Processamento de Linguagem Natural (PLN) – aula IV

por: Rafael Stoffalette João

Data: 24/10/2020

UNIversidade Paulista (UNIP) – Araçatuba
Pós-graduação em Data Science e Machine Learning

Agenda da disciplina

Módulo 01 - Introdução ao Processamento de Linguagem Natural
Módulo 02 - Análise Semântica e Morfológica

Módulo 03 - Processo de Mineração de Texto
Módulo 04 - Modelagem Estatística da Linguagem

Módulo 05 - Word Embeddings
Módulo 06 - Classificação de Texto

Módulo 07 - Extração de Informação
Módulo 08 - Geração de Resumo

Módulo 09 - Análise de Sentimentos
Módulo 10 - Deep Learning aplicado ao Processamento de Linguagem Natural

Antes, entretanto...

Por que é tão difícil criar um robô que se comporta como uma pessoa?

Talvez a resposta mais direta é que nós (humanos) agimos com tanta "naturalidade" que é difícil identificar os passos realizados.

Antes, entretanto...

Por que é tão difícil criar um robô que se comporta como uma pessoa?

Semântica e sintaxe devem ser analisadas
Botar a bota significa o que pra você?

Talvez a resposta mais direta é que nós (humanos) agimos com tanta "naturalidade" que é difícil identificar os passos realizados.

Antes, entretanto...

As frases:

Vou viajar para a praia no final de semana

E

Viajo pra praia no fim de semana

são facilmente reconhecidas por um ser humano, mas o computador precisa reduzir as palavas à sua forma mais simples para compreender.

Viajo e viajar são reduzidos para “viaj”

Fim e final são sinônimos de “fim”

...

Antes, entretanto...

Outro passo é identificar a gramática (análise léxica) das palavras.

Dessa forma, é possível identificar o contexto das frases:

Ligar para China;	verbo + pronome (entidade)
Ligar o carro;	verbo + substantivo
A liga metálica é resistente.	substantivo + substantivo

Antes, entretanto...

Outro passo é identificar a relação entre as palavras

O Corpus é uma estratégia muito importante nas tarefas de PLN pois relaciona frases de contextos que são comparadas às frases analisadas.

"Ligar para China" é identificada como do contexto de telecomunicações pois o Corpus deste assunto contém sentenças parecidas (gramaticalmente e por contagem de ocorrências)

Antes, entretanto...

Enfim...

...Qualquer atividade que visa processar o texto para facilitar a compreensão pelo computador é bem vinda.

Normalização, stop words removal, tokenização, stemmização, lemmatização, bag of words, pattern matching, etc...

No encontro de hoje...

Outras bibliotecas para Processamento de Linguagem Natural.

Vamos conhecer outras bibliotecas que permitem a manipulação de linguagem natural e que combinadas compõem ferramentas extremamente potentes.

Bibliotecas Python

Discutimos, então>

NLTK;

Spacy;

Scrapy;

BeautifulSoup;

Selenium;

Chatterbot;

...

Entretanto existem outras bibliotecas que merecem ser exploradas...

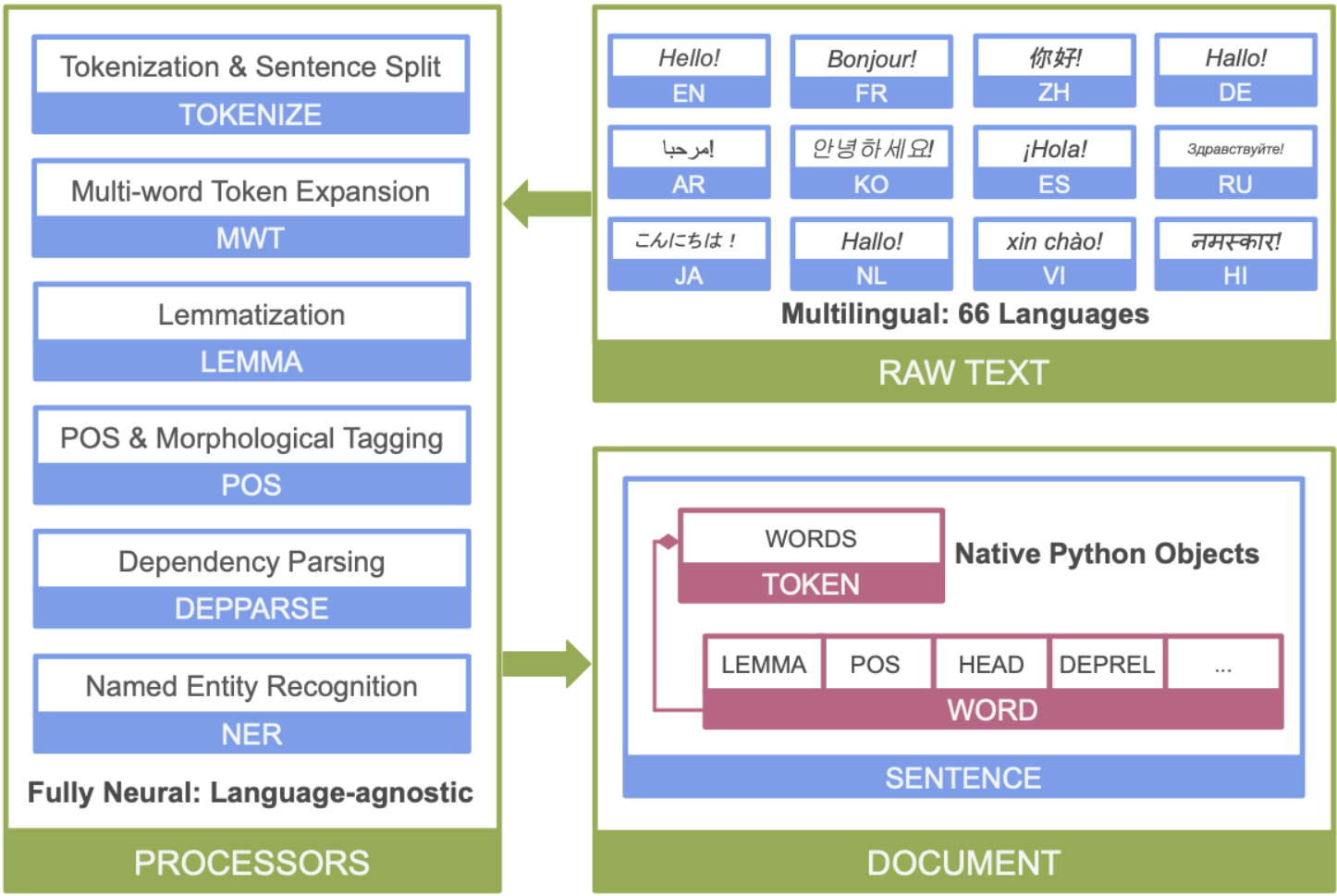
Bibliotecas Python - Stanza

Stanza é uma biblioteca que contém uma grande quantidade de linguagens (60) e que realiza tarefas simples como a análise sintática e reconhecimento de entidades.

Utiliza a estrutura pipeline (tarefas sequenciais) e redes neurais para o processamento dos textos

https://stanfordnlp.github.io/stanza/available_models.html

Bibliotecas Python - Stanza



Bibliotecas Python - textBlob

https://textblob.readthedocs.io/en/dev/index.html

textBlob é outra biblioteca para processamento de texto que independe da linguagem que está sendo considerada.

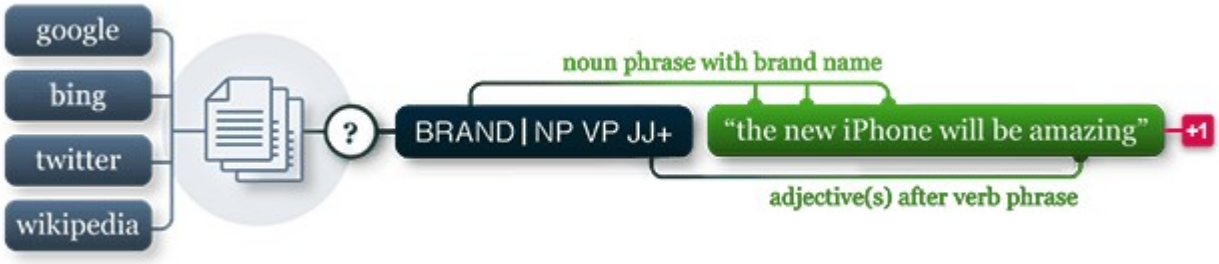
Por que independe da linguagem?

Bibliotecas Python - Pattern

https://github.com/clips/pattern/

Pattern é, talvez, a biblioteca mais completa que se pode encontrar em Python.

É reconhecido como um módulo de programação para mineração web completo.



Bibliotecas Python - Pattern

Pattern tem suporte para

- mineração de dados (Google, Twitter e Wikipedia API, um web crawler, ...);
- processamento de linguagem natural;
- machine learning (agrupamento, classificação); e
- visualization.

Bibliotecas Python - Pattern

É composto por vários módulos:

pattern.web

pattern.db

pattern.en | es | de | fr | it | nl

pattern.search

pattern.vector

pattern.graph

Bibliotecas Python - Pattern

É composto por vários módulos:

pattern.web

Contém vínculo com APIs Google, Twitter, Facebook, Gmail, Bing, Wikipedia, Flickr, …

pattern.db

Implementa um web crawler

pattern.en | es | de | fr | it | nl

pattern.search

pattern.vector

pattern.graph

Bibliotecas Python - Pattern

É composto por vários módulos:

pattern.web

pattern.db

pattern.en | es | de | fr | it | nl

pattern.search

pattern.vector

pattern.graph

Contém módulos de conexão a bancos de dados

Bibliotecas Python - Pattern

É composto por vários módulos:

pattern.web

pattern.db

pattern.en | es | de | fr | it | nl

pattern.search

pattern.vector

pattern.graph

Implementa um processamento de linguagem natural.

Infelizmente, somente para essas linguagens

Bibliotecas Python - Pattern

É composto por vários módulos:

pattern.web

pattern.db

pattern.en | es | de | fr | it | nl

pattern.search

pattern.vector

pattern.graph

Implementa algoritmos de busca e aquisição de informação em textos.

Como casamento de padrões

Bibliotecas Python - Pattern

É composto por vários módulos:

pattern.web

pattern.db

pattern.en | es | de | fr | it | nl

pattern.search

pattern.vector

pattern.graph

Implementa algoritmos vetoriais, como o SVM para classificação;

Valores de TF-IDF e distâncias entre tokens.

Bibliotecas Python - Pattern

É composto por vários módulos:

pattern.web

pattern.db

pattern.en | es | de | fr | it | nl

pattern.search

pattern.vector

pattern.graph

Estrutura de dados em grafos para representar as relações e ligações entre elementos.

Bibliotecas Python - SpeechRecognition

Uma das tarefas que podem ser executadas com o PLN é o reconhecimento de fala.

A biblioteca mais conhecida é a SpeechRecognition

https://github.com/Uberi/speech_recognition#readme

Watson speech to text

https://speech-to-text-demo.ng.bluemix.net/

Bibliotecas Python - SpeechRecognition

Mas até agora só trabalhos com textos...

A linguagem natural pode ser representada por diversas formas...

Vimos isso láaa no começo da disciplina...

Libras;

Fala;

Escrita;

Texto digitado;

...

Outras bibliotecas - Recomendação

Polyglot

CoreNLP

Gensim

PyNLPI

Quepy

...

Nossa atividade final

Conclusão da disciplina:

Divisão em trios para por em prática o nosso aprendizado!

Via microfone, ou áudio gravado, é possível reconhecer o texto de uma fala e...?

- Normalizar o texto;
- Construir um analisador de sentimentos;
- Tokenizar;
- Realizar o stop word removal;
- Construir a bag of words;
- Identificar entidades e tokens mais importantes;
- Realizar uma busca no Twitter/crawler? OU buscar links a partir de um link inicial (crawler) e armazenar em um banco de dados.