**Different ways to read pdf:**

1. Normal pdf :

mypdf = open(f, mode='rb')

   pdf\_document = PyPDF2.PdfFileReader(mypdf)

   first\_page = pdf\_document.getPage(0)

   page = first\_page.extractText()

1. From scanned image in pdf form:

from wand.image import Image

from PIL import Image as PI

import pyocr

import pyocr.builders

import io

path = "your path directory\demo.pdf"

tool = pyocr.get\_available\_tools()[0]

lang = tool.get\_available\_languages()[0] // 0 is eng

req\_image = []

final\_text = []

image\_pdf = Image(filename=path, resolution=300)

image\_jpeg = image\_pdf.convert('jpeg')

for img in image\_jpeg.sequence:

    img\_page = Image(image=img)

    req\_image.append(img\_page.make\_blob('jpeg'))

for img in req\_image:

    txt = tool.image\_to\_string(

        PI.open(io.BytesIO(img)),

        lang=lang,

        builder=pyocr.builders.TextBuilder()

    )

    final\_text.append(txt)

**Using different feature in word embeddings:**

1. We can append new data as columns to the end of the matrix.
2. As our input to CBOW will be sparse matrix and our new feature will be dense and continuous so it’s better to bin the dense variable into binary and if categorical then hot encoded into binary.