Przetwarzanie języka naturalnego
Ćwiczenia 2
Zajęcia 6

Zadanie 1. Przypomnij, co to jest perplexity? Rozważmy język, w którym słowa są ciągami cyfr, który:

a) składa się z bloków k-cyfrowych (w bloku są te same cyfry)

b) po danym bloku może nastąpić blok z dowolną cyfrą (z tym samym prawdopodobieństwem)

Rozważamy bardzo duży tekst z tego języka. Oblicz dla tego tekstu perplexity dla następujących modeli: unigramowego, bigramowego oraz optymalnego n-gramowego (jakie jest n?) Podaj wartości liczbowe dla $k = 10$.

Zadanie 2. Opowiedz, jak zrobiłeś (lub jak zrobiłbyś, gdybyś miał więcej czasu) zadanie z generowaniem Pana Tadeusza. Opisz użyte struktury oraz sposób zapewniania „jakości poezji". Zastanów się nad (hipotetycznym lub rzeczywistym) czasem działania Twojego algorytmu.

Zaproponuj jakiś sposób poprawy jakości generowanych tekstów.

Zadanie 3. Słownik morfosyntaktyczny jest programem, który dla danego wyrazu zwraca listę możliwych opisów gramatycznych tego wyrazu. Przykładowo dla wyrazu *mam* ta lista mogłamy wyglądać tak:

```
mam
  [mama], rzecz. Rodz=z,Liczba=mn,Przyp=mian
  [mieć], czas. Czas=ter, Osoba=o1, Tryb=ozn
  [mamić], czas. Osoba=o2, Tryb=roz
```

Jak wyglądałaby taka lista dla wyrazów: *nogi*, *rad*, *musi*, *raczy*, *szkoda*, *jak*? (nie musisz używać prawdziwych oznaczeń ze strony sgjp.pl, skoncentruj się na miejscach, w których pojawiają się wieloznaczności).

Zadanie 4. Język polski jest jezykiem typu SVO (sprawdź co to znaczy?). Niemniej jednak jest on mniej „ortodoksyjnym" reprezentantem tej klasy niż np. angielski (dlaczego?)[1] Zaproponuj eksperyment, który przy danym dużym zbiorze tekstów i słowniku morfosyntaktycznym pozwoli:

a) ocenić, czy rzeczywiście j.polski jest typu SVO,

b) wyznaczyć jego „stopień przynależności" do pozostałych pięciu klas.

Zadanie 5. Znajdź sytuacje (dwie conajmniej), w których założenie o typie języka (SVO, patrz poprzednie zadanie) powinno pomóc w wyborze interpretacji jakiegoś dwuznacznego zdania. Jeżeli masz taką możliwość, to zapytaj parę osób (najlepiej niewiedzących nic o SVO) o pierwszą możliwą interpretację tych dwuznacznych zdań. Czy wyniki eksperymentu są takie, jak przewidywałeś?

Zadanie 6. 1 Niektóre polskie słowa mają cztery różne lematy. Przykładowo:

rysie, wole, woli, mole, dziwa, macie, ślepi, kurze

Wybierz z powyższej listy 3 słowa i następnie

1. Zaznacz część mowy, odpowiadającą poszczególnym lematom
2. Przypisz lematowi jakiś numer

Następnie przygotuj kilka zdań (minimum 2), w których występują wyżej wymienione słowa wieloznaczne. Jeżeli nie jest to oczywiste ze względu na treść, dla każdego wystąpienia zaznacz numer lematu z nim związanego. Mile widziane zdania, w którym jest wiele tych samych słów o różnych znaczeniach, czyli np.

[1] Wskazówkozagadka: podaj poularny dziecięcy wierszyk potwierdzający tę nieortodoksyjność.

Kucharz soli/3 solę, czyli sypie szczyptę soli/4 do soli/1.

przy założeniu, że słowo *soli* opisałeś wcześniej w następujący sposób:

sola/1 [subst]
sol/2 [subst]
solić/3 [verb]
sól/4 [subst]

Zadanie 7. Opisz test χ^2 (Chi kwadrat). Jak wykorzystać go do znajdowania kolokacji?

Zadanie 8. ($\star$) Przeczytaj rozdział 6.3 z książki o korpusie NKJP (`http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf`). Odpowiedz na pytania:

a) Ile jest rodzajów przymiotników i co one oznaczają?

b) Co to jest pseudoimiesłów i odsłownik?

c) Co to jest predykatyw i burkinostka?

d) Jakie części mowy nie składają się z liter?

e) Co to jest aspekt i jakie są aspekty?

f) Jakie kategorie gramatyczne ma kublik?

g) Czym różni się zestaw kategorii dla części mowy subst oraz ger.

Pamiętaj, że możesz korzystać ze strony `http://sgjp.pl/morfeusz/demo/`, aby sprawdzać hipotezy odnośnie tagów wybranych wyrazów.

Zadanie 9. Wyszukiwarka IPI-PAN korpusu NKJP (`http://nkjp.pl/poliqarp/`) obsługuje specjalny język zapytań[2]. Poniżej kilka przykładowych zapytań do korpusu:

```
[orth="miał" & base="miał"]
[orth="woli" & pos="adj"]
[orth="woli" & tag="adj:sg:acc:m3:pos"]
[tag="adj:sg:acc:f:pos"] [base="niewiasta"]
pojechał [case="inst"] do [case="gen"]
```

Wykonaj te zapytania i wyjaśnij zaprezentowany fragment składni języka zapytań do korpusu. Wykorzystaj język zapytań, żeby:

a) Znaleźć (co najmniej) 2 błędy w tagowaniu w korpusie (inne niż te, które można znaleźć wykonując powyższe zapytania).

b) Wymyśl krótkie zdanie. Wskaż w tym zdaniu podmiot i znajdź wszystkie zdania w korpusie mające ten podmiot i pasujące do schematu Twojego zdania. Przykładowo dla zdania:

> Piękna dziewczyna spacerowała po tajemniczym parku.

w korpusie znajdujemy między innymi:

> Jakaś dziewczyna uklękła przy Nocnym Śpiewaku.
> Szesnastoletnia dziewczyna usiadła na szerokim tapczanie.
> Śliczna dziewczyna opowiadała o swojej wędrówce
> Zadowolona dziewczyna ruszyła w szeroki świat.
> (...)

[2]Opis jest dostępny po kliknięciu na HELP