Big data analytics :

Data science Bookcamp - Case study 4

VALDIVIA Kimberley - BLICHER Cloé-lys - ABDUL AZIM Hina

1. Introduction (kim)

- Data Science Bookcamp Five Real World Python Projects written by Leonard Apeltsin
- Purpose: provide readers all the knowledge to deepen their skills in data science
- What will they learn through the book ?
- Techniques for computing and plotting probabilities
- Statistical analysis using SciPy
- How to organize datasets with clustering algorithms,
- How to visualize complex multi-variable datasets
- How to train a decision tree machine learning algorithm
- Book is structured in case studies based on real-world situations in order for the readers to gain the ability of resolving problems

Leonard Apeltsin, the author

2. Use case presentation & problem definition (kim)

- Our case study: **Using online job postings to improve your data science resume.**
- The situation: we are data science professional who just came out of school and we are looking for our first job in the field. We have drafted our resume but we are wondering which skills we should to our resume in order to find a job accordingly.
- With job postings online platforms it is difficult to find results that match with our keywords when looking for a job.
- This case study shows how by using data science approach we can create the most complete resume and find all the missing skills not mentioned in our resume but seeked by job offers.

3. Data Analytics approach definition and explanation (hina)

- Pourquoi utiliser data analytics pour ce problème ?
- quoi : data searching (données à analyser = vient d'ou ⇒ voir doc projet khalfallah)
- Comment :

❏ measuring Text similarity

⇒ text vectorization = term frequency vectors / one dimensional arrays / scikit-learn

■ Efficiently cluster large text datasets

■ Visually display multiple text clusters

■ Parse HTML files for text content

4. Solution development and illustration (cloé)

résultats / solution du case study

Algorithmes utilisés pour le case study et il faut les définir je pense

K-means Clustering

DBSCAN Clustering

Jaccard similarity computation

Text comparison

Cosine similarity computation

Parsing text from HTML

Computing text similarities

Clustering and exploring large text datasets

5. Evaluation and feedback (kim)

- Est ce que c’est useful for the problem ? = mieux de faire ça avec big data ou sans ?
- Est ce que cela répond à la problématique ?
- Easy ou difficile à faire la démonstration ?

6. Conclusion (kim)

· We discovered that unique job skills are market by bullet points in each HTML file

· Text clustering is hard. Ideal cluster count rarely exist because of language fluidity and boundaries between topics But despite the uncertainty, certain

· topics consistently appear across multiple cluster counts. So, even if our elbow

· plot does not reveal the exact number of clusters, the situation is salvageable:

· sampling over multiple clustering parameters can reveal stable topics in the text.

7. Appendix: software and datasets (cloé)

- annexes
-

Webography

https://www.manning.com/books/data-science-bookcamp

https://livebook.manning.com/book/data-science-bookcamp/welcome/v-5/9