Processamento

- **Representação vetorial de textos (*word embedding*)**:
  - **Binário**:
    - Cada palavra é atribuído um valor 1 ou 0 de acordo com sua presença ou ausência na sentença.

Documento 1 (D1)	Primeira sentença do corpus
Documento 2 (D2)	A segunda sentença é curta
Documento 3 (D3)	A terceira é curta
Documento 4 (D4)	A quarta sentença é a maior do corpus

Termos	"primeira"	"quarta"	"a"	"corpus"	"curta"	"do"	"maior"	"segunda"	"sentença"	"terceira"	"é"
D1	1	0	0	1	0	1	0	0	1	0	0
D2	0	0	1	0	1	0	0	1	1	0	1
D3	0	0	1	0	1	0	0	0	0	1	1
D4	0	1	1	1	0	1	1	0	1	0	1

Processamento

- ***Bag Of Words***:
  - O texto é simplificado para um vetor de palavras distintas e a respectiva contagem de ocorrência de cada uma delas.

Termos	"primeira"	"quarta"	"a"	"corpus"	"curta"	"do"	"maior"	"segunda"	"sentença"	"terceira"	"é"
D1	1	0	0	1	0	1	0	0	1	0	0
D2	0	0	1	0	1	0	0	1	1	0	1
D3	0	0	1	0	1	0	0	0	0	1	1
D4	0	1	2	1	0	1	1	0	1	0	1

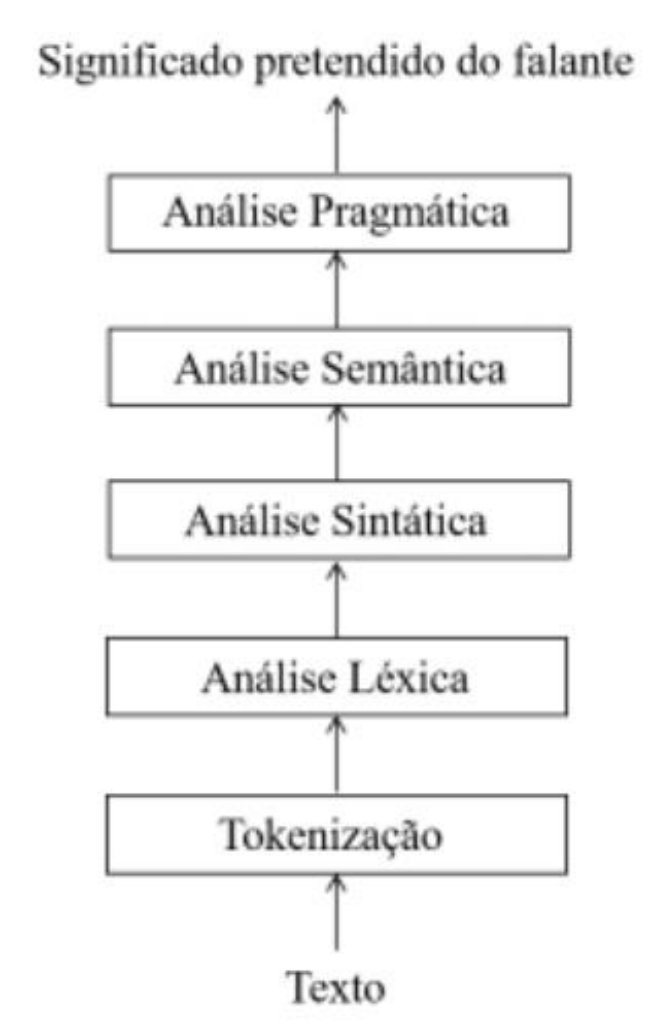
Processamento

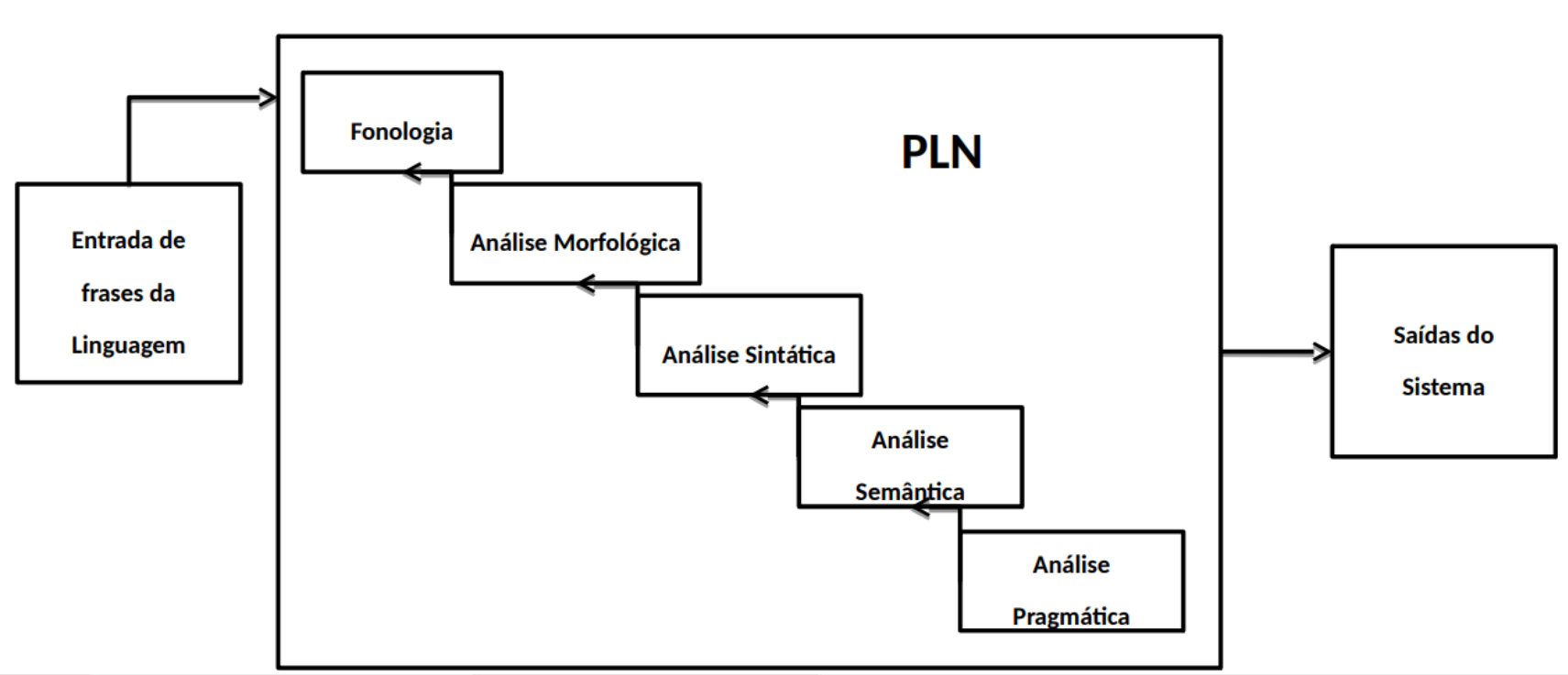
- **TF-IDF**:
  - *Term Frequency* mede o quão frequente um termo ocorre em um documento.
  - *Inverse Document Frequency* mede a raridade do termo para o documento.

$$tfidf_t = tf_{t,d} \times idf_t$$

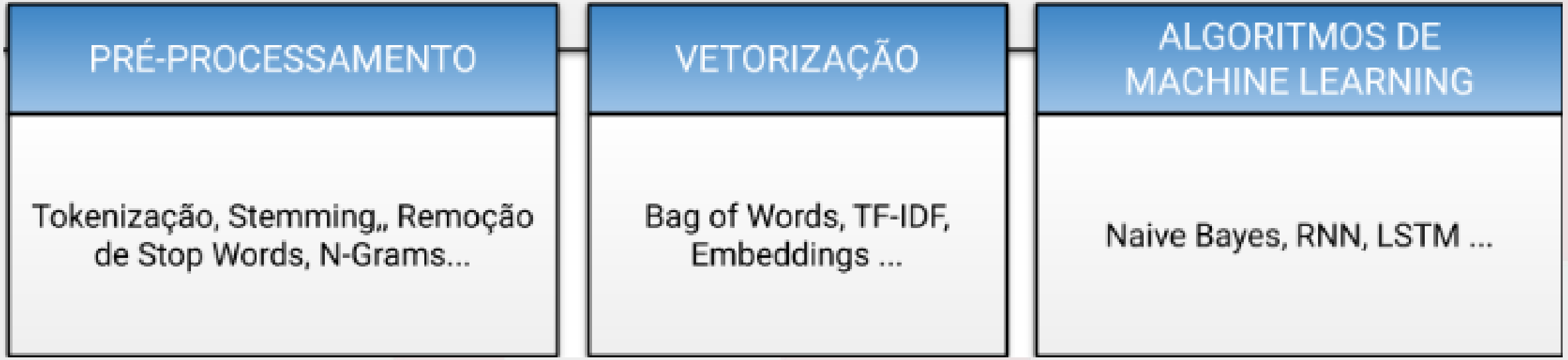
Termos	"primeira"	"quarta"	"a"	"corpus"	"curta"	"do"	"maior"	"segunda"	"sentença"	"terceira"	"é"
D1	0.614	0	0	0.484	0	0.484	0	0	0.392	0	0
D2	0	0	0.378	0	0.467	0	0	0.592	0.378	0	0.378
D3	0	0	0.408	0	0.505	0	0	0	0	0.640	0.408
D4	0	0.419	0.535	0.330	0	0.330	0.419	0	0.267	0	0.267

Processamento

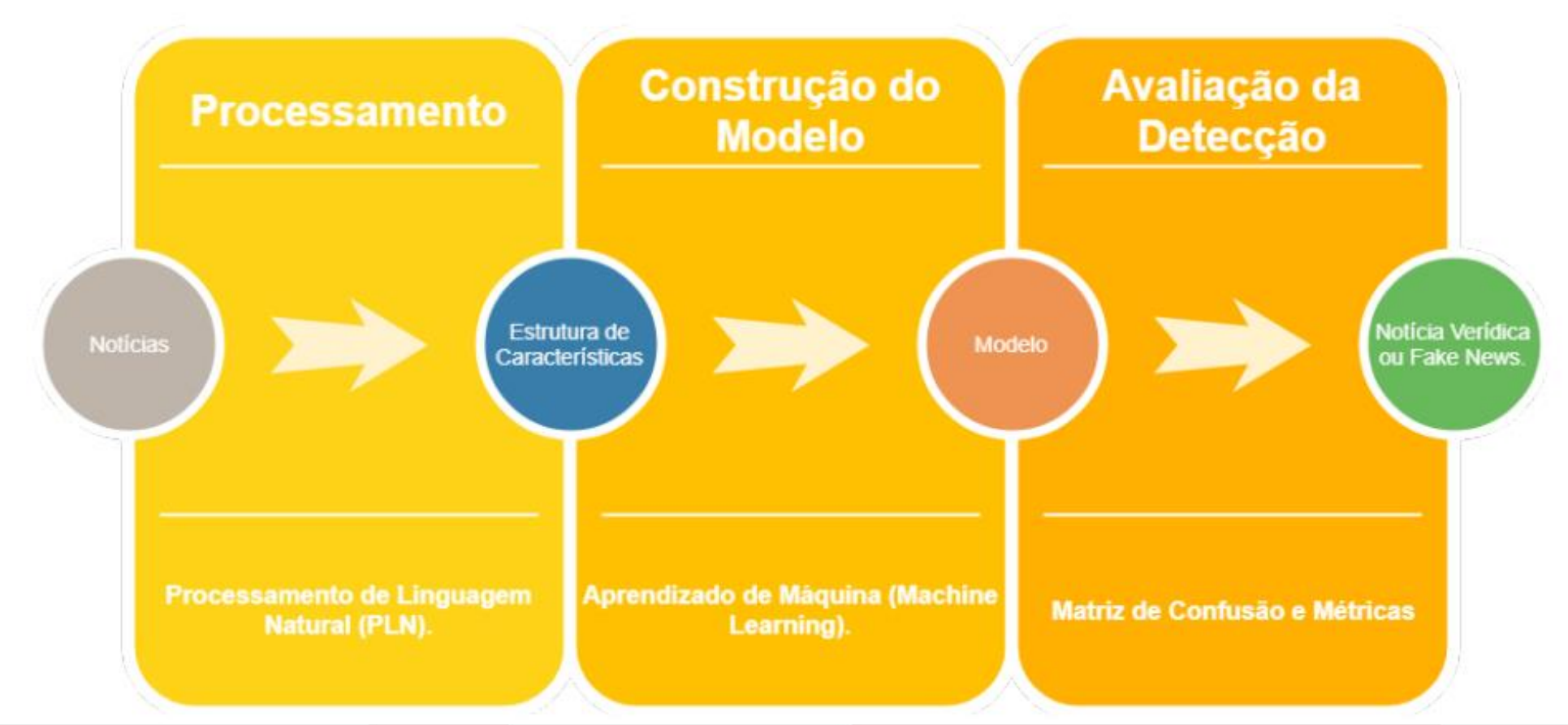




Processamento



Processamento



Questões de Concurso

- Prova: FGV - 2022 - MPE-SC - Analista de Dados e Pesquisa
- A atividade de classificação de documentos envolve um grande número de tarefas de processamento de linguagem natural, o que pode levar a dúvidas quanto a sua aplicação.
- A alternativa que contém apenas tarefas que sejam exemplos de classificação de documentos é:
- A análise de sentimento, tokenização;
- B POS-tagging, reconhecimento de entidades nomeadas;
- C filtragem de SPAM, análise de sentimento;
- D análise sintática, POS-tagging;
- E filtragem de stopwords, reconhecimento de linguagem.

Questões de Concurso

- Prova: FGV - 2022 - MPE-SC - Analista de Dados e Pesquisa
- A atividade de classificação de documentos envolve um grande número de tarefas de processamento de linguagem natural, o que pode levar a dúvidas quanto a sua aplicação.
- A alternativa que contém apenas tarefas que sejam exemplos de classificação de documentos é:
- A análise de sentimento, tokenização;
- B POS-tagging, reconhecimento de entidades nomeadas;
- C filtragem de SPAM, análise de sentimento;
- D análise sintática, POS-tagging;
- E filtragem de stopwords, reconhecimento de linguagem.

Questões de Concurso

- Prova: FGV - 2022 - CGU - Auditor Federal de Finanças e Controle - Tecnologia da Informação
- Durante a elaboração de um sistema de busca de informações biomédicas, foi construído um modelo de linguagem vetorial não contextual para estimar relações de similaridade semântica necessárias para comparação entre queries e documentos. Entretanto, verificou-se nos testes iniciais que o desempenho do modelo ficou insatisfatório, devido a muitos termos técnicos presentes nos documentos testados, que não haviam sido incorporados ao modelo. Para aliviar esse problema, uma tarefa de processamento do texto e seu estágio correspondente no processamento de linguagem natural que poderiam ser aplicados na construção do modelo são, respectivamente:
- A Word embedding; Análise léxica;
- B Lematização; Análise sintática;
- C Decomposição morfológica; Análise léxica;
- D Word embedding; Análise semântica;
- E Decomposição morfológica; Análise sintática.

Questões de Concurso

- Prova: FGV - 2022 - CGU - Auditor Federal de Finanças e Controle - Tecnologia da Informação
- Durante a elaboração de um sistema de busca de informações biomédicas, foi construído um modelo de linguagem vetorial não contextual para estimar relações de similaridade semântica necessárias para comparação entre queries e documentos. Entretanto, verificou-se nos testes iniciais que o desempenho do modelo ficou insatisfatório, devido a muitos termos técnicos presentes nos documentos testados, que não haviam sido incorporados ao modelo. Para aliviar esse problema, uma tarefa de processamento do texto e seu estágio correspondente no processamento de linguagem natural que poderiam ser aplicados na construção do modelo são, respectivamente:
- A Word embedding; Análise léxica;
- B Lematização; Análise sintática;
- C Decomposição morfológica; Análise léxica;
- D Word embedding; Análise semântica;
- E Decomposição morfológica; Análise sintática.