É POSSÍVEL ENCONTRAR UMA AGULHA NUM PALHEIRO?

INSTITUTO UNIVERSITÁRIO DE LISBOA

BPSTAT: PESQUISA 2.0

yabstat.shinyapps.io/bportugal

github.com/NCdJ/yabpstat

Henrique Luís[1] Maria Ribeiro[2] Nelson Jesus[3] Pedro Ferreira[4] Telma Garção[5]

[1] *halsa@iscte-iul.pt; Licenciatura Ciência de Dados, ISCTE-IUL*
[2] *mcroo2@iscte-iul.pt; Licenciatura Ciência de Dados, ISCTE-IUL*
[3] *nelson_carvalho_jesus@iscte-iul.pt; Licenciatura Ciência de Dados, ISCTE-IUL*
[4] *pvfaa@iscte-iul.pt; Licenciatura Ciência de Dados, ISCTE-IUL*
[3] *telma_garcao@iscte-iul.pt; Licenciatura Ciência de Dados, ISCTE-IUL*

Introdução

No BPstat, o Banco de Portugal divulga um vasto conjunto de estatísticas sobre a economia portuguesa e da área euro, que podem ser consultadas por domínios (`BPstat > Domínios`), exploradas de forma personalizada (`BPstat > Dados > Exploração Livre`) ou exploradas com recurso a séries e quadros pré-definidos (`BPstat > Dados > Exploração em Árvore`). Com mais de 270 mil séries à disposição, os utilizadores podem ainda fazer uso da pesquisa exclusiva, isto é, o acesso ao conteúdo é feito através de uma *query* na barra de navegação. O desafio insere-se num problema de ***information retrieval***.

Objetivos

Os objetivos estão dirigidos à pesquisa exclusiva e ao refinamento das queries.

1. Pluralização
2. *Stemming*
3. Expansão por sinónimos
4. Lógica axiomática
5. Pesquisa avançada
6. *Autocomplete*

Metodologia

A abordagem ao desafio implica compreender como a informação está organizada. Os dados encontram-se estruturados no BPstat, contudo as *queries* podem retornar informação não relevante para os utilizadores. A primeira fase de testes identifica a oportunidade de mapear termos procurados com frequência a sinónimos e palavras relacionadas. As séries estatísticas são recolhidas através da API do Banco de Portugal, em formato JSON, e carregadas num *cluster* do MongoDB Atlas.

As diferentes fases do projeto utilizam várias técnicas.

- Tokenização
- Webscraping
- ETL - Extract, Transform & Load

A prova de conceito pode ser descrita pelo fluxograma:

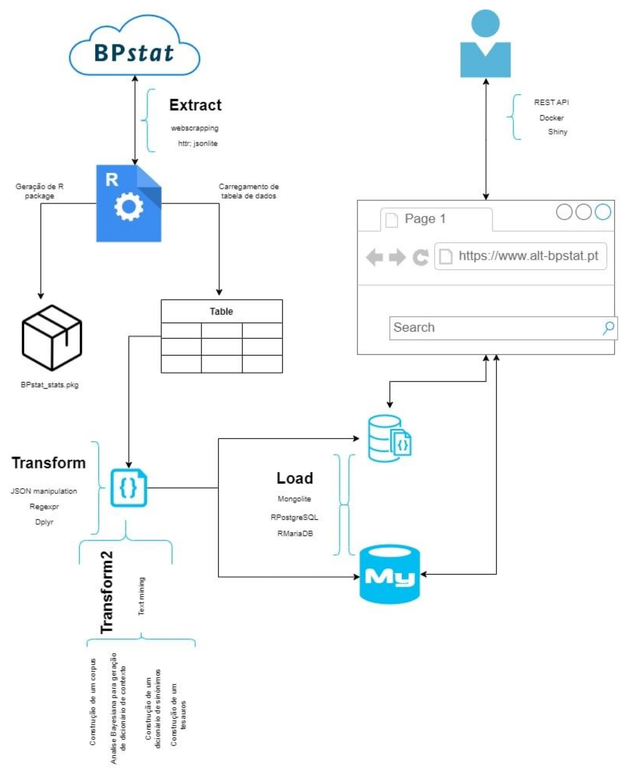

Figura 1: Visão geral do projeto

Produtos Intermédios

- Dicionário de sinónimos
- *Thesaurus*
- REST API

Data Science

- *Forecasting* com séries temporais
- TF-IDF

A funcionalidade de FTS (*Full-Text Search*) no MongoDB Atlas recorre ao scoring do Apache Lucene, que tem por base o Vector Space Model. Perante uma query do utilizador, é determinado o grau de relevância face à totalidade de documentos. Cada documento, d_j, e a query, q, são tratados como vectores e a similaridade por cosseno é calculada.

$$\cos(d_j, q) = \frac{d_j \times q}{\|d_j\| \cdot \|q\|}$$

Tabela 1: Scoring da query «euribor 3M»

Título	Score
Montante de empréstimos à habitação própria permanente com taxa variável indexados à Euribor 3M, em percentagem	1.3782494
Montante de novos empréstimos à habitação própria permanente com taxa variável indexados à Euribor 3M, em percentagem	1.3217597
TBA calculada a partir das EURIBOR - diária	0.5913645

A métrica **tf-idf** faz corresponder ao termo t um peso no documento d. Seja $tf_{t,d}$ o *term frequency* e idf_t o *inverse document frequency*:

$$tf\ idf_{t,d} = tf_{t,d} \times idf_t$$

O cálculo do **tf-idf** do termo `empresas`, por documento, permite identificar `Empresas da central de balanços` como o domínio onde o termo tem maior relevância.

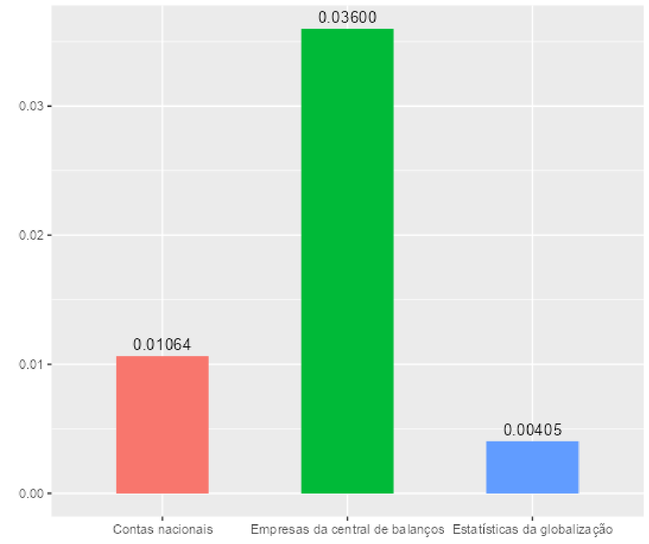

Figura 2: Domínios por grau de relevância do termo «empresas»

Produtos Finais

Foram desenvolvidos dois produtos finais.

Aplicação Shiny

Interface gráfica para pesquisa de séries.

Figura 3: Versão em shiny do BPstat

Package

Construído em R, permite aceder às séries estatísticas através da *BPstat Data API*.

- Designação: YABPstat: Yet Another BPstat
- Versão: 0.0.0.9000
- Licença: MIT

INSTALAÇÃO

```
# install.packages("devtools")
library(devtools)
devtools::install_github("NCdJ/yabpstat")
library(yabpstat)
```

Agradecimentos

Ao Professor José Dias e ao Professor Jorge Sinval por nos terem auxiliado a ultrapassar os obstáculos com que nos deparámos ao longo deste projeto e a transformá-los, assim, em oportunidades.

SCAN