NewsRanking

Protótipo de um sistema de ranqueamento para notícias. As buscas são feitas por meio da API disponível em `https://newsapi.org/`.

Requisitos

- Python 2.7.x
- Biblioteca python -> pandas
- Biblioteca python -> requests
- Biblioteca python -> xlwt
- Biblioteca python -> xlrd
- Biblioteca python -> matplotlib
- Biblioteca python -> XlsxWriter
- S.O. Windows com PowerShell para utilizar os scripts `.bat`.

Instalando

EXECUTE TODOS OS ARQUIVOS EM MODO DE ADMINISTRADOR!!

O arquivo `setup.bat` instala todas as dependências do projeto e copia a pasta com o projeto para `C:\NewsRanking\`.

O arquivo `folder.bat` faz apenas a cópia da pasta, sem instalar as dependências.

Utilizando

Para utilizar o programa é necessário ter os termos de busca definidos no arquivo `termos.xlsx` e os critérios de classificação no arquivo `doencas_info.xlsx` contidos na pasta `pyNews/source`. Também é necessário um cadastro no site `https://newsapi.org/`. Com o cadastro efetuado, o site liberará uma chave de acesso, que deve ser inserida no arquivo `apiKey.txt`, também contido na pasta `pyNews/source`.

Para executar, rode o script `run.bat` contido na pasta `pyNews` ou execute por meio do atalho.

Limitações

- A implementação do cálculo das notas, utilizando os valores da planilha e pesos, ainda não foi implementado, mas ao usuário é fortemente recomendado o preenchimento de todos os dados no arquivo `doencas_info.xlsx` contido na pasta `pyNews/source`, pois a próxima etapa de implementação será com base nisso.
- Ainda no arquivo `doencas_info.xlsx`, a lista de sinônimos ainda não é percorrida na busca por termos no título e na descrição.
- Termos em inglês ainda não são suportados e os portais de notícias são todos brasileiros.

A ser desenvolvido

- Suporte à busca de termos em inglês;
- Suporte à lista de sinônimos do arquivo `doencas_info.xlsx`;
- Criação de uma fórmula para gerar notas de acordo com os valores de `doencas_info.xlsx`;
- Identificação dos locais das notícias, para ser usado como critério de classificação;
- Visualização, na planilha de saída, dos critérios que fizeram a notícia pontuar.

Pasta com os arquivos de entrada

Os arquivos contidos nesta pasta são responsáveis pelos dados de entrada do programa.

Arquivos

Os arquivos no formato `.xlsx` somente serão lidos na primeira planilha, caso contenham mais de uma.

apiKey.txt

O arquivo `apiKey.txt` recebe a chave utilizada para autenticação na API. O arquivo deve conter apenas a chave e nada mais. Para obter uma chave, acesse `https://newsapi.org` e faça um cadastro. Após o cadastro será disponibilizada uma chave que deve ser copidada para o arquivo.

doencas_info.xlsx

O arquivo `doencas_info.xlsx` contém todas as palavras-chave para os critérios de classificação do algoritmo. Por conta da forma como o algoritmo está estruturado, é necessário que os nomes escritos não contenham caracteres especiais (ç, á, ã ...) e estejam apenas em letras minúsculas.

Nome da doenca	**Sinonimos**	**Int. CieTec**	**Int. Politicos**	**Int. Economicos**	**Potencial de disseminacao**	**Impacto Saude Publ.**	**Gravidade**	**Interesse atual**
exemplo	example	0-1	0-1	0-1	1-5	1-5	1-5	1-3
doenca	disease	0	1	1	4	2	3	1

- Nome da doenca -> Nome da doenca, agravo ou palavra-chave. Em letras minúsculas e sem caracteres especiais.
- Sinonimos -> Sinônimos ou traduções para outras línguas relevantes. Em letras minúsculas e sem caracteres especiais. No caso de mais de uma entrada, separar por vígulas.
- Int. CieTec -> Interesse Científico e Tecnológico, 1 se possui, 0 se não possui.
- Int. Politicos -> Interesse Político, 1 se possui, 0 se não possui.
- Int. Economicos -> Interesse Econômico, 1 se possui, 0 se não possui.
- Potencial de disseminacao -> Potencial de disseminação, graduado de 1 (baixo potencial) a 5 (alto potencial).
- Impacto Saude Publ. -> Impacto na Saúde Pública, graduado de 1 (pouco impacto) a 5 (muito impacto).
- Gravidade -> Gravidade, graduada de 1 (menos grave) a 5 (mais grave)
- Interesse atual -> Interesse atual na palavra correspondente, 1 - algum interesse, 2 - interesse médio, 3 - muito interesse.

termos.xlsx

O arquivo `termos.xlsx` contém os termos que serão buscados. Por conta da forma como o algoritmo está estruturado, é necessário que todos os termos sejam escritos em letras minúsculas e sem caracteres especiais (ç, á, ã ...). O arquivo está estruturado em apenas uma coluna, onde os novos termos a serem buscados devem ser escritos apenas na primeira coluna.

Pasta que contém os arquivos de saída

yyyy-mm-dd_saida.xlsx

Arquivo que contém a saída "crua". Esta planilha mostra os dados na forma como eles são recebidos pela API.

yyyy-mm-dd_organizado.xlsx

Este arquivo é uma forma organizada do arquivo `yyyy-mm-dd_saida.xlsx`. Esta planilhas acrescenta as seguintes colunas:

- Score -> Pontuação atribuída pelo algoritmo.
- Termos no Titulo -> Lista com os termos encontrados no título da notícia.
- Termos na Desc. -> Lista com os termos encontrados na descrição da notícia.

Para que os termos sejam listados, é necessário que eles estejam na planilha do arquivo `doencas_info.xlsx` presente na pasta `source`.