1 Zasady ogólne

1. Projekt polega na zaproponowaniu rozwiązania dla wybranego zadania INL. Może to być albo zadanie sprecyzowane w punkcie 2, albo jedno z zadań z konkursów Poleval z tego roku (http://poleval.pl/) lub z lat poprzednich. Można wybrać także inne, własne zadanie, ale w takiej sytuacji należy uzyskać zgodę na jego realizację.
2. Zadanie można rozwiązać metodami statystycznymi (CRF lub inne) lub wykorzystując sieci neuronowe. W każdym przypadku należy podać dwa warianty rozwiązania – na przykład dwa różne zestawy cech, dwie metody statystyczne lub dwie różne architektury sieci neuronowej.
3. Podstawowy warunek – nie można użyć gotowych rozwiązań w wersji niezmienionej. Jeśli użyta zostanie gotowa sieć wytrenowana do dokładnie tego zadania, należy wprowadzić jakieś modyfikacje, na przykład dotrenować inaczej sieć, dodać jakieś dane, nową warstwę i porównać z oryginalnym rozwiązaniem.
4. Pracę należy wykonać samodzielnie lub we dwie osoby
5. Rozwiązania należy zaprezentować na ostatnich ćwiczenia (najlepiej przygotować parę slajdów; można też pokazywać kod; w prezentacji trzeba zawrzeć wyniki)
6. Rozwiązanie należy przekazać przez MS Teams(w grupie ćwiczeniowej).

6.1. Należy umieścić kod (notatnik ipynb+ plik.py, ewentualnie inne pliki źródłowe i wykonywalne, jeśli program nie jest w Pythonie)

6.2. Notatnik musi zawierać wyniki z uruchomienia. Powinien dać się też uruchomić mnie - należy napisać jakich zmian trzeba dokonać by uruchomić program w innym środowisku

6.3. Bezpośrednio w notatniku, albo w pliku dodatkowym należy umieścić opis wybranego zadania, ilościową analizę danych, opis przyjętego sposobu rozwiązania, wyniki i ich analizę (tam gdzie to jest adekwatne - precyzję pełność i miaręF_1, zarówno ogółem jak i dla każdej etykiety odrębnie).

7. Opóźnienie przysłania rozwiązania będzie karane stopniowo coraz większą utratą punktów. Rozwiązania nie nadesłane i nie gotowe do prezentacji na ostatnich zajęciach mogą otrzymać co najwyżej 70% punktów.
8. Ocena projektu dotyczyć będzie:

- spełnienia wymagań formalnych
- wykazania się kreatywnością
- skuteczności działania – więcej punktów dla rozwiązań najlepszych

2 Temat podstawowy (NER)

Podstawowym tematem jest rozpoznawanie nazw własnych w tekstach polskich. Dane treningowe to NKJP –w wersji dostarczonej na wykładowych MsTeams, plik NKJP_org..zip, spakowany plik csv z uproszczonym formatem kolumnowym (forma, lemat, tag, etykiety). Przykładowe dwie linie tego pliku są poniżej:

tamtym	adj:sg:inst:m3:pos		
Krzemieńcem	subst:sg:inst:m3	placeName	settlement

W pliku tym etykiety są w 3 i 4 kolumnie. Należy użyć tylko tych z kolumny 4 (chyba, że jest etykieta w kolumnie 3, a nie ma w 4, to wtedy użyć tej z kolumny 3). W przypadku pustych kolumn 3 i 4 należy wstawić etykietę ‘O’.

Dane te należy podzielić na dane treningowe (50%), walidacyjne (30%) i testowe (20%).