MINERAÇÃO DE REPOSITÓRIOS PARA ANÁLISE DE CICLOS DE SOFTWARE

Aluno - Ronaldo Rubens Gesse Junior - 201026937
Orientador - Prof. Dr. Higor Amario de Souza

Ciência da Computação - UNESP/Bauru

unesp

Introdução

Introdução

- Mineração de repositórios como Github e Gitlab e API's como o Google Trends.
- Foco em *frameworks* e bibliotecas.
- Detectar tendências de alta ou baixa na manutenibilidade e interesse em ferramentas.
- Fornecer *insights* valiosos para desenvolvedores, gerentes de projeto e pesquisadores que utilizam essas ferramentas, decidindo pela adoção ou substituição de determinados softwares. Além de disponibilizar um repositório para replicação de análises.

Problema

1. Novos softwares surgem constantemente, apresentando diferentes abordagens e funcionalidades para solucionar problemas específicos.

2. A escolha de frameworks e bibliotecas é crucial para a base de qualquer projeto, sendo uma decisão importante e de grande risco.

3. Ferramentas mais antigas, sem o devido acompanhamento, pouco atualizadas e utilizadas podem comprometer a sequência e manutenção de um desenvolvimento.

unesp

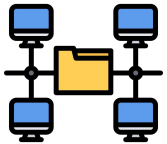
Justificativa

1. Dados para esse tipo de análise são fáceis de encontrar, ainda mais de projetos de código aberto.
2. Repositórios contêm diversos projetos relevantes, possibilitando o acesso ao histórico de desenvolvimento de softwares em uma ampla quantidade de linguagens.
3. Comparar e decidir ferramentas dentro de um projeto se torna um processo facilitado com uma análise prévia de dados confiáveis.

Metodologia

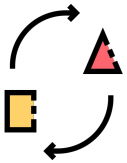
Mineração de Dados

Seleção dos Softwares



Obtenção dos Dados

Transformação



Análise

unesp

Seleção dos Softwares

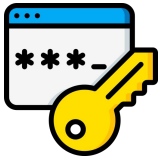
- A seleção dos softwares foi baseada em suas áreas de atuação, linguagens de programação e períodos ativos.
- Foram levados em consideração tanto projetos atuais e ainda ativos, quanto aqueles que já são legado, que não tem manutenção ativa e são pouco utilizados.
- No total foram escolhidos 85 softwares, sendo 60 atuais e 25 legados.

Seleção dos Softwares

Áreas escolhidas:

Machine Learning

Segurança



API *Rest*

Ciência de Dados

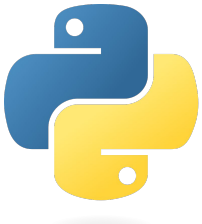
Web

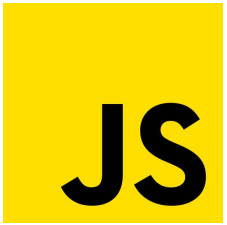
Teste de Software

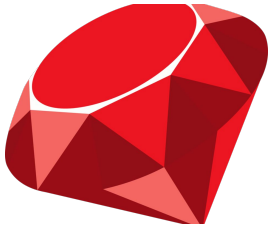
unesp

Seleção dos Softwares

Linguagens escolhidas:







Obtenção dos Dados

- Com a seleção de *frameworks* e bibliotecas definida, foi utilizada a ferramenta Jupyter para criar dois *notebooks* de código iniciais com as seguintes finalidades:
    - Minerar repositórios git para obter informações sobre *commits*.
    - Adquirir informações sobre a quantidade de pesquisas por projeto via Google Trends.

unesp

Obtenção dos Dados

Parte de código para minerar repositórios :

```
from pydriller import Repository
import pandas as pd
import datetime as dt
```

```
planilha_repos = pd.read_csv("CSV/links.csv")
```

```
repos_list = planilha_repos['Links'].tolist()
```

```
response = []
for commit in Repository(repos_list).traverse_commits():
    print(commit.project_name)
    response.append(f"{commit.author.name},,,,{commit.committer},,,,{commit.hash},,,,{commit.committer_date},,,,
                    {commit.project_name},,,,{commit.deletions},,,,{commit.insertions},,,,{commit.msg}")
data = pd.DataFrame(response, columns = ["Author"])
```

unesp

Obtenção dos Dados

Parte de código para minerar *trends* :

```
from pytrends.request import TrendReq
import pandas as pd

# Lista de termos que você quer pesquisar
termos = pd.read_csv("CSV/nomes_de_softwares.csv")
termos = termos["softwares_names"].tolist()
# Divida a lista em grupos de 5
grupos_de_termos = [termos[i:i+1] for i in range(0, len(termos), 1)]
# Crie uma instância da TrendReq
pytrends = TrendReq(hl='pt-BR', tz=360)  # Defina a linguagem (hl) e o fuso horário (tz)
resultados = pd.DataFrame()
# Para cada grupo de termos
for grupo in grupos_de_termos:
    # Configure os parâmetros da busca
    pytrends.build_payload([grupo], cat=5, timeframe='all', geo='', gprop='')
    # Obtenha os dados
    dados_novos = pytrends.interest_over_time()
    resultados = pd.concat([resultados, dados_novos], axis=0, join='outer')
```

unesp

Transformação dos Dados

- A transformação nos dados ocorreu com um foco maior na disposição de linhas e colunas.
- Tabela de *commits* é obtida com as informações concatenadas em uma única coluna, impossibilitando a visualização.
- Tabela de *trends* dispõe cada projeto como uma coluna, não distribuindo as informações em linhas.

unesp

Transformação dos Dados

Commits

	Commits
0	Wes McKinney,,,,<pydriller.domain.developer.De...
1	Wes McKinney,,,,<pydriller.domain.developer.De...
2	Wes McKinney,,,,<pydriller.domain.developer.De...
3	Wes McKinney,,,,<pydriller.domain.developer.De...
4	Wes McKinney,,,,<pydriller.domain.developer.De...
...	...
33437	jbrockmendel,,,,<pydriller.domain.developer.De...
33438	jbrockmendel,,,,<pydriller.domain.developer.De...
33439	jbrockmendel,,,,<pydriller.domain.developer.De...
33440	jbrockmendel,,,,<pydriller.domain.developer.De...
33441	Natalia Mokeeva,,,,<pydriller.domain.developer...

33442 rows × 1 columns

Trends

	Pandas	isPartial
date		
2004-01-01	0	False
2004-02-01	1	False
2004-03-01	0	False
2004-04-01	0	False
2004-05-01	4	False
...	...	...
2023-06-01	69	False
2023-07-01	64	False
2023-08-01	69	False
2023-09-01	75	False
2023-10-01	62	True

238 rows × 2 columns

unesp

Análises e Resultados

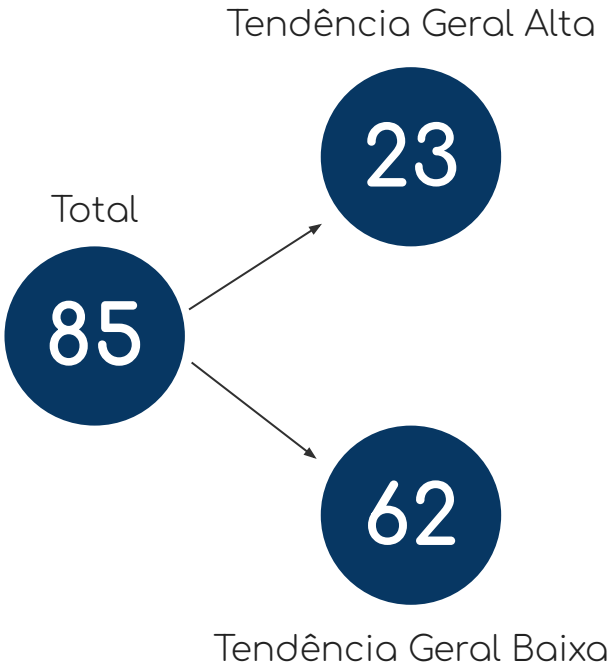
unesp

Análises e Resultados

- Para análise dos dados tratados, foram utilizadas três métricas principais:
    - Tendência baseada em média móvel exponencial (MME) de curto e longo prazo.
    - Correlação entre os dados pelo método de Spearman.
- A tendência de alta ou baixa e a correlação foram utilizadas para 3 dados principais: número de commits, número de autores e interesse relativo por software escolhido.

unesp

Tendência Geral



Dos 23 projetos com tendência geral alta, 18 são atuais e 5 legados.

Dos 62 projetos com tendência geral baixa, 27 deles tem tendência alta em interesse, com 23 projetos atuais e 4 legados.

unesp

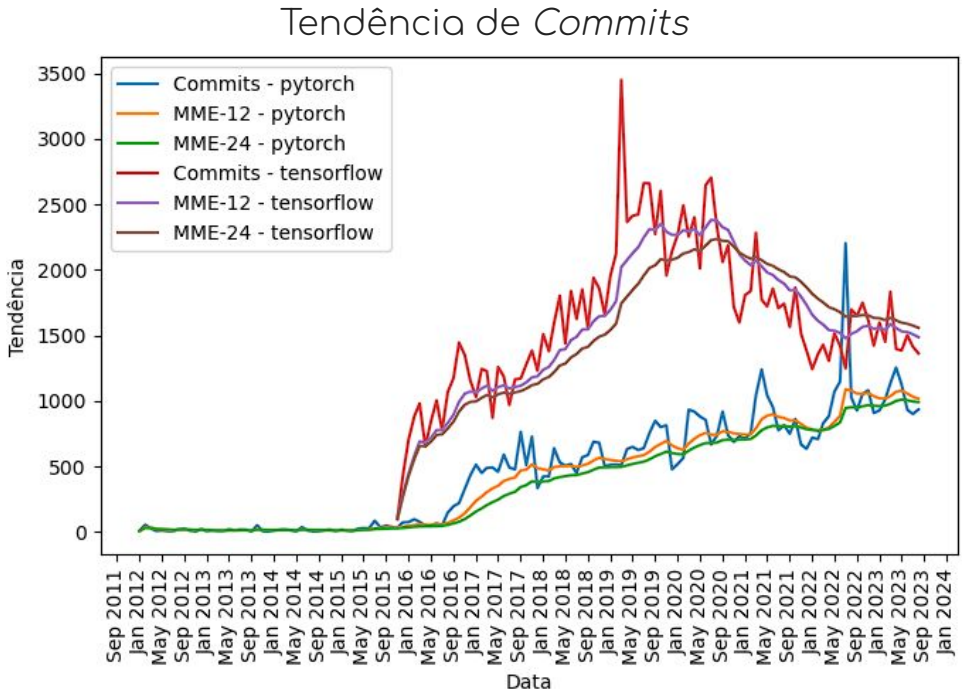
Correlação Geral

	Commit/Interesse	Autor/Interesse	Commit/Autor
> 0,5	16	25	50
< -0,5	15	8	0
Entre -0,5 e 0,5	43	41	35
Sem dados	11	11	0

É possível identificar um padrão entre *commits* e autores, que não apresentam valores negativos e tem o maior percentual de correlações fortes entre as comparações. Isso é natural pois uma maior quantidade de autores normalmente implica em uma maior quantidade de *commits*.

unesp

PyTorch X TensorFlow



PyTorch

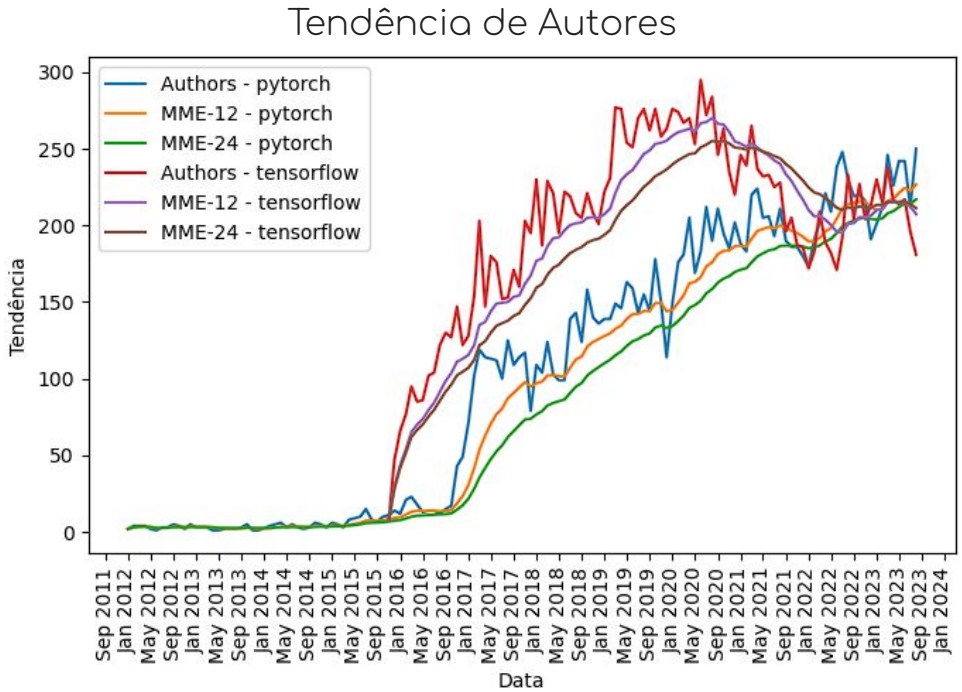
- MME de curto prazo = 1017,15
- MME de longo prazo = 1004,10

TensorFlow

- MME de curto prazo = 1486,54
- MME de longo prazo = 1533,46

unesp

PyTorch X TensorFlow



PyTorch

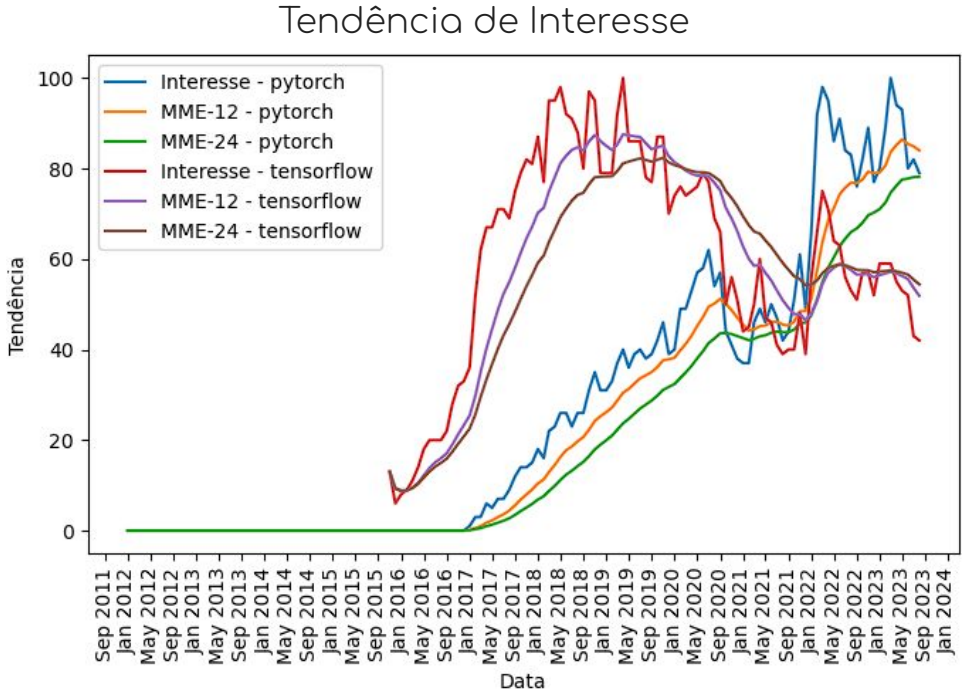
- MME de curto prazo = 226,74
- MME de longo prazo = 220,21

TensorFlow

- MME de curto prazo = 207,49
- MME de longo prazo = 209,97

unesp

PyTorch X TensorFlow



PyTorch

- MME de curto prazo = 83,99
- MME de longo prazo = 80,37

TensorFlow

- MME de curto prazo = 51,85
- MME de longo prazo = 53,65

unesp

PyTorch X TensorFlow

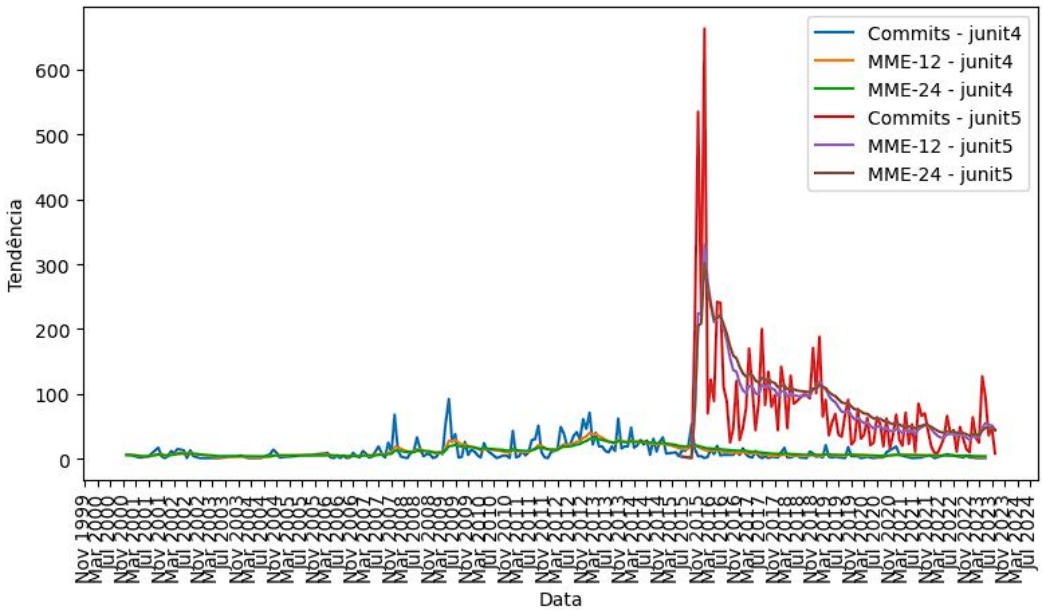
Correlações

Projeto	Commit/Interesse	Autor/Interesse	Commit/Autor
pytorch	0,952	0,953	0,972
tensorflow	0,659	0,609	0,982

unesp

Junit 4 X Junit 5

Tendência de *Commits*



Junit 5

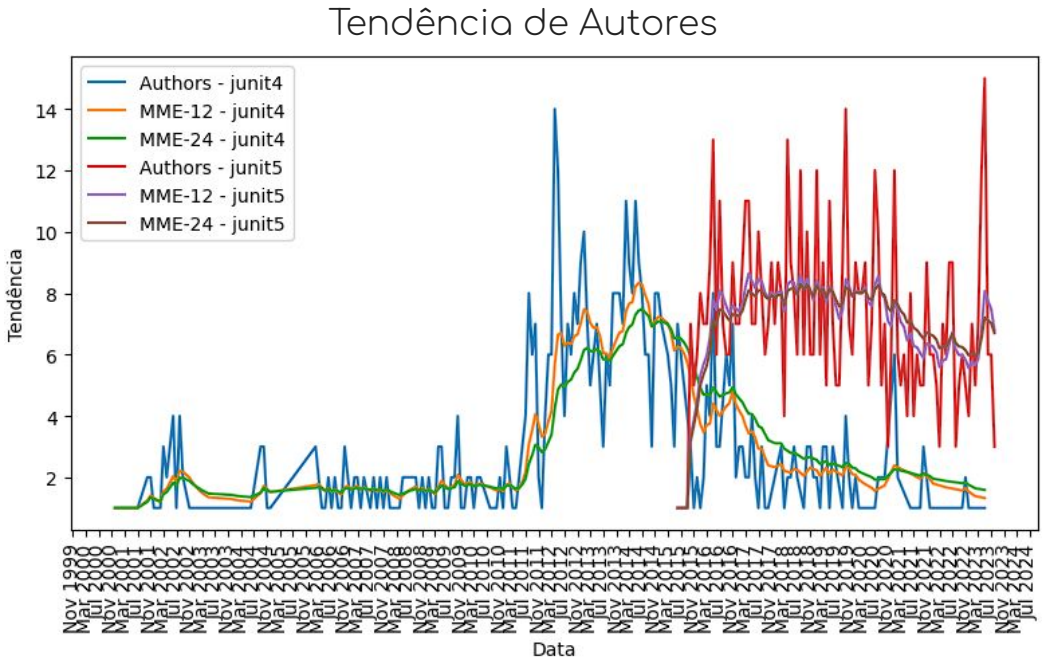
- MME de curto prazo = 44,74
- MME de longo prazo = 43,90

Junit 4

- MME de curto prazo = 2,95
- MME de longo prazo = 3,84

unesp

Junit 4 X Junit 5



Junit 5

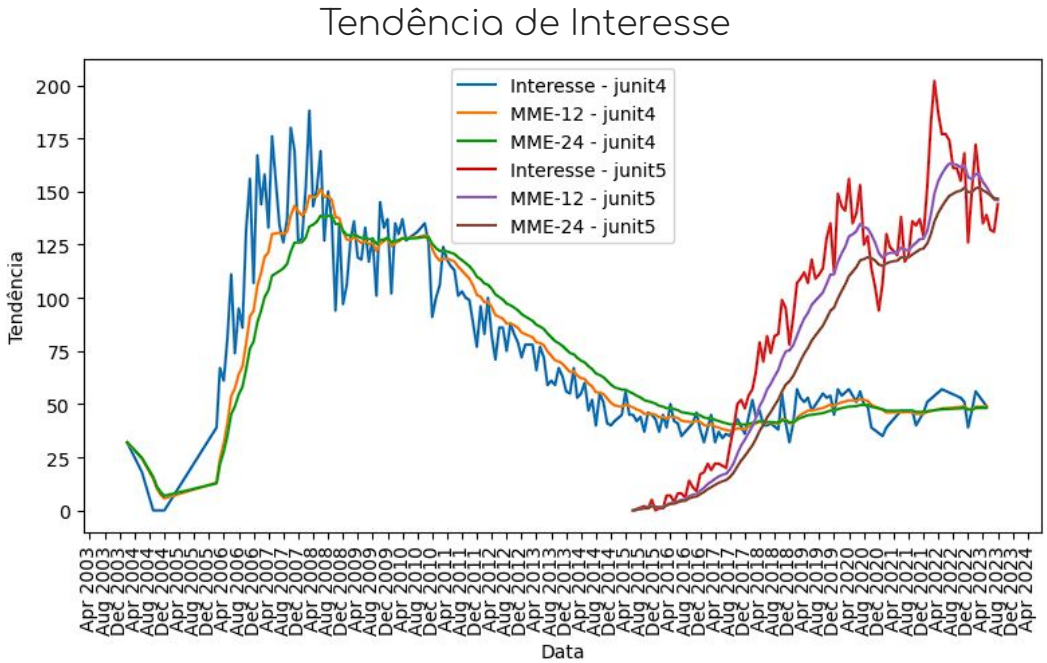
- MME de curto prazo = 6,79
- MME de longo prazo = 6,69

Junit 4

- MME de curto prazo = 1,32
- MME de longo prazo = 1,59

unesp

Junit 4 X Junit 5



Junit 5

- MME de curto prazo = 145,90
- MME de longo prazo = 146,64

Junit 4

- MME de curto prazo = 48,79
- MME de longo prazo = 48,20

unesp

Junit 4 X Junit 5

Correlações

Projeto	Commit/Interesse	Autor/Interesse	Commit/Autor
Junit4	0,443	-0,259	0,602
Junit5	-0,839	0,216	0,391

Considerações Finais

Considerações Finais

- As análises realizadas apresentam um cenário real que auxilia na escolha de um novo software em um projeto.
- Além dos resultados de tendências e correlações sobre quantidade de *commits*, autores e interesse relativo, é possível observar pontos importantes nos gráficos como cada quantidade e distribuição dos dados ao longo do tempo.
- Pesquisas auxiliares são utilizadas para fundamentar e complementar a análise a fim de obter resultados mais coesos.

unesp

Referências

- DAI, H.; PENG, X.; SHI, X.; HE, L.; XIONG, Q.; JIN, H. Reveal training performance mystery between tensorflow and pytorch in the single gpu environment. Science China Information Sciences, Springer, v. 65, p. 1–17, 2022.

- ELDER, A. Aprenda a operar no mercado de ações. Rio de Janeiro: Editora Campus, 2006.

- GARCIA, B. Mastering Software Testing with JUnit 5: Comprehensive guide to develop high quality Java applications. [S.l.]: Packt Publishing Ltd, 2017.

- SOUSA, Á. Coeficiente de correlação de pearson e coeficiente de correlação de spearman: o que medem e em que situações devem ser utilizados? Correio dos Açores, Gráfica Açoreana, Lda, p. 19–19, 2019.

unesp

MUITO OBRIGADO

unesp