Hard PDF Text for Extraction (Hyphenation + Line Breaks)

English (with hyphenated line breaks)

This docu- ment is designed to stress-test PDF text ex- traction. It contains hyphen- ation at line breaks, plus irregu- lar spacing and newlines.

The course structure was gener- ally coherent, but the assign- ments felt under- specified. Some students re- ported that feed- back arrived too late, while others found it excep- tionally helpful.

In week three we covered token- ization, normal- ization, and sentence seg- mentation. The lecture notes were well- written, yet a few sections were need- lessly verbose. The project was motivat- ing, but the evaluation rubric was un- clear.

A more realistic example (multiple line breaks):

We learned about embed- dings
and vector re- trieval,
then built a small RAG pipe- line to reduce halluci- nations.

Danish (med trunke-

rede ord og linje- skift)

Dette doku- ment er lavet til at teste ud- træk af tekst fra PDF. Det inde- holder ord, der bliver delt ved linje- skift, og det kan øde- lægge tokeni- sering, hvis man ikke ryd- der op først.

Kurset var over- ordnet set godt struk- tureret, men nogle krav var u- klare. Flere stude- rende savnede hurtig feed- back, især tidligt i for- løbet, og det gjorde plan- lægning frustre- rende.

I uge tre arbejdede vi med forbe- handling: tokeni- sering, nor- malisering og sætnings- segmentering. Undervi- seren var dygtig til at for- klare, men tempoet var til tider u- jævnt.

Et “paper-agtigt” eksempel med ekstra linje- skift:

Vi talte om embed- dings
og vektor- søgning,
og bagefter byggede vi en lille RAG- pipeline for at mindske halluci- nationer.