UNIVERSIDADE ESTADUAL DE MARINGÁ
DEPARTAMENTO DE INFORMÁTICA - C56

INTRODUÇÃO A INTELIGENCIA ARTIFICIAL

PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

Gabriel Thiago - RA10774
Sergio Alvarez - RA115735

Prof. Dr. Wagner Igarashi

RESUMO DE TÓPICOS

O QUE VAMOS DISCUTIR

- Descrição do problema
- Informações
- Extração dos top 10 termos
- Extração das informações
- Execuções
- Analise dos resultados
- Conclusão

Universidade Estudual de Maringá
Departamento de Informática - C56
Introdução a Inteligencia Artificial

PROBLEMA

INFORMAÇÕES

- Área: Processamento de linguagem natural (PLN)
- Acesso a Artigos Científico do portal
  - https://ieeexplore.ieee.org/Xplore/home.jsp
- Extração de informações
- Escrita resultantes em arquivo
- Linguagem utilizada: python3

EXTRAÇÃO DOS 10 TERMOS MAIS CITADOS

LEITURA DO PDF

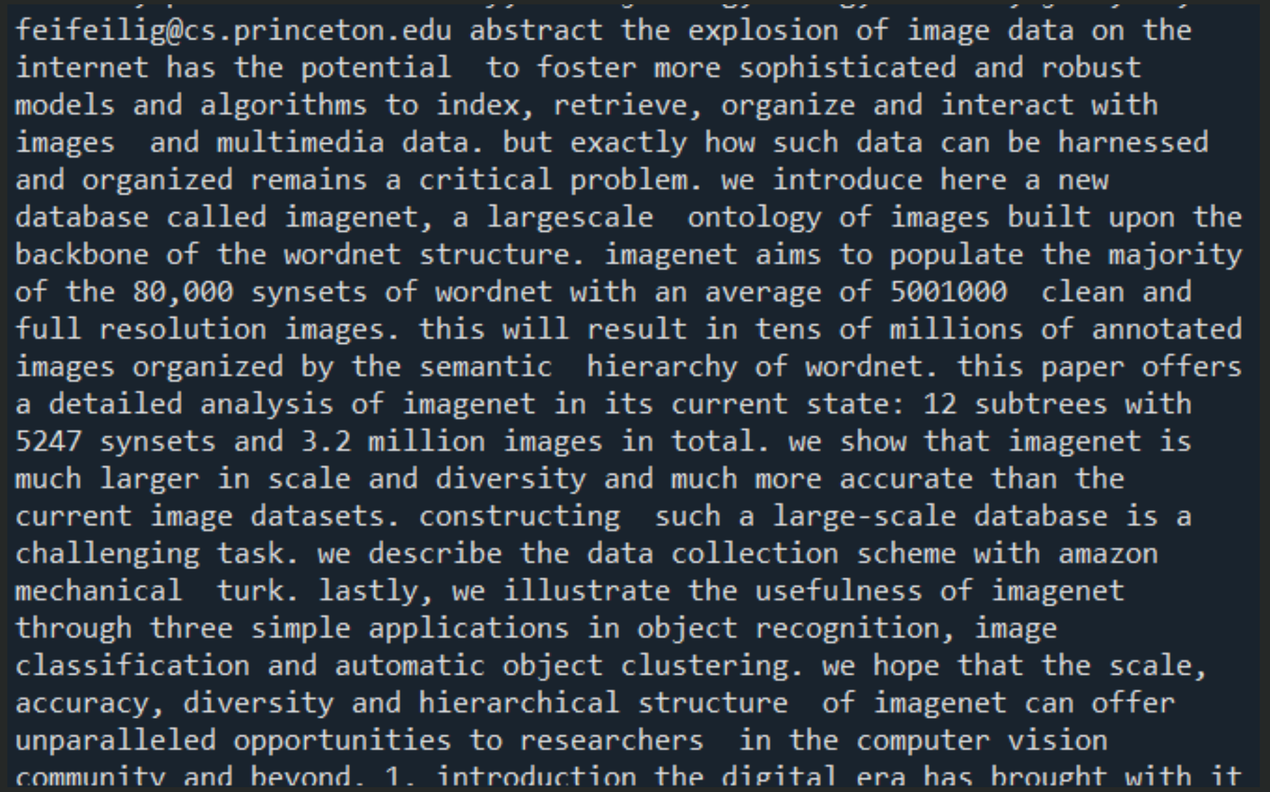
- Leitura do pdf, por meio da biblioteca pyPDF2
- Organizar o texto para não termos palavras cortadas
- Remoção de vírgulas e pontuações por meio de regex: "text = re.sub(r'[^a-zA-Z\s]', '', text)"
- Ignoramos a referência por meio de regex
  referencias_regex = re.compile(r"References[\s\S]*$", re.IGNORECASE | re.MULTILINE)

```
The explosion of image data on the Internet has the po-
tential to foster more sophisticated and robust models and
algorithms to index, retrieve, organize and interact with im-
ages and multimedia data. But exactly how such data can
be harnessed and organized remains a critical problem. We
introduce here a new database called “ImageNet”, a large-
scale ontology of images built upon the backbone of the
WordNet structure. ImageNet aims to populate the majority
of the 80,000 synsets of WordNet with an average of 500-
1000 clean and full resolution images. This will result in
tens of millions of annotated images organized by the se-
mantic hierarchy of WordNet. This paper offers a detailed
analysis of ImageNet in its current state: 12 subtrees with
5247 synsets and 3.2 million images in total. We show that
ImageNet is much larger in scale and diversity and much
more accurate than the current image datasets. Construct-
ing such a large-scale database is a challenging task. We
describe the data collection scheme with Amazon Mechan-
ical Turk. Lastly, we illustrate the usefulness of ImageNet
```

EXTRAÇÃO DOS 10 TERMOS MAIS CITADOS

PRÉ-PROCESSAMENTO

- Remoção dos hifens
  - Palavras em que separadas com a quebra de linha
- Remoção de espaços duplos
- Normalização
  - palavras todas em minuscula
  - unicodedata -> utf-8
  - Substituicao de abreviação
    - Ex.: Fig. -> figure



EXTRAÇÃO DOS 10 TERMOS MAIS CITADOS

PRÉ-PROCESSAMENTO

- Remoção de '\n' (quebras de linahs
- Tokenização do texto
- Remoção das stop-words utilizando a biblioteca de processamento de linguagem natural nltk
- Alguns exemplos de stop-words são: "the", "and", "a", "an", "in".

21	str	8	abstract
22	str	6	deeper
23	str	6	neural
24	str	8	networks
25	str	3	are
26	str	4	more
27	str	7	difcult
28	str	2	to
29	str	5	train
30	str	2	we

19	str	8	abstract
20	str	6	deeper
21	str	6	neural
22	str	8	networks
23	str	7	difcult
24	str	5	train
25	str	7	present
26	str	8	residual
27	str	8	learning
28	str	9	framework

EXTRAÇÃO DOS 10 TERMOS MAIS CITADOS

CONTAGEM

- Contagem simples dos tokens já filtrados

```
word_counts = collections.Counter(filtered_tokens)
most_common_words = word_counts.most_common(10)
```

```
risk: 49
cyber: 46
manufacturing: 39
model: 36
energy: 35
attack: 31
quantication: 31
layer: 31
impact: 30
wafer: 29
```

EXTRACAO DE INFORMAÇÕES

STEMMING DO TEXTO

- Para facilitar a montagem da ontologia
- Transforma a palavra deixando apenas sua raiz

```
IPdb [2]: !next
the main conclusions are summarized in .


IPdb [2]: !next
the main conclus are summar in .
```

```
("objective, objectives, aim, aims, purpose, purposes, article, articles, study")
```

```
object , object , aim , aim , purpos , purpos , articl , articl , studi
```

EXTRACAO DE INFORMAÇÕES

MODELAGEM - OBJETIVOS

Após a stemmização:

- object + studi or research or articl
- purpose + studi or research or articl
- aim + studi or research or articl
- goal + studi or research or articl

A ideia é encontrar frases como:

- "This paper aims to..."
- "The goal of this article is..."
- "The purpose of this research is.."

```
Objetivo: the objective of this study is to establish and assess performance metrics by
exploring the modeling and simulation of semiconductor wafer fab manufacturing processes,
with the goal of providing an energy quantication framework for semiconductor manufacturing
in the presence of a cyber attack.
```

The objective of this study is to establish and assess performance metrics by exploring the modeling and simulation of semiconductor wafer fab manufacturing processes, with the goal of providing an energy quantification framework for semiconductor manufacturing in the presence of a cyber attack. This study also aims to provide a cyber risk assessment framework for semiconductor manufacturing. This analysis in-

EXTRACAO DE INFORMAÇÕES

MODELAGEM - PROBLEMA

Após a stemmização:

- problem + studi or research or articl
- issue + studi or research or articl
- challeng + studi or research or articl

A ideia é encontrar frases como:

- "This paper address the challenge of ..."
- "This research investigates the problem of ..."
- "The issue investigated in this article is.."

```
Problema:   the task of actually fabricating large memristor
arrays is still very much a research challenge for instance,
prezioso and others use a fabricated memristor array to
demonstrate a linear classifier
```

tions. The task of actually fabricating large memristor arrays is still very much a research challenge; for instance, Prezioso *et al.* [112] use a fabricated 12 × 12 memristor array to demonstrate a linear classifier.

EXTRACAO DE INFORMAÇÕES

MODELAGEM - METODOLOGIA

Após a stemmização:

- methodolog + studi or research or model or example
- conduct + studi or research or model or exampl
- employ+ studi or research or model or exampl
- util + studi or research or model or exampl

A ideia é encontrar frases como:

- "The methodology of this study ..."
- "In this research, we employed..."
- "In this model, we utilized..."

```
Metodologia:  in order to solve these problems, we conduct a study of
the threelayered framework for recommending security requirements
through goaloriented risk assessment using a problem domain ontology
pdo
```

holistic analysis of STS due to heterogeneity characteristics. In order to solve these problems, we conduct a study of the three-layered framework for recommending security requirements through goal-oriented risk assessment using a Problem Domain Ontology (PDO). By using this framework, we demonstrate how

EXTRACAO DE INFORMAÇÕES

MODELAGEM - CONTRIBUIÇÃO

Após a stemmização:

- contribut + studi or research or articl

A ideia é encontrar frases como:

- "This study makes a significant contribution to..."
- "This article makes a contribution to..."
- "In this paper, we contribut with..."

```
Contribuição: this study contributes to the literature by integrating cyberattack into
semiconductor manufacturing simulation modeling to analyze the impact of cyberattack on
energy quantication.
```

energy quantification. This study contributes to the literature by integrating cyber-attack into semiconductor manufacturing simulation modeling to analyze the impact of cyber-attack on energy quantification.

EXEMPLOS DE EXECUÇÕES

ARTIGOS SELECIONADOS

Utilizando::

- Artigos IEEE

Temas

- Machine learning
- Networking
- Computing

-- Exemplos práticos

- ANALYSIS_OF_THE_IMPACT.pdf
- Analytical_Study_of_Cognitive_Layered_Approach_for_Understanding_Security_...
- Attributes_and_Entrepreneurial.pdf
- Efficient_Processing_of_Deep_Neural_Networks_A_Tutorial_and_Survey.pdf
- ImageNet_A_large-scale_hierarchical_image_database.pdf
- Internet_of_Things_Platform.pdf
- Natural_Rivers_Longitudinal_Dispersion_Coefficient_Simulation_Using_Hybrid_S...
- Person_Transfer_GAN_to_Bridge_Domain_Gap_for_Person_Re-identification.pdf
- Search_and_Evaluation_of_Coevolving_Problem_and_Solution_Spaces_in_a_Co...
- Technology_Roadmapping.pdf

EXECUÇÕES

E EXTRAÇÃO PARA ARQUIVO

Execução:

- $ python3 top_termos.py
- $ python3 extracao_info.py

- Arquivo de saida: "output.txt"
- Arquivo com as referencias: "referencias.txt"

D: > UEM > 4_ano > IA > IA1_IA2 > IA1 > Trabalho2 > output.txt

```
1  Search_and_Evaluation_of_Coevolving_Problem_and_Solution_Spaces_in_a_Complex_Healthcare_Design_Science_Research_Project.pdf
2  introduction over the last decade, the authors have been conducting design science research dsr with the goal of creating a smartphone app to support patients with advan
3  march search and evaluation of coevolving problem and solution spaces in a complex healthcare design science research project diane m;;
4  abstract this article employs design ethnography to study the design process of a design science research dsr project conducted over eight years;;
5  studying the dsr design process generates the knowledge that research project managers need for managing and guiding a dsr project, and contributes to our knowledge of t
```

CONCLUSÃO

- Realizamos a combinação de técnicas de processamento de linguagem natural para realizar a contagem dos top 10 termos mais frequentes de um artigo
- Através de técnicas de processamento de linguagem natural e definição de ontologias, conseguimos extrair as informações de objetivo, problema, contribuição e metodologia de textos científicos
- A prática executada permitiu ampliar nosso conhecimento em processamento de linguagem natural

REFERENCIAS

Documentação natural language toolkit NKLT
Documentação PyPDF2 <https://pypdf2.readthedocs.io/en/latest/user/cropping-and-transforming.html>
Processamento de linguagem natural para iniciantes - https://www.insightlab.ufc.br/pln-processamento-de-linguagem-natural-para-iniciantes/
Testes regex <https://regex101.com/>
Processamento de linguagem natural: o que é e para que serve <https://blog.xpeducacao.com.br/processamento-de-linguagem-natural-pnl/>
Artigos cientificos IEEE <https://ieeexplore.ieee.org/Xplore/home.jsp>
Slides IA - Processamento de linguagem natual

Universidade Estudual de Maringá
Departamento de Informática - C56

OBRIGADO!

GABRIEL THIAGO
SERGIO ALVAREZ
RA<107774, 115735>@UEM.BR

INTRODUÇÃO A INTELIGENCIA ARTIFICIAL

PROF. DR. WAGNER IGARASHI