Przetwarzanie języka naturalnego
Pracownia 4a

Wszystkie zadania są ważne do końca semestru, zadania oznaczone literką **E** można oddawać również podczas dodatkowego terminu przed egzaminem.

Część zadań dotyczy materiału, który pojawi się na kolejnych wykładach (co będzie wyraźnie opisane). Na tej liście znajdą się również punkty za zadania z rekonstrukcją (permutacje, ogonki i małe/wielkie litery) oraz nowe dane do zadania ze zgadywaniem autorstwa tekstu (wraz z informacją o punktacji).

Osoba, która rozwiąże zadania z literką **E** za co najmniej 10 punktów może uzyskać zwolnienie z egzaminu i przepisanie oceny z ćwiczeń, jeżeli ta ocena jest co najmniej 4. Jeśli chodzi o zaliczanie ćwiczeń, to zadania z E działają jak normalne zadania.

Zadanie 1. (**4p**) Zmodyfikuj program `np_original.pl` tak, by był w stanie rozpoznać 8000 fraz nominalnych (w programie tym jest zaszyty warunek, że fraza nominalna ma mniej niż K wyrazów, jego (w tym zadaniu) nie powinieneś zmieniać.

Przykładowe polecenie, które zlicza frazy poprawnie rozpoznane przez gramatykę, to:

```
swipl -c np_original.pl | grep -c GOOD
```

(powinno dać około 6000 pozytywnych odpowiedzi)

Zadanie 2. (**4p**) Napisz program wykorzystujący bibliotekę NLTK, by był w stanie rozpoznać 8000 fraz nominalnych. To, jak tego typu gramatyki są wykorzystane w NLTK można przeczytać w rodziale Building Feature Based Grammar, (`http://www.nltk.org/book/ch09.html`). Temu zagadnieniu poświęcone będzie też 3 część Wykładu 11, który pojawi się 23.12.2020.

Zadanie 3. (**1-6, ⋆, Ep**) W zadaniu tym możesz korzystać z dowolnego języka, w którym jesteś w stanie napisać parser taki, jak w poprzednich zadaniach (w szczególności możesz korzystać rozwiązań poprzednich zadań. Możesz również korzystać z informacji „wyciągniętych" ze słownika walencyjnego (Walentego), które znajdziesz na stronie wykładu. W zadaniu chodzi o to, żeby poprawić jakość parsera. Mały punkcik otrzymujemy za sparsowanie poprawnej frazy. Dodatkowo na stronie wykładu jest lista fraz, które nie są frazami **np**, zaakceptowanie którejś z nich oznacza 100 punkcików kary[1].

Punkciki na punkty przekładają się w następujący sposób:

```
8300   1p
9000   2p
9500   3p
10000  4p
10500  5p
11000  6p
```

Przekroczenie liczby 11000 będzie dodatkowo wynagradzane, wg wzoru $\sum_{i=1}^{K} 0.2 \times 0.9^{i-1}$, gdzie K jest liczbą pełnych setek przekraczających 11000. Uwaga: reguły gramatyki powinny opisywać albo ogólne prawidłowości, albo wyjątki, nie jest dozwolone opisywanie jako wyjątkowych tych zjawisk, które da się opisać w sposób bardziej ogólny[2].

Zadanie 4. (7p, E) W zadaniu będziemy ponownie losować wersy Pana Tadeusza, ale tym razem nie korzystając z oryginału, lecz tylko z korpusu PolEvala oraz z zanurzeń wektorowych. Dla ułatwienia przygotowany został zbiór *poprawnych rytmicznie*[3] zdań z PolEvala (zawierających tylko

[1] Lista fraz negatywnych jest utworzona automatycznie i potencjalnie mogą znajdywać się na niej błędy – na SKOS będzie miejsce do zgłaszania takich fraz

[2] Chodzi o to, że nie wolno stworzyć reguły, która mówi, że frazą nominalną jest wszystko to, co znajduje się w pliku phrases.pl, ewentualnie to, co ma te same tagi jak frazy w phrases.pl

[3] Mają one 26 sylab z przerwami po 7, 13 i 20 sylabie, na końcu każdego wersu i przed średniówkami nie ma słów jednosylabowych

słowa z pliku supertags). Większość z nich się nie rymuje (ale dla ułatwienie w pliku z tymi zdaniami zawarte są tylko takie, które da się zrymować, z zachowaniem liczby sylab i tagów gramatycznych ostatnich słów).

a) Napisz program losujący dwuwers i modyfikujący ostatnie wyrazy (być może nie oba) w wersach w ten sposób, by się rymowały (z zachowaniem tagu i liczby sylab). W wyborze powinieneś premiować sytuację, w których wyrazy po zrymowaniu są podobne do wyrazów oryginalnych (czyli wektory ich form bazowych mają możliwie duży iloczyn skalarny). Nie dla każdego dwuwersu to da się zrobić, powinieneś to uwzględniać przy wyborze dwuwersu. Przykładowy wynik działaniu programu (potencjalnie użyteczny do testów):

   ORYGINAŁ: po zjednoczeniu niemiec dotychczas strzeżony [*] obszar został otwarty i przebudowany .
   POEZJA: po zjednoczeniu niemiec dotychczas strzeżony [*] obszar został otwarty i podpiwniczony .

b) Dodaj do powyższego programu możliwość zamiany wybranych słów na inne. Zamieniać powinieneś tylko czasowniki, rzeczowniki, przysłówki, imiesłowy i przymiotniki. Oczywiście w zamianie zachowujemy tag i staramy się zachować podobieństwo do oryginału. Nie wolno nam też zepsuć rymu. Przykładowe działanie:

   ORYGINAŁ: seria dziecięcych skarpet antypoślizgowych [*] z motywami mieszkańców obszarów polarnych .
   POEZJA: seria dziecięcych skarpet antypoślizgowych [*] z motywami mieszkańców obszarów szelfowych .
   ZMODYFIKOWANA: seria przedszkolnych skarpet antypoślizgowych [*] z pejzażami parafian terenów magmowych .