Na sequência das atividades propostas em sala de aula e do exercício individual pedido, disponibilizamos este breve guião como material de formação contínua ao longo do curso e para a vossa autocorreção em vista da avaliação final.

O Reconhecimento Ótico de Caracteres (mais comummente conhecido como OCR, usando o acrônimo inglês) é um procedimento, que faz uso de ferramentas digitais, para reconhecer caracteres a partir de um arquivo de imagens não editáveis. Resumindo, os programas de OCR desbloqueiam o acesso ao texto dos ficheiros que precisamos usar, de forma a torná-los manipuláveis, arquiváveis e compatíveis para outros programas.

✍	Como escolher a ferramenta mais adequada para transformar um ficheiro não editável (JPEG, PNG ou PDF) para um ficheiro editável (.doc, .txt..)?

A escolha da ferramenta mais adequada depende de uma série de fatores:

- do tipo de ficheiros de partida que temos (manuscrito ou impresso, nem todas as ferramentas aceitam qualquer tipo de ficheiro de *input*);
- do tipo de ficheiro final como *output* que precisamos;
- da acessibilidade das ferramentas (licenças pagas), possibilidade de usar versões demo (ex. Abby Fine Reader, Adobe, etc.);

- da quantidade de ficheiros;
- da língua dos ficheiros de partida.

Devemos testar diversas ferramentas e avaliar os resultados obtidos para podermos escolher o melhor *output* que facilite a revisão manual em curto espaço de tempo.

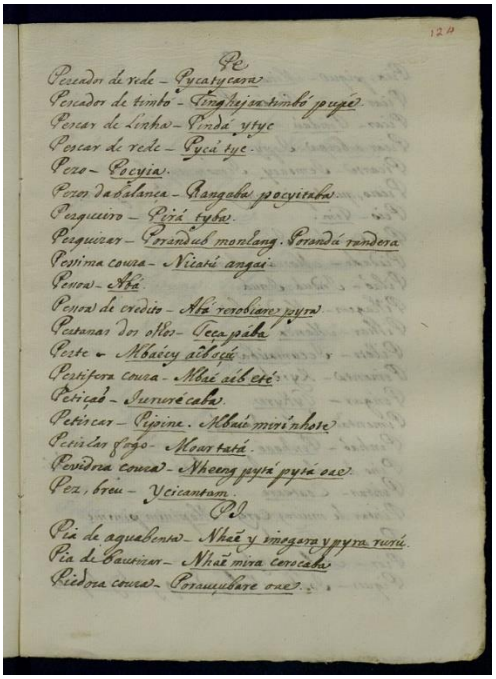
Apesar de cada ferramenta ter uma interface própria, de um modo geral todas apresentam:

- um espaço para fazer o *upload* do ficheiro que queremos transformar: é prática corrente existir um espaço para fazer o *upload* do ficheiro, através da função *arrastar e soltar* (drag and drop) numa caixa da página do software ou , em alternativa, de um botão para selecionar a pasta do seu computador e carregar o ficheiro em questão.
- uma série de configurações internas que podem (e devem) ser utilizadas para adaptar e calibrar o processo de OCR ao ficheiro de partida, por exemplo: a língua a ser usada (é importante selecionar a língua pois melhora substancialmente o desempenho do *software*); a possibilidade de escolher o formato de ficheiro final de *output;*
- um comando para iniciar a conversão do ficheiro;
- um comando para o *download* do ficheiro final como *output* (quando não é realizado em automático).

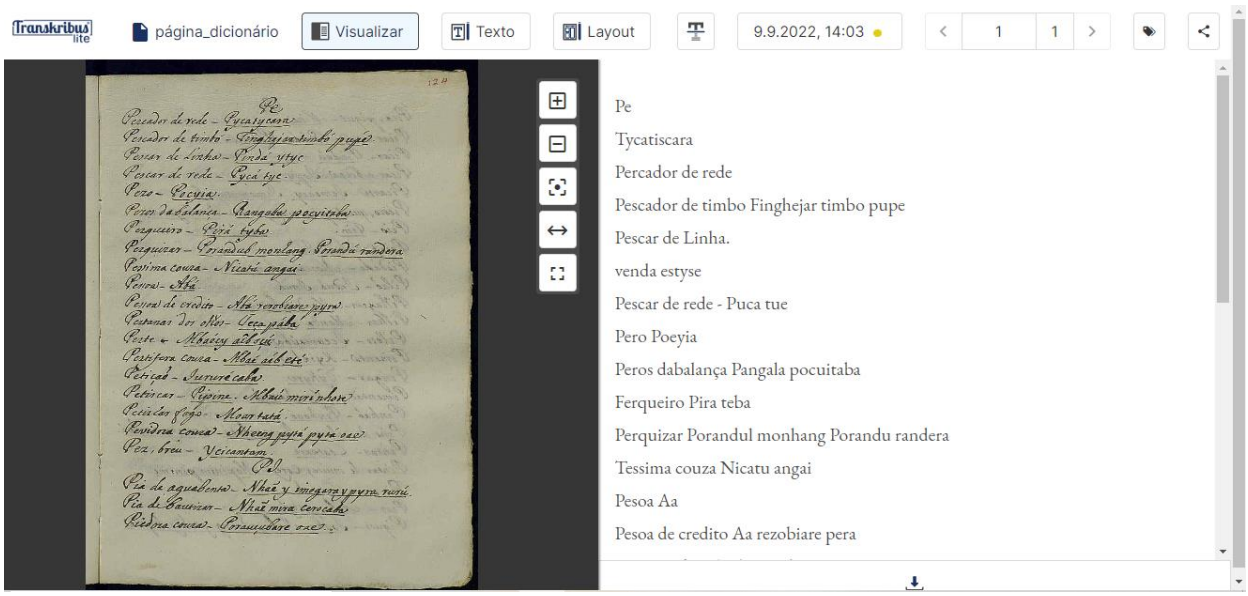
De seguida, apresentamos alguns exemplos de ficheiros de imagens com a respetiva metodologia a aplicar consoante a ferramenta selecionada para o efeito, que pode ser replicada para outros ficheiros.

Exemplo 1 (de JPEG para DOCX)

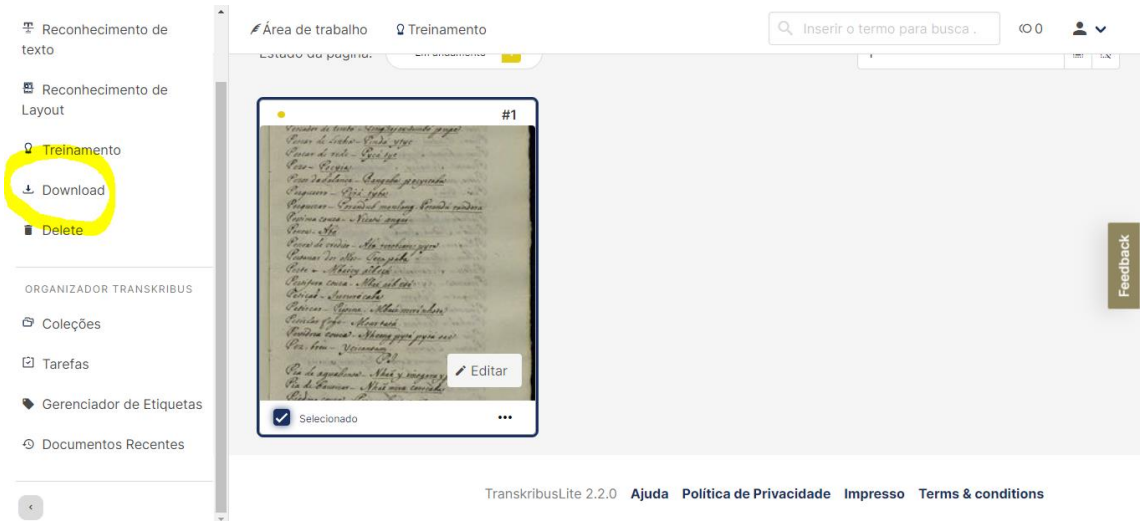
- Descrição: Ficheiro de uma imagem de uma página com texto manuscrito retirada de um dicionário "Diccionario da lingua geral do Brasil".
- Esta imagem encontra-se disponível em https://digitalis-dsp.uc.pt/bg3/UCBG-Ms-81/UCBG-Ms-81_item1/P275.html e está no formato JPEG.



- Ferramenta utilizada: Transkribus (disponível em https://transkribus.eu/lite). Apesar de ser uma ferramenta gratuita é necessário registar-se e fazer *login*.
- Esta ferramenta aceita como ficheiros de partida: JPEG, PNG ou PDF.
- Esta ferramenta disponibiliza como ficheiros finais: XML, Alto, PDF, TEI, DOCX.



- Esta ferramenta possibilita a extração do texto através de (i) a função de copiar e colar manualmente, ou (ii) de fazer o *download*.



- Relativamente à imagem submetida obtivemos os seguintes resultados:
  - Ao nível da disposição do texto na página, na interface *online* a visualização é alinhada entre texto original e transcrição, no *download* temos desformatação quase total do ficheiro de partida;
  - Ao nível do reconhecimento de caracteres obtivemos alguns erros (i.e. transcrição de letras erradas)

- De modo geral as palavras em português foram bem reconhecidas (ferramenta usada e calibrada para o português), os equivalentes em língua indígena (não é especificada qual) não o foram.

- Para saber mais como utilizar esta ferramenta consulte o guia de utilização (https://readcoop.eu/transkribus/howto/getting-started-with-transkribus-lite/).

Exemplo 2 (de PDF para DOCX)

- Descrição: Ficheiro de uma página do jornal *Público*, em formato digital.
- Esta imagem encontra-se disponível em


azul.

SUSTENTABILIDADE

Há quase uma década que "gente bonita come fruta feia"

Projecto que vende a preços baixos cabazes de fruta da época e legumes que agricultores não conseguem vender a supermercados diz já ter evitado o desperdício de 3685 toneladas de alimentos. Estivemos num dos espaços da Fruta Feia, no Bonfim (Porto).

Nelson Garrido e Daniel Dias · 3 de Maio de 2022, 9:04

Olá! Em que é que podemos ajudar-te?

NELSON GARRIDO

Liberdade faz-se com verdade

Começou em Novembro de 2013 como um pequeno projecto cooperativo interessado em combater o desperdício alimentar e, no último dia 20 de Abril, abriu, no Mercado de São Domingos de Benfica, em Lisboa, o seu 15.º posto de venda de fruta e produtos hortícolas com um aspecto "imperfeito". Mais de oito anos e algumas práticas europeias depois, a Fruta Feia diz já ter evitado o desperdício de 3685 toneladas de alimentos, ajudando agricultores espalhados pelo país a escoar as pegas de fruta e os legumes que, por não terem o aspecto desejado, não conseguem vender aos supermercados.

https://www.publico.pt/2022/05/03/azul/fotogaleria/fruta-feia-sao-ja-15-postos-venda-cooperativa-combate-desperdicio-alimentar-407910e está no formato PDF.

- Ferramenta utilizada: ONLINE OCT (https://www.onlineocr.net/pt/). Apesar de ser uma ferramenta gratuita há restrições quanto ao tamanho do ficheiro de partida.

- Esta ferramenta aceita como ficheiros de partida: JPEG, PNG ou PDF.
- Esta ferramenta disponibiliza como ficheiros finais: WORD, EXCEL, TXT.


azul.

SUSTENTABILIDADE

Há quase uma década que "gente bonita come fruta feia"

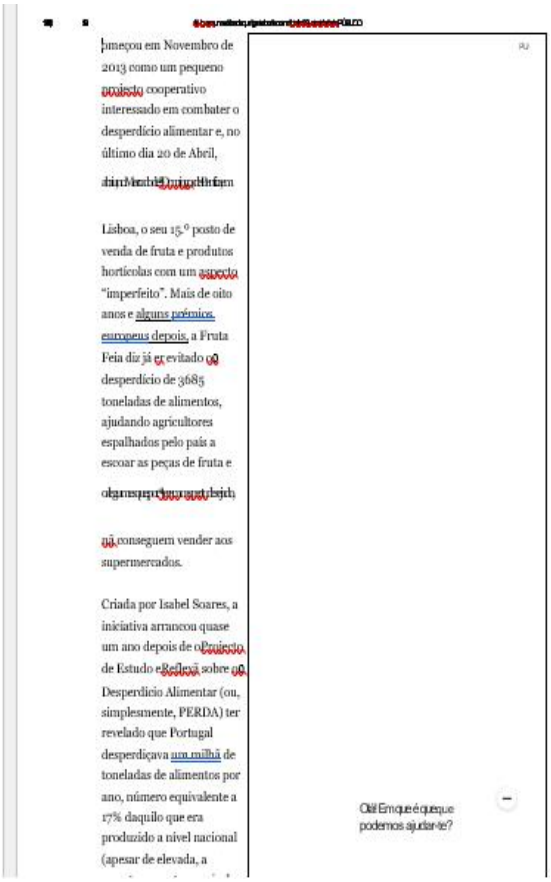
Projecto que vende a preços baixos cabazes de fruta da época e legumes que agricultores não conseguem vender a supermercados diz já ter evitado o desperdício de 3685 toneladas de alimentos. Estivemos num dos espaços da Fruta Feia, no Bonfim (Porto).

Nelson Garrido e Daniel Dias - 3de Maio de 2004 9:04

NELSON GARRIDO

Liberdade faz-se com verdade

Olá! Em que é que podemos ajudar-te? ?


começou em Novembro de 2013 como um pequeno projecto cooperativo interessado em combater o desperdício alimentar e, no último dia 20 de Abril,

Lisboa, o seu 15.º posto de venda de fruta e produtos hortícolas com um aspecto "imperfeito". Mais de oito anos e alguns prémios europeus depois, a Fruta Feia diz já ter evitado o desperdício de 3685 toneladas de alimentos, ajudando agricultores espalhados pelo país a escoar as peças de fruta e

não conseguem vender aos supermercados.

Criada por Isabel Soares, a iniciativa arrancou quase um ano depois de o Projecto de Estudo e Reflexão sobre o Desperdício Alimentar (ou, simplesmente, PERDA) ter revelado que Portugal desperdiçava um milhão de toneladas de alimentos por ano, número equivalente a 17% daquilo que era produzido a nível nacional (apesar de elevada, a

Olá! Em que é que podemos ajudar-te?

- Relativamente à imagem submetida obtivemos os seguintes resultados:
  - Ao nível da disposição do texto na página, temos desformatação parcial, nomeadamente no tipo de letra e na disposição das linhas que não estão totalmente emparelhadas com o ficheiro de partida.
  - Mantém algumas marcas de formatação do texto de partida (e.g. imagens, *links*, títulos etc.)
  - Ao nível do reconhecimento de caracteres obtivemos alguns problemas:
    - caracteres diferentes do original (acrescenta ou perde caracteres que não existem no original)
    - não reconhece caracteres especiais (e.g. parêntesis)
    - Erros de reconhecimento ortográfico (i.e. diacríticos, espaços a mais ou a menos)
    - Exemplos dos erros listados:
      - nã conse guem → não conseguem
      - P(orto) → (Porto)
      - abriu, no Mercado de São Domingos de Benfica, em → abriu, no Mercado de São Domingos de Benfica, em

Exemplo 3 (de PDF para DOCX)

- Descrição: Ficheiro de uma página do jornal *O Século Ilustrado*, em formato papel.
- Esta imagem encontra-se disponível em https://purl.pt/39776/2/ e está no formato PDF.


O SECULO ILUSTRADO

A SEMANA DE PORTUGAL NO RIO

O êxito da iniciativa do "Seculo" foi notavel em todas as cerimonias

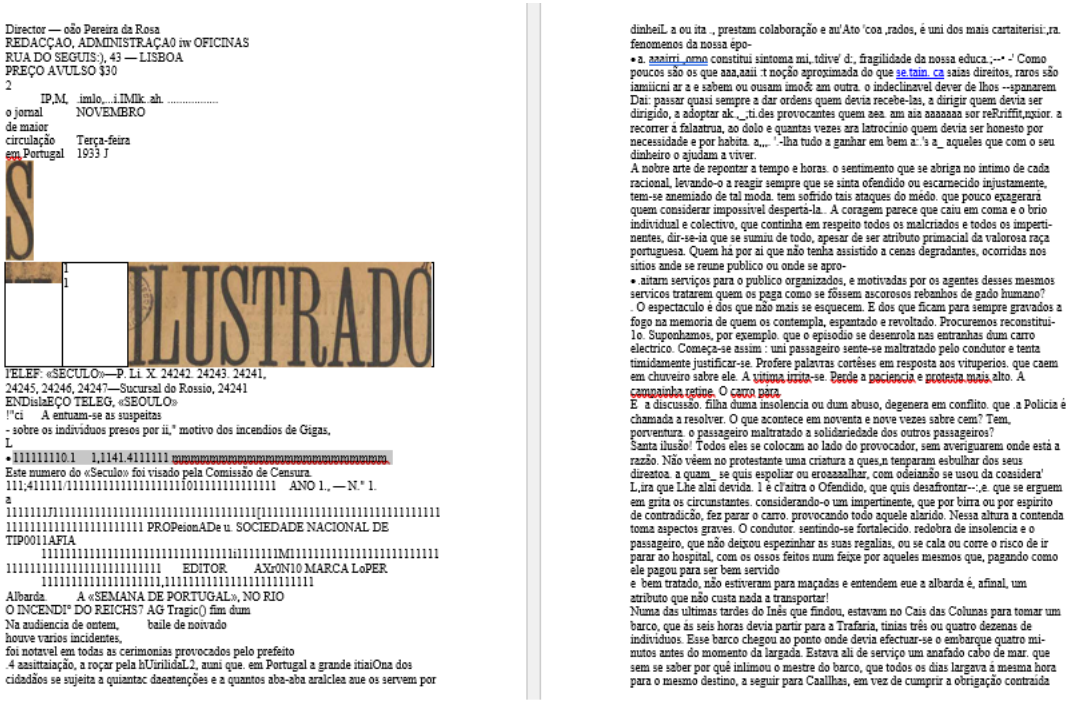
Na audiencia de ontem, houve varios incidentes, provocados pelo prefeito da Policia de Breslau

O ARMISTICIO

Italo Balbo

O Bandido das Olhos de Ouro

- Ferramenta utilizada: ONLINE OCR (https://www.onlineocr.net/pt/). Apesar de ser uma ferramenta gratuita tem restrições quanto ao tamanho do ficheiro de partida.
- Esta ferramenta aceita como ficheiros de partida: JPEG, PNG ou PDF.
- Esta ferramenta disponibiliza como ficheiros finais: WORD, EXCEL, TXT.



Director — oão Pereira da Rosa
REDACÇAO, ADMINISTRAÇA0 iw OFICINAS
RUA DO SEGUIS:), 43 — LISBOA
PREÇO AVULSO $30
2
IP,M, .imlo,...i.IMlk .ah. .....
o jornal NOVEMBRO
de maior
circulação Terça-feira
em Portugal 1933 J

ILUSTRADO

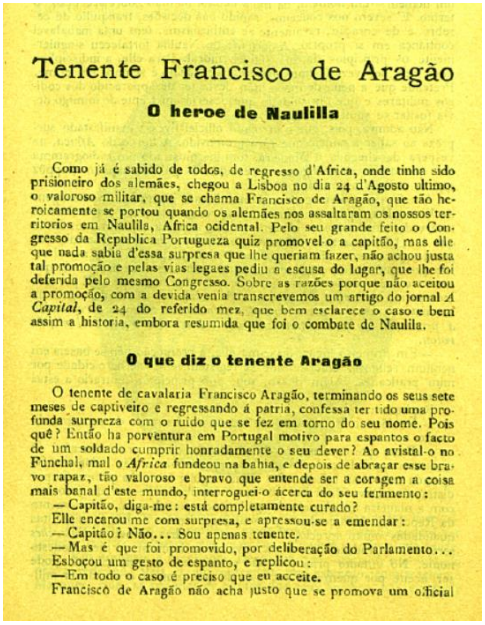
TELEF. «SECULO»—P. Li. X. 24242. 24243. 24241,
24245, 24246, 24247—Sucursal do Rossio, 24241
ENDislaEÇO TELEG, «SEOULO»
!"ci A entuam-se as suspeitas
- sobre os individuos presos por ii," motivo dos incendios de Gigas,
L
• 111111110.1 1,1141.4111111
Este numero do «Seculo» foi visado pela Comissão de Censura.
111;411111/1111111111111111111011111111111 ANO 1., — N.° 1.
a
1111111/1111111111111111111111111111111111[11111111111111111111111111111
11111111111111111111 PROPeionADe u. SOCIEDADE NACIONAL DE
TIP0011AFIA
11111111111111111111111111111111111111111M1111111111111111111111
1111111111111111111111 EDITOR AXr0N10 MARCA LoPER
11111111111111111,11111111111111111111111
Albarda. A «SEMANA DE PORTUGAL», NO RIO
O INCENDI° DO REICHS? AG Tragic() fim dum
Na audiencia de ontem, baile de noivado
houve varios incidentes,
foi notavel em todas as cerimonias provocados pelo prefeito
.4 asaittaiação, a roçar pela bUirilidaL2, auni que. em Portugal a grande itiaiOna dos
cidadãos se sujeita a quiantac daeatenções e a quantos aba-aba aralclea aue os servem por

dinheiL a ou ita ., prestam colaboração e au'Ato 'coa ,rados, é uni dos mais cartaiterisi:,ra.
fenomenos da nossa épo-
• a. aaairri, .omo constitui sintoma mi, tdive' d:, fragilidade da nossa educa.;--• -' Como
poucos são os que aaa,aaii :t noção aproximada do que se tain. ca saias direitos, raros são
iamiicni ar a e sabem ou ousam imoõ: am outra. o indeclinavel dever de lhos --spanarem
Dai: passar quasi sempre a dar ordens quem devia recebe-las, a dirigir quem devia ser
dirigido, a adoptar ak.,_.ti.des provocantes quem aea. am aia aaaaaaa sor reRriffit,nxior. a
recorrer à falaatrua, ao dolo e quantas vezes ara latrocinio quem devia ser honesto por
necessidade e por habita. a,,.. '.-lha tudo a ganhar em bem a.:'s a_ aqueles que com o seu
dinheiro o ajudam a viver.
A nobre arte de repontar a tempo e horas. o sentimento que se abriga no intimo de cada
racional, levando-o a reagir sempre que se sinta ofendido ou escarnecido injustamente,
tem-se anemiado de tal moda. tem sofrido tais ataques do médo. que pouco exagerara
quem considerar impossivel desperta-la.. A coragem parece que caiu em coma e o brio
individual e colectivo, que continha em respeito todos os malcriados e todos os imperti-
nentes, dir-se-ia que se sumiu de todo, apesar de ser atributo primacial da valorosa raça
portuguesa. Quem ha por ai que não tenha assistido a cenas degradantes, ocorridas nos
sitios ande se reune publico ou onde se apro-
• .aitarn serviços para o publico organizados, e motivadas por os agentes desses mesmos
servicos tratarem quem os paga como se fôssem ascorosos rebanhos de gado humano?
. O espectaculo é dos que não mais se esquecem. E dos que ficam para sempre gravados a
fogo na memoria de quem os contempla, espantado e revoltado. Procuremos reconstitui-
1o. Suponhamos, por exemplo. que o episodio se desenrola nas entranhas dum carro
electrico. Começa-se assim : uni passageiro sente-se maltratado pelo condutor e tenta
timidamente justificar-se. Profere palavras cortêses em resposta aos vituperios. que caem
em chuveiro sabre ele. A vitima irrita-se. Perde a paciencia e protesta mais alto. A
campainha retine. O carro para.
E a discussão. filha duma insolencia ou dum abuso, degenera em conflito. que .a Policia é
chamada a resolver. O que acontece em noventa e nove vezes sabre cem? Tem,
porventura. o passageiro maltratado a solidariedade dos outros passageiros?
Santa ilusão! Todos eles se colocam ao lado do provocador, sem averiguarem onde está a
razão. Não vêem no protestante uma criatura a ques,n tenparam esbulhar dos seus
direatoa. a quam_ se quis espoliar ou eroaaaalhar, com odeianão se usou da coasidera'
L,ira que Lhe alai devida. 1 è cl'aitra o Ofendido, que quis desafrontar--';.e que se erguem
em grita os circunstantes. considerando-o um impertinente, que por birra ou por espirito
de contradicão, fez parar o carro. provocando todo aquele alarido. Nessa altura a contenda
toma aspectos graves. O condutor. sentindo-se fortalecido. redobra de insolencia e o
passageiro, que não deixou espezinhar as suas regalias, ou se cala ou corre o risco de ir
parar ao hospital, com os ossos feitos num feixe por aqueles mesmos que, pagando como
ele pagou para ser bem servido
e bem tratado, não estiveram para maçadas e entendem eue a albarda é, afinal, um
atributo que não custa nada a transportar!
Numa das ultimas tardes do Inês que findou, estavam no Cais das Colunas para tomar um
barco, que às seis horas devia partir para a Trafaria, tinias três ou quatro dezenas de
individuos. Esse barco chegou ao ponto onde devia efectuar-se o embarque quatro mi-
nutos antes do momento da largada. Estava ali de serviço um anafado cabo de mar. que
sem se saber por quê inlimou o mestre do barco, que todos os dias largava a mesma hora
para o mesmo destino, a seguir para Caallhas, em vez de cumprir a obrigação contraida

- Relativamente à imagem submetida obtivemos os seguintes resultados:
  - Ao nível da disposição do texto na página, temos desformatação parcial. A ferramenta propõe uma reorganização do texto de partida para a modalidade texto corrido, que é problemática, pois, em muitos casos emparelha as linhas de diferentes colunas.
  - Apresenta erros de reconhecimento de caracteres que produzem palavras inexistentes (e.g. aaairri, .tdive' d:, 50IreU)
  - Apresenta erros ortográficos (apostrofes, diacríticos, maiúsculas e minúsculas)
  - Não corrige a grafia antiga das palavras.

Exemplo 4 (de PDF para TXT)

- Descrição: Ficheiro de uma página de um livro *O Diário da Grande Guerra*, em formato papel.
- Esta imagem encontra-se disponível em https://grandeguerra.bnportugal.gov.pt/pdf/1916/jan/1916-01-01_PP-14-V.pdf e está no formato PDF.



Tenente Francisco de Aragão

O heroe de Naulilla

Como já é sabido de todos, de regresso d'Africa, onde tinha sido prisioneiro dos alemães, chegou a Lisboa no dia 24 d'Agosto ultimo, o valoroso militar, que se chama Francisco de Aragão, que tão heroicamente se portou quando os alemães nos assaltaram os nossos territorios em Naulila, Africa ocidental. Pelo seu grande feito o Congresso da Republica Portugueza quiz promovel-o a capitão, mas elle que nada sabia d'essa surpresa que lhe queriam fazer, não achou justa tal promoção e pelas vias legaes pediu a escusa do lugar, que lhe foi deferida pelo mesmo Congresso. Sobre as razões porque não aceitou a promoção, com a devida venia transcrevemos um artigo do jornal *A Capital*, de 24 do referido mez, que bem esclarece o caso e bem assim a historia, embora resumida que foi o combate de Naulila.

O que diz o tenente Aragão

O tenente de cavalaria Francisco Aragão, terminando os seus sete meses de captiveiro e regressando á patria, confessa ter tido uma profunda surpreza com o ruido que se fez em torno do seu nome. Pois quê? Então ha porventura em Portugal motivo para espantos o facto de um soldado cumprir honradamente o seu dever? Ao avistal-o no Funchal, mal o *Africa* fundeou na bahia, e depois de abraçar esse bravo rapaz, tão valoroso e bravo que entende ser a coragem a coisa mais banal d'este mundo, interroguei-o ácerca do seu ferimento:

— Capitão, diga-me: está completamente curado?

Elle encarou-me com surpresa, e apressou-se a emendar:

— Capitão? Não... Sou apenas tenente.

— Mas é que foi promovido, por deliberação do Parlamento...

Esboçou um gesto de espanto, e replicou:

— Em todo o caso é preciso que eu acceite.

Francisco de Aragão não acha justo que se promova um official

- Ferramenta utilizada: OCR.SPACE (https://ocr.space/). Apesar de ser uma ferramenta gratuita há restrições quanto ao tamanho do ficheiro de partida.

- Esta ferramenta aceita como ficheiros de partida: JPEG, PNG ou PDF.
- Esta ferramenta disponibiliza como ficheiros finais: TXT para download (apesar de permitir a visualização *pre-view* numa caixa de texto na página *online*.

```
1
2  ****** Result for Image/Page 1 ******
3  Tenente Francisco de Aragão
4  O heroe de Naulilla
5  Como já é sabido de todos, de regresso d'Africa, onde tinha sido
6  prisioneiro dos alemães, chegou a Lisboa no dia 24 d'Agosto ultimo,
7  o valoroso -militar, que se chama Francisco de Aragão, que tão he-
8  roicamentese portou quando os alemães nos assaltaram os nossos ter-
9  ritorios cm Naulila, Africa ocidental. pelo seu grande feito o Con•
10 gresso da Republica Portugueza quiz promovel•o a capitão, mas elle
11 que nada sabia d'essa surpresa que lhe queriam fazer, não achou justa
12 tal: promoção e pelas vias legaes pediu a escusa do lugar, que lhe foi
13 deferida pelo mesmo Congresso. Sobre as razões porque não aceitou
14 a promoçãoj com a devida venia transcrevemos um artigo do jornal A
15 Capital, de 24 do referido mez, que bem esclarece o caso e bem
16 assim historia, embora resumida que foi o combate de Naulila.
17 O que diz o tenente Aragão
18 O tenente de cavalaria Francisco Aragão, terminando os seus sete
19 meses de captiveiro e regressando á patriá, confessa ter tido uma pro.
20 funda -surpreza com o ruido que se fez em torno do seu nome. Pois
21 uê ? Então ha porventura em Portugal motivo para espantos o facto
22 e um soldado cumprir honradamente o seu dever? Ao avistal-o no .
23 Funchals mal o Africa fundeou na bahia, e depois de abraçar esse bras
24 vo ra aza, tão valoroso e bravo que entende ser a coragem a coisa
25 mais anal d'este mundo, T interroguei-o ácerca do seu ferimento :
26 —Capitão, diga-me : está completamente curado?
27 Elle encaroune com surpresa, e apressou-se a emendar :
28 Não.
29 Sou apenas tenente.
30 .—Mase é que foi promovido, por deliberação do Parlamento.
31 Esboçou um gesto de espanto, e replicou :
32 '—Em todo o caso é preciso que eu acceite.
33 Francisco de Aragão não acha nustO que se promova um oficial
34
```

- Relativamente à imagem submetida obtivemos os seguintes resultados:
  - Ao nível da disposição do texto na página, a ferramenta acrescenta linhas de metadados que não existem no texto de partida – e.g. “Result for image/Page 1). Mantém parcialmente a formatação do texto de partida (e.g. marca de fim de linha) mas perde-se a identificação dos títulos.
  - Apresenta erros de reconhecimento ortográfico (e.g. hifenização, diacríticos, acrescenta/elimina espaços que não existem no texto original)
  - Não corrige a grafia antiga (e.g. d’Agosto, mez).
  - Exemplos do erros listados:
    - he – roicamente → heroicamente
    - Com gresso → (acrescenta espaços no meio das palavras e hiper-corrige a preposição “com”) congresso
    - patriá → pátria

Através da apresentação dos diferentes exemplos e da utilização das diferentes ferramentas esperamos que seja mais fácil em trabalhos futuros saber escolher qual a ferramenta que melhor se adequa ao seu trabalho e qual apresenta melhores resultados para que possa executar esta tarefa da melhor forma.

Equipa Hugod

Erasmus+

NOVAFCSH

EN	The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.
PT	O apoio da Comissão Europeia à produção desta publicação não constitui um aval do seu conteúdo, que reflete unicamente o ponto de vista dos autores, e a Comissão não pode ser considerada responsável por eventuais utilizações que possam ser feitas com as informações nela contidas.