# POS tagging using modified Viterbi

This file contains 3 modifications to the Vanilla Viterbi algorithm as follows:

* **Transition Based Technique** : This sets the tag of the word based on the maximum of the Transition probabilities.
* **Transition and Reverse Transition Based Technique** : This sets the tag of the word based on the maximum of Transition and Reverse Transition probabilities. Reverse Transition probability is probability of t2 given t1 **follows** t2.
* **Rule Based Technique** : This sets the tag of the word based on certain rules

#### Method 1: A transition probability based approach

This flavour considers the maximum of transition probability for tagging the unknown words.

Steps: \* Both Emission and Transition probability is calculated for a word \* If a word has the max probability as 0, this means that the Emission probability is 0 because none of the Transition probabilities are 0 as per the tags-matrix. Effectively, the word is unknown. \* In this case, tag having the maximum Transition probability is considered as the POS tag for the unknown word.

#### Method 2: Another transition probability based approach

The Transition probability considers the probability of a **tag given a previous tag**. This is another flavour which considers the probability of a **tag given a following tag** also. Maximum of transition probabilities and reverse transition probabilities is considered here for tagging the unknown words.

Steps: \* Both Emission and Transition probability is calculated for a word \* If a word has the max probability as 0, this means that the Emission probability is 0 because none of the Transition probabilities are 0 as per the tags-matrix. Effectively, the word is unknown. \* After this the words are scanned again(disadvantage, but this method depends on knowing the next tag) and the probabilities are calculated for a tag given a following tag. Maximum of these is taken as the tag.

#### Method 3: A rule based approach

Another set of modifications to the Vanilla Viterbi algorithm where it applies some rules for tagging the unknown words The rules are applied in the following order:

* Rule 1: All words having one or more digits are tagged as NUM
* Rule 2: All words starting with one or more star are tagged as X
* Rule 3: All words starting with numbers and ending with st,nd,rd,th are tagged as ADJ
* Rule 4: All words starting with caps in the middle of a sentence are tagged as NOUN

The next 4 rules are taken from https://dictionary.cambridge.org/grammar/british-grammar/word-formation/suffixes

* Rule 5: All words ending with ly|wards|wise are tagged as ADV
* Rule 6: All words ending with ate|en|ify|ise|ize|ed|ing are tagged as VERB
* Rule 7: All words ending with able|ible|al|en|ese|ful|i|ic|ish|ive|ian|less|ous are tagged as ADJ
* Rule 8: All words ending with age|al|ance|ence|dom|ee|er|or|hood|ism|ist|ity|ty|ment|ness|ry|ship|sion|tion|xion are tagged as NOUN
* Rule 8: If a word does not belong to any of the above rules, then it probably is some Name and hence default tagging done is a NOUN