**

Pre-processing II (POS tagging/Dependency Parsing) Lab

# Objective

1. To be able to use Python libraries for **POS Tagging**.
2. To be able to do **Dependency parsing** with Python libraries.
3. To be able to use **evaluating tools** using python libraires.

# Task

1. Go to any online newspaper NZHerald or another newspaper website and extract **a paragraph** from a news article of your choice.
2. Use the tags from end of the lecture 1 to manually tag all the sentences in the paragraph. You can do this in pairs so that you can discuss any tags that might be ambiguous.
3. Use NLTK package in Python to POS tag all the sentences in the paragraph.

You can use the code segment sample below as a guide.

#################################

import nltk

tokens = nltk. word\_tokenize("AUT is in New Zealand")

postags = nltk.pos\_tag(tokens)

print(postags)

#################################

In the code above , we first tokenize the text and then apply POS tagging.

The code will give you a result similar to one shown below.

[('AUT', 'NNP'), ('is', 'VBZ'), ('in', 'IN'), ('New', 'NNP'), ('Zealand', 'NNP')]

Note that these tokens are taken from Penn Treebank guidelines. <https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html>

1. Manually determine the accuracy of the tagger.
2. Use the code example from lecture to determine the Precision, Recall and F1 value of the paragraph under study. Note, that you will need to manually label the POS tags which will form the reference for comparison.
3. Use the code from lectures on Dependency parsing and clause extraction for this part.
   1. Download about 5 text articles from the web on your favourite sport that have names of players scoring goals, tries, points etc.
   2. Read the articles from the directory.
   3. Extract all the clauses in which some type of point scoring is reported. For example from the test “*John scored a goal*” you should output **[John] [scored] [goal]**