Universidade Federal do Rio Grande do Norte

Instituto Metrópole Digital

DIM0135 - Processamento de Linguagem Natural

# Part-of-Speech Tagger: Hidden Markov Model

João Vitor Venceslau Coelho

Natal/RN

2022

## Introdução

Este relatório tem como objetivo apresentar as particularidades da abordagem selecionada para a tarefa de **Part-of-Speech Tagging**, utilizando de **Hidden Markov Model (HMM)**. Ao fim são apresentados os resultados da aplicação do algoritmo desenvolvido no corpus Floresta disponibilizado pela Linguateca (link nas referências) e uma breve discussão sobre os resultados obtidos. Foi utilizada uma versão no formato do Penn Treebank, que foi pré-processada para um formato de pares tag-word antes de ser passada para o algoritmo desenvolvido para PoS tagging.

## Estratégia Geral

O primeiro passo do algoritmo é ler os dados disponibilizados para o treinamento, isto é, para cada par de tag-word, armazenar a sequência de palavras, separando o início de uma frase com um token especial (“\_\_SS\_\_”), assim como para as tags, também é determinado nesse passo o conjunto de palavras (vocabulário) e tags conhecidas (estados possíveis), além da probabilidade de cada tag iniciar uma frase. Em seguida a sequência de palavras e tags que serão usadas para avaliar o modelo são armazenadas para uso futuro.

Após ler os dados que serão utilizados, é computada a matriz de emissão, isto é, qual a probabilidade de um estado (tag) ser emitido por uma observação (word), e a matriz de transição, ou seja, para cada tag, qual a probabilidade dela ser seguida por cada uma das possíveis tags. Ambas as matrizes são calculadas utilizando os dados da sequência de treino.

Em seguida, para cada frase utilizada no teste, a sequência de palavras é pré-processada, para que cada palavra conhecida seja trocada pelo índice dela no vocabulário visto no treino, assim pode-se identificar qual probabilidade utilizar da matriz de emissão. Caso a palavra não faça parte do vocabulário conhecido, são utilizadas algumas regras para estimar qual a tag daquela palavra (o conjunto de regras utilizadas será apresentado na próxima seção), e é armazenado o negativo do índice da tag estimada mais 1 *-(tag\_idx+1)* no array de sequência das palavras, assim ao final deste pré-processamento, a sequência de palavras é substituída por uma sequência de números inteiros, que, quando negativos, indicam uma palavra desconhecida, que provavelmente está emitindo uma tag codificada no valor daquele número negativo.

Após codificar a sequência de palavras, é iniciado o algoritmo Viterbi para determinar qual a sequência de tags tem maior probabilidade de estar correta, dadas as informações coletadas durante o treino, sendo elas as matrizes de emissão e transição, e probabilidades de início de frase de cada tag. Foi utilizada a versão logarítmica do Viterbi, para substituir as multiplicações por somas, reduzindo o tempo de computação necessário, um ponto que vale a pena destacar na implementação do Viterbi é que ao encontrar um valor negativo na sequência de observações, em vez de acessar o índice indicado pelo valor da observação, na matriz de emissão na linha da tag que está sendo analisada, é atribuída uma emissão de 0 (mas não exatamente, na verdade se usa um valor muito pequeno) se for uma tag diferente da codificada no valor negativo e 1 caso seja a tag codificada.

Após computar todos os elementos da matriz de probabilidades acumuladas e da matriz de índices que será usada para o backtracking da sequência de maior probabilidade é feito de fato o backtracking e é retornada a sequência encontrada.

## Conjunto de regras utilizadas

Antes de aplicar as regras é feito um pequeno pré-processamento na palavra desconhecida, caso ela termine com o caractere ‘-’ (hífen), o mesmo é removido da string.

A primeira regra aplicada é a verificação se a palavra caracteriza um numeral, isto é identificado quando a palavra é formada apenas por números, com a presença ou não de ‘.’ ou ‘,’ entre os números.

A segunda regra a ser verificada é se a palavra termina com ‘ndo’, se for o caso, então é atribuída a tag de 'v-ger', relativa a verbos no gerúndio, em seguida é verificado se a palavra termina com algum dos seguintes sufixos: ‘ada’, ‘adas’, ‘ado’, ‘ados’, ‘ida’, ‘idas’, ‘ido’, ‘idos’, ‘ída’, ‘ídas’, ‘ído’ ou ‘ídos’, sendo atribuída em qualquer um dos casos a tag 'v-pcp', associada a verbos no particípio.

A próxima regra a ser verificada também analisa os sufixos, no caso, se a palavra termina com: ‘ar’, ‘er’, ‘ir’, ‘arem’, ‘erem’ ou ‘irem’, é atribuída a tag ‘v-inf’ relativa a verbos no infinitivo. A quinta regra é a última relacionada a verbos, atribuindo a tag ‘v-fin’ de verbo finito, quando a palavra termina com os seguintes sufixos: ‘ava’, ‘imos’, ‘aram’, ‘emos’, ‘avam’, ‘amos’, ‘ou’ ou ‘ia’.

A regra seguinte atribui a tag ‘adv’, relativa a advérbios, quando a palavra termina com o sufixo ‘mente’, a sétima regra, atribui a tag ‘adj’, relativa a adjetivos, quando a palavra termina com algum dos seguintes sufixos: ‘ento’, ‘entos’, ‘ável’, ‘áveis’, ‘ante’, ‘antes’, ‘esco’, ‘escos’, ‘ível’, ‘íveis’, ‘ano’ ou ‘ino’.

A penúltima regra, é direcionada a substantivos/nomes, a tag ‘n’, sendo atribuída quando as palavras terminam em ‘ismo’, ‘ismos’, ‘idade’ ou ‘idades’, e a última regra apenas verifica se a primeira letra é maiúscula, atribuindo a tag ‘prop’ relacionada a nomes próprios.

É interessante notar que a ordem de aplicação dessas regras interfere nos resultados obtidos, porém é importante deixar claro que não foram feitos os devidos testes para determinar a melhor ordem, nem quais regras de fato valem a pena serem utilizadas.

## Resultados Obtidos

Como mencionado anteriormente, foi utilizado o corpus Floresta da Linguateca para realizar os testes. Durante o pré-processamento o corpus foi separado em três conjuntos, treino, validação e teste. Sendo que o script deste pré-processamento aceita um inteiro como seed para o sorteio de quais frases vão para cada conjunto. Os conjuntos de treino e validação juntos correspondem a 80% das frases do corpus e os testes aos 20% restantes.

|  | **False Positive** | **False Negative** | **True Positive** | **True Negative** |
| --- | --- | --- | --- | --- |
| **pred\_!** | 0 | 0 | 15 | 33.139 |
| **pred\_'** | 3 | 0 | 9 | 33.142 |
| **pred\_\*** | 0 | 0 | 1 | 33.153 |
| **pred\_,** | 0 | 0 | 2.022 | 31.132 |
| **pred\_-** | 1 | 0 | 5 | 33.148 |
| **pred\_--** | 0 | 0 | 96 | 33.058 |
| **pred\_.** | 0 | 0 | 1.221 | 31.933 |
| **pred\_...** | 0 | 0 | 11 | 33.143 |
| **pred\_/** | 0 | 1 | 3 | 33.150 |
| **pred\_;** | 0 | 0 | 89 | 33.065 |
| **pred\_?** | 2 | 0 | 32 | 33.120 |
| **pred\_[** | 1 | 0 | 2 | 33.151 |
| **pred\_]** | 0 | 0 | 3 | 33.151 |
| **pred\_adj** | 436 | 150 | 1.319 | 31.249 |
| **pred\_adv** | 161 | 87 | 1.344 | 31.562 |
| **pred\_art** | 95 | 111 | 4.543 | 28.405 |
| **pred\_conj-c** | 7 | 2 | 803 | 32.342 |
| **pred\_conj-s** | 89 | 56 | 288 | 32.721 |
| **pred\_in** | 7 | 1 | 0 | 33.146 |
| **pred\_n** | 256 | 1.026 | 6.096 | 25.776 |
| **pred\_num** | 55 | 34 | 587 | 32.478 |
| **pred\_pp** | 21 | 11 | 35 | 33.087 |
| **pred\_pron-det** | 81 | 46 | 652 | 32.375 |
| **pred\_pron-indp** | 62 | 74 | 445 | 32.573 |
| **pred\_pron-pers** | 16 | 43 | 389 | 32.706 |
| **pred\_prop** | 198 | 97 | 1.496 | 31.363 |
| **pred\_prp** | 88 | 102 | 4.946 | 28.018 |
| **pred\_v-fin** | 406 | 51 | 2.068 | 30.629 |
| **pred\_v-ger** | 7 | 4 | 130 | 33.013 |
| **pred\_v-inf** | 21 | 72 | 797 | 32.264 |
| **pred\_v-pcp** | 38 | 83 | 680 | 32.353 |
| **pred\_{** | 0 | 0 | 151 | 33.003 |
| **pred\_}** | 0 | 0 | 151 | 33.003 |
| **pred\_«** | 0 | 0 | 337 | 32.817 |
| **pred\_»** | 0 | 0 | 337 | 32.817 |
| **total** | 2.051 | 2.051 | 31.103 | 1.125.185 |

**Tabela 1**: Acertos e erros para cada tag utilizada no conjunto de validação

|  | **False Positive** | **False Negative** | **True Positive** | **True Negative** |
| --- | --- | --- | --- | --- |
| **pred\_!** | 0 | 0 | 12 | 41.433 |
| **pred\_"** | 1 | 0 | 1 | 41.443 |
| **pred\_'** | 0 | 0 | 38 | 41.407 |
| **pred\_\*** | 2 | 0 | 0 | 41.443 |
| **pred\_,** | 0 | 1 | 2.597 | 38.847 |
| **pred\_-** | 2 | 0 | 13 | 41.430 |
| **pred\_--** | 1 | 0 | 136 | 41.308 |
| **pred\_.** | 0 | 0 | 1.533 | 39.912 |
| **pred\_...** | 1 | 0 | 21 | 41.423 |
| **pred\_/** | 1 | 0 | 8 | 41.436 |
| **pred\_;** | 0 | 0 | 137 | 41.308 |
| **pred\_?** | 0 | 0 | 56 | 41.389 |
| **pred\_[** | 1 | 0 | 9 | 41.435 |
| **pred\_]** | 2 | 0 | 8 | 41.435 |
| **pred\_adj** | 543 | 228 | 1.485 | 39.189 |
| **pred\_adv** | 171 | 95 | 1.510 | 39.669 |
| **pred\_art** | 95 | 218 | 5.568 | 35.564 |
| **pred\_conj-c** | 5 | 5 | 962 | 40.473 |
| **pred\_conj-s** | 88 | 80 | 341 | 40.936 |
| **pred\_ec** | 2 | 0 | 0 | 41.443 |
| **pred\_in** | 5 | 0 | 3 | 41.437 |
| **pred\_n** | 329 | 1.321 | 7.411 | 32.384 |
| **pred\_num** | 54 | 34 | 743 | 40.614 |
| **pred\_pp** | 27 | 13 | 43 | 41.362 |
| **pred\_pron-det** | 180 | 53 | 841 | 40.371 |
| **pred\_pron-indp** | 74 | 59 | 561 | 40.751 |
| **pred\_pron-pers** | 31 | 56 | 523 | 40.835 |
| **pred\_prop** | 245 | 110 | 2.178 | 38.912 |
| **pred\_prp** | 129 | 116 | 6.238 | 34.962 |
| **pred\_v-fin** | 538 | 44 | 2.493 | 38.370 |
| **pred\_v-ger** | 6 | 6 | 146 | 41.287 |
| **pred\_v-inf** | 28 | 95 | 923 | 40.399 |
| **pred\_v-pcp** | 62 | 87 | 805 | 40.491 |
| **pred\_{** | 0 | 0 | 223 | 41.222 |
| **pred\_}** | 0 | 0 | 224 | 41.221 |
| **pred\_«** | 0 | 0 | 516 | 40.929 |
| **pred\_»** | 0 | 2 | 516 | 40.927 |
| **total** | 2.623 | 2.623 | 38.822 | 1.489.397 |

**Tabela 2**: Acertos e erros para cada tag utilizada no conjunto de teste

|  | **Positive Predictive value** | **False Discovery rate** | **Accuracy** | **F1 Score** |
| --- | --- | --- | --- | --- |
| **pred\_!** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_'** | 75,00% | 25,00% | 99,99% | 85,71% |
| **pred\_\*** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_,** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_-** | 83,33% | 16,67% | 100,00% | 90,91% |
| **pred\_--** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_.** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_...** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_/** | 100,00% | 0,00% | 100,00% | 85,71% |
| **pred\_;** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_?** | 94,12% | 5,88% | 99,99% | 96,97% |
| **pred\_[** | 66,67% | 33,33% | 100,00% | 80,00% |
| **pred\_]** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_adj** | 75,16% | 24,84% | 98,23% | 81,82% |
| **pred\_adv** | 89,30% | 10,70% | 99,25% | 91,55% |
| **pred\_art** | 97,95% | 2,05% | 99,38% | 97,78% |
| **pred\_conj-c** | 99,14% | 0,86% | 99,97% | 99,44% |
| **pred\_conj-s** | 76,39% | 23,61% | 99,56% | 79,89% |
| **pred\_in** | 0,00% | 100,00% | 99,98% | 0,00% |
| **pred\_n** | 95,97% | 4,03% | 96,13% | 90,49% |
| **pred\_num** | 91,43% | 8,57% | 99,73% | 92,95% |
| **pred\_pp** | 62,50% | 37,50% | 99,90% | 68,63% |
| **pred\_pron-det** | 88,95% | 11,05% | 99,62% | 91,13% |
| **pred\_pron-indp** | 87,77% | 12,23% | 99,59% | 86,74% |
| **pred\_pron-pers** | 96,05% | 3,95% | 99,82% | 92,95% |
| **pred\_prop** | 88,31% | 11,69% | 99,11% | 91,03% |
| **pred\_prp** | 98,25% | 1,75% | 99,43% | 98,12% |
| **pred\_v-fin** | 83,59% | 16,41% | 98,62% | 90,05% |
| **pred\_v-ger** | 94,89% | 5,11% | 99,97% | 95,94% |
| **pred\_v-inf** | 97,43% | 2,57% | 99,72% | 94,49% |
| **pred\_v-pcp** | 94,71% | 5,29% | 99,64% | 91,83% |
| **pred\_{** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_}** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_«** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_»** | 100,00% | 0,00% | 100,00% | 100,00% |
| **total** | 93,81% | 6,19% | 99,65% | 93,81% |

**Tabela 3**: Valores arredondados de *Positive Predictive value*, *False Discovery rate*, *Accuracy* e o *F1-Score* de cada tag utilizada no conjunto de validação.

|  | **Positive Predictive value** | **False Discovery rate** | **Accuracy** | **F1 Score** |
| --- | --- | --- | --- | --- |
| **pred\_!** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_"** | 50,00% | 50,00% | 100,00% | 66,67% |
| **pred\_'** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_\*** | 0,00% | 100,00% | 100,00% | 0,00% |
| **pred\_,** | 100,00% | 0,00% | 100,00% | 99,98% |
| **pred\_-** | 86,67% | 13,33% | 100,00% | 92,86% |
| **pred\_--** | 99,27% | 0,73% | 100,00% | 99,63% |
| **pred\_.** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_...** | 95,45% | 4,55% | 100,00% | 97,67% |
| **pred\_/** | 88,89% | 11,11% | 100,00% | 94,12% |
| **pred\_;** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_?** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_[** | 90,00% | 10,00% | 100,00% | 94,74% |
| **pred\_]** | 80,00% | 20,00% | 100,00% | 88,89% |
| **pred\_adj** | 73,22% | 26,78% | 98,14% | 79,39% |
| **pred\_adv** | 89,83% | 10,17% | 99,36% | 91,91% |
| **pred\_art** | 98,32% | 1,68% | 99,24% | 97,27% |
| **pred\_conj-c** | 99,48% | 0,52% | 99,98% | 99,48% |
| **pred\_conj-s** | 79,49% | 20,51% | 99,59% | 80,24% |
| **pred\_ec** | 0,00% | 100,00% | 100,00% | 0,00% |
| **pred\_in** | 37,50% | 62,50% | 99,99% | 54,55% |
| **pred\_n** | 95,75% | 4,25% | 96,02% | 89,98% |
| **pred\_num** | 93,22% | 6,78% | 99,79% | 94,41% |
| **pred\_pp** | 61,43% | 38,57% | 99,90% | 68,25% |
| **pred\_pron-det** | 82,37% | 17,63% | 99,44% | 87,83% |
| **pred\_pron-indp** | 88,35% | 11,65% | 99,68% | 89,40% |
| **pred\_pron-pers** | 94,40% | 5,60% | 99,79% | 92,32% |
| **pred\_prop** | 89,89% | 10,11% | 99,14% | 92,46% |
| **pred\_prp** | 97,97% | 2,03% | 99,41% | 98,07% |
| **pred\_v-fin** | 82,25% | 17,75% | 98,60% | 89,55% |
| **pred\_v-ger** | 96,05% | 3,95% | 99,97% | 96,05% |
| **pred\_v-inf** | 97,06% | 2,94% | 99,70% | 93,75% |
| **pred\_v-pcp** | 92,85% | 7,15% | 99,64% | 91,53% |
| **pred\_{** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_}** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_«** | 100,00% | 0,00% | 100,00% | 100,00% |
| **pred\_»** | 100,00% | 0,00% | 100,00% | 99,81% |
| **total** | 93,67% | 6,33% | 99,66% | 93,67% |

**Tabela 4**: Valores arredondados de *Positive Predictive value*, *False Discovery rate*, *Accuracy* e o *F1-Score* de cada tag utilizada no conjunto de teste.

Outra métrica computada, porém não presente nas tabelas mostradas, é o quão precisa foram as sequências de tags atribuídas para cada frase, essa informação foi computada considerando a sequência de tags esperada e a retornada pelo Viterbi, contando o número de tags corretamente atribuídas e dividindo o resultado da soma pelo total de tags da sequência. Essa métrica foi calculada para todas as frases do conjunto de validação e de teste, e estão presentes nos arquivos ‘sentences\_precision\_hmm\_val.csv’ e ‘sentences\_precision\_hmm\_test.csv’, a média dessa métrica no conjunto de validação foi de *93,03%* e de *92,78%* no de teste, essas médias estão presentes no final de cada arquivo.

Todos os resultados aqui apresentados podem ser conferidos nos arquivos de métricas gerados pelo algoritmo, assim como é possível conferir qual a tag atribuída e qual a tag esperada em cada conjunto de teste.

## Discussões dos Resultados

Considerando inicialmente os resultados das **Tabelas 1 e 2**, sobre o número de acertos e erros nos dois conjuntos testados, temos que, das 33.154 palavras do conjunto de validação, 31.103 tiveram a tag corretamente estimada, enquanto 2.051 não. Ao passo que no conjunto de teste, das 41.445 palavras, 38.822 tiveram a tag corretamente estimada e 2.623 não. Porém essas informações são melhor sintetizadas nas **Tabelas 3 e 4**, onde as métricas de *Positive Predictive Value***,***Accuracy* e *F1-Score* são apresentadas. No conjunto de validação elas foram respectivamente: *93,81%, 99,65%* e *93,81%* enquanto que no conjunto de treino foram: *93,67%, 99,66%* e *93,67%*. A diferença entre as métricas obtidas entre os dois conjuntos testados é bem pequena, mesmo que durante a validação a performance tenha sido melhor.

Porém vale destacar nas **Tabelas 1 e 2** que alguns símbolos de pontuação tiveram a tag errada atribuída, o que não era esperado por mim, dado que esses símbolos sempre assumem a mesma tag, o da própria pontuação que representam. Porém cogita-se que seja devido a estratégia utilizada durante o viterbi, para o *acúmulo de probabilidade*, mesmo que esses símbolos de pontuação não emitam nenhuma outra tag além da própria pontuação, durante o acúmulo da probabilidade para a transição, outras sequências de tags acabaram sendo mais prováveis ao não atribuir a tag de pontuação característica desse símbolo de pontuação.

Ainda nas **Tabelas 1 e 2** é interessante mencionar que a tag que teve mais palavras erroneamente atribuídas é a ‘n’, isto é, nomes e/ou substantivos, o que pode ser devido ao fato de ser a tag com mais amostras ou ao conjunto de regras, que “chuta” essa tag quando nenhuma das outras regras se aplica. Enquanto que a tag ‘in’ relacionada a interjeições, não teve nenhum único acerto. Esse fenômeno relacionado aos erros, pode ser visualizado por cada tag nas **Tabelas 3 e 4**, por meio da métrica de *False Discovery rate*, onde a tag ‘in’ obteve *100%* na validação e *62,50%* no teste, enquanto a tag ‘n’ ficou com menos de *5%* em ambos os conjuntos, mesmo sendo a que teve maior contagem de erros. Porém a tag ‘in’ é muito pouco utilizada, assim, considerando as tags com uma quantidade considerável de amostras, digamos, pelo menos 100 amostras, a tag com maior taxa de erro foi a dos adjetivos ‘adj’ com *24,84%* na validação e *26,78%* no teste.

Diante disso, vale destacar que duas tags que não apareceram na validação surgiram durante o teste, sendo elas as aspas (“) e a tag “ec”, que aparece apenas 5 vezes em todo o conjunto de dados, 3 vezes no treino e 2 no teste, sendo associada aos tokens: “anti-”, “ex-”, “pós” e “pré-”, onde “ex-” aparece duas vezes. Essa tag não foi encontrada no Glossário de etiquetas florestais, então não sei o seu significado, mas aparenta ser relacionado com prefixos.

Por fim, a título de comparação, ao aplicar as técnicas de bigramas e de unigramas para PoS tagging nos mesmos conjuntos de validação e teste, obteve-se as seguintes métricas gerais:

| **Abordagem** | **Conjunto** | **False Positive** | **False Negative** | **True Positive** | **True Negative** |
| --- | --- | --- | --- | --- | --- |
| bigramas | teste | 4.873 | 4.873 | 36.572 | 1.487.147 |
| bigramas | validação | 3.670 | 3.670 | 29.484 | 1.123.566 |
| unigramas | teste | 5.357 | 5.357 | 36.088 | 1.486.663 |
| unigramas | validação | 4.041 | 4.041 | 29.113 | 1.123.195 |
| hmm | teste | 2.623 | 2.623 | 38.822 | 1.489.397 |
| hmm | validação | 2.051 | 2.051 | 31.103 | 1.125.185 |

**Tabela 5**: Acertos e erros para cada técnica em cada conjunto de dados

| **Abordagem** | **Conjunto** | **Positive Predictive value** | **False Discovery rate** | **Accuracy** | **F1 Score** |
| --- | --- | --- | --- | --- | --- |
| bigramas | teste | 88,24% | 11,76% | 99,36% | 88,24% |
| bigramas | validação | 88,93% | 11,07% | 99,37% | 88,93% |
| unigramas | teste | 87,07% | 12,93% | 99,30% | 87,07% |
| unigramas | validação | 87,81% | 12,19% | 99,30% | 87,81% |
| HMM | teste | 93,67% | 6,33% | 99,66% | 93,67% |
| HMM | validação | 93,81% | 6,19% | 99,65% | 93,81% |

**Tabela 6**: Valores arredondados de *Positive Predictive value*, *False Discovery rate*, *Accuracy* e o *F1-Score* para cada técnica em cada conjunto de dados

Como pode-se constatar, a utilização do HMM para PoS tagging obteve resultados superiores a da utilização de bigramas ou de unigramas, mas vale ressaltar que não foram aplicadas as mesmas regras para palavras desconhecidas que foram aplicadas ao algoritmo com HMM, essa diferença é relevante. Para palavras desconhecidas foi utilizada a mesma abordagem descrita no relatório anterior sobre bigramas e unigramas.

## Referências

Capítulo 8 do livro Speech and Language Processing de Daniel Jurafsky & James H. Martin

<https://web.stanford.edu/~jurafsky/slp3/8.pdf>

Site de download do corpus Floresta:

<http://corpora.di.uminho.pt/linguateca/FS/fs.html>

Glossário de etiquetas florestais:

<https://www.linguateca.pt/Floresta/BibliaFlorestal/anexo1.html>

Tutorial de PoS Tagging em Python

<https://www.mygreatlearning.com/blog/pos-tagging/>

Tutorial do algoritmo Viterbi em Python

<https://www.audiolabs-erlangen.de/resources/MIR/FMP/C5/C5S3_Viterbi.html>

Consulta sobre sufixos no português

<https://www.normaculta.com.br/sufixos/>