TIMKoD – Lab 2 – Przybliżanie języka naturalnego – kontynuacja

Daniil Martsich, 136766, L1

27 listopada 2020

Spis treści

1 Wstęp

Wszystkie wygenerowane ciągi znaków w tym sprawozdaniu mają długość nie mniejszą, niż 200 znaków.

Podobnie, jak w poprzednim sprawozdaniu, wszystkie metody, opisane w tym sprawozdaniu należą do klasy *Generator*. Kunstruktor oraz metody tej klasy zostały dostosowane do nowych potrzeb.

2 Kod klasy Generator

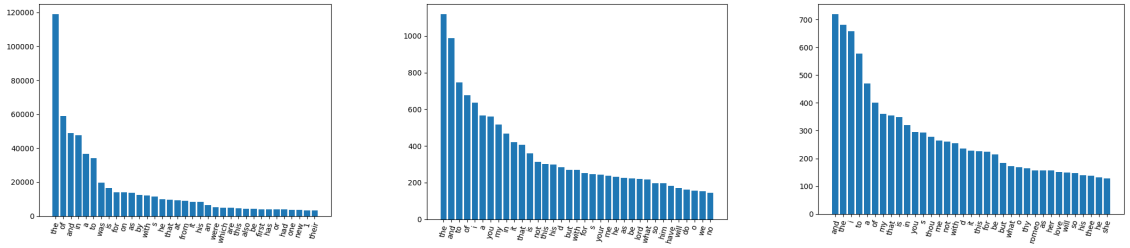
Kod został umieszczony w pliku, dołączonym do sprawozdania.

3 Częstość słów

Z plików wejściowych została policzona częstość występowania słów w tekscie. Poniżej są przedstawione wykresy, pokazujące top-40 słów w tych tekstach.

Z tych wykresów widać, że w tych kopusach danych dość często występują krótkie słowa:

the, a, to, for i t.d.



(a) *norm_wiki_sample.txt* (b) *norm_hamlet.txt* (c) *norm_romeo.txt*

Rysunek 1: Częstość występowania słów w tekstach

Nie ma w tym nic dziwnego, ponieważ z gramatyki języka angielskiego wynika, że są one bardzo często wykorzystywane.

Ponieważ ostatnie dwa korpusy są utworami, nie artykułami z wikipedii, w zbiór najczęściej wykorzystywanych słów wchodzą również zaimki:

i, you, my, he, she i t.d.

Ponadto, z tych wykresów widać, że słowa mają rozkład Pareto. Im większy jest korpus danych, tym wyraźniej to widać.

4 Przybliżenie pierwszego rzędu

Listing 1: Wywołanie metody

```
1 generator.basic_approximation(length=200)
```

Na podstawie przedstawionych korpusów danych wygenerowane zostało przybliżenie języka naturalnego pierwszego rzędu.

- *norm_wiki_sample.txt*

  noted maimonides daughter were label 50207 its and mexico than developed defeated in episode band cities the of members season twins in 19591968 darren which departed doctorate center of also of forest major s over ireland study following

- *norm_hamlet.txt*

  s nothing my the delight it i the received what this a can would follow tis to beast man and love sense forgiveness speak upon a in me which being own cov christian smiles villanous thus i further hill sultry hamlet by admiration first as of before

- *norm_romeo.txt*

  speeding both that banish unsubstantial died eye sadness room the smoke and heavens an yielding me supported rest say do i his and minute thee still pestilent may well and her they slewest a romeo dead you then waverer haste thou might s him them

5 Przybliżenia na podstawie źródła Markova, gdzie prawdopodobieństwo...

5.1 ... zależy od 1 poprzedniego słowa

- *norm_wiki_sample.txt*

  3 34 6 3 cm in 2001 together that he has indicated on pakhtunkhwa national defense and to this the freight operations gao huan had taken from the region which in the long as the railroad enthusiast who produced in the first magical swords called

- *norm_hamlet.txt*

  extinct in the prenominate crimes the writ and him to give thy name and heaven nor mine arm and sings white as much for passion that but where my lord to know what it is not so far from the queen o i most sacred bands player tis gone and guildenstern

- *norm_romeo.txt*

  will speak anything you to do what i verona ladies lips and all the purpose signior romeo is my old capulet his twisted gyves and gentlewomen to sinners minds tybalt from fearful man as a glove upon you to an old tiberio what said my foe enter paris

5.2 ... zależy od 2 poprzednich słów

- *norm_wiki_sample.txt*

  fourteen songs for her fifth studio album bangerz was released on april 28 2010 fantasy flight games although the origin of cantinflas characteristic speech entertainment career before gaining popularity as a free improvisation with the

- *norm_hamlet.txt*

  t express his love or no to guildenstern what say you then would heart of heart as i do it wrong being so majestical to offer you service he that plays the king mark it horatio looks it not perdie come some music come the recorders for if the gods look

- *norm_romeo.txt*

  i was your mother madam i am too fond and therefore women being the weaker vessels are ever thrust to the prince s doom a gentler judgment vanish d from his grave with tears distill d by you send for the goose thou wast not there for the matter nurse

5.3 ... zależy od 2 poprzednich słów oraz zaczyna się od słowa probability

- *norm_wiki_sample.txt*

  probability bayes theorem likelihood functions markov chains odds ratio proportional hazards models sensitivity and other home electronics motorcycles artificial flowers and michael stewart in the informal idea of kbcss is one of the

- *norm_hamlet.txt*

  probability kisses laying lot owl miching occasion believe otherwise cannons exactly sweet divine fiction husbands recount seiz shipp happiness annexment immortal meed burden ratifiers dies craves dish wake recoveries steward prison

- *norm_romeo.txt*

  probability injuries wills affords ghost joints sit chid masterless joy stair drum blessing dainty kinsman humours without children instant chid mannerly detestable lo locks chapless dearly hearest deny opens ry dealing pride matron