## **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**  **2**  **3**  **4**  **5**  **6**  **7**  **8**  **9**  **10**  **11**  **12**  **13**  **14**  **15**  **16**  **17**  **18**  **19**  **20**  **21**  **22**  **23**  **24**  **25**  **26**  **27**  **28**  **29**  **30**  **31**  **32**  **33**  **34**  **35**  **36**  **37**  **38**  **39**  **40**  **41**  **42**  **43**  **44**  **45**  **46**  **47**  **48**  **49**  **50**  **51**  **52**  **53**  **54**  **55**  **56**  **57**  **58**  **59**  **60** | Data Professional Skills by Skill Domain  └─ Education  | ├─ Advanced degree in a quantitative discipline  | ├─ Mathematics, Linguistics, Computer Science  | ├─ Enrolled in an M.S./Ph.D. program in Comp. Science or Elect. Engineer  └─ Experience  | ├─ Industry or academic experience in applied NLP - 2+ years  | ├─ Research experience in fields such as machine learning, languages  | └─ program synthesis, software eng., or human-computer interaction  | ├─ Research or practical experience in applying deep learning  | └─ on large-scale and real-world data - 3+years  └─ Programming and Technical Skills  | ├─ Familiarity with OCR libraries like Tesseract, PyOCR, OpenCV, .NET, SDK  | ├─ Extracting, cleaning, and preprocessing data sets using NumPy and Pandas  | ├─ Knowledge of supervised and unsupervised machine learning techniques  | └─ regression models, decision tree models, clustering, deep learning  | └─ with tools like Scikit-learn, Tensorflow, Keras, or PyTorch  | ├─ Data visualization skills using tools such as Matplotlib, Tableau, etc  | ├─ Familiarity with rule-based NLP like CFG, constituency, and parsing  | └─ and related libraries including NLTK, spaCy, Stanford NLP  | ├─ Specialization in OCR and familiarity with Transformers, ELMo, and BERT  | ├─ Experience with Python NLP packages like Spacy, NLTK, and  | └─ Statistical packages familiarity like R, Python, SPSS, SAS, STATA  | ├─ 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 | dfdf |

**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