Disciplina:

Processamento de linguagem natural

Professor: Gabriel Assunção

PUC Minas

Apresentação do curso

Módulos

1. Introdutório
2. Expressão regular
3. Processamento de texto
4. Representação textual
5. Classificação de texto
6. **Modelos de NLP**

Revisão:

- o que é regex?
- Quando utilizar regex?
- De três exemplos de etapa de pré processamento
- Porque é importante pré processar um dado textual ?
- O que é stemming ?
- O que é lemmatization?
- De um exemplo de classificação textual (Diferente de análise de sentimentos)
- Qual o processo para se treinar um modelo de classificação ?
- Quais as dificuldades na análise de sentimentos?
- Uma análise de sentimentos pode ser construída através de regras?

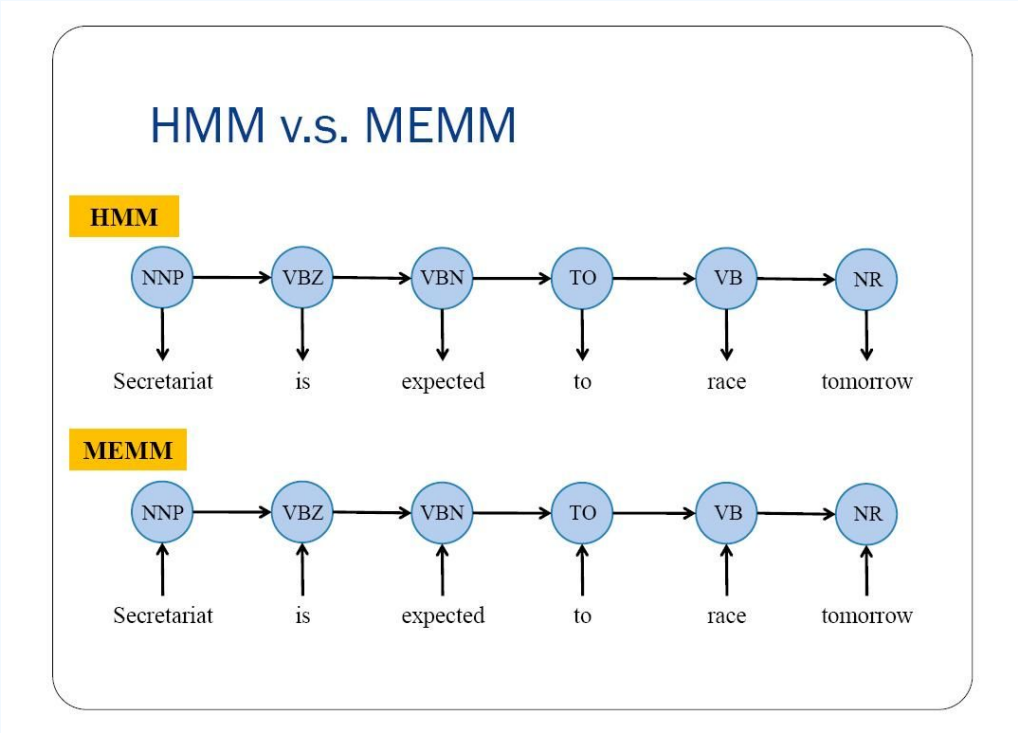
POS-Tagging

Part-of-speech Tagging

- Anotação de cada palavra em uma sentença, com um part-of-speech (marcador)
- Nível mais baixo da análise sintática
  - substantivo, verbo, pronome, preposição, advérbio, conjunção, artigos…
- Abordagens: utilizando conjunto de regras (dicionários) e modelos do tipo sequencial HMM ou LSTM.
- Desafios:
  - Português: morro (substantivo) e morro (verbo)
  - Inglês: object (substantivo) e object (verbo)
- Exemplo em Português: MacMorpho

PUC Minas

Part-of-speech Tagging



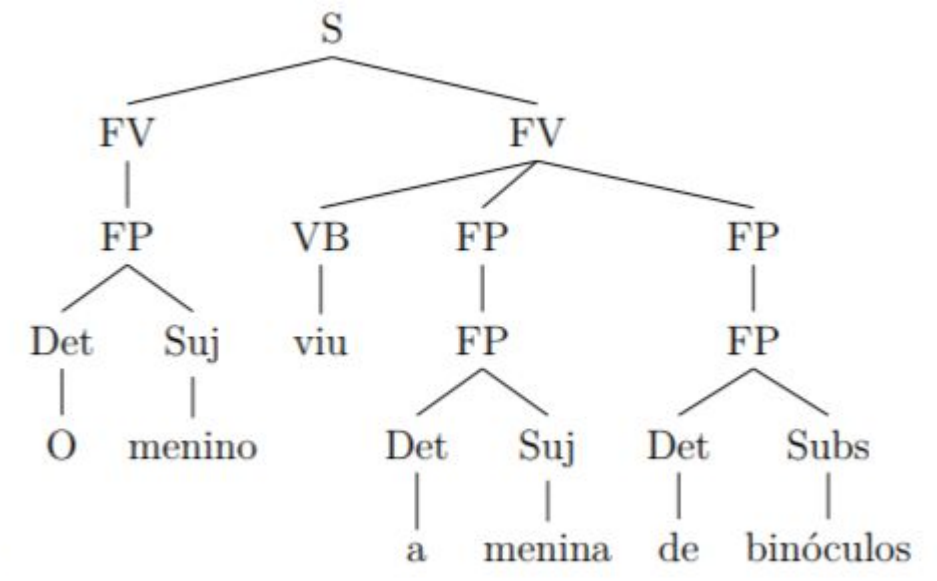
Parsing

Parsing

- É a análise automática de uma sentença com relação à sua estrutura
- Cria uma estrutura do tipo árvore
    - Tokens são representados como folhas, os nós internos agrupam tokens e a raiz define a sentença como um todo.

PUC Minas

Parsing



- S = Sentença
- FV = Frase Verbal
- FP = Frase preposicional
- VB = Verbo
- Det = Determinante
- Suj = Sujeito
- Subs = Substantivo

NER

Named Entity Recognition

- Tarefa importante para encontrar e classificar nomes no texto
- Reconhecimento de entidade nomeada é a tarefa de identificar nomes de pessoas, lugares, organizações etc. no texto.
- De grosso modo: qualquer coisa que possa ser referida com um nome próprio
- Normalmente, é estendido para incluir itens que não são entidades em si, incluindo **datas**, **horas** e outros tipos de **expressões temporais** e até expressões **numéricas**, como preços
- Exemplo para uma área NER Juridico LeNERBr
- Rotulos: ‘O’, ‘B-Rotulo’, ‘I-Rotulo’

Extração de tópicos

Extração de tópicos

- Modelagem de tópicos é o processo de identificação de tópicos em um conjunto de documentos
- Isso pode ser útil para o quê?
  - mecanismos de pesquisa
  - automação de atendimento ao cliente
  - seleção de features
  - recuperação informação de dados não estruturado
- Existem vários métodos, LSI (Latent Semantic Indexing); HDP (Hierarchical Dirichlet Process); LDA (Latent Dirichlet Allocation)
- Aprendizado não supervisionado.

Sumarização

Sumarização

- Produzir uma versão resumida de um texto e que contenha informações importantes ou relevantes para o usuário
- Realizar resumo de textos para reduzir tempo de leitura
- Cria resumos menos tendenciosos em relação a resumos feitos por humanos
- Úteis em sistemas de respostas às perguntas
- Entrada:
  - Um único documento
  - Múltiplos documentos
- Saída
  - Texto abstrato - criação de um novo texto
  - Extração do texto - extrai partes importantes do texto original

PUC Minas

Q&A

Question Answering

- Responder uma pergunta baseado em documentos que contêm respostas
  - automação de atendimento ao cliente
  - seleção de features
  - recuperação informação de dados não estruturado
- Pode ser separado por tipo de contexto usado:
  - Específico
  - domínio fechado
  - domínio aberto.
- Ou por técnica:
  - Sistemas de regras
  - Estatístico
  - Híbrido