m.8.Exercises.an.templates → Navigating Data Employment

Overview	Use Apache Spark and machine learning to determine sentence authorship labels.
Data	https://www.kaggle.com/competitions/spooky-author-identification/code

Exercise -

Jupyter Not <decscriptio

Dataset Description: The spooky author

Data Professional Skills by Skill Domain
Skills Ontology Self-Assessment Template

ID	Focus & Medium	YesNo (Yn)
1	Data Professional Skills by Skill Domain	dfdf
2	└─ Education	
3	├─ Advanced degree in a quantitative discipline	
4	├─ Mathematics, Linguistics, Computer Science	
5	├─ Enrolled in an M.S./Ph.D. program in Comp. Science or Elect. Engineer	
6		
7	└─ Experience	
8	├─ Industry or academic experience in applied NLP - 2+ years	
9	├─ Research experience in fields such as machine learning, languages	
10	└─ program synthesis, software eng., or human-computer interaction	
11	├─ Research or practical experience in applying deep learning	
12	└─ on large-scale and real-world data - 3+years	
13		
14	└─ Programming and Technical Skills	
15	├─ Familiarity with OCR libraries like Tesseract, PyOCR, OpenCV, .NET, SDK	
16	├─ Extracting, cleaning, and preprocessing data sets using NumPy and Pandas	
17	├─ Knowledge of supervised and unsupervised machine learning techniques	
18	└─ regression models, decision tree models, clustering, deep learning	
19	└─ with tools like Scikit-learn, Tensorflow, Keras, or PyTorch	
20	├─ Data visualization skills using tools such as Matplotlib, Tableau, etc	
21	├─ Familiarity with rule-based NLP like CFG, constituency, and parsing	
22	└─ and related libraries including NLTK, spaCy, Stanford NLP	
23	├─ Specialization in OCR and familiarity with Transformers, ELMo, and BERT	
24	├─ Experience with Python NLP packages like Spacy, NLTK, and	
25	└─ Statistical packages familiarity like R, Python, SPSS, SAS, STATA	
26	├─ Experience with deep learning techniques and publishing in related	

```
|       └─ conferences (ICML, CVPR, NeurIPS)
|    ├─ Handling and analyzing data at scale w Hadoop, Dask, Spark, MapReduce
|    ├─ Working knowledge of data store tools like SQL, Elasticsearch

  └─ Analytical and Problem-Solving Skills
|    ├─ Proficiency in quantitative and qualitative analytical techniques rooted
|       └─ in business, economic, and statistical analysis
|    ├─ Ability to perform business analysis of market competitiveness,
|       └─ financial analysis, social media monitoring
|    ├─ Expertise in statistical analysis (linear regression, logistic regression,
|       └─ nonparametric statistics, probabilistic modeling, spatial modeling
|    ├─ Ability to tell stories using data
|    ├─ Strong problem-solving abilities

  └─ Additional Skills and Preferences
|    ├─ Knowledge of healthcare industry practices and medical coding (a plus)
|    ├─ Experience with computational imaging, cyber security, dist systems,
|       └─ logistics, next-generation networking, quantum information processing,
|       └─ sensor systems, speech and language processing, etc.
|    ├─ Security Clearance (for specific positions)
|    ├─ Experience managing, coding, and analyzing qualitative data using
|       └─ content analysis software
|    ├─ Time series analysis expertise (Prophet, ARIMA, LSTMs)
|    ├─ Writing maintainable, testable, production-grade Python code
|    ├─ Understanding of different machine learning and deep learning algorithm
|       └─ families and their tradeoffs
|    ├─ Experience with Selenium and SeleniumGrid
|    ├─ Data analytics, data mining, or other data science skills
|    ├─ Database experience, preferably working with Mongo databases
|    ├─ Experience working with data in Information Security, Cybersecurity,
|       └─ or Threat Intelligence
|    ├─ Experience working with bulletin boards and forums
```

Dataset Description: The spooky author identification dataset contains text from works of fiction written by spooky authors of the public domain: Edgar Allan Poe, HP Lovecraft and Mary

TASK SUMMARY

Jupyter Not <decscription here>

- **a.:** Jupyter Notebook
- **B:** Jupyter Notebo