NLP Extraction & Tagging Test Pack

Document de test conçu pour valider : extraction PDF, segmentation layout, détection de langue, tokenisation, POS tagging (NN/VB/JJ...), lemmatisation et gestion de la ponctuation.

Mode d’emploi : exécute ton pipeline sur ce PDF. Compare la sortie (sentences_layout / chunks / tokens) avec les attentes ci-dessous. Le header/footer est volontairement répétitif pour tester la détection de bruit.

Couverture de tests

• Easy: phrases simples + dates, emails, URLs, nombres, ponctuation standard.
• Medium: listes numérotées, sections (1.1 / a. / (iv)), camelCase, ALLCAPS collés, mots collés chiffres/lettres, hyphenation sur saut de ligne.
• Complex: micro-table header multi-colonnes, tableau de lignes (facture), colonnes alignées, séparateurs, blocs adresse.
• Noise/Edge: répétitions (header/footer), variations de chiffres, longues séquences de ponctuation, token , caractères arabes.

1) EASY — Tokens & POS basiques

Objectif : vérifier tokenisation WORD/PUNCT, offsets, et tags Penn simples.

```
Hello world!
Price: $12.50; Tax: 19.0%. Total = $14.88.
Date: 2026-02-10. Time: 09:15.
Email: test.user+nlp@example.com; alt: support@example.co.uk.
URL: https://example.com/a/b?x=1&y=2#anchor
Quotes: 'single' and “double” and Mercy Corps’ apostrophe.
Parentheses (like this), brackets [ok], braces {ok}.
Math-ish: 3/4, 10-20, 1,000.00, 2.002,00 (locale mix).
Punctuation mix: , . : ; ! ?
Code-ish: from __future__ import annotations
```

Attendu: tokeniser correctement les symboles (__, :, ;, ., ?, !), et gérer les nombres + emails/URLs.

2) MEDIUM — Cas réalistes (contrat) + pièges de normalisation

Ce bloc contient des motifs pour tester la normalisation: ALLCAPS collés, camelCase, lettres/chiffres collés, et hyphenation sur saut de ligne.

```
THISPURCHASECONTRACT is entered into as of _____ by and between MERCYCORPSNONPROFITCORPOR
1. Defined Terms. Each term has the meaning given in ScheduleIattachedHereto.
1.1 DeliveryDate and DeliveryLocation must match Invoice2026 and Section5.
a. The Supplier will provide thisContractNumber within 60days.
(iv) NonconformingGoods may be rejected; acceptance occurs only with GRN.
Hyphenation demo: inter-
national shipments and multi-
lingual documents.
ALLCAPS glued demo: THISISALONGALLCAPSWORDWITHNOBREAKS.
Letters/digits glued: VAT20percent, item123ABC, RefNo0002.
PUNCT runs: !!!! ????? .... ---- ____ :::::
French sample: La société s’engage à livrer les biens à la date convenue, conformément à l’Art
French list:
  (1) Quantité: 1000
  (2) Prix unitaire: 1,00
  (3) Montant total: 1 000,00
<SPx12> markers should become spaces if you apply SP_RE.
```

Attendu: normaliser inter-\n national -> international (si ton extraction conserve \n). Segmenter ALLCAPS si wordset fourni. Séparer thisContractNumber et VAT20percent selon tes règles.

3) COMPLEX — Header multi-colonnes + tableau de facture

Objectif: déclencher tes heuristiques layout (header/table/multicol) et vérifier que tes spans restent cohérents.

Facturé à	Envoyé à	Facture n°	Date	Échéance
Cendrillon Ayot 69 rue Nations 22000 Paris	Cendrillon Ayot 46 Rue St Ferreol 92360 Ile-de-France	FR-001	29/01/2019	24/05/2019

Ligne pseudo multi-colonnes (espaces) :

```
CODE          CLIENT                        NUMERO
FC001         SARL EL HANA                  0002
FC002         MA PETITE ENTREPRISE          0003
```

Table: lignes de facture

Référence	Description Produit	Quantité	P.Unitaire	Valeur
cl001	Produit 1 (grand brun)	1000	1.00	1,000.00
c1002	Produit 2 | avec barre verticale	1001	2.00	2,002.00
c1003	Produit 3 — texte long pour tester le retour à la ligne dans une cellule de tableau (wrap).	1002	3.00	3,006.00
c1004	Produit 4	1003	4.00	4,012.00
	TOTAL HT			145.00
	TVA 20.0%			29.00
	TOTAL TTC			174.00 €

Séparateurs (devraient être filtrés par _SEP_RE si tu les utilises) :

```
--------------------------------------------------------------------------------
________________________________________________________________________________
----------------------------------------  ----------------------------------------
```

4) NOISE & EDGE CASES — répétitions, ponctuation, arabe

Objectif: déclencher la suppression de bruit (lignes répétées) et valider la détection de langue arabe.

```
CONFIDENTIAL
Page 1 of 99
PAGE 12 / 99
Invoice No: 0001
Invoice No: 0002
Invoice No: 0003
SAMPLE SAMPLE SAMPLE
..................
!!!!!!!!!!!!!!!!
----- ----- -----
((((( ))))) 
Mix: A--B, C...D, E???F, G!!!!H
```

Arabe (pour détecter lang='ar' sur cette page si ta détection est par page) :

مرحبا بكم في اختبار معالجة اللغة. رقم الفاتورة: ١٢٣٤٥. التاريخ: ٢٠٢٦٠٢١٠.

Remarque: l’affichage peut être imparfait (direction/ligatures), mais les codepoints sont présents pour l’extraction.

Fin du pack — Utilise ce PDF comme baseline et ajoute tes propres cas réels ensuite.