

Análise de Redes Sociais

Processamento de Língua Natural

Prof. Luciano Antonio Digiampietri
EACH-USP

Processamento de Língua Natural

Processamento de Língua Natural (PLN) (*Natural Language Processing - NLP*) é um campo interdisciplinar envolvendo ciência da computação, inteligência artificial e linguística focado na interação entre computadores e línguas humanas (naturais), em especial no processamento ou produção de dados relacionados a estas línguas (textos, áudios, vídeos etc.).

PLN e ARS

Apesar de não ser algo muito utilizado no início da ARS, atualmente diversos problemas de ARS são resolvidos utilizando métodos de Processamento de Língua Natural (PLN).

PLN e ARS

Apesar de não ser algo muito utilizado no início da ARS, atualmente diversos problemas de ARS são resolvidos utilizando métodos de Processamento de Língua Natural (PLN).

Em ARS, a aplicação de PLN atual mais típica é no processamento de textos (*posts*) publicados em Redes Sociais Online.

PLN e ARS

No contexto de nossa disciplina, o enfoque de PLN será para auxiliar nas atividades de mineração de texto (*Text Mining*), ou seja, para auxiliar na descoberta de conhecimento utilizando, também, informações textuais.

Descoberta de Conhecimento

A maioria dos algoritmos de mineração de dados recebe como entrada dados estruturados, no formato de uma tabela em que cada linha representa uma instância e cada coluna representa um atributo (características ou variável).

Descoberta de Conhecimento

A maioria dos algoritmos de mineração de dados recebe como entrada dados estruturados, no formato de uma tabela em que cada linha representa uma instância e cada coluna representa um atributo (características ou variável).

Para a utilização de textos por estes algoritmos é necessário que os textos sejam convertidos/transformados para um formato estruturado.

Atividades/Tarefas Básicas

POS tagging (*Part-of-speech tagging*):
substituição das palavras por suas classes gramaticais (substantivos, verbos, adjetivos, advérbios etc.).

Atividades/Tarefas Básicas

POS tagging (*Part-of-speech tagging*):

substituição das palavras por suas classes gramaticais (substantivos, verbos, adjetivos, advérbios etc.).

Remoção de *stop-words* (ou “palavras vazias”):

remoção de palavras que são insignificantes no contexto da análise que será realizada (por exemplo, artigos e preposições).

Atividades/Tarefas Básicas

Conversão de todas as letras para caixa baixa,
remoção de acentos e caracteres especiais.

Atividades/Tarefas Básicas

Conversão de todas as letras para caixa baixa, remoção de acentos e caracteres especiais.

Lematização (*lemmatization*): ato de retirar flexões/conjugações de palavras de forma a retorná-las a um formato padronizado (de dicionário).

Atividades/Tarefas Básicas

Conversão de todas as letras para caixa baixa, remoção de acentos e caracteres especiais.

Lematização (*lemmatization*): ato de retirar flexões/conjugações de palavras de forma a retorná-las a um formato padronizado (de dicionário).

Radicalização (*stemming*): ato de reduzir palavras à sua raiz/radical.

Atividades/Conceitos Básicos

Tokenização (*Tokenization*): separação do texto em unidades básicas (palavras, fonemas, caracteres etc.).

Atividades/Conceitos Básicos

Tokenização (*Tokenization*): separação do texto em unidades básicas (palavras, fonemas, caracteres etc.).

n-gramas: sequência contínua de itens (por exemplo, *n tokens*).

Atividades/Conceitos Básicos

Tokenização (*Tokenization*): separação do texto em unidades básicas (palavras, fonemas, caracteres etc.).

n-gramas: sequência contínua de itens (por exemplo, *n tokens*).

BOW (*bag-of-words*) - saco de palavras: equivalente ao unigrama de palavras.

Representação

Os textos podem ser representados, de forma estruturada, como conjuntos de n-gramas de *tokens* (caracteres, fonemas, palavras, classes gramaticais etc.).

Representação

Os textos podem ser representados, de forma estruturada, como conjuntos de n-gramas de *tokens* (caracteres, fonemas, palavras, classes gramaticais etc.).

Cada linha corresponde a um texto (por exemplo, uma postagem) e cada coluna representa um n-grama (por exemplo, uma palavra).

Representação

Cada célula indica que o texto contém ou não o respectivo *token* (ou conjunto de tokens):

Representação

Cada célula indica que o texto contém ou não o respectivo *token* (ou conjunto de tokens):

De forma binária;

Representação

Cada célula indica que o texto contém ou não o respectivo *token* (ou conjunto de tokens):

- De forma binária;

- Contagem (valor absoluto ou relativo);

Representação

Cada célula indica que o texto contém ou não o respectivo *token* (ou conjunto de tokens):

- De forma binária;

- Contagem (valor absoluto ou relativo);

- TF-IDF (*Term frequency–inverse document frequency*).

Representação

Exemplos:

Representação

Exemplos:

Introdução à Análise de Algoritmos

Análise de Redes Sociais

Representação

Exemplos: **Letras Minúsculas**

introdução à análise de algoritmos

análise de redes sociais

Representação

Exemplos: Remoção de Acentos

introducao a analise de algoritmos

analise de redes sociais

Representação

Exemplos: Remoção de Stopwords

introducao analise algoritmos

analise redes sociais

Representação

Exemplos: **Lematização**

introducao analise algoritmo

analise rede social

Representação

Exemplos: POS tagging

substantivo substantivo substantivo

substantivo substantivo substantivo

Representação

Exemplos:

introducao analise algoritmo

analise rede social

Representação

Exemplos: unigramas de palavras

('introducao', 'analise', 'algoritmo')

('analise', 'rede', 'social')

Representação

Exemplos: bigramas de palavras

('introducao analise', 'analise algoritmo')

('analise rede', 'rede social')

Representação

Exemplos: unigramas de palavras

('introducao', 'analise', 'algoritmo')

('analise', 'rede', 'social')

Representação

BOW / Unigramas de palavras

	algoritmo	analise	introducao	rede	social
Introdução à Análise de Algoritmos					
Análise de Redes Soci- ais					

Representação

BOW / Unigramas de palavras

	algoritmo	analise	introducao	rede	social
Introdução à Análise de Algoritmos	true	true	true	false	false
Análise de Redes Soci- ais	false	true	false	true	true

Representação Binária

Representação

BOW / Unigramas de palavras

	algoritmo	analise	introducao	rede	social
Introdução à Análise de Algoritmos	1	1	1	0	0
Análise de Redes Soci- ais	0	1	0	1	1

Representação por Contagem

Representação

BOW / Unigramas de palavras

	algoritmo	analise	introducao	rede	social
Introdução à Análise de Algoritmos	0,632	0,449	0,632	0	0
Análise de Redes Sociais	0	0,449	0	0,632	0,632

Representação usando TF-IDF

Representação

Word embedding: estratégia para representar cada palavra, tipicamente, na forma de um arranjo de números reais de forma a tentar representar o “significado” da palavra.

Isto é, espera-se que palavras com significados semelhantes tenham representações vetoriais próximas.

Word embedding

Os modelos são tipicamente construídos com base em Redes Neurais.

Word embedding

Os modelos são tipicamente construídos com base em Redes Neurais.

CBOW (Continuous Bag of Words Model): tenta prever uma palavra com base em seus vizinhos.

Word embedding

Os modelos são tipicamente construídos com base em Redes Neurais.

CBOW (Continuous Bag of Words Model): tenta prever uma palavra com base em seus vizinhos.

Skip Gram: tenta prever os vizinhos de uma palavra.

PLN - Algumas aplicações

PLN - Algumas aplicações

Reconhecimento de Entidade Nomeada (*Named Entity Recognition (NER)*): identificação de quais itens do texto se referem a entidades nomeadas (nomes de pessoas, lugares, organizações etc.) e quais são os tipos dessas entidades.

PLN - Algumas aplicações

Reconhecimento de Entidade Nomeada (*Named Entity Recognition (NER)*): identificação de quais itens do texto se referem a entidades nomeadas (nomes de pessoas, lugares, organizações etc.) e quais são os tipos dessas entidades.

Análise de Sentimentos (*Sentiment Analysis*): pode envolver a identificação dos sentimentos presentes numa sentença e/ou de sua polaridade (positiva, negativa ou neutra).

PLN - Algumas aplicações

Detecção de Posicionamento (*Stance Detection*):

identificação do posicionamento do texto em relação a algo (a favor, contra, neutro/sem posicionamento).

PLN - Algumas aplicações

Detecção de Posicionamento (*Stance Detection*):

identificação do posicionamento do texto em relação a algo (a favor, contra, neutro/sem posicionamento).

Extração de Termos/Tópicos (*Terminology/Topic extraction/Topic Modeling*): identificação dos principais termos de uma sentença ou do tópico/assunto que está sendo discutido.

PLN - Algumas aplicações

Desambiguação (*Word-Sense Disambiguation (WSD)*):

identificação de significado/sentido correto de uma da palavra com base em seu contexto.

PLN - Algumas aplicações

Desambiguação (*Word-Sense Disambiguation (WSD)*):

identificação de significado/sentido correto de uma da palavra com base em seu contexto.

Sumarização Automática (*Text Summarization*): resumo automático de um texto.

PLN - Algumas aplicações

Desambiguação (*Word-Sense Disambiguation (WSD)*): identificação de significado/sentido correto de uma da palavra com base em seu contexto.

Sumarização Automática (*Text Summarization*): resumo automático de um texto.

Tradução Automática (*Machine Translation*): tradução automática de texto entre idiomas.

PLN - Algumas aplicações

Geração de Língua Natural (*Natural-Language Generation* (NLG)): produção de textos em língua natural com base, por exemplo, em informações estruturadas.

PLN - Algumas aplicações

Geração de Língua Natural (*Natural-Language Generation (NLG)*): produção de textos em língua natural com base, por exemplo, em informações estruturadas.

Respostas a Perguntas (*Question Answering*): fornecer respostas a questões formuladas em língua natural (usada em *chatbots*, por exemplo).

Referências

- 1 JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing. Prentice Hall, 2008.
- 2 RIBEIRO, Filipe N. et al. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science, v. 5, n. 1, p. 1-29, 2016.

Análise de Redes Sociais

Processamento de Língua Natural

Prof. Luciano Antonio Digiampietri
EACH-USP