Natural Language Processing (viterbi tesk- 1) Report Given the POS tags in the train dataset, We need to calculate the transition and emission probabilities of each word and tag, and create the Viterbi matrix from scratch using these values. We will have to show classification reports on both the train and test data.

1. calculate_metrics:

- train Accuracy: 0.8044 Precision: 0.8204 Recall: 0.8044 F1-score: 0.7939
- test Accuracy: 0.6830 Precision: 0.7245 Recall: 0.6830 F1-score: 0.6580
- Predictions saved in viterbi_predictions_train.tsv viterbi_predictions_test.tsv in format of (sent_idtoken number =i ith wordith pred_tag)
- Smoothed words- train = 0 test = 260 {'adjusted', 'bqj', 'color', 'Palmeri', 'workplace', 'demographic', 'Afiniam', 'Fennig', '-□æk', 'Bassène', 'species', 'reduce', 'section', 'exploratory', 'theorists', 'bearing', 'east', 'fullest', 'trainings', 'sex', 'neighbouring', 'analytical', 'counts', 'Bolon', 'river', 'discriminatory', 'Gújjolaay', 'respondents', 'independence', 'briefly', 'bordered', 'estimates', 'neurobiological', 'conscious', 'beyond', 'prevalence', 'interviewed', '1873', 'courtesy', 'rare', 'River', 'Kessler', 'idiom', 'rediscovered', 'Búluf', 'build', 'endangered', 'subtracting', 'options', 'Romantic', '□dvora□k', 'shade', 'Indian', 'phylum', 'Essil', 'facing', '639-3', 'Black', 'respects', 'BAK', 'implement', 'adjustments', 'Niger', 'recommendation', 'estimating', 'Sállagi', 'unreturned', 'underpin', 'Smetana', 'collapsed', 'weighted', 'Atlantic', 'distribution', 'Guinea', 'codes', 'Congo', 'Kuluunaay', 'recreation', 'Native', 'composer', 'refer', 'Gubëeher', 'commands', 'Bohemia', 'homeland', 'transcription', 'adults', 'discriminated', 'recoded', 'minority', 'SPEAKERS', 'villages', 'wealth', '76', 'Islander', 'proportions', 'Rao', 'ITS', 'statistic', 'Prague', 'apt', 'income', 'Bandial', '1904', 'Symphony', 'Gazio', 'Bissau', 'separates', 'symphonies', 'adverse', 'logic', 'employing', 'Fogny', 'EEGIMAA', 'Basse', 'Ziguinchor', 'Djibonker', 'coincides', 'classified', 'examined', 'policy', 'noninteger', 'code', 'Austrian', 'd(□)-VOR-zha(h)k', 'similarity', 'Brin', 'Dvořák', 'Bedřich', 'cohabit', 'varieties', 'χ2', 'GUJJOLAAY', 'pattern', 'steps', 'underpinnings', 'Scott', 'Gufiñamay', '1995', '11,200', 'Gusiilay', 'dichotomously', 'administrative', 'Djilapaor', 'worldwide', 'absorbing', 'competition', 'd(□)□v□□r□□□k', 'Decree', 'ISO', 'bivariate', 'THEIR', 'Composition', 'points', 'Bourofaye', 'Baïnounk', 'Kamobeul', 'suite', 'prevention', 'Jóola', 'Djifanghor', 'settings', '1841', 'respondent', 'bsl', 'Czech', 'musical', 'sampling', '2005-981', 'kingdom', 'peoples', 'Seleki', 'folk', 'arrow', 'rooted', 'African', 'urban', 'stratification', 'mainly', 'Kusiilay', 'Bajjat', '74', 'nationally', 'referred', 'mutually', 'preserve', 'migrant', 'generations', 'Kaasa', 'analysed', 'Mof-Ávvi', 'Moravia', 'influences', '7,000', 'Hispanic', 'Pearson', 'Pacific', 'outside', 'category', 'clustering', '§', 'calculations', 'strategy', 'rhythms', 'north', 'differential', 'NEIGHBOURS', 'Banjal', 'examine', 'approximate', 'Ésuulaalu□', 'attrition', '□anto□i□n', 'nationalist', 'Sapir', 'speech', 'Dakar', 'appreciable', 'discrimination', 'occurrence', 'declining', 'ritual', 'orthography', '□l□opolt', 'questionnaire', 'speakers', 'capturing', 'Bayot', 'violin', 'Ethnologue', 'Kujireray', 'Gibbons', 'exclusive', 'peninsula', 'push', 'Senegal', 'transmission', 'disability', 'symphonic', 'seeks', 'svy', 'Casamance', 'race', 'Lewis', 'spelling', 'mixed', 'usual', 'Thionk',

'corrects', 'Gambia', 'Stata', 'Butam', 'assumption', 'select', 'Eegimaa',
'Húluf', '1874', 'Gulapaoray', 'eds.', 'west'}

- challenges

1. implementing viterbi algo and backtrack (used DP pointer)
2. smoothing (used dict iinstead of matrix)
3. calculte smooth words (checking if max emission of a word is 1/V)

- Assumption

1. train.csv and test.csv is in same folder as the codes (.ipynb /.py) files