03646613694522938
'l': 0.03505543315140939
'd': 0.03160977523187864
'c': 0.027571009981424498
'm': 0.02152852629372985
'u': 0.021310247224449554
'f': 0.017617762484751748
'p': 0.017076930905452165
'g': 0.016282506318275353
'b': 0.013455630167965513
```

```
'w': 0.012853532149262843
'y': 0.012442741136502646
'v': 0.008546343890470808
'k': 0.006031361187349157
'1': 0.005869806869830876
'0': 0.004674786895210568
'9': 0.0035601269855864443
'2': 0.003480693795619051
'j': 0.002127734316092747
'8': 0.0019228022472270448
'3': 0.0017645846798124117
'5': 0.0016506717387739908
'x': 0.001634080675758631
'4': 0.0016072939874265694
'7': 0.001531475610071461
'6': 0.001527860797459176
'z': 0.0012914149776145777
'q': 0.0008531884640021667
```

Tak jak zostało wspomniane to na zajęciach laboratoryjnych, widać silną zależność pomiędzy częstością występowania znaków w języku angielskim, a kodem Morse'a. Najczęściej używanym znakom przypisane zostały najkrótsze sekwencje sygnałów. Dzięki temu średnia długośc wiadomości Morse'a (średnia ilość sygnałów na jedną wiadomość) została zminimalizowana.

Zad 3

```
r o les loppne0hdceny g ni tp   nt eai itc   y   eddt revnvns oqwyns
rikeshsi u d ie clio 8gio
```

Średnia długość słowa dla przybliżenia pierwszego rzędu (10 mln znaków): **5.87**
Średnia długość słowa dla pliku *norm_wiki_sample.txt*: **4.86**

Różnica w średnich długościach słów wynika z faktu wstawiania kilku znaków spacji koło siebie w generatorze pierwszego rzędu. W korpusie występuje zawsze jedna spacja pomiędzy wyrazami. W przypadku przybliżenia pierwszego rzędu, znaki spacji zamiast stać obok siebie mogłby by rozdzielać inne słowa, powodując tym samym zmniejszenie ich długości. Jednak generator ze względu na swoją losowość i fakt, że spacja jest najczęściej występującym znakiem, w większości przypadków generuje ciągi stojących koło siebie spacji.

Zad 4

Z tych prawdopodobieństw można wywnioskować, że litera „t” jest najczęściej występującym znakiem na początku wyrazu oraz, że znak „e” najczęściej stawiany jest na końcu słowa.

Prawdopodobieństwa dla znaku spacji	Prawdopodobieństwa dla znaku „e”
' t': 0.12933943165629452	'e ': 0.3078863765634321
' a': 0.11310802572770749	'er': 0.14449868107868144
' s': 0.07196118893391273	'en': 0.0873540119584842
' i': 0.061578718026455766	'ed': 0.08666928270895143
' o': 0.06128423379222887	'es': 0.08210706351235386
' c': 0.05381237550977829	'ea': 0.04793798394304955
' w': 0.04862358503585427	'el': 0.03874219894208836
' b': 0.04601180327583827	'ec': 0.02975450821377822
' f': 0.04183143113904546	'et': 0.026254560732808935
' p': 0.04123485606675555	'em': 0.02234833395761615
' m': 0.03796238643068808	'ee': 0.021248407087889113
' h': 0.03732669168152671	'ev': 0.01480144833613765
' r': 0.03174127115043363	'ep': 0.011132052661723933
' d': 0.03173746784851557	'ei': 0.010917021913317835
' l': 0.026836098333827763	'ex': 0.010108427025302283
' 1': 0.023873869468504855	'eg': 0.010039062267751928
' e': 0.023580471891968838	'ef': 0.009343432842032665
' n': 0.02132511385455956	'ew': 0.009250285881893618
' g': 0.018139033504916583	'ey': 0.008770678129688314
' 2': 0.013206150917193423	'eo': 0.005252893996777512
' u': 0.012151006299354634	'eb': 0.004097475321010189
' v': 0.00927951335121972	'eu': 0.003297798758965395
' j': 0.00923115708397582	'eh': 0.002186971713051871
' k': 0.00824990518911647	'eq': 0.0020492331230590253
' y': 0.005034485081819891	'ek': 0.0020244599953624706
' 3': 0.0038962111506292293	'ez': 0.0010047980593722687
' 4': 0.002736204065621085	'ej': 0.00042312502105715857
' 5': 0.0024313965833308885	'e1': 0.00024178572631837632
' 0': 0.002389016933386797	'e5': 8.323770906042464e-05
' 6': 0.0020755161895696073	'e2': 4.95462553931099e-05
' q': 0.0017734253515065965	'e3': 3.963700431448792e-05
' 8': 0.0016305298651566472	'e0': 3.2700528559452535e-05
' 7': 0.0015647870748587618	'e4': 2.972775323586594e-05
' 9': 0.0014642712384529038	'e6': 7.927400862897584e-06
' z': 0.0010051583640585796	'e8': 6.936475755035386e-06
' x': 0.0005732119319361089	'e7': 5.945550647173188e-06
	'e9': 1.981850215724396e-06

Zad 5

Przykładowy ciąg znaków na podstawie źródła Markova pierwszego rzędu:

```
nis t m jalecols beafof s rk 13 af trioradduinde alilig ame otathoche
waifomialsen ablid
```

Średnia długość słowa (10 mln znaków): **4,86**

Przykładowy ciąg znaków na podstawie źródła Markova trzeciego rzędu:

```
votectury threement tead clainly it in willeged atting to faults witch
receiventifying ram
```

Średnia długość słowa (10 mln znaków): **4,86**

Przykładowy ciąg znaków na podstawie źródła Markova piątego rzędu:

```
account of certain and lette the known the introduce a major copenhagen
day concertain standings
```

Średnia długość słowa (10 mln znaków): **4,86**

Przekazując źródła Markowa na wejściu algorytmu generującego ciągi znaków, rozwiązany został problem wielokrotnych, leżących koło siebie znaków spacji, gdyż w korpusie (w pliku norm_wiki_sample.txt) takowe nie występowały. Prawdopodobieństwo wystąpienia znaku spacji po spacji wynosiło w źródłach Markova 0. Dzięki temu, można zauważyć, że średnia długość słowa dla każdego przybliżenia Markova jest bliska średniej korpusu.
W pierwszym ciągu widać, że słowa są podobne do tych z języka angielskiego tylko pod względem długości. Zdecydowana większość słów to zlepki losowych znaków. Jedynie liczby prezentują się zazwyczaj w znany nam sposób. Wynika to z tego, że prawdopodobieństwo wystąpienia liczby po liczbie, jest zdecydowanie większe od tego, że po liczbie pojawi się litera.
Dla źródła Markova 3-ciego rzędu można już znaleźć wiele znanych słów. Należą do nich jednak zazwyczaj krótkie ciągi. Wynika to z faktu prawdopodobieństwa warunkowego, które jest obliczane na podstawie tylko trzech poprzednich znaków. Dla dłuższych wyrazów zauważyć można tendencję tworzenia podobnych słów do tych z języka angielskiego. Często są to połączenia różnych słów, takie jak na przykład threement, które zawiera słowo „three", ale także dołączoną do niego końcówkę „ment", która jest znanym zakończeniem wielu rzeczowników („environment", „government").
W ostatnim przykładzie, większość wygenerowanych wyrazów znajdziemy w słowniku języka angielskiego. Prawdopodobieństwo warunkowe, wyliczane na podstawie pięciu poprzednich liter, generuje zadowalające wyniki. Jednak, co nie było zaskoczeniem, ciężko znaleźć sens w połączeniach tych słów.